HAL
open science

Catégorisation visuelle rapide des scènes naturelles : limites du parallélisme et spécificité des visages.Une étude comportementale et électrophysiologique chez l'humain

Guillaume A. Rousselet

▶ **To cite this version:**

Guillaume A. Rousselet. Catégorisation visuelle rapide des scènes naturelles : limites du parallélisme et spécificité des visages.Une étude comportementale et électrophysiologique chez l'humain. Neurosciences [q-bio.NC]. Ecole des Hautes Etudes en Sciences Sociales (EHESS), 2003. Français. NNT : . tel-00071015

HAL Id: tel-00071015

https://theses.hal.science/tel-00071015v1

Submitted on 22 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ECOLE DES HAUTES ETUDES EN SCIENCES SOCIALES

Année : 2003

THESE

pour obtenir le grade de

DOCTEUR DE L'EHESS

***Discipline* : Sciences cognitives**

***Spécialité* : Neurosciences computationnelles**

présentée et soutenue publiquement

par

Guillaume Alexis Rousselet

le 03 novembre 2003

Titre :

Catégorisation visuelle rapide des scènes naturelles : limites du parallélisme et spécificité des visages.

Une étude comportementale et électrophysiologique chez l'humain

JURY

Pr. Michel Imbert – Cerco – CNRS, Université Paul Sabatier, Toulouse	**Président**
Dr. Marie-Hélène Giard – Inserm unité 280, Lyon	**Rapporteur**
Dr. Philippe G. Schyns – Glasgow University, Glasgow	**Rapporteur**
Dr. Nathalie George – LENA – CNRS, Paris	
Dr. Muriel Boucart – Hôpital Salengro – CNRS, Lille	
Dr. Michèle Fabre-Thorpe – Cerco – CNRS, Toulouse	**Directeur de thèse**

Centre de Recherche Cerveau & Cognition (CerCo)
UMR 5549 CNRS – Université Paul Sabatier
Faculté de Médecine de Rangueil, 133 route de Narbonne
31062 Toulouse, Cedex

à mes parents, à mes soeurs

à Roxane

Remerciements

Je voudrais remercier très sincèrement Jean Bullier pour son accueil, son soutien ainsi que pour la richesse des discussions que nous avons pu avoir. Tu resteras toujours très haut dans mon estime. Salut Capitaine !

Je tiens ici à exprimer toute ma reconnaissance et mon amitié à Michèle Fabre-Thorpe. Travailler avec toi a été extraordinairement enrichissant tant au niveau scientifique qu'au niveau humain. Merci pour ta franchise et ton souci du détail. Merci de m'avoir laissé autant de libertés dans le déroulement de cette thèse. Mais l'aventure ne s'arrête pas là, à bientôt !

Mille mercis à Simon Thorpe pour sa gentillesse, son écoute et ses conseils. Il faudra bien quelques décennies pour venir à bout de toutes tes idées d'expériences. Merci de m'avoir transmis ce virus. Et encore merci à Michèle de savoir fixer les limites du raisonnable !

Je remercie Michel Imbert de bien vouloir évaluer mon travail de thèse. Merci aussi de m'avoir accueilli lors de mes tout premiers passages au laboratoire. Merci à Marie-Hélène Giard et à Philippe Schyns d'avoir accepté d'être les rapporteurs de ce travail. Merci à Nathalie George et Muriel Boucart d'avoir accepté de faire partie du jury. J'espère que mon travail les satisfera.

Marc Macé a été un collègue et un ami extrêmement précieux pendant cette thèse. Il m'aurait été difficile de réaliser autant de choses sans lui. Comment imaginer qu'après une discussion animée un vendredi soir, le samedi une nouvelle expérience était mise au point, le dimanche matin on se testait mutuellement pour tout analyser l'après-midi même et présenter les résultats le lundi matin à Michèle & Simon ? Et s'il n'y avait que ça… Je tiens à te remercier chaleureusement pour ton aide et tous ces moments passés ensemble.

Merci à Nadège Bacon-Macé pour son aide inestimable dans la mise au point de certaines expériences et surtout pour son amitié, son dynamisme à toute épreuve et son petit grain de folie, une qualité que semblent partager tous les étudiants de l'équipe... Courage, tu vas faire une super thèse !

Merci à Rufin VanRullen pour ses conseils et sa perspicacité. C'est un plaisir de travailler avec toi.

Merci à mes parents pour… tant de choses… Vous m'avez donné l'essentiel pour réaliser cette thèse : le goût du travail.

Les mots me manquent pour exprimer ma gratitude à Roxane Itier. Ton soutient quotidien et infaillible m'ont donné des ailes. Tu es ma boussole et mon oasis.

Je voudrais remercier tous les membres de l'équipe de Michèle et Simon avec qui j'ai eu la chance d'interagir, tout particulièrement Arnaud Delorme et Denis Fize qui m'ont mis le pied à l'étrier, Catherine Marlot et Ghislaine Richard, qui vont regretter les cris dans les couloirs, ainsi que Rudy Guyonneau, Nicolas Guilbaud et Jong-Mo Allegraud, les gladiateurs du neurone à spike… Vous avez tous participé à la bonne ambiance qui règne dans l'équipe.

Merci enfin à tous les membres du Cerco qui participent à la vie de ce lieu unique où il fait bon travailler, mais aussi rire, échanger et boire un coup de temps en temps… Merci à tous !

Publications

Articles publiés en anglais

Rousselet, G.A., Fabre-Thorpe, M. & Thorpe, S.J. (2002) Parallel processing in high-level categorization of natural images. *Nature Neuroscience* 5 : 629-630.

Rousselet, G.A., Thorpe, S.J. & Fabre-Thorpe, M. (2003) Taking the MAX from neuronal responses. *Trends in Cognitive Sciences* 7 : 99-102.

Rousselet, G.A., Macé, M.J.-M. & Fabre-Thorpe M. (2003) Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *Journal of Vision* 3 : 440-455.

Delorme, A., Rousselet, G.A., Macé, M. J.-M. & Fabre-Thorpe, M. Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Cognitive Brain Research* (sous presse).

Rousselet, G.A., Macé, M.J.-M. & Fabre-Thorpe M. The N170 ERP component: specificity, effects of task status and inversion for animal and human faces in natural scenes. *Journal of Vision* (sous presse).

Rousselet, G.A., Thorpe, S.J. & Fabre-Thorpe, M. Processing of one, two or four natural scenes in humans: the limits of parallelism. *Vision Research* (sous presse).

Articles publiés en français

Rousselet, G.A. (2003) L'analyse visuelle des scènes naturelles : rapide et parallèle ! *La lettre du neurologue* 1 : 33-35.

Rousselet, G.A. & Fabre-Thorpe, M. (2003) Les mécanismes de l'attention visuelle / Visual attention : underlying mechanisms. *Psychologie française* 48 : 29-44.

Articles soumis

Rousselet, G.A. & Fabre-Thorpe, M. How long to get to the "gist" of real-world natural scenes. *Visual Cognition*.

Rousselet, G.A. Macé, M.J.-M & Fabre-Thorpe, M. Comparing animal and face processing in the context of natural scenes using a fast categorization task. *Neurocomputing*.

Articles en préparation

Rousselet, G.A., Macé, M.J.-M., Thorpe, S.J. & Fabre-Thorpe, M. ERP studies of object categorization in natural scenes: in search for category specific differential activities.

Rousselet, G.A., Macé, M.J.-M. & Fabre-Thorpe, M. N170 evoked by faces in natural scenes: specificity, effects of size, task status and inversion.

Macé, M. J-M., Rousselet, G.A., C.R., Thorpe, S.J., & Fabre-Thorpe, M. Very early ERP effects in rapid visual categorisation of natural scenes: a reflect of low-level visual properties?

Résumés de conférences publiés

Fabre-Thorpe, M., Delorme, A., Rousselet, G. & Thorpe, S., 2000. The speed of processing of natural scenes: detection, categorization and the role of top-down knowledge. Proceedings of the Fourth International Conference on Cognitive and Neural Systems. Boston, MA, USA, 25-27 May.

Rousselet, G., Delorme, A. & Fabre-Thorpe, M., 2001. Reconnaissance et catégorisation rapides de scènes naturelles : effets de la diagnosticité sur la dynamique temporelle du traitement chez le sujet humain, une étude en potentiels évoqués. Comptes-rendus du 5ème Colloque de la Société des Neurosciences.

Rousselet, G.A., Fabre-Thorpe, M. & Thorpe, S.J., 2001. Two unrelated natural scenes can be processed as fast as one. *Perception* supplement, volume 30, 107.

Rousselet, G.A., Fabre-Thorpe, M. & Thorpe, S.J., 2002. Two natural images can be processed as fast as one in a superordinate visual categorization task. *Journal of Cognitive Neuroscience* supplement, A110, 40.

Rousselet, G.A., Macé, M. J-M., Sternberg, C. R., Fabre-Thorpe, M. & Thorpe, S.J., 2002. Rapid categorization of faces and animals in upright and inverted natural scenes: no need for mental rotation and evidence for a selective visual streaming of upright faces. *Perception* supplement, volume 31, p132a.

Macé, M. J-M., Rousselet, G.A., Sternberg, C.R., Fabre-Thorpe, M. & Thorpe, S.J., 2002. Very early ERP effects in rapid visual categorisation of natural scenes: Distinguishing the role of low-level visual properties and task requirements. *Perception* supplement, volume 31, p132b.

Thorpe, S.J., Bacon, N.M., Rousselet, G.A., Macé, M. J-M. & Fabre-Thorpe, M., 2002. Rapid categorization of natural scenes: feedforward vs. feedback contribution evaluated by backward masking. *Perception* supplement, volume 31, p150.

Rousselet, G.A. & Fabre-Thorpe, M., 2003. Processing speed of natural scenes: categorization of the global context. *Journal of Cognitive Neuroscience* supplement, B298, p84.

Rousselet G.A., Macé M.J-M. & Fabre-Thorpe M., 2003. Comparing animal and face processing in the context of natural scenes using a fast categorization task. *12th Annual Computational Neuroscience Meeting, Alicante (Spain), Proceedings.*

Table des matières

Chapitre 1 :
Le traitement des informations visuelles au sein des scènes naturelles 1

Partie A :
Quel est le degré de parallélisme dans le traitement visuel des scènes naturelles ? 2

1 Sériel vs. parallèle : ce que nous apprend le comportement 3
1.1 Le point de vue classique : le modèle sériel 3
1.2 La dichotomie parallèle/sériel remise en cause 8
1.3 Un modèle hybride : le guided search model 10
1.4 Les modèles parallèles 12
1.5 Sériel vs. parallèle : apports d'autres paradigmes expérimentaux 14
1.5.1 Présentations rapides de photographies de scènes naturelles 14
Le paradigme RSVP 14
La mémoire conceptuelle à court terme 15
Paradigme go/no-go : catégorisation animal/non-animal 16
La catégorisation rapide d'un animal s'effectue t'elle en parallèle ? 17
1.5.2 Repetition Blindness 19
Paradigme 19
Interprétation 20
1.5.3 Attentional Blink 21
Paradigme 21
Interprétation 21
1.5.4 Inattentional Blindness 23
Paradigme 24
Interprétation 24
1.5.5 Change Blindness 25
Paradigme 25
Première interprétation 27
La vision post-attentive 28
Nouvelles interprétations 31
2 Sériel vs. parallèle : au cœur du cerveau 33
2.1 L'organisation neuronale du système visuel : des éléments en faveur du modèle sériel ? 33
2.2 Quelques données chez l'humain 40
2.2.1 Apports de la neuropsychologie 40
Syndrome de Balint 40
Négligence spatiale unilatérale et extinction visuelle 42
2.2.2 Apports de l'imagerie fonctionnelle 45
2.2.3 Apports de la TMS 47
2.2.4 Apports des potentiels évoqués 49
2.3 Codage neuronal dans la voie ventrale : au-delà des apparences 54
2.3.1 Les briques de base de la perception visuelle 55
2.3.2 Compétition et codage spatial dans la voie ventrale 57
Codage spatial 58
Convergence ventrale-dorsale dans les cortex perirhinal et entorhinal 59
Le code neuronal 60
L'hypothèse des MAX locaux 62

Le biais fovéal 63
Liage perceptif par synchronisation des décharges neuronales 66

3 Attention et conscience 67
3.1 De l'attention à la conscience 68
L'importance du cortex préfrontal 71
3.2 De la conscience à l'attention 72

Résumé des travaux antérieurs de l'équipe 74

Article 1 : Parallel processing in high-level categorization of natural images 75
*** données complémentaires non publiées** 85
*** poster présenté au congrès Cognitive Neuroscience Society 2002** 93

Article 2 : Processing of one, two or four natural scenes in humans: the limits of parallelism 94

Partie B :
Catégorisation visuelle des scènes naturelles : le contexte et son influence sur la perception des objets 125

1. L'influence du contexte sur la catégorisation des objets 126

2. Mécanismes de la catégorisation du contexte 129
2.1 Bases anatomiques de la catégorisation des scènes naturelles 129
2.2 Bases fonctionnelles de la catégorisation des scènes naturelles 130

Article 3 : How long to get to the "gist" of real-world natural scenes? 134
*** poster présenté au congrès Cognitive Neuroscience Society 2003** 158

Article 4 : Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes 159

Chapitre 1 : conclusion générale 176

Chapitre 2 :
Les visages ont-ils un statut particulier au sein du système visuel ? 180

1. Y a t'il une ségrégation anatomique des mécanismes de traitement des visages ? 181
1.1 Données neuropsychologiques 182
Les cas « purs » n'existent pas 182
Des modules plastiques 183
Des patients virtuels aux patients fantômes 184
1.2 Données IRMf et TEP 185
1.3 Données issues de l'électro- et de la magnéto-encéphalographie 190

2. Le traitement des visages est-il différent de celui des objets ? 193
2.1 L'effet d'inversion au niveau comportemental 194
2.2 L'effet d'inversion au niveau électrophysiologique 196

3. Le traitement des visages est-il spécifiquement plus rapide que celui des objets ? 198
3.1 Données électro- et magnéto-encéphalographiques 198
3.2 Données neuropsychologiques 202

4. Conclusion 204

Article 5 : Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes 205
*** poster présenté au congrès ECVP 2002** 211

Article 6 : ERP studies of object categorization in natural scenes: in search for category specific differential activities 212

Article 7 : The N170 ERP component: specificity, effects of task status and inversion for animal and human faces in natural scenes 236

Chapitre 2 : conclusion générale 242

Perspectives 243

Bibliographie 245

"C'est dans l'épreuve que je fais d'un corps explorateur voué aux choses et au monde, d'un sensible qui m'investit jusqu'au plus individuel de moi-même et m'attire aussitôt de la qualité à l'espace, de l'espace à la chose et de la chose à l'horizon des choses, c'est-à-dire à un monde déjà là, que se noue ma relation avec l'être."

Merleau-Ponty (1945)

Chapitre 1

Le traitement des informations visuelles au sein des scènes naturelles

Notre environnement visuel est d'une extraordinaire richesse. Pourtant, de nombreuses études révèlent que les mécanismes visuels implémentés dans notre cerveau sont par essence limités, incapables de saisir toute la richesse du monde extérieur à chaque instant. Notre perception complète et instantanée d'un monde riche autour de nous serait-elle alors une illusion ? Comment percevons nous réellement notre environnement ?

Le premier chapitre de cette thèse ne prétend pas répondre de manière complète et décisive à ces questions. Il s'attachera plus particulièrement à comprendre quelles sont les stratégies mises en place par le système visuel pour faire face à la complexité des scènes naturelles, ces stimuli qui nous entourent en permanence. Deux axes principaux seront abordés.

Nous verrons tout d'abord comment le système visuel parvient à sélectionner certains objets dans le but d'interagir avec eux dans des environnements qui contiennent typiquement de nombreux éléments non pertinents pour le comportement. Cette importante revue de la littérature dévoile les limites du codage en parallèle des objets de notre environnement. Cependant, en opposition à un courant important de recherche, des capacités insoupçonnées de traitement simultané seront mises en évidence.

Dans une deuxième partie, nous tenterons de comprendre le fonctionnement du système visuel quand celui-ci n'a plus pour tâche de se focaliser sur quelques objets, mais au contraire d'assigner une catégorie visuelle à la scène elle-même, ce qui requiert indéniablement la prise en compte simultanée d'un maximum d'informations dans la scène.

Partie A : Quel est le degré de parallélisme dans le traitement visuel des scènes naturelles ?

Lorsque nous ouvrons les yeux, nous avons sans aucun effort de notre part la sensation immédiate d'un monde riche, spatialement structuré, texturé, ombré, et peuplé de nombreux objets de tailles et de formes très variées. Cette sensation d'une perception simultanée de nombreux objets pourrait reposer sur des mécanismes visuels opérant massivement en parallèle. Cependant, la perception visuelle est plus qu'une simple prise d'information passive sur le monde, elle fait appel à un acte sensori-moteur, puisque nous explorons de manière active notre environnement en bougeant sans arrêt nos yeux (Yarbus, 1967). Entre la sensation de tout percevoir devant nous de manière simultanée et notre besoin d'exploration oculomotrice sérielle, quelle est la véritable nature des mécanismes visuels ?

Cette première partie a pour objectif de fournir une description relativement vaste, bien que non exhaustive, des arguments en faveur de la présence de mécanismes sériels ou parallèles dans le système visuel. Le survol de la littérature comportementale nous entraînera des modèles sériels aux modèles parallèles, tout d'abord au travers du paradigme dit de recherche visuelle, puis au travers de nombreuses études plus récentes ayant mis en jeu d'autres paradigmes expérimentaux. La revue de la littérature en neurosciences cognitives permettra d'élargir considérablement le débat, au travers de données acquises chez le primate humain et non humain. La prise en compte de ces données a amené certains auteurs à proposer une explication moderne du fonctionnement du système visuel, allant bien au-delà de la dichotomie sériel/parallèle, dont je tenterai une synthèse.

1. Sériel vs. parallèle : ce que nous apprend le comportement

1.1 Le point de vue classique : le modèle sériel

En première approximation, le traitement visuel des scènes naturelles a été étudié abondamment par le biais du paradigme de recherche visuelle (« visual search », voir e.g., Pashler, 1998 ; Treisman, 1998a, 1998b ; Wolfe, 1998 ; Wolfe & Cave, 1999). La tâche impliquée est largement utilisée dans la vie de tous les jours. A chaque instant nous cherchons des objets (les mots de ce texte, un stylo sur la table, une tasse…) soit de manière explicite, en bougeant les yeux, soit de manière implicite, en déplaçant la cible

de notre attention. De manière générale, dans une tâche de recherche visuelle, les sujets ont pour consigne de rechercher un objet cible prédéterminé parmi un ensemble d'objets non cibles, dits aussi distracteurs. D'un essai à l'autre le nombre de distracteurs utilisés dans la stimulation varie. La cible apparaît en général dans 50% des essais. Dans les autres essais seuls des distracteurs sont présentés. Les sujets répondent par exemple en pressant un bouton pour indiquer qu'ils ont trouvé la cible et sur un autre bouton pour indiquer que la cible est absente. Dans une telle tâche, la mesure dépendante peut être le temps de réaction (TR), dans ce cas les stimuli restent affichés jusqu'à la réponse du sujet. Dans le cas de présentations brèves, parfois suivies d'un masque pour bloquer la persistance de l'image, on mesure la précision des sujets. Dans la plupart des études seuls les TR sont mesurés. Ainsi, il a été montré que le temps nécessaire pour trouver une cible dépend du type de cibles utilisées et du nombre de distracteurs. Quand les sujets doivent trouver une cible qui diffère des distracteurs selon une seule dimension (e.g., chercher une barre horizontale parmi des barres verticales, Figure 1), leurs TR sont relativement courts et indépendants du nombre de distracteurs (Figure 2). Cette configuration de résultats a été interprété comme étant la marque de mécanismes pré-attentifs opérant en parallèle sur l'ensemble des stimuli (Treisman & Gelade, 1980). Les stimuli traités en parallèle semblent « surgir » de la masse des distracteurs (ils « pop out », selon l'expression consacrée). Au contraire, quand les sujets cherchent une cible définie par une conjonction d'éléments eux-mêmes partagés par les distracteurs (e.g., chercher un X rouge parmi des X bleus et des A rouges), leurs TR sont plus longs et augmentent avec le nombre de distracteurs (Figure 2). En variant systématiquement le nombre de distracteurs présentés, il est possible de construire une fonction de recherche décrivant l'évolution des TR en fonction du nombre de distracteurs dans une recherche de conjonction d'éléments. La pente d'une telle fonction est utilisée comme outil pour inférer la nature des mécanismes de recherche visuelle. Ainsi, la présence de pentes non nulles, contrairement aux pentes nulles obtenues lors de la recherche d'éléments uniques, a été considérée comme une preuve en faveur de l'implication de l'attention dans la recherche de conjonctions d'éléments (Treisman & Gelade, 1980). Plus précisément, l'attention spatiale serait nécessaire pour assembler les divers éléments de base dont est constitué un

objet à un endroit donné dans l'espace. Les éléments d'un objet ne seraient correctement assemblés que lorsque l'attention est focalisée sur cet objet.

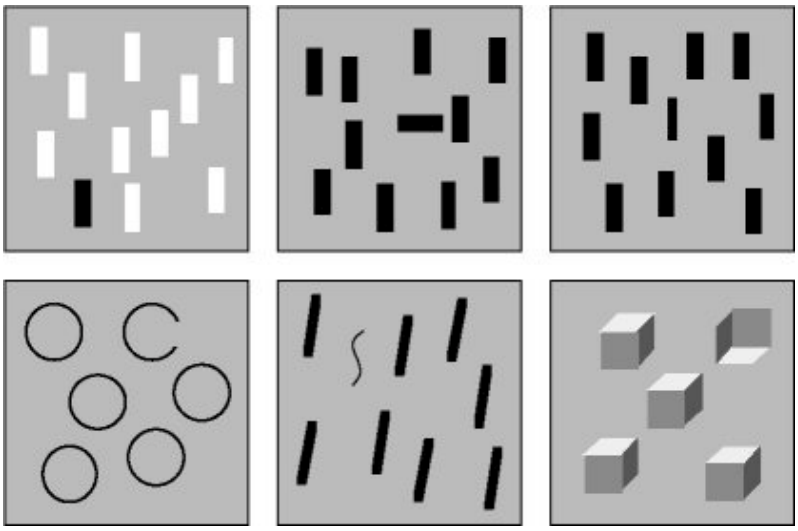


Figure 1. Exemples "d'éléments de base" donnant lieu à des recherches simples. Tiré de Wolfe (2001).

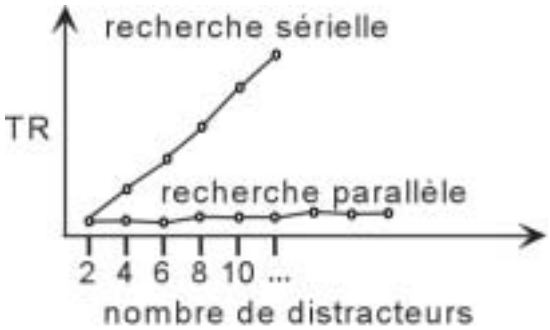


Figure 2. Deux exemples caractéristiques de fonctions de recherche. Le temps de réaction, en unités arbitraires, est exprimé en fonction du nombre de distracteurs présentés en même temps que la cible. Lorsque la recherche prend le même temps quelque soit le nombre de distracteurs, c'est un cas dit de recherche parallèle. Si le temps de recherche augmente avec le nombre de distracteurs, c'est alors un cas dit de recherche sérielle.

L'existence de conjonctions illusoires est en parfait accord avec cette théorie. En effet, quand plusieurs objets sont présentés centralement (e.g. une croix bleue et un « T » rouge) pendant un temps bref mais que les sujets doivent simultanément effectuer une tâche sur d'autres objets présentés en périphérie, les sujets rapportent souvent avoir vu des objets qui n'étaient pas présents, mais composés d'éléments appartenant aux objets réellement présentés (dans notre exemple, une croix rouge, Treisman & Schmidt, 1982) (Figure 3). De telles erreurs ne surviennent pas lorsque les sujets peuvent focaliser leur

attention sur les éléments centraux, sans être distraits par les éléments périphériques. L'existence des conjonctions illusoires démontrerait ainsi que (1) les caractéristiques pré-attentives d'un objet sont codées séparément, sinon elles ne pourraient pas se recombiner ; (2) le problème du liage perceptif est tout à fait réel ; (3) l'attention focalisée est impliquée dans la résolution de ce problème (Robertson, 2003 ; Treisman, 1998a ; Wolfe & Cave, 1999).

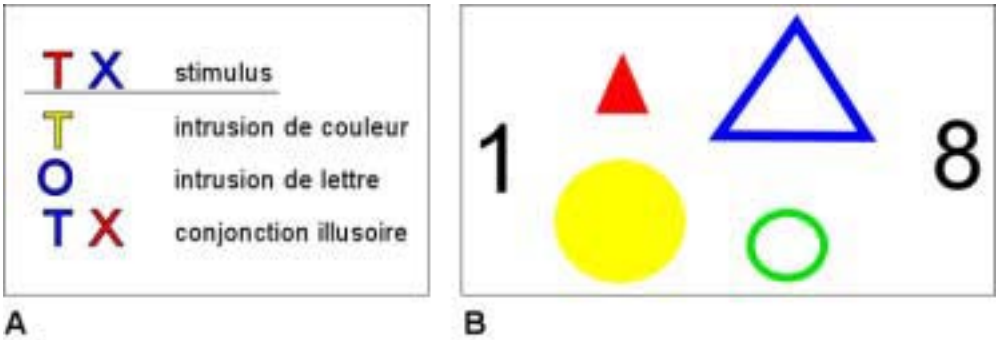


Figure 3. Conjonctions illusoires. A Exemples de stimuli utilisés pour tester des sujets sains et des sujets cérébrolésés comme le patient étudié par Friedman-Hill et al. (1995) (voir plus bas). Des conjonctions de caractéristiques dites de base comme la couleur et la forme (des lettres dans cette figure) peuvent être recombinées pour former une conjonction illusoire. Les conjonctions illusoires peuvent être isolées des réponses où les sujets devinent en les comparant aux erreurs d'intrusions (par exemple quand les sujets rapportent une couleur ou une lettre non présentées. **B** Illustration d'une tâche permettant de mettre en évidence des conjonctions illusoires. La tâche consistait à rapporter l'identité des deux chiffres en priorité et ensuite du plus possible d'éléments vus entre les deux chiffres. Adapté de Treisman, 1998a.

Pour expliquer la dichotomie entre traitements pré-attentifs et traitements nécessitant une attention focalisée, la « théorie d'intégration des traits » (FIT, pour « Feature Integration Theory », Treisman & Gelade, 1980), stipule que les mécanismes visuels de la perception des scènes naturelles sont divisés en deux étapes. Tout d'abord, des mécanismes pré-attentifs agiraient en parallèle dans l'ensemble de la scène visuelle pour extraire des éléments dits de base (comme la couleur, la texture, les contours locaux, le mouvement, la taille…). Ces différents types d'éléments simples seraient encodés dans des cartes neuronales séparées. Selon ce modèle, la recherche d'un élément simple pourrait être réalisée aisément en vérifiant la présence d'une quelconque activité dans la carte codant cet élément. Deuxièmement, l'attention est impliquée d'une manière sérielle dans l'assemblage des différents éléments de base constituant un objet complexe afin d'en former une représentation de haut niveau. Les objets complexes (faits d'un ensemble d'éléments de base) ne pourraient être représentés dans le système visuel sans faire appel

à l'attention. Il existerait selon Treisman & Gelade (1980) une carte de contrôle (« master map ») enregistrant la position de tous les éléments présents dans le champ visuel. Porter son attention sur une position donnée se traduirait par la mise en place de liens dynamiques entre la représentation de cette position dans la carte de contrôle et chacune des caractéristiques codées à cette position dans les différentes cartes d'éléments. Ce lien dynamique permettrait d'assembler de manière explicite les éléments présents à cette position en une représentation complexe tout en excluant les éléments se trouvant à d'autres positions. Avant l'arrivée de l'attention, le monde visuel serait composé d'une « soupe » d'attributs sans organisation spatiale (Figure 4). Etant donné qu'une même position ne peut être occupée que par un objet à la fois, la FIT propose un mécanisme simple pour faire face à la richesse de notre environnement visuel tout en évitant les conjonctions illusoires. La limitation d'un tel mécanisme réside dans le fait que le processus d'assemblage des éléments simples par l'attention ne peut être réalisé que pour un objet à la fois, i.e. il fonctionne de manière sérielle, à la manière d'une fenêtre attentionnelle « illuminant » les cibles les unes après les autres (voir aussi Posner, 1980).

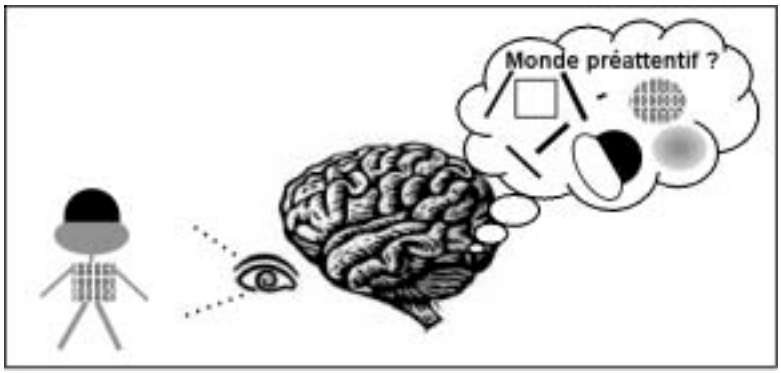


Figure 4. La vision préattentive d'un objet : une collection d'éléments simples, sans organisation spatiale ?

La notion d'attention, ou plus particulièrement d'attention visuelle, avait déjà une longue histoire à l'époque où Treisman proposa son modèle (Hatfield, 1998). L'idée d'une séparation entre traitements pré-attentifs et attentifs n'était pas nouvelle non plus (Kinchla, 1992). L'élément majeur apporté par Treisman était cette dichotomie stricte entre un traitement pré-attentif (parallèle) d'éléments simples et un traitement attentionnel (sériel) de conjonctions d'éléments, l'ensemble étant intégré dans un modèle

très simple de fonctionnement du système visuel. Par son originalité et sa simplicité, le modèle FIT a suscité de très nombreuses recherches qui l'ont très rapidement invalidé en faveur de modèles plus réalistes.

1.2 La dichotomie « parallèle/sériel » remise en cause

De nombreux éléments remettent en cause le modèle original de Treisman et sa dichotomie stricte entre traitements parallèles et sériels. Ces éléments s'articulent autour de quatre points principaux.

* *La dichotomie entre traitements parallèle et sériel n'est pas toujours retrouvée*. De nombreuses recherches ont tenté de déterminer les éléments de bases traités en parallèle par le système visuel (voir revue dans Wolfe, 1998). Autrement dit, quelles sont les « briques de base » de la perception visuelle ? Ces études ont dévoilé trois résultats importants :

(a) La recherche de certains éléments dits « de base » s'est révélée sensible au nombre de distracteurs présentés conjointement à la cible, c'est-à-dire qu'ils seraient traités de manière sérielle selon la nomenclature introduite par Treisman (e.g., Verghese & Nakayama, 1994 ; Nagy & Sanchez, 1990 ; Wolfe et al. 1992 ; Treisman & Gormican, 1988). De manière concomitante, les TR absolus associés à une recherche dite parallèle peuvent varier considérablement d'une expérience à l'autre, alors que d'après le modèle de Treisman ils ne devraient pas varier.

(b) La recherche de certaines conjonctions (e.g. disparité et couleur ; mouvement et forme ; couleur, taille et forme…) s'avère peu sensible ou insensible au nombre de distracteurs, c'est-à-dire qu'elles seraient traités de manière parallèle ou quasi parallèle (e.g., Egeth et al., 1984 ; Nakayama & Silverman, 1986 ; Steinman, 1987 ; McLeod et al., 1988 ; Wolfe et al., 1989 ; Duncan & Humphreys, 1989 ; Treisman & Sato, 1990 ; Ellison & Walsh, 1998).

(c) Certaines propriétés dites de haut niveau telles que l'occlusion, la profondeur, la structure tridimensionnelle et la forme, qui supposent l'intégration de plusieurs éléments de bas niveau, semblent aussi être traitées en parallèle (Enns & Rensink, 1991, 1990 ; He & Nakayama, 1992, 1994 ; Kimchi, 2000 ; Kleffner & Ramachandran, 1992 ; Rensink & Enns, 1995; Verghese & Nakayama, 1994). Il n'y aurait donc pas deux grands

types de recherche visuelle mais plutôt un continuum allant de celles présentant des pentes de recherche plates à celles présentant des pentes fortes. Cela ne veut pas dire qu'il n'y ait pas de mécanismes sériels et parallèles à l'œuvre dans le système visuel, mais plutôt qu'il n'est pas possible de classer une tâche de recherche donnée comme étant strictement sérielle ou parallèle. Certains auteurs ont ainsi proposé d'abandonner la dichotomie sériel/parallèle en faveur d'une échelle de performance allant d'efficace (pente de fonction de recherche nulle) à très inefficace (pente de fonction de recherche très forte).

* *La variation des temps de traitement par objet*. Le modèle sériel strict prévoit un temps de traitement fixe pour chaque objet. La première chose que l'on peut constater est que ce temps de traitement varie considérablement d'une expérience à l'autre, allant de 5 ms par objet, jusqu'à parfois plusieurs centaines de ms (Duncan, Ward & Shapiro, 1994). Ces variations sont difficiles à concilier avec un mécanisme unique et sériel. De plus, avec des pentes de recherche de parfois moins de 10 ms par objet, il devient impossible de réconcilier les hypothétiques mécanismes de recherche visuelle avec la réalité physiologique. Si l'on suppose que le temps minimal nécessaire à un neurone pour opérer une sommation de dépolarisations et décharger est d'environ 5-10 ms, cela est clairement insuffisant pour la réalisation des quatre étapes du mécanisme d'assemblage des représentations de haut niveau : (1) focaliser l'attention sur la zone de l'espace où se trouve un objet, (2) assembler ses éléments de base, (3) comparer la représentation de haut niveau ainsi formée à une représentation en mémoire, (4) désengager l'attention puis la déplacer vers le prochain groupe d'éléments à assembler.

* *La recherche strictement sérielle n'existe pas*. En effet, si un mécanisme sériel était à l'œuvre dans notre système visuel, il devrait parcourir tous les objets d'une scène de manière systématique lorsque celle-ci ne contient pas la cible recherchée. Or la plupart des articles sur la recherche visuelle sérielle rapportent que les sujets font 5 à 10% d'erreurs, contrairement au 0% attendu si tous les objets étaient examinés. De plus, pour expliquer les recherches très rapides, avec par exemple un objet examiné toutes les 10

ms, certains chercheurs ont proposé que plusieurs items pourraient être examinés en une seule fixation de l'attention (e.g. Treisman, 1998a).

* *L'apport d'autres outils d'analyses*. Il existe un problème fondamental à interpréter des fonctions de recherche. Bien qu'il soit tentant d'inférer la nature sérielle ou parallèle d'un mécanisme d'après l'étude des TR, ceux-ci sont particulièrement inadéquats pour une telle entreprise. Pour s'en rendre compte il faut souligner que tous les résultats montrant une augmentation du TR avec le nombre de distracteurs, classiquement interprétés dans le cadre des modèles sériels, pourraient très bien être dus à la mise en œuvre d'un modèle parallèle ayant des capacités de traitement limitées (e.g. Kinchla, 1992, voir plus bas). Enfin, certains ont développé des outils d'analyse différents des pentes de fonction de recherche, tels que la procédure SAT (pour speed accuracy tradeoff, ou échange precision-vitesse, McElree & Carrasco, 1999) ou des modélisations mathématiques sophistiquées (Palmer, 1998). Ces analyses, que je n'ai pas la place de décrire ici, ont permis de montrer que certaines conjonctions d'éléments semblant être traitées de manière sérielle selon les critères classiques, sont en fait traitées de manière parallèle.

1.3 Un modèle hybride : le guided search model

Pour rendre compte de cet ensemble de données contradictoires, la FIT a eu à prendre en considération certaines propriétés des modèles parallèles suggérées par d'autres (voir plus bas). Pour expliquer les recherches efficaces de conjonctions de caractéristiques élémentaires, beaucoup de chercheurs écartent l'idée de détecteurs pré-attentifs de telles conjonctions qui poseraient un problème d'explosion combinatoire des possibilités de codage. Le modèle de recherche guidée (GSM, pour « guided search model », Wolfe et al., 1989 ; Wolfe & Gancarz, 1996) met en avant le fait que des mécanismes parallèles pourraient restreindre la recherche sérielle aux endroits les plus probables dans la scène visuelle (il faut signaler que Treisman (1998a) a proposé une version révisée de son propre modèle qui est très similaire au GSM, voir aussi Cave (1999) qui propose une version alternative du GSM). Ceci serait réalisé par un amorçage descendant (« top-down ») des cartes d'éléments de base. Un tel amorçage serait rendu

possible par la connaissance en avance de la composition en éléments de base de la cible. Concrètement, le GSM est composé, comme la première version de la FIT, de cartes d'éléments simples et d'une carte d'activation similaire à la carte de contrôle de la FIT. Le mode de fonctionnement du GSM est le suivant. L'attention se dirigerait d'abord vers l'objet qui a envoyé l'activité la plus forte à la carte d'activation. Pour chaque position dans la carte d'activation, i.e. pour chaque objet dans le champ visuel, la somme des activations des différentes cartes d'éléments de base est calculée. Dans chacune de ces cartes, le degré d'activation est proportionnel au degré de similarité entre l'élément encodé dans une carte donnée et les éléments de la cible, spécifiés par un amorçage descendant (dans la version révisée de la FIT la sélection descendante se fait par une inhibition des distracteurs, plutôt que par une activation de la cible dans le cas du GSM). La carte d'activation classe tous les items du champ visuel par ordre, de celui qui a le plus de chance d'être une cible à celui qui a le moins de chance d'être une cible. La recherche visuelle consisterait a parcourir cette liste, un item après l'autre jusqu'à ce que la cible soit trouvée. Ainsi, selon le GSM, il n'y a pas de différence intrinsèque entre les recherches d'éléments de base et de conjonctions d'éléments. Les sujets se comportent différemment dans les deux tâches parce que dans la recherche d'une conjonction, les distracteurs reçoivent aussi une activation descendante, ayant pour conséquence un niveau de bruit plus important dans la carte d'activation par rapport à la situation d'une recherche d'un élément non partagé par les distracteurs. Par conséquent, plusieurs déplacements de l'attention sont déclenchés par la carte d'activation dans le premier cas, pas dans le dernier.

L'information visuelle pré-attentive bénéficie d'un statut particulier dans les modèles hybrides tels que le GSM. Au lieu de la « soupe » d'attributs élémentaires originellement envisagée par Treisman & Gelade (1980), il semblerait plutôt que le monde visuel pré-attentif soit découpé en fichiers d'objets (Wolfe & Bennett, 1997, voir aussi Rensink, 2000a,b). Selon cette hypothèse, avant l'arrivée de l'attention, le système visuel découperait le monde en objets potentiels, ou proto-objets, les éléments appartenant à chaque objet étant regroupés sous la forme d'un « fichier »[1]. Au sein de

[1] Cette conclusion provient notamment du fait que les fonctions de recherche indiquent que les espaces non occupés par des objets dans une scène ne sont pas visités (Wolfe, 1994). On pourrait aussi voir dans cette

chaque fichier, les éléments qui composent un objet potentiel ne sont pas reliés entre eux, c'est-à-dire qu'il serait impossible de connaître leur organisation spatiale avant le déploiement de l'attention à cet endroit (Figure 3). Cette absence totale de structuration spatiale pré-attentive a cependant été contestée récemment sur la base de nouvelles expériences de recherche visuelle (Donnelly et al., 2000). Le codage spatial dans le système visuel et tout particulièrement dans les aires représentant les objets est abordé en détail plus loin dans ce chapitre.

1.4 Les modèles parallèles

Les modèles parallèles à capacités limitées stipulent que tous les items dans notre champ visuel sont traités en même temps par un mécanisme compétitif. Ce mécanisme est fondé sur des interactions mutuellement inhibitrices entre les représentations activées par les différents éléments d'une scène (Duncan & Humphreys, 1989). Des « évidences» s'accumulent à chaque position spatiale occupée par un item en faveur de la présence d'une cible ou d'un distracteur. Une réponse est déclenchée quand un certain seuil de réponse en faveur du « oui » ou du « non » est atteint. Ces modèles ayant une quantité de « ressources attentionnelles » limitées, plus il y a d'items à traiter, moins il y a de ressources par item et donc plus la recherche est lente. Il existe une vaste panoplie de modèles parallèles qu'il est impossible de présenter ici. Certains emploient la métaphore d'une « course » entre représentations pour décrire le fonctionnement du système visuel, les items les plus rapidement catégorisés gagnant cette course (e.g., Bundesen, 1998). Au sein de cette classe de modèles parallèles, différents degrés d'interactions compétitives entre les stimuli en course ont été proposés (Kinchla, 1992). Certains ont également formulé des modèles parallèles dans le cadre de la théorie de la détection des signaux (Eckstein, 1998). Selon ce type de modèles, le bruit intrinsèque aux mécanismes de traitement visuels serait amplifié par la présence de distracteurs, chacun étant susceptible

---

interprétation un exemple de cette tendance à vouloir systématiquement « icônographier» les représentations dans notre cerveau. Une explication bien plus simple au phénomène consiste à dire que l'attention est dirigée vers les objets parce que ceux-ci présentent généralement des forts contrastes locaux dans les images, produisant ainsi plus de décharges neuronales. Quand il n'y a pas de contrastes locaux, il n'y a pas ou peu de décharges, donc pas de traitement.
A noter également que la notion de fichier objet développée par Wolfe & Bennett (1997) est différente de celle de Treisman (1998b) pour qui un fichier est une représentation épisodique et dynamique obtenue *après* le déploiement de l'attention.

d'être catégorisé comme une cible. Cette proposition est à mettre en relation avec le fait que dans les modèles parallèles, les principaux facteurs influençant la performance sont notamment la similarité entre cibles et distracteurs ainsi que l'hétérogénéité des distracteurs (Duncan & Humphreys, 1989). Ainsi, le niveau de bruit engendré par l'activation des distracteurs dans une tâche de recherche d'une conjonction d'éléments serait responsable de la baisse de performance avec le nombre de distracteurs, tout comme dans les modèles hybrides, mais sans pour cela faire appel à un stade de traitement sériel. Cependant, dans des tâches de recherche visuelle généralement considérées comme mettant en jeu des mécanismes sériels, il a aussi été démontré qu'un modèle parallèle à *ressources illimitées* parvenait très bien à expliquer les performances alors qu'un modèle sériel en était incapable (Palmer, 1998), comme nous l'expliquerons plus loin.

La récente remise en cause de l'existence des conjonctions illusoires constitue un autre élément à porter aux crédits des modèles parallèles. En effet, celles-ci sont très souvent considérés comme synonyme d'un codage indépendant de certaines propriétés visuelles simples nécessitant d'être assemblées par un mécanisme attentionnel, formant ainsi une des clés de voûte des systèmes sériels et hybrides. De récents résultats suggèrent cependant que les conjonctions illusoires seraient dues à des confusions entre cibles et non-cibles sans erreurs de liage perceptif proprement dit, elles pourraient donc constituer une illusion expérimentale (Donk 1999, 2001 ; mais voir Prinzmetal et al., 2001).

Des arguments plus directs existent en faveur des modèles parallèles. Notamment, toute une littérature montre que l'attention se déploie d'objets en objets et non pas d'une position spatiale à une autre (Duncan et al., 1997). De plus, il est possible d'extraire les propriétés visuelles de deux objets en même temps sans coût additionnel par rapport au traitement d'un objet isolé (Davis et al., 2000). Ceci étant valable dans la mesure où les deux objets ont la même taille que l'objet isolé, il semble que le traitement en parallèle des éléments d'une scène visuelle soit tout de même contraint par des limitations spatiales, renforçant l'idée que les modèles parallèles pertinents sont à ressources limitées.

Malheureusement, malgré un ensemble grandissant d'indices en faveur des modèles parallèles, il est toujours impossible de trancher en faveur d'un type de modèle particulier. Il n'y a pas un mais une multitude de modèles possibles permettant d'expliquer les résultats des études portant sur la recherche visuelle. Malgré la richesse des études réalisées dans ce domaine (le 02 septembre 2003, j'ai trouvé 1184 résultats sous Medline à la requête « visual search »), d'autres outils sont donc nécessaires pour comprendre l'architecture fonctionnelle du système visuel. Différents paradigmes expérimentaux ont été développés en parallèle du paradigme de recherche visuelle. Dans la section qui suit, les apports les plus intéressants seront discutés.

1.5 Sériel vs. parallèle : apports d'autres paradigmes expérimentaux

Parmi les nombreux paradigmes expérimentaux utilisés pour essayer de comprendre le fonctionnement de notre système visuel, certains ont apporté des éléments très intéressants, source de nouveaux axes de recherche et également de nouvelles controverses.

1.5.1 Présentations rapides de photographies de scènes naturelles

Si le paradigme de recherche visuelle capture certaines propriétés essentielles des scènes naturelles comme la présence de plusieurs objets ayant une certaine organisation spatiale, les stimuli utilisés sont généralement très artificiels (voir cependant Wolfe, 1994, Wolfe et al., 2000, 2002, pour de notables exceptions). Quelles sont les performances des sujets lorsqu'ils doivent catégoriser des stimuli plus « écologiques », comme par exemple des photographies de scènes naturelles ?

Le paradigme RSVP. Lorsque des photographies de scènes naturelles sont présentées selon un paradigme RSVP (« rapid serial visual presentation » ou « présentation visuelle sérielle rapide») les images sont présentées très rapidement les unes après les autres (Figure 5). Les facteurs critiques sont la durée de présentation de chaque image et le délai entre les images (ISI, « inter stimulus interval » ou intervalle inter-stimulus). L'impression subjective d'une telle stimulation est celle d'un train d'images sans rapports les unes avec les autres. Des sujets humains testés dans de telles conditions n'ont aucune

difficulté à trouver une image cible placée n'importe où dans la séquence d'images avec des taux de présentation allant jusqu'à 10 images par seconde et parfois plus (Intraub, 1981 ; Potter, 1975, 1976). Ces très bonnes performances sont atteintes même lorsque la cible n'est désignée que par son nom (« bateau » par exemple). Biederman et ses collaborateurs ont également montré que des dessins de scènes naturelles peuvent être interprétés à partir de présentations très brèves de quelques dizaines de millisecondes (Biederman, 1972, 1981 ; Biederman et al., 1973, 1974). Ces expériences montrent que l'information sémantique relative à un stimulus est activée et sélectionnée très rapidement. Cependant, cette information de haut niveau est également très volatile, pouvant être aussi vite oubliée qu'elle a été acquise. En effet, si une image peut être comprise en « un clin d'œil », il faut plus de temps pour la mémoriser lorsqu'elle est présentée dans un train d'autres images (environ une seconde selon Potter & Levy, 1969 ; plutôt 400 ms selon Potter, 1976). Par contre, la présentation d'une image cible suivie d'un masque visuel bloquant la persistance de l'image n'altère pas la capacité à la catégoriser (Bacon-Macé et al., soumis ; Thorpe et al., 2002) ou à la mémoriser (Potter, 1976). Il semble donc que les images qui suivent l'image cible agissent comme des « masques conceptuels », bloquant la formation d'une représentation sémantique à long terme de l'image cible (Intraub, 1984 ; Potter, 1976).

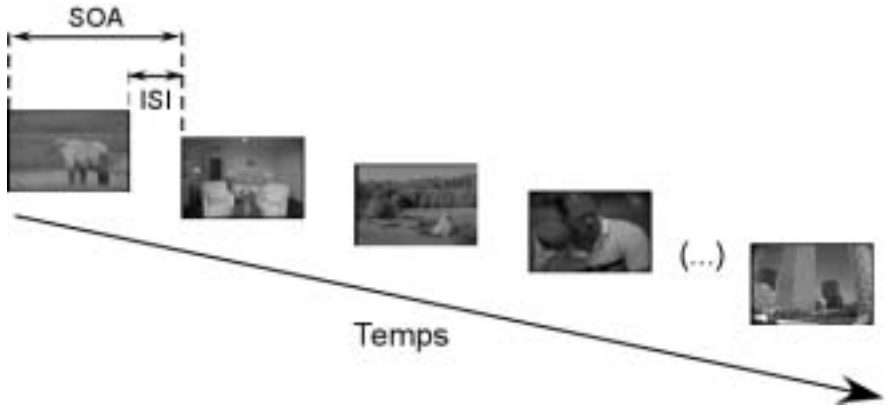


Figure 5. Illustration du paradigme RSVP.

La mémoire conceptuelle à court terme. Ces résultats ont amené Potter (1999, voir aussi Chun & Potter, 1995) a formuler l'hypothèse de la « conceptual short term memory » (CSTM ou « mémoire conceptuelle à court terme »). Ce modèle propose que lorsqu'un

stimulus est identifié, sa signification est très rapidement activée et maintenue très brièvement en CSTM. Si l'information en CSTM n'est pas structurée, mise en relation avec d'autres informations en mémoire à long terme, elle est immédiatement oubliée et ne donne pas lieu à une perception consciente. Ainsi, la plupart des représentations de haut niveau activées dans le système visuel seraient inconscientes, seul un sous ensemble correctement structuré donnerait lieu à une perception consciente. La CSTM est un stade représentationnel qui précèderait la classique mémoire à court terme, cette dernière correspondant à un stade d'intégration permettant la perception consciente.

Le modèle de Potter introduit une distinction nouvelle et fondamentale par rapport aux modèles précédents, qu'ils soient sériels, parallèles ou hybrides : l'incapacité ou la difficulté des sujets à rapporter la présence d'une cible ne reflèterait pas nécessairement une limite attentionnelle au niveau du traitement visuel, mais plutôt à un niveau post catégoriel, avant la prise de décision. Ce modèle implique que pour répondre à un stimulus, il faut en avoir une perception consciente (une conclusion remise en cause par d'autres, e.g. Thorpe et al., 2001a). De plus, le mécanisme de compréhension du sens d'une image est extrêmement rapide, pouvant fonctionner même avec 10 images par seconde, ce qui implique qu'une image pourrait être traitée en 100 ms. Cette durée est très courte mais pourrait cependant être compatible avec un modèle sériel dans lequel les informations sont traitées en « pipeline » - voir plus loin.

Paradigme go/no-go : catégorisation animal/non-animal. Cependant, il faut être prudent avec l'interprétation des données issues du paradigme RSVP. En effet, 10 images par seconde constituent un taux de traitement, pas un temps effectif. Par exemple, il est tout à fait possible que plusieurs images soient traitées en même temps, chacune à un niveau différent de complexité. Pour évaluer plus directement le temps nécessaire pour comprendre une scène naturelle, Thorpe et al., 1996 ont utilisé un paradigme go/no-go de catégorisation rapide. Dans cette étude, les sujets devaient détecter des animaux dans des photographies de scènes naturelles flashées pendant seulement 20 ms. Une telle tâche de catégorisation était réalisée à la fois très précisément (94% de bonnes réponses) et rapidement (TR médian de 445 ms). Ce résultat a été répliqué plusieurs fois et il a aussi été montré que non seulement les sujets sont rapides en moyenne mais aussi que leurs

réponses présentent un biais vers les bonnes réponses apparaissant dès environ 300 ms (Fabre-Thorpe et al., 2001 ; Van Rullen & Thorpe, 2001a). Cette grande précision et cette rapidité sont obtenues alors même que les images ne sont vues qu'une seule fois et qu'il est impossible de prédire par avance le nombre d'animaux qui apparaîtront, le type d'animal dont il s'agira, ni sa position dans l'image. Etant donnée la complexité de la tâche et le parcours de l'information dans le système visuel, le fait que les sujets soient capables de relâcher un bouton de réponse en moins de 400 ms laisse vraiment peu de temps à des mécanismes attentionnels lents pour se mettre en place. Il a été suggéré qu'un tel traitement mettrait en jeu essentiellement une propagation unidirectionnelle et en parallèle de l'information allant de la rétine au système moteur (Thorpe & Imbert, 1989 ; Thorpe & Fabre-Thorpe, 2001 ; Thorpe et al., 1996). Cette conclusion est renforcée par la découverte que dans cette tâche particulièrement exigeante, le système semble fonctionner d'emblée de façon optimale avec les stimuli nouveaux puisque les sujets ne sont pas capables de traiter plus rapidement des images avec lesquelles ils se sont longuement familiarisés (Fabre-Thorpe et al., 2001). De plus, ce mécanisme de catégorisation rapide n'implique pas nécessairement la vision fovéale puisqu'il est possible de réaliser très efficacement cette tâche en vision périphérique (Fabre-Thorpe et al., 1998) avec un coût en précision proportionnel à la baisse d'échantillonnage rétinien (Thorpe et al., 2001a). Ces résultats, associés au fait que les scènes contiennent typiquement plusieurs objets, laissent supposer que le système visuel pourrait fonctionner vite et en parallèle, traitant non seulement plusieurs objets à la fois, mais pourquoi pas plusieurs scènes naturelles en même temps ! Cette hypothèse a fait l'objet de deux expériences décrites dans les deux premiers articles constituant cette thèse.

La catégorisation rapide d'un animal s'effectue t'elle en parallèle ? Si des sujets humains se révèlent parfaitement capables de catégoriser des scènes naturelles en périphérie sur la base de la présence d'un animal alors que leur attention est simultanément occupée à résoudre une tâche centrale (protocole de double tâche, Li et al., 2002) (Figure 6), cette même tâche s'avère nécessiter un temps de traitement qui croît avec le nombre de scènes distracteurs dans un protocole classique de recherche visuelle (VanRullen et al., 2003). De manière surprenante, des stimuli simples qui apparaissent

être traités en parallèle (‘pop-out’) selon le protocole de recherche visuelle (e.g., un ‘L’ parmi des ‘+’) ne sont pratiquement pas discriminables les uns des autres dans un protocole de double tâche.

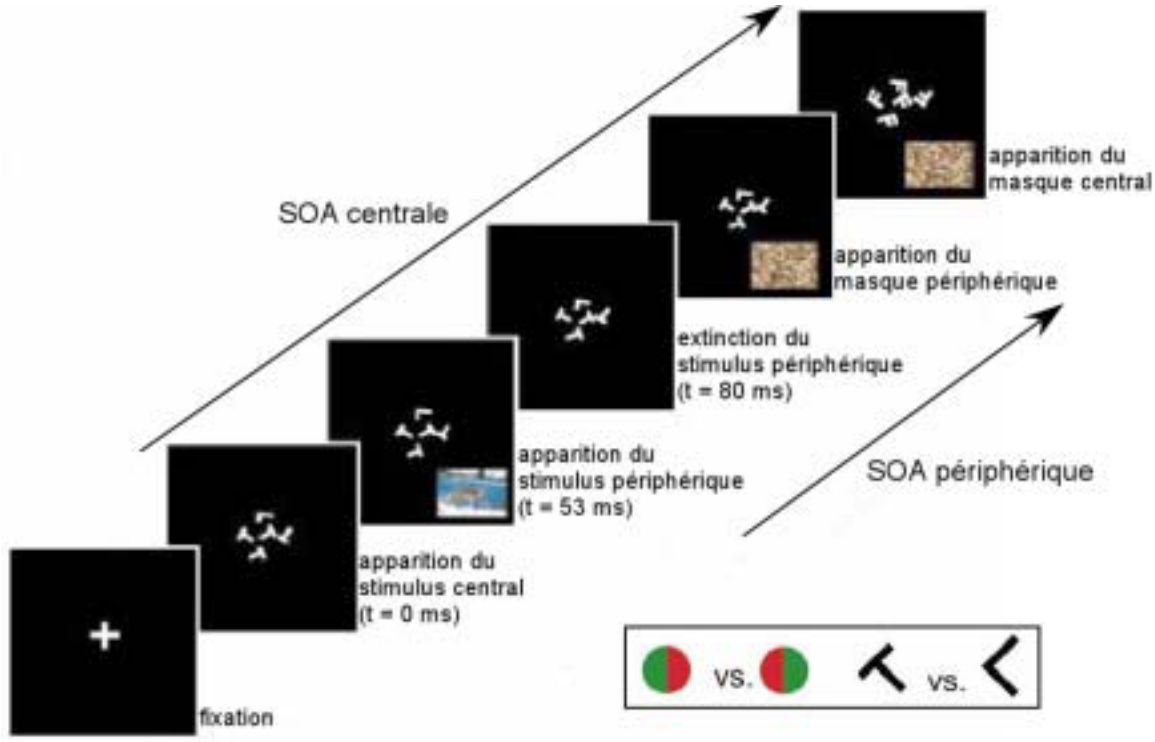


Figure 6. Protocole expérimental utilisé par Li et al. (2002). Au cours d’un essai, après l’apparition d’un point de fixation, les sujets réalisent une tâche difficile centralement (les lettres sont-elles toutes les mêmes ?). En périphérie, un stimulus est flashé pendant qu’ils réalisent la tâche centrale : soit une photographie de scène naturelle, soit un stimulus simple (encadré). Dans le premier cas, une tâche consiste à répondre le plus rapidement possible quand l’image contient un animal, l’autre quand elle contient un moyen de transport (distracteur = autres scènes naturelles). Dans le second cas une tâche consiste à répondre quand un « L » est présenté (distracteur = « T »), l’autre quand un disque avec du rouge sur la droite est présenté (distracteur = disque avec du rouge sur la gauche). Les SOA étaient différentes pour chaque sujet de telle sorte que leurs performances pour les tâches périphériques réalisées seules étaient équivalentes pour les scènes naturelles et pour les stimuli simples. Cette procédure permet d’évaluer le coût attentionnel sur la tâche périphérique engendré par l’ajout de la tâche centrale. Alors que les scènes naturelles peuvent être catégorisées tout en réalisant la tâche centrale, cela s’avère impossible pour les formes « simples ». Tiré de Li et al. (2002).

VanRullen et al. (2003), synthétisant les apports de différents travaux antérieurs, ont expliqué ces phénomènes en proposant que les performances obtenues dans les paradigmes de recherche visuelle et de double tâche ne nous renseignent pas sur les mêmes mécanismes visuels. Dans le paradigme de recherche visuelle, une recherche ‘parallèle’ (par opposition à une recherche ‘sérielle’) refléterait soit (1) une absence

d'interaction inhibitrice entre les représentations de la cible et des distracteurs, soit (2) la mise en jeu de mécanismes de structurations des distracteurs en une sorte de texture, la cible apparaissant alors comme un élément incongru, sans pour autant nécessiter une véritable discrimination. En revanche, dans le paradigme de double tâche, une recherche 'pré attentive' (par opposition à une recherche 'attentive') montrerait qu'il existe dans le système visuel des détecteurs des éléments cibles (par exemple des neurones répondant sélectivement à des animaux). Une recherche 'pré attentive' ne serait donc pas synonyme de 'parallèle'. Selon cette perspective très intéressante, l'attention serait impliquée dans deux mécanismes majeurs consistant (1) à générer ou affiner des représentations de haut niveau quand celles-ci ne sont pas disponibles, comme le proposent FIT et GSM, (2) à résoudre la compétition entre stimuli à différents niveaux au sein du système visuel. Ce second point met en avant l'existence de représentations de haut niveau des stimuli en l'absence d'attention (il sera approfondi plus loin dans ce chapitre). Il apparaît donc que le monde visuel pré attentif pourrait être beaucoup plus riche qu'une simple collection de paquets d'éléments simples sans structuration spatiale (Wolfe & Bennett, 1997). Ainsi, contrairement aux hypothèses des modèles sériels et des modèles hybrides, l'attention focalisée ne serait pas indispensable pour former des représentations complexes.

1.5.2 Repetition Blindness

S'il parait possible d'extraire très rapidement le sens d'une scène visuelle, probablement sur la base d'informations pré-attentives, il existe tout de même des limites à ce type de traitement sémantique et à la capacité à rapporter les représentations ainsi formées. Le phénomène appelé « repetition blindness » (RB, ou « répétition aveuglante ») fournit un exemple de ces limites (Kanwisher et al., 1999).

Paradigme. Dans un paradigme de RB les sujets voient une séquence rapide de 3 images précédée et suivie d'une séquence de 3 masques perceptuels abstraits. Leur tâche consiste à décrire verbalement les 3 images à la fin de la séquence. Dans un essai contrôle, les 3 images sont différentes et les sujets n'ont aucun problème à décrire les 3 images. Dans un essai critique, les images 1 et 3 sont les mêmes. Dans ce cas les sujets détectent et rapportent moins facilement l'occurrence de la troisième image (Figure 7). Après l'identification correcte d'une image, il existerait donc une sorte de période réfractaire

pendant laquelle les représentations mises en jeu par l'apparition de la première occurrence d'une image ne sont plus disponibles pour individualiser la seconde occurrence. C'est un nouveau cas de compétition dans le temps pendant une séquence RSVP.

Interprétation. Kanwisher et al. (1999) ont examiné l'effet de variations physiques et sémantiques dans les relations entre les deux images répétées. Leur raisonnement était que si l'effet RB était présent et de même amplitude malgré ces variations, il serait possible d'inférer que les représentations mises en jeu dans le phénomène sont invariantes à ces transformations.

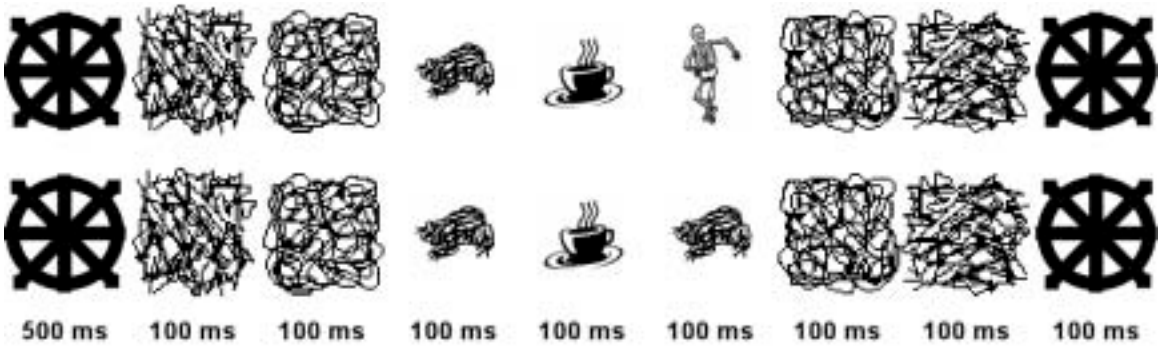


Figure 7. Illustration des deux séquences types utilisées dans un paradigme RB. Chaque séquence est composée de 9 images dont la durée de présentation est indiquée en millisecondes. Dans la première séquence trois images différentes d'objets sont présentées entre deux séquences de masques visuels. Les sujets sont généralement capables de rapporter la présence des trois objets dans la séquence. La seconde séquence contient une répétition d'image, dans ce cas les sujets rapportent souvent avoir vu une grenouille et une tasse de café mais pas deux grenouilles. Tiré de Kanwisher et al. (1999).

Il apparaît qu'un effet RB est présent pour deux images variant en taille, en position, en orientation dans le plan de l'image et en point de vue. L'effet était cependant moins fort pour des exemplaires différents de la même catégorie comme un piano à queue et un piano droit et encore plus réduit pour des images présentant seulement une relation sémantique catégorielle comme un hélicoptère et un avion.

L'analyse de nombreux résultats comportementaux a conduit Kanwisher et ses collaboratrices à conclure que les représentations mises en jeu lors de présentations rapides d'images pourraient être relativement abstraites. De plus, elles ont suggéré que la sensibilité à la répétition manifestée dans un paradigme RB impliquerait l'extraction de l'identité de l'image répétée en l'absence de perception consciente.

Ces expériences montrent indéniablement que les « ressources » de traitement visuel sont limitées. Cependant, le phénomène pourrait également s'expliquer par un effet d'amorçage des propriétés bas niveau de l'image lors de sa première présentation, conduisant à une sorte de « fatigue neuronale » diminuant la saillance perceptive de l'image répétée.

1.5.3 Attentional Blink

Une autre illustration des limites de notre système visuel à former des représentations explicites, durables, lors de présentations rapides d'images est illustrée par notre difficulté à détecter et rapporter la seconde cible au sein d'une séquence RSVP lorsque celle-ci apparaît au cours des 500 ms qui suivent la détection d'une première cible (Duncan et al., 1994). Cette difficulté a été nommée « attentional blink » (AB, ou « clignement attentionnel »). Contrairement au RB, l'effet AB a lieu alors que la seconde cible n'est pas identique à la première.

Paradigme. Dans une expérience typique d'AB, les sujets observent une séquence de lettres noires. Dans la condition de double cible, ils doivent identifier la lettre blanche (C1) qui apparaît dans la séquence (Figure 8). Dans la condition avec cible unique, les sujets doivent ignorer C1. Après la présentation de C1, huit lettres noires apparaissent, et dans les conditions simple et double cible les participants doivent rapporter si la lettre X (C2) était présente. C2 peut apparaître à n'importe laquelle des huit positions qui suivent C1, dans 50% des essais. La condition double cible est supposée mettre en évidence l'effet de l'identification correcte de C1 sur la capacité des sujets à réallouer leur attention sur C2 après le traitement de C1. Dans la condition à une cible, les sujets rapportent l'identité de C2 dans 90% des essais, quelle que soit sa position dans la série. Dans la condition à deux cibles, les sujets rapportent C2 dans 50% à 80% des cas selon qu'elle apparaît entre 100 et 400 ms après C1.

Interprétation. L'AB laisse supposer qu'une certaine ressource, probablement attentionnelle, était disponible pour la première cible mais pas pour la seconde (Shapiro et al., 1997a,b ; Shapiro & Terry, 1998 ; Shapiro & Luck, 1999). Cette incapacité à accomplir un traitement complet des stimuli proviendrait d'une interférence entre les deux cibles et entre chacune des cibles et les stimuli qui les suivent et les précèdent,

agissant comme autant de masques visuels ou conceptuels (selon les modèles). Les cibles entreraient dans une mémoire à court terme visuelle (MCTV) à partir de laquelle elles seraient ensuite sélectionnées. La MCTV étant surchargée d'informations, la sélection deviendrait beaucoup plus difficile. Cet effet est d'autant plus important que les items entre C1 et C2 appartiennent à la même catégorie que C2.

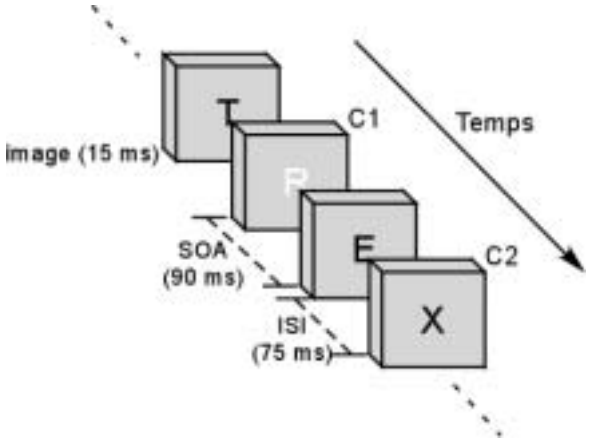


Figure 8. Illustration d'une séquence RSVP utilisée dans une expérience d'AB. Tiré de Shapiro et al. (1997b).

Une expérience récente a permis de mieux cerner le type de limitations mises en évidence par le paradigme AB. Lorsqu'un stimulus du type employé dans une tache de recherche visuelle est présenté pendant le 'blink' ('clignement attentionnel'), les sujets sont incapables de détecter une cible de type 'pop out' qui devrait être traitée en parallèle, sans attention (Joseph et al., 1997). Pendant le blink, toutes les ressources attentionnelles seraient consommées, empêchant la perception consciente du stimulus. Ainsi, comme le suggéraient déjà les travaux sur le RB, l'attention serait indispensable pour rapporter de manière explicite la présence d'un stimulus, mais peut être pas pour le traiter au niveau implicite. Ce point a d'ailleurs été souligné par Treisman elle-même : « Preattentive processing cannot directly affect responses or experience ; it is an inferred stage of early vision, which I attribute to the separate feature modules. Before any conscious visual experience is possible, some form of attention is required, since information from the different feature maps must be combined. We can never be aware of a free-floating orientation, colour, or shape » (Treisman, 1998b, p.33). En accord avec cette théorie, une

série d'expériences a permis de montrer que bien qu'elle ne soit pas rapportée par les sujets, la seconde cible (un mot) d'un protocole RB est tout de même traitée jusqu'au niveau sémantique (Luck et al., 1996 ; Maki et al., 1997 ; Shapiro et al., 1997a). Il semble donc que l'interférence entre les traitements des deux cibles ait plutôt lieu à un niveau décisionnel post-perceptuel, en accord avec le modèle en deux étapes de traitement de Chun & Potter (1995 ; Potter, 1999).

L'AB nous montre aussi que l'attention ne semble pas pouvoir être distribuée très rapidement au sein d'une scène mais qu'elle se focalise sur un objet pendant un temps relativement long. Ceci constitue indéniablement un point en faveur des modèles qui proposent une sélection tardive des stimuli. Par exemple, Duncan et al. (1994) ont mis en avant le fait que les modèles sériels dérivés des expériences de recherche visuelle devraient être réinterprétés sur la base des résultats obtenus avec le paradigme AB. Etant donné que l'attention semble être allouée de manière exclusive à un seul item pendant environ 400 ms, des pentes de recherche de par exemple 50 ms par item obtenues dans de nombreuses expériences seraient dues à un mécanisme parallèle à ressources limitées. Dans une expérience de recherche visuelle classique, les distracteurs pourraient agir à la manière de masques visuels ou conceptuels, comme dans une séquence RSVP. Ainsi, le même phénomène pourrait être à l'œuvre dans les tâches d'attention spatiale et d'attention temporelle, à savoir une étape contraignante et compétitive de sélection de représentations de haut niveau. Cette interprétation est cependant remise en cause par certains, les *temps* de traitement dérivés des pentes de recherche pouvant être interprétés comme des *taux* de traitement (Treisman, 1998a ; Wolfe et al., 2000). Selon cette perspective, un nouvel item pourrait par exemple être pris en charge par des mécanismes attentionnels toutes les 50 ms, le temps effectif de traitement étant cependant de 300-400 ms (Wolfe et al., 2000).

1.5.4 Inattentional Blindness

Mack & Rock (1998) ont mis au point un astucieux paradigme qui leur a permis de mener une longue série d'investigations décrites dans leur livre *Inattentional Blindness* (IB ou cécité d'inattention. Dans l'optique de la présente discussion, ces expériences apportent un argument de plus en faveur des modèles de sélection tardive.

Paradigme. Les sujets voient une croix (un '+') présentée brièvement au centre de l'écran et doivent évaluer laquelle de la composante horizontale ou verticale de la croix est la plus longue (Figure 9). Pendant la stimulation, un stimulus non pertinent pour la tache est flashé dans un des quatre quadrants formés par la croix. Après l'essai, l'expérimentateur demande aux sujets s'ils ont vu quelque chose d'anormal. Les sujets sont très souvent incapables de rapporter la présence de stimuli qu'ils détectent pourtant très facilement quand ils savent par avance que de tels stimuli peuvent apparaître. C'est pourquoi il n'y a qu'un seul essai critique par sujet dans un paradigme IB. De manière très frappante, les sujets ne voient pas le stimulus critique même lorsque celui-ci est présenté au niveau du point de fixation, centralement, la croix à analyser apparaissant en périphérie.

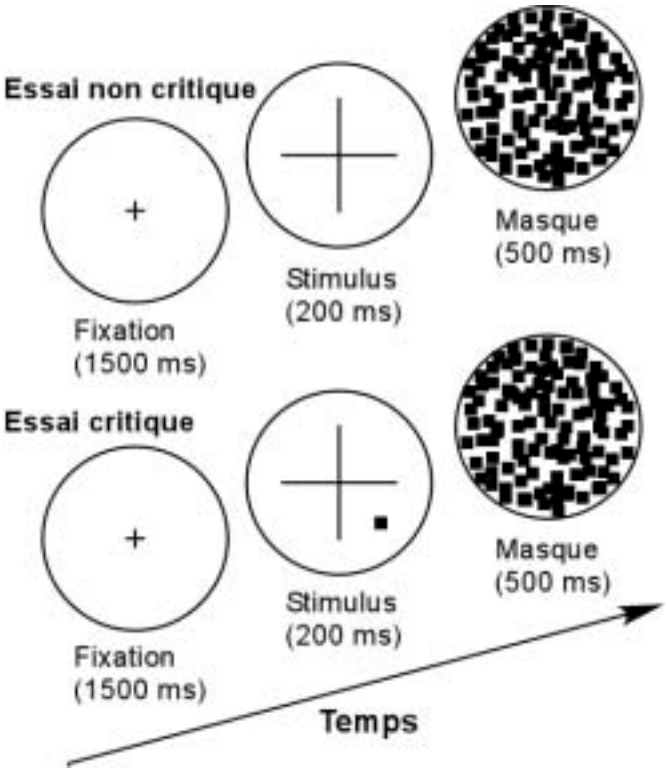


Figure 9. Illustration du paradigme d'inattentional blindness. Tiré de Mack & Rock (1998).

Interprétation L'hypothèse avancée par Mack & Rock pour expliquer le phénomène d'IB est que nous ne percevons pas consciemment les objets sur lesquels nous n'avons pas porté notre attention. Ainsi, tout comme l'ont montré Joseph et al. (1997) dans le cadre du paradigme AB, il n'y pas ici de perception de formes simples quand l'attention est

concentrée sur une autre tâche. En l'absence d'attention, le stimulus n'atteint jamais le niveau de la perception consciente. Mais encore une fois, cela ne veut pas dire que les stimuli ne sont pas traités du tout, qu'ils restent dans un état pré-attentif non structuré (Wolfe & Bennett, 1997). Certaines conditions expérimentales ont en effet révélé une reconnaissance des objets flashés. Par exemple, si le prénom de la personne testée est flashé, ou un visage schématique souriant ou encore une photographie de scène naturelle, le taux de non détection chute considérablement. Par contre, les lettres du prénom réarrangées, un visage triste, ainsi que divers stimuli contrôles sont beaucoup moins facilement détectés. Ceci démontre qu'une analyse sémantique des stimuli est réalisée, même en l'absence d'attention focalisée ; l'attention serait par contre nécessaire pour pouvoir rapporter explicitement un événement (Mack & Rock, 1998). Cette dernière idée est parfaitement illustrée dans une expérience étendant les travaux de Mack & Rock à une situation très écologique : pendant une séquence vidéo ininterrompue, les sujets comptent les échanges de ballons entre des joueurs et un gorille traverse la scène, totalement inaperçu par de nombreux sujets (Simons & Chabris, 1999).

1.5.5 Change Blindness

Si les paradigmes passés en revue jusqu'à présent tendent à montrer que l'analyse visuelle va, de manière pré-attentive, beaucoup plus loin qu'une simple description physique de bas niveau, le paradigme de « change blindness » (CB, ou « cécité au changement »), utilisant des photographies de scènes naturelles, semble montrer que cette conclusion ne s'applique peut être pas à des conditions réelles de perception.

Paradigme. Notre capacité à détecter un changement dans une image est remarquablement faible. Ce résultat provient d'un ensemble d'études dans lesquelles un changement parfois très important intervient dans une image après une brève interruption de la scène sans que les sujets soient capables de rapporter ce changement, ou seulement après une longue série d'alternances entre la scène d'origine et la scène modifiée (d'où le terme « change blindness », pour une revue voir Henderson & Hollingworth, 1999 ; Rensink, 2002 ; Simons & Levin, 1997). Les changements apportés aux images peuvent être l'addition, la soustraction, la substitution, le déplacement ou encore la modification d'un objet de la scène. Ce changement peut intervenir pendant le mouvement des yeux

(Grimes, 1996), pendant le déplacement d'une image à l'écran (Blackmore et al., 1995), ou encore après la présentation d'un écran gris entre les deux versions successives de l'image (Rensink et al., 1997) (Figure 10). Ces interruptions sont nécessaires car un changement sans une telle interruption est immédiatement détecté à cause de l'impression de mouvement qu'il crée. Ce résultat reste valable même lorsque la scène est en permanence présente à l'écran mais que des 'taches' à fort contraste ('mud splashes') apparaissent brièvement superposées à la scène au moment où le changement apparaît (O'Regan et al., 1999). De manière encore plus surprenante, l'incapacité à rapporter un changement existe aussi lors de la perception d'une séquence vidéo lorsque le changement intervient pendant un mouvement de caméra (Simons, 1997). Ce phénomène s'étend même à une interaction réelle entre deux personnes. Dans cette expérience, une personne (sujet d'une expérience sans le savoir) était interpellée dans un campus américain par une autre personne lui demandant son chemin (un expérimentateur). Pendant la discussion, deux ouvriers transportant une porte interrompaient la conversation, et une personne cachée derrière la porte prenait la place de la personne qui demandait son chemin. Cela peut paraître incroyable, mais dans 50% des cas la personne interrogée ne se rendait pas compte qu'elle ne parlait plus à la même personne (Simons & Levin, 1997).

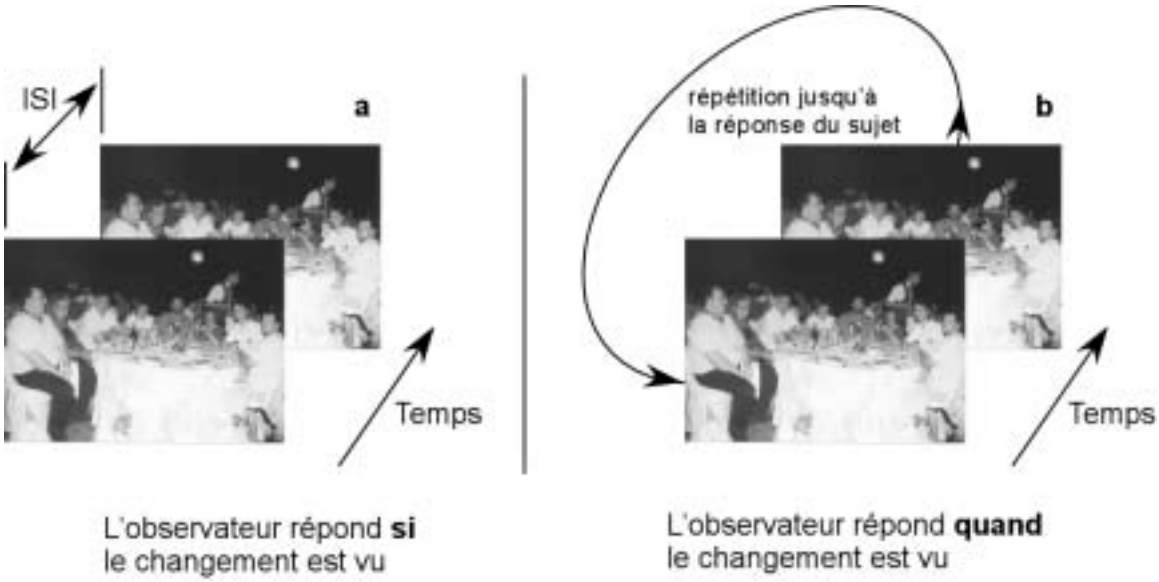


Figure 10. Illustration de deux paradigmes de change blindness très utilisés. Adapté de Rensink (2000b).

Première interprétation. Les résultats apportés par les expériences de CB prolongent les premières découvertes montrant que la mémoire trans-saccadique est faible voire inexistante, le but de la saccade semblant seul être mémorisé (pour une revue voir Henderson & Hollingworth, 1999). En effet, la logique utilisée pour interpréter les expériences de CB est généralement la suivante : si nous construisons une riche représentation au fur et à mesure que nous explorons une scène visuelle, alors un changement apporté à cette scène devrait être détecté ; or nous sommes très mauvais pour détecter des changements, cela veut donc dire que nous ne formons pas de représentations précises de notre environnement visuel. La plupart des interprétations vont même un peu plus loin en postulant que nous percevons en réalité beaucoup moins de choses dans notre environnement visuel que nous ne le supposons subjectivement. Il est tout a fait envisageable que malgré notre impression de percevoir un monde visuel riche, nous n'ayons pas besoin de reconstruire une riche représentation de cette réalité. Selon cette perspective, le monde extérieur servirait de mémoire externe, de telle sorte que chaque fois que nous voulons voir, il suffirait de diriger notre attention vers la partie de l'environnement qui nous intéresse, nous donnant l'impression d'avoir un accès conscient à l'ensemble du champ visuel (O'Regan, 1992 ; Rensink, 2000a,b). Ainsi, il n'y aurait pas de perception consciente sans attention, comme beaucoup d'auteurs l'ont aussi conclu (O'Regan & Noë, 2001). En accord par exemple avec les explications apportées par Mack & Rock au phénomène d'IB, les changements ne seraient pas vus parce que les stimuli qui ont changé n'étaient pas cibles de l'attention, et ainsi non consciemment perçus. Mais l'interprétation du CB va plus loin. Entre deux images ou entre deux saccades, seuls l'organisation spatiale grossière de la scène et son sens général seraient retenus, permettant le déploiement efficace de l'attention et donnant l'impression d'une continuité dans la perception visuelle d'une scène (la théorie la plus détaillée à ce sujet est décrite dans Rensink, 2000a,b, 2002). En ce sens, la littérature concernant le CB est en parfait accord avec les modèles sériels de la perception visuelle : seul ce qui se trouve dans la fenêtre attentionnelle est structuré, reconnu, perçu consciemment et disponible de manière très volatile en mémoire à court terme, permettant la comparaison de deux représentations d'un objet séparées par une interruption. En dehors de cette fenêtre

attentionnelle aucune information structurée n'est disponible et il n'y a pas de perception consciente. Ceci expliquerait pourquoi, si l'attention n'est pas dirigée sur un objet de manière continue, un changement apporté à cet objet n'est pas détecté, puisque sa représentation cesserait tout simplement d'exister en dehors de la fenêtre attentionnelle.
Cependant, il faut noter que les stimuli sont présentés pendant relativement longtemps dans les expériences de Change Blindness (CB), et il ne fait aucun doute que l'interlocuteur dans l'expérience de Simons & Levin (1997) était perçu consciemment par le sujet. De plus, l'explication de CB semble en totale contradiction avec notre expérience subjective, malgré les très séduisants modèles développés par O'Regan & Noë et par Rensink. Nous avons la sensation de percevoir quelque chose dans l'ensemble de notre champ visuel. Mais le problème de la perception sans attention est très difficile à aborder dans la mesure où si l'on demande à quelqu'un de rapporter la présence d'un stimulus, il portera obligatoirement son attention sur lui. Malgré cette limitation intrinsèque, une proposition récente a remis en cause l'idée selon laquelle le paradigme CB montrerait qu'en l'absence d'attention il n'y a pas de perception consciente.

La vision post-attentive. A ce stade de l'exposé, il est bon de rappeler un point essentiel : l'incapacité à rapporter un événement n'est pas synonyme d'une incapacité à voir. Ceci fut démontré en 1960 par Sperling, qui mit en évidence l'existence de mémoires sensorielles, ou iconiques, qui correspondent à un stockage très bref d'information. Dans la tâche utilisée par Sperling (Figure 11), le sujet voit un ensemble de lettres pendant 50 ms. Après un délai variable, le sujet doit restituer les lettres présentées. En moyenne, un sujet rapporte correctement 3 ou 4 lettres. Dans une variante de ce protocole, la fréquence d'un son émis (aigu, grave ou intermédiaire) indique au sujet qu'il doit restituer les lettres du haut, du bas ou de la rangée intermédiaire, respectivement. Cette fois-ci les sujets sont capables de rapporter 9 à 10 lettres sur 12. En revanche, si le son est émis plus d'une seconde après la présentation des lettres la performance est à nouveau de 4 lettres. La capacité de la mémoire iconique serait donc relativement élevée mais de courte durée (environ 1 seconde). Cette découverte pourrait laisser penser que les déficits de détection de changement dans le paradigme de CB seraient dus à des limitations mnésiques et non pas à des problèmes de perception visuelle.

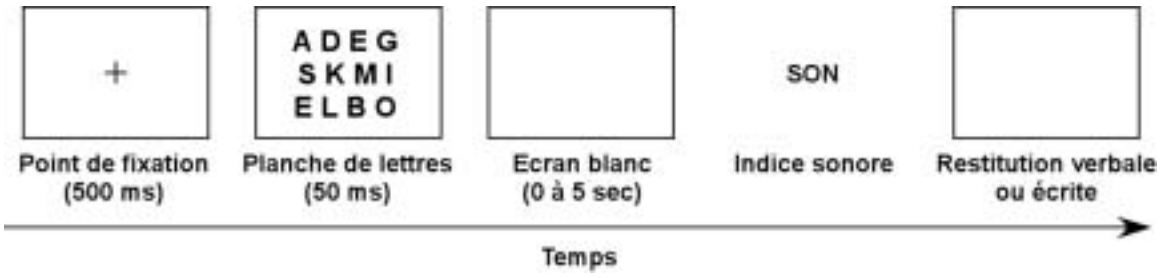


Figure 11. Illustration du protocole ayant permis à Sperling de mettre en évidence l'existence d'une mémoire iconique. Adapté de Lemaire (1999).

Cette hypothèse a été explorée par Wolfe et ses collaborateurs (2000, Horowitz & Wolfe, 1998, 2001) qui ont mis au point la technique de recherche visuelle répétée. Au cours d'une tâche de recherche répétée, le stimulus pouvant contenir la cible reste le même pendant plusieurs essais. Par exemple, des lettres sont présentées en cercle autour d'un point de fixation. Au niveau du point de fixation lui-même, une lettre cible est présentée (Figure 12). Les sujets doivent déterminer si la lettre cible est présente parmi les lettres en périphérie. De multiples variantes de ce test ont été utilisées et les résultats montrent tous clairement que la performance des sujets ne s'améliore pas avec les répétitions. Ce résultat reste valable même lorsque les stimuli utilisés sont des dessins très réalistes d'objets de la vie de tous les jours et que les sujets réalisent 100 essais de suite (Figure 13). A l'appui de ces résultats, Wolfe (1999) a fourni une explication alternative au phénomène de CB, l'étendant d'ailleurs aux phénomènes d'IB et d'AB présentés plus haut. Ces phénomènes, interprétés comme résultant de problèmes attentionnels, pourraient en fait s'expliquer en termes de déficits mnésiques. Dans un paradigme de CB, les stimuli seraient vus mais oubliés avant d'être rapportés. Selon cette hypothèse appelée « inattentional amnesia » (« amnésie d'inattention »), (1) nous percevons consciemment des éléments visuels pré-attentifs (qui restent à définir) partout dans notre champ visuel ; (2) à l'endroit où se porte notre attention, l'information visuelle est restructurée, permettant la reconnaissance d'objets et leur mémorisation ; (3) à un instant donné, une représentation visuelle consciente serait composée d'éléments pré-attentifs et de la représentation structurée d'un objet ; (4) cette représentation visuelle serait instantanée et ne resterait pas en mémoire. Selon l'hypothèse de Wolfe, l'attention étant indispensable à la mémorisation, il s'ensuit que les stimuli n'étant pas cibles de l'attention pourraient être vus mais seraient instantanément oubliés.

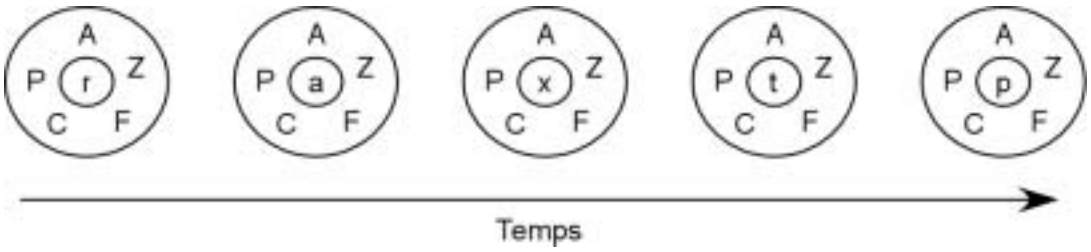


Figure 12. Exemple d'une des tâches de recherche répétée utilisées par Wolfe et al. (2000), pour étudier la vision post-attentive. Adapté de Wolfe et al. (2000).

Figure 13. Exemples de stimuli utilisés dans une expérience de recherche répétée avec des objets réalistes. Les stimuli restaient identiques et à la même place pendant un bloc de 100 essais. A chaque essai, les sujets entendaient le nom de la cible à trouver. Tiré de Wolfe (2003).

Cette proposition est en partie en accord avec la théorie de la mémoire conceptuelle à court terme (Potter, 1999, voir pp.15-16), selon laquelle nous percevons beaucoup d'objets lorsque nous regardons une scène mais qu'une petite partie seulement de cette information est encore disponible lorsque nous regardons une nouvelle scène, par exemple après une saccade. La nouvelle scène remplacerait la scène précédente et seul le sens général de la scène serait mémorisé.

Si un système visuel amnésique parait pouvoir expliquer les résultats de CB, l'extension de ce raisonnement à tous les phénomènes qui ont été passés en revue plus haut parait très difficile. Dans une tâche de CB, on a clairement la sensation de voir toute la scène tout en ressentant une incapacité à se souvenir des détails, alors qu'ils se

volatilisent à chaque interruption. En revanche, dans les autres tâches, la sensation est toute autre ; les paradigmes de RB, AB & IB sont en effet associés à une réelle incapacité à voir. Dans tous les cas, ces quatre paradigmes semblent indiquer, comme l'avaient déjà montré les travaux de Li et collaborateurs (2002), qu'un traitement de haut niveau est possible sans attention focalisée.

Nouvelles interprétations. Avant de conclure, notons que de récents travaux ont poussé encore plus loin la réévaluation des conclusions tirées des expériences de CB en remettant en cause l'absence d'une représentation riche du monde. Ces travaux montrent notamment que le taux de non détection d'un changement augmente significativement lorsque les yeux n'étaient pas dirigés vers l'objet au moment du changement ou avant celui-ci (Hollingworth et al., 2001b). Une partie du phénomène de CB pourrait donc s'expliquer par le fait qu'une représentation détaillée de l'objet cible avant modification n'a en réalité jamais été formée. Lorsque les fixations sont contrôlées, par exemple quand le changement intervient après que l'objet critique ait été observé, les sujets se révèlent capables de détecter beaucoup plus de changements que ne le prédit un modèle totalement amnésique (Hollingworth et al., 2001a ; Henderson & Hollingworth, 2003). La sémantique de la scène guide aussi l'attention des sujets qui détectent plus facilement les changements apportés aux objets les plus intéressants (Kelley et al., 2003 ; Rensink et al., 1997). La détection du changement est par ailleurs sensible aux relations sémantiques entre l'objet cible et son contexte, un changement intervenant sur un objet incongru étant plus facilement détecté (Hollingworth & Henderson, 2000). De plus, il semble qu'une représentation implicite du changement existe parfois, qui n'apparaît donc pas dans les tâches qui demandent une réponse explicite du sujet (Fernandez-Duque & Thornton, 2000 ; Hollingworth et al., 2001a ; Henderson & Hollingworth, 2003). Cette détection implicite, caractérisée par exemple par une plus longue durée de fixation oculaire sur l'objet modifié sans réponse explicite, peut avoir lieu plusieurs secondes après le changement, impliquant bien l'existence d'une représentation relativement riche de la scène visuelle (Hollingworth et al., 2001a). En accord avec cette hypothèse, la mémoire explicite des objets dans une scène naturelle brièvement flashée a été évaluée à environ trois ou quatre objets, plus l'existence d'une mémoire implicite de deux ou trois autres

objets (VanRullen & Koch , 2003). Enfin, il a été montré que l'effet CB proviendrait au moins en partie de problèmes d'accès en mémoire malgré l'existence d'une représentation de l'objet cible (Hollingworth, 2003), un résultat s'accordant très bien avec l'existence de représentations implicites ayant probablement été inhibées par les représentations d'autres objets en mémoire de travail (VanRullen & Koch, 2003).

L'ensemble de ces travaux semble montrer que les déficits de détection dans les paradigmes de CB ne dérivent pas nécessairement du déploiement sériel de l'attention dans les scènes visuelles, ni de problèmes de mémorisation de représentations de haut niveau. Cependant, même si les performances s'améliorent dans ces nouveaux protocoles, il reste à expliquer pourquoi dans de nombreux cas nous sommes toujours incapables de détecter ces changements. Il est bien possible que notre système visuel soit en grande partie amnésique, comme le suggère Wolfe (1999), et que nous ayons recours, par défaut, à la richesse du monde qui nous entoure plutôt qu'à une hypothétique représentation interne, comme le suggèrent O'Regan (1992) et Rensink (2000a,b). Ceci n'exclut pourtant pas le fait que nous formions effectivement des représentations de haut niveau en l'absence d'attention focalisée et que celles-ci soient mémorisées. Il reste cependant à analyser la nature de ces traces mnésiques qui pourraient bien être beaucoup plus sémantiques que visuelles. Il reste également possible que de nombreuses représentations visuelles puissent être activées très vite en parallèle et au moins partiellement mémorisées. Cependant, ce type de perception rapide n'a probablement qu'un accès limité à la conscience, seule une petite partie des représentations visuelles formées en parallèle pouvant atteindre ce stade explicite. En accord avec cette hypothèse, il semble que nous ayons du mal à prendre des décisions reposant sur autre chose que des représentations explicites. Ce raisonnement s'appuie en partie sur une découverte troublante de Wolfe et al. (2000) qui ont montré que même lorsque des sujets réalisent une tâche de recherche visuelle avec un stimulus identique pendant 350 essais leurs performances ne s'améliorent pas ! Pourtant, lorsqu'ils doivent réaliser la même tâche sur le même matériel mais en l'ayant mémorisé au préalable, leurs performances s'améliorent. Ainsi, quand un stimulus visuel est disponible, nous n'adoptons pas de stratégie mnésique, peut être trop risquée, mais nous préférons systématiquement vérifier l'information puisqu'elle est disponible en permanence. Cette stratégie semble également

s'appliquer dans des conditions réelles d'interaction avec l'environnement, l'acquisition de l'information visuelle étant réalisée en temps réel au fur et à mesure des gestes à accomplir, ce qui constitue une définition plus opérationnelle de la perception visuelle (Hayhoe et al., 2003 ; Triesch et al., 2003). Une grande partie des résultats de la littérature portant sur l'incessant débat sériel/parallèle pourrait ainsi trouver sa source non pas dans des contraintes fonctionnelles purement visuelles, mais plutôt dans des contraintes cognitives de plus haut niveau, probablement indépendantes de la modalité sensorielle.

2. Sériel vs. parallèle : au cœur du cerveau

La littérature comportementale portant sur le degré de parallélisme des traitements visuels est extrêmement vaste. La mise en œuvre de nouveaux paradigmes comportementaux à permis récemment d'enrichir le débat de nouvelles idées. Il semble notamment que l'apparent mode sériel de fonctionnement de la vision pourrait trouver son explication à un niveau d'intégration sensori-moteur plutôt que purement perceptif. Mais un même résultat pouvant recevoir des explications parfois totalement opposées, il est quasiment impossible de trancher en faveur d'une hypothèse sur la seule base des résultats comportementaux. Pour aller plus loin, cette seconde partie a pour objectif de porter le débat au cœur du cerveau, en intégrant des données empiriques issues des neurosciences cognitives. Ce faisant, nous tenterons de déterminer s'il existe une preuve tangible de l'existence d'un mécanisme sériel de sélection visuelle.

2.1 L'organisation neuronale du système visuel : des éléments en faveur du modèle sériel ?

L'architecture du système visuel impose des contraintes physiologiques aux modèles fonctionnels. Pour comprendre le mode de fonctionnement du système visuel des primates humains, de nombreux résultats présentés dans la suite de cet exposé proviennent d'études menées chez le singe macaque dont le système visuel est anatomiquement et fonctionnellement très similaire au nôtre. Il sera fait référence indifféremment aux travaux chez le singe macaque et chez l'humain (même si cela est

critiquable dans la mesure où il n'existe pas une correspondance parfaite entre les aires corticales des deux espèces). De manière schématique, l'image du monde extérieur se forme sur la rétine par transduction d'un motif photonique et les potentiels d'action émis sont transmis au corps genouillé latéral (structure thalamique) avant d'atteindre le cortex au niveau de l'aire visuelle primaire V1. Cette aire V1 contient une représentation rétinotopique détaillée du champ visuel (les relations spatiales sur la rétine y sont préservées). Les neurones de V1 codent de nombreuses propriétés visuelles simples comme l'orientation, la texture, le mouvement, la couleur… Ces nombreuses cartes de propriétés élémentaires pourraient constituer les briques de base de la perception visuelle et sont souvent considérées comme analogues, voire homologues des cartes visuelles des modèles tels que le GSM et la FIT. Au-delà des aires primaires V1 et V2, un élément clé de l'organisation du système visuel (Figure 14) est sa sub-division en deux grandes voies (Ungerleider & Mishkin, 1982 ; Haxby et al., 1991). La voie "dorsale" se dirige vers le cortex pariétal et serait impliquée dans la représentation du mouvement, la localisation spatiale des objets et la programmation de l'action vers l'objet. La voie "ventrale" se dirige vers le cortex inféro-temporal (IT) et jouerait un rôle direct dans l'identification des objets, la représentation de leur forme et de leur couleur...

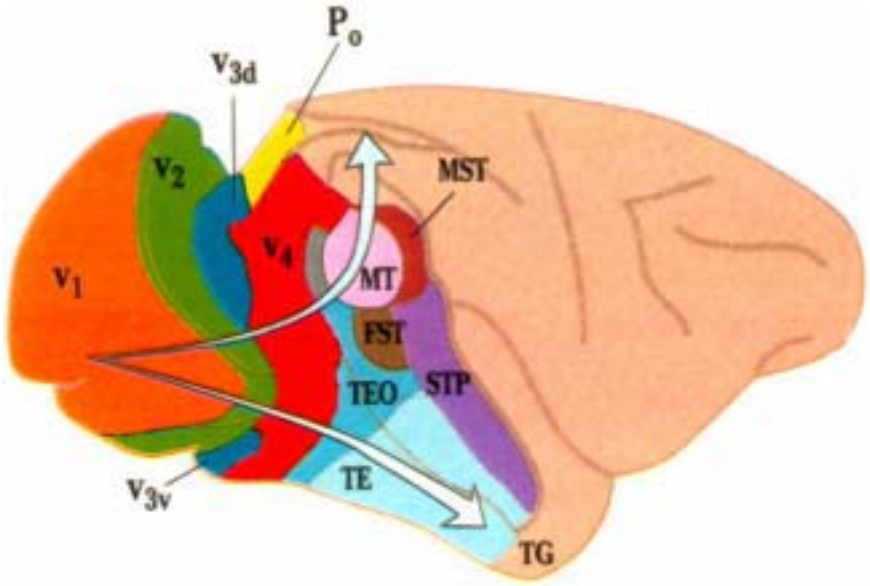


Figure 14. Les voies visuelles corticales chez le singe. A partir de V1, les informations visuelles sont ensuite traitées par V2 puis peuvent suivre la voie dorsale et être traitées dans V3, PO, MT et MST avant d'atteindre le cortex pariétal ; alternativement elles peuvent suivre la voie ventrale et subir des traitements dans V3, V4 et différentes aires temporales (TEO puis TE). Avec l'aimable autorisation de J. Bullier.

Il y a donc un éclatement spatial des représentations des différentes caractéristiques d'objets. Comment dès lors le système visuel peut-il savoir quelles caractéristiques appartiennent à un objet donné d'une scène naturelle ? Si tous les neurones codaient de manière sélective une position spatiale, le problème de l'intégration des éléments de base pourrait être résolu en combinant les caractéristiques séparément pour chaque position du champ visuel. Mais ce n'est pas le cas dans le système visuel dont l'organisation apparaît d'abord hiérarchique (Figure 15). Au fur et à mesure de la progression des traitements au travers des différentes étapes de la voie visuelle ventrale, les propriétés d'objets qui sont codées deviennent de plus en plus complexes. Alors qu'au niveau de V1 les neurones répondent pour des barres orientées, dans IT, souvent décrit comme l'étape ultime de la voie ventrale, la réponse neuronale peut être spécifiquement liée à la présentation de stimuli aussi complexes que des visages, des animaux, des voitures, des scènes naturelles... (chez l'humain : Epstein et al., 1999 ; Gauthier et al., 2000 ; Grill-Spector et al., 2001 ; Haxby et al., 2001 ; chez le singe : Baylis et al., 1987 ; Gross et al., 1969 ; Logothetis & Sheinberg, 1996 ; Perrett et al., 1982 ; Sheinberg & Logothetis, 2001 ; Tanaka, 1996). La sélectivité des neurones semble encore se raffiner davantage au niveau du cortex perirhinal qui reçoit des projections massivement divergentes de IT, mais ce cortex à la fois perceptif, associatif et mnésique est beaucoup moins connu que les autres aires de la voie ventrale (Murray & Richmond, 2001). Parallèlement à l'augmentation de la complexité des réponses neuronales, la seconde caractéristique clé du système visuel est une augmentation de la taille des champs récepteurs (CR) des neurones, à savoir la zone du champ visuel à laquelle ils s'intéressent. Dans V1, où une représentation fine de l'espace est disponible, les CR sont très petits (1-2° d'angle visuel), alors que dans IT, ils sont parfois décrits comme pouvant couvrir l'ensemble du champ visuel. Ces larges CR seraient nécessaires pour appréhender de grands objets et pourraient expliquer notre capacité à reconnaître un objet quelles que soient sa taille et la zone de la rétine qu'il a stimulé. Cependant, ce mode de fonctionnement pose problème lorsque plusieurs objets sont présents simultanément dans le champ récepteur d'un neurone car ces objets entrent alors en compétition pour définir la réponse du neurone qu'ils stimulent (Figure 16). Par exemple, l'intense réponse qu'un neurone produit à la présentation d'un stimulus de référence est fortement diminuée

lorsque, dans le CR de ce neurone, ce stimulus de référence est présenté simultanément avec d'autres stimuli n'induisant que peu ou pas de réponse (Luck et al., 1997a ; Moran & Desimone, 1985 ; Reynolds et al., 1999). Il semble donc que les différents stimuli entrent en compétition pour pouvoir être représentés dans les neurones ayant des CR suffisamment larges pour pouvoir contenir plusieurs stimuli en même temps.

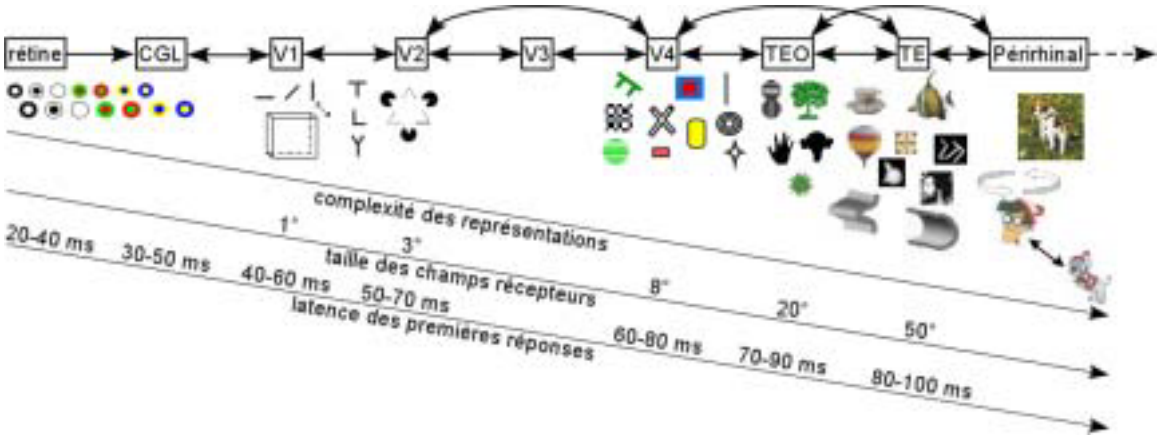


Figure 15. Illustration simplifiée du flot d'information allant de la rétine à la voie ventrale spécialisée dans la reconnaissance des objets chez le singe macaque. Sous chacune des aires se trouve une sélection des stimuli pour lesquels des réponses sélectives ont été enregistrées. Tout au long de la voie ventrale, la complexité des représentations augmente, ainsi que la taille des champs récepteurs et la latence des réponses neuronales (le premier nombre indique la latence approximative des réponses les plus courtes enregistrées, le second la latence moyenne d'activation). Les propriétés visuelles les plus complexes sont codées dans TE où des neurones peuvent répondre sélectivement à des visages, des arbres, des formes tridimensionnelles… Dans le cortex périrhinal, des neurones codent des objets indépendamment du point de vue, une propriété qu'ils partagent avec une petite partie des neurones de l'aire TE. Certains neurones codent aussi les relations entre deux objets ou entre un objet et le contexte dans lequel il apparaît, constituant un support indispensable à la perception d'une scène visuelle. Abréviations : CGL = corps genouillé latéral, TEO = cortex inféro-temporal postérieur, TE = cortex inféro-temporal antérieur.

Comment déterminer à quel objet appartient une propriété visuelle donnée alors que la résolution spatiale est médiocre dans les aires de haut niveau du système visuel ? Cette ambiguïté, conséquence directe de la grande taille des CR des neurones de IT, pourrait être résolue par des mécanismes attentionnels. En effet, quand l'attention spatiale est dirigée vers une zone spécifique au sein d'un champ récepteur comportant deux stimuli, le neurone tend à se comporter comme si seul le stimulus sur lequel l'attention se porte était présent (Figure 16). L'attention spatiale semble agir en rétrécissant la taille des CR de telle sorte qu'un seul stimulus y soit présent, éliminant toute ambiguïté dans sa réponse (Luck et al., 1997a ; Moran & Desimone, 1985 ; Reynolds et al., 1999), un

phénomène également mis en évidence chez l'humain (Kastner et al., 1998). Un même mécanisme serait à l'œuvre lorsque l'attention ne sélectionne pas une zone de l'espace mais un objet, comme dans une tâche de recherche visuelle (Chelazzi et al., 1993, 1998, 2001).

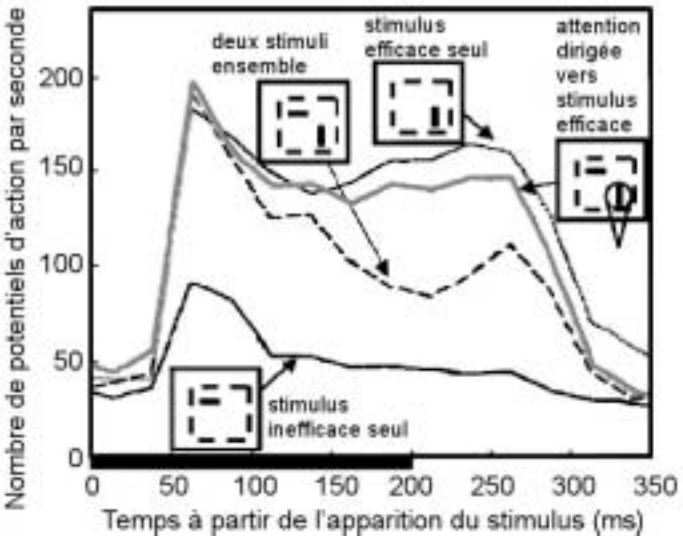


Figure 16. Réponses d'un neurone de l'aire V2 quand un ou deux stimuli sont présents dans son champ récepteur. La barre noire sur l'axe des abscisses indique la durée de présentation du stimulus. Le champ récepteur de la cellule est symbolisé par un carré en pointillés. En haut, la ligne en pointillés fins montre la réponse d'un neurone de V2 à un stimulus efficace, entraînant une très forte réponse. En bas, la ligne continue montre la réponse à un stimulus inefficace. Au milieu, la ligne en pointillés larges montre la réponse associée à la présentation simultanée des deux stimuli. L'addition du stimulus inefficace diminue fortement la réponse du neurone. Quand l'attention (symbolisée par un cercle) est dirigée vers le stimulus efficace, cette suppression est éliminée, le neurone répondant comme si le stimulus inefficace était absent. Adapté de Reynolds & Desimone (1999).

Cette description semble tout à fait compatible avec les modèles hybrides décrits plus haut (p.10): un ensemble de cartes bas niveau fourniraient une représentation détaillée et rétinotopique de l'espace visuel, l'intégration de ces éléments de base se faisant progressivement en impliquant une perte d'information spatiale. La structuration correcte de représentations de haut niveau se feraient donc par l'intervention de l'attention spatiale, réduisant la taille des CR afin de ne traiter qu'un seul stimulus à la fois (e.g., Treisman, 1998a). Notons que les travaux chez le singe ont permis d'apporter des contraintes supplémentaires aux modèles hybrides, puisque l'attention n'est nécessaire que si les stimuli se trouvent dans le même CR (Luck et al., 1997b, pour une version mise à jour du modèle FIT).

Cependant, une autre école de pensée s'est développée à partir de ces travaux. Un point important mis en avant par certains chercheurs est le fait que bien que la réponse d'un neurone soit détériorée quand plusieurs stimuli se trouvent dans son CR, cette détérioration prend effet *après* un certain délai (tout à fait clair sur la Figure 16). Plus particulièrement, dans une tâche de recherche visuelle d'objets complexes associée à une pente de 25 ms par stimulus (recherche « sérielle »), les réponses des neurones de IT et V4 pour les stimuli cibles et distracteurs étaient confondues pendant les 200 ms suivant le début de la présentation (Chelazzi et al., 1993, 1998, 2001) (Figure 17). Il est donc possible que pendant cette première phase les différents stimuli d'une scène visuelle soient encodés en parallèle avant d'entrer en compétition les uns avec les autres. Appliquant ce raisonnement, Desimone & Duncan (1995) ont proposé un modèle de compétition biaisée (voir aussi Desimone 1996, 1998 ; Duncan, 1998 ; Duncan et al., 1997) dans lequel les différents stimuli d'une scène visuelle activent des populations de neurones qui s'engagent dans des interactions compétitives dans l'ensemble du système visuel, aussi bien dans la voie ventrale que dans la voie dorsale. Ces processus de compétition seraient l'essence même de l'attention. A un instant donné, un stimulus gagnerait la compétition pour accéder à des ressources de traitements limitées et serait ainsi représenté de manière explicite à travers tout le système visuel. Typiquement, la compétition pour l'accès aux représentations de haut niveau serait biaisée (d'où le nom du modèle) par des modulations descendantes et ascendantes. Les premières seraient dépendantes de la tâche dans laquelle est impliquée le sujet (comme la description par avance de la cible à trouver) et plus largement de son état cognitif, alors que les secondes dépendraient de la saillance intrinsèque des caractéristiques des stimuli (comme le contraste figure/fond). Les modulations descendantes proviendraient de l'action du cortex pariétal et du cortex frontal sur le système visuel, tout particulièrement sur la voie ventrale lors de la recherche d'un objet ; les modulations ascendantes proviendraient quant à elles de mécanismes de comparaison propres à la voie ventrale (Kastner & Ungerleider, 2000). De manière générale, l'attention agirait en modulant la saillance des représentations (Reynolds et al., 2000). L'idée de compétition est illustrée et renforcée de manière frappante par des résultats neurophysiologiques utilisant la rivalité binoculaire. En effet, lorsqu'un stimulus différent est présenté à chaque œil, nous ne percevons pas

deux objets superposés mais alternativement un objet puis l'autre. A chaque instant, les informations véhiculées par un des yeux gagnent (parfois de manière partielle) la compétition dans la voie ventrale et par là même le contrôle de la réponse comportementale, puis sont inhibées en faveur de celles provenant de l'autre œil (Sheinberg & Logothetis, 1997 ; Logothetis, 1998). Selon cette perspective, tous les phénomènes attentionnels rapportés plus haut au niveau comportemental pourraient se comprendre dans un cadre unique mettant en jeu un vaste réseau d'interactions compétitives à la fois spatiales et temporelles (Keysers & Perrett, 2002 ; Blake & Logothetis, 2002).

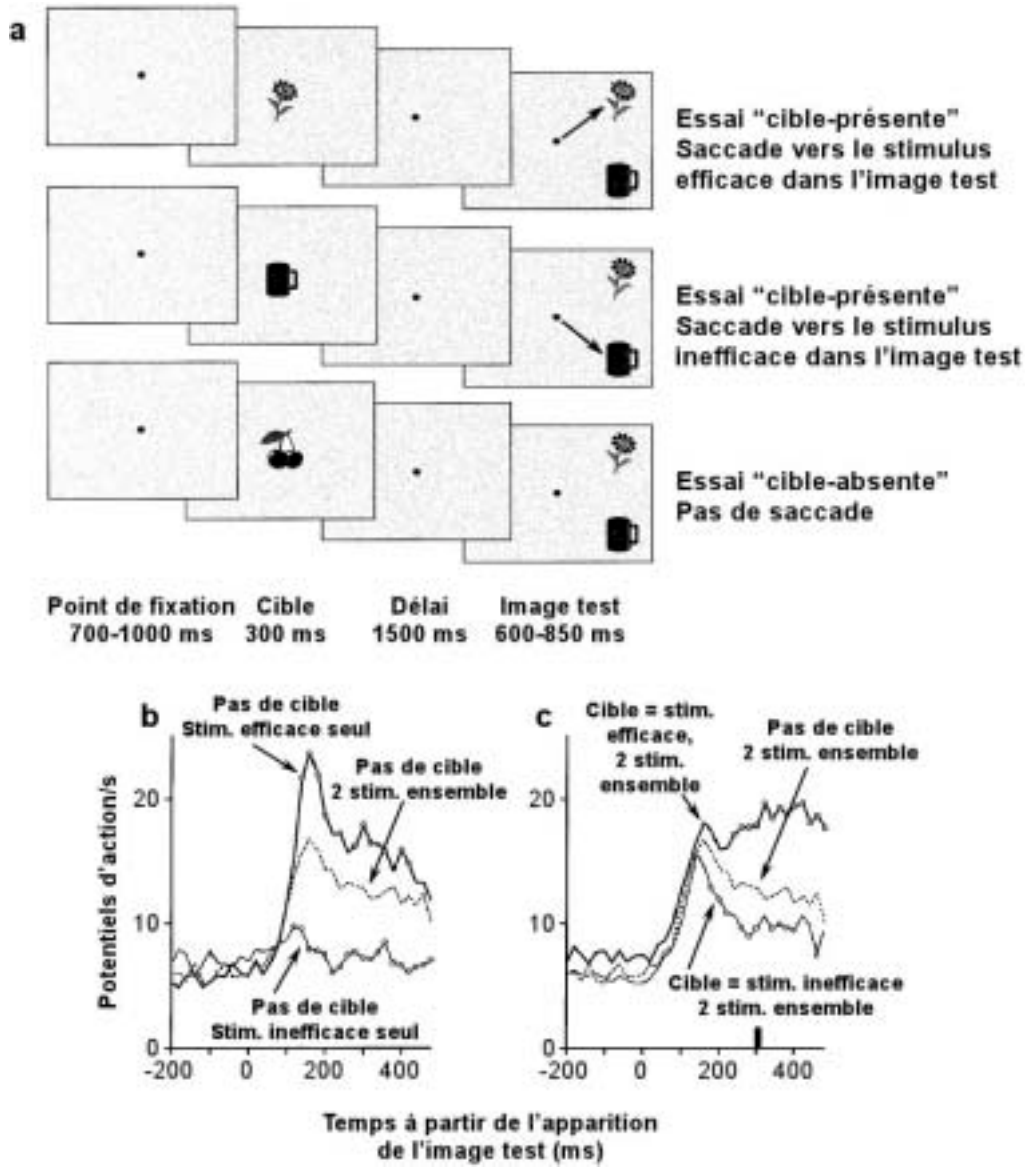


Figure 17. Recherche d'un objet et réponses neuronales dans IT. Dans une tâche d'appariement différé (a) un singe doit maintenir son regard sur un point de fixation puis mémoriser un objet-cible qui lui est présenté sur un écran d'ordinateur. Après un délai, le stimulus-test (composé ici de 2 objets) apparaît, le singe doit effectuer une saccade vers l'objet-cible. Pendant la réalisation de la tâche, des neurones du cortex inféro-temporal sont enregistrés. Pour un neurone donné, on choisit un stimulus « efficace » (ici la fleur) qui induit une forte réponse neuronale lorsqu'il est présenté seul et un stimulus « inefficace » (ici la tasse) qui induit une faible réponse du neurone lorsqu'il est présenté seul. En b) ces réponses aux stimuli efficace et inefficace présentés seuls sont illustrées et comparées à la réponse du neurone lorsque les deux stimuli sont présentés simultanément. La réponse induite par les 2 stimuli est plus faible que celle induite par le seul stimulus efficace. Notez que ces réponses sont enregistrées lorsque ces stimuli ne doivent pas être traités comme cible (3ème essai présenté en a). En c), la réponse aux 2 objets présentés simultanément lorsqu'ils ne doivent pas être traités comme cible est comparée à celle enregistrée quand le statut de cible est accordé soit au stimulus efficace (1er essai en a) soit au stimulus inefficace (2ème essai en a). stim. = stimulus. (Adapté de Chelazzi et al., 1998).

Malgré le poids croissant du modèle parallèle de compétition biaisée, les modèles présentant une composante sérielle conservent toujours une place importante dans la littérature. De nombreux arguments issus de travaux réalisés chez l'humain semblent en effet indiquer qu'un traitement sériel serait à l'œuvre dans le système visuel. Ces arguments vont être exposés et discutés ci-dessous.

2.2 Quelques données chez l'humain

Il faut envisager la possibilité que les travaux chez le singe, focalisés sur la voie ventrale et mettant essentiellement en jeu des enregistrements unitaires, ne parviennent pas à capturer le véritable mode de fonctionnement du système visuel. La perception visuelle nécessitant l'intégration d'informations en provenance de très nombreux sites corticaux et sous corticaux, une approche à un niveau plus intégratif est sans doute justifiée (Varela et al., 2001). Ainsi, de nombreuses études récentes chez l'humain ont utilisé différentes techniques pour tenter de percer la nature des mécanismes de la perception des scènes naturelles. Certains résultats tendent à montrer que des mécanismes sériels de déplacement de l'attention seraient mis en jeu, en impliquant notamment le cortex pariétal. Nous verrons que les arguments en faveur d'une telle hypothèse sont en réalité plutôt ténus.

2.2.1 Apports de la neuropsychologie

Syndrome de Balint. L'intérêt pour l'étude des patients cérébrolésés dans le cadre de la compréhension de la perception des scènes naturelles n'est pas récent mais fut

particulièrement avivé par la découverte d'un patient présentant des troubles très forts de liage perceptif (Friedman-Hill et al., 1995). Suite à une importante lésion occipito-pariétale bilatérale, ce patient présentait un syndrome de Balint, caractérisé par une simultanagnosie, i.e. une incapacité à percevoir plus d'un objet à la fois dans son environnement. Lorsqu'on lui présentait simultanément deux lettres colorées en lui demandant de décrire la première perçue, il réalisait de nombreuses erreurs de type « conjonctions illusoires » (13% des réponses), rapportant avoir vu une lettre portant la couleur de la seconde lettre (voir Figure 3 p.6). Les mêmes erreurs apparaissaient avec des présentations allant jusqu'à 10 secondes. La mise en évidence de ce problème de liage perceptif fut répliqué dans l'hémichamp visuel droit d'un groupe de 8 patients atteints de troubles attentionnels controlésionnels suite à diverses lésions de l'hémisphère gauche (taux de conjonctions illusoires d'environ 25%, Arguin et al., 1994).

Se plaçant dans la perspective adoptée par les tenants des conjonctions illusoires (mais voir p.13), il a été conclu d'après l'étude du patient simultanagnosique que la « carte de contrôle » de l'attention spatiale dépendrait directement du cortex pariétal (Friedman-Hill et al., 1995). En effet, des expériences complémentaires ont montré que ce patient ne souffre pas de problèmes intrinsèques de liage perceptif puisqu'il est capable de discriminer correctement deux lettres colorées lorsqu'elles sont présentées successivement. De plus, il n'a pas non plus de problème à détecter une cible définie par une seule caractéristique mais présente d'énormes difficultés dès que la cible est définie par une conjonction d'une forme et d'une couleur dans une tâche de recherche visuelle. Ce patient est aussi incapable de localiser un objet. Il semble avoir presque complètement perdu toute représentation visuelle de l'espace. Cependant, il présente d'étonnantes capacités spatiales à un niveau implicite (Robertson, 2003 ; Treisman, 1998a). Ces capacités résiduelles, peut-être dues à ses voies ventrales intactes, ne lui permettent toutefois pas de répondre volontairement lorsqu'on le lui demande.

L'étude de cas tels que ce syndrome de Balint suggère que de vastes lésions pariétales (et occipitales) peuvent être associées à une perte de représentations explicites affectant (1) le traitement de conjonctions de formes et de couleurs, (2) la perception *consciente* de plusieurs objets et (3) la capacité à localiser les objets.

En effet, l'ensemble des travaux sur les conjonctions illusoires soutient l'idée selon laquelle la voie dorsale et la voie ventrale doivent interagir pour lier entre elles les représentations structurales des objets et les représentations de surface comme la couleur. Or il est fort probable que notre système visuel ne contient pas une représentation en mémoire de toutes les associations possibles entre formes et couleurs. De telles associations doivent probablement être « construites » à chaque fois qu'elles se présentent. Que se passerait-il si des patients atteints du syndrome de Balint étaient testés avec des objets complexes très familiers pour lesquels des représentations de haut niveau sont sans doute disponibles en mémoire à long terme ? S'ils étaient capables de traiter plusieurs objets complexes de ce type, ceci amènerait à réviser considérablement l'importance du cortex pariétal dans la sélection attentionnelle. Il est également indéniable que la perception consciente est fortement altérée chez les patients simultanagnosiques. Cependant, une telle conclusion laisse complètement ouverte la possibilité d'un traitement normal d'une scène visuelle à un niveau inconscient. En outre, l'incapacité à localiser les objets malgré des capacités spatiales résiduelles suggère que le syndrome de Balint pourrait être dû à des problèmes touchant la sphère des actes moteurs volontaires. Ces questions vont être approfondies à la lumière des nombreuses données concernant des patients atteints d'une lésion unilatérale.

Négligence spatiale unilatérale et extinction visuelle La négligence spatiale unilatérale (NSU) est un trouble qui apparaît souvent après une attaque vasculaire cérébrale touchant les lobes pariétaux (des lésions frontales et sous-corticales peuvent aussi être à l'origine du trouble mais les effets sont généralement moins forts) (Farah, 1990). Ce trouble est plus sévère après lésion du lobe pariétal droit. Les patients atteints de NSU souffrent d'une perte partielle ou totale de perception consciente des stimuli apparaissant dans la zone de l'espace opposée à la lésion cérébrale, donc habituellement dans l'hémichamp visuel gauche. Notons cependant que, comme le souligne Baylis et al. (2002), la sous représentation des patients à lésion gauche pourrait être due à une atteinte conjointe des aires du langage chez ces patients, diminuant fortement la possibilité de les tester. Dans les cas d'héminégligence partielle, certains patients présentent souvent un trouble additionnel d'extinction visuelle : lors de présentations simultanées de deux stimuli ils ne

perçoivent pas un stimulus sur la gauche quand celui-ci est associé à un autre stimulus sur la droite (après une lésion du cortex pariétal droit), alors qu'ils peuvent percevoir ce même stimulus à gauche quand il est présenté seul (voir l'excellente revue de Driver & Vuilleumier, 2001). En outre, chez trois patients présentant des problèmes d'orientation de l'attention dans l'hémichamp droit suite à des lésions de l'hémisphère gauche, la recherche d'une conjonction d'éléments est deux fois plus longue dans l'hémichamp droit que dans le gauche, alors que la recherche d'un élément simple n'est pas perturbé (Arguin et al., 1993). Cette dichotomie qualitative entre les deux types de recherches visuelles, associée à l'extinction visuelle lors de présentations simultanées, semble indéniablement être en faveur de la théorie de Treisman. Cependant, certaines études ont montré qu'un traitement résiduel relativement sophistiqué des stimuli non perçus avait toujours lieu, allant de divers mécanismes de structuration de la forme jusqu'à l'extraction d'informations sémantiques (Driver & Vuilleumier, 2001). De plus, Ashbridge et al. (1999) ont montré que le cortex pariétal ne semble pas nécessaire au liage perceptif, mais plutôt au déplacement de l'attention une fois qu'une représentation de haut niveau est construite. Ces capacités résiduelles, en l'absence de perception consciente, mettraient en jeu la voie ventrale, souvent intacte chez les patients négligents, comme ont pu l'attester plusieurs investigations en imagerie cérébrale fonctionnelle (Rees et al., 2000 ; Vuilleumier et al., 2001). Ce qui caractérise la perception consciente d'un stimulus semble être l'interaction des régions frontales et pariétales avec la voie ventrale, dont l'activation à elle seule ne permet pas de traitement conscient. De telles activations de la voie ventrale semblent cependant suffisantes pour former une trace mnésique implicite des stimuli perçus inconsciemment, mais pas pour en former une mémoire explicite (Vuilleumier et al., 2002). Il faut ici préciser que les traitements implicites concernent souvent un objet isolé présenté dans l'hémichamp controlésionnel. Quand deux stimuli sont présentés simultanément dans l'hémichamp controlésionnel, la perception d'un objet cible semble dépendre de la forme du second objet. En effet, deux étoiles présentées simultanément entraînent beaucoup plus d'extinction qu'une étoile et un triangle (Vuilleumier & Rafal, 2000). Cet effet reste valable pour un stimulus par hémichamp. Ce résultat suggère l'implication directe du cortex pariétal dans la résolution de la compétition au sein de la voie ventrale entre objets de formes similaires,

indépendamment de l'hémichamp visuel dans lequel ils se trouvent. Ceci renforce également l'idée d'une intervention tardive du cortex pariétal dans la perception d'une scène, les stimuli non attendus faisant l'objet d'un traitement relativement sophistiqué. Dans la même veine, des études ont montré que la compétition semble avoir lieu à une étape où les stimuli sont sélectionnés d'après leur significativité en fonction de la tâche à accomplir (Humphreys & Riddoch, 2001 ; Rafal et al., 2002). Cette modulation de la perception des objets en fonction du contexte comportemental implique que des informations sémantiques sont extraites à propos des stimuli qui ne sont pas cibles de l'attention. La relation entre le cortex pariétal, la perception consciente et la capacité à agir sur l'environnement est également renforcée par la découverte d'extinctions bimodales, par exemple entre modalités visuelle et tactile (Mattingley et al., 1997). Ce phénomène est inexplicable si l'on suppose que les représentations atteintes après certaines lésions du cortex pariétal sont entièrement dédiées à la perception visuelle, mais il prend tout son sens si l'on suppose l'existence d'une vaste compétition de toutes les représentations d'objets à travers l'ensemble du cerveau pour pouvoir contrôler la sphère motrice (Duncan, 1998 ; Duncan et al., 1997). Ces conclusions sont compatibles avec la découverte que même les stimuli qui "pop-out" dépendent de ressources attentionnelles pour être consciemment rapportés (Joseph et al., 1997). De plus, il apparaît que l'attention spatiale est automatiquement dirigée vers la position d'une cible, même quand les sujets n'ont pas intérêt à déplacer leur attention vers la cible (Kim & Cave, 1995). Toutes ces données sont en accord avec le modèle de double codage suggéré par Humphreys (1998), selon lequel une connaissance implicite des positions spatiales des objets serait disponible dans la voie ventrale, une représentation spatiale explicite étant prise en charge par la voie dorsale. Ces deux types de représentations seraient combinés pour accéder à une perception explicite d'une scène visuelle. Il y aurait donc tout de même un lien particulier entre l'attention spatiale et le cortex pariétal. Mais toute conclusion sur le rôle précis du cortex pariétal dans la formation des représentations d'objets apparaissant simultanément est pour l'instant prématurée. Certains résultats compliquent encore le tableau en montrant par exemple que des patients négligents présentent aussi un gros déficit d'attention *temporelle*, puisqu'ils sont particulièrement sensibles au paradigme d'AB, leur perception consciente d'un stimulus étant

significativement altérée pendant les 1,6 secondes après l'identification d'un stimulus (Husain et al., 1997). Baylis et al. (2002) rapportent aussi que chez des patients atteints de lésions hémisphériques droites ou gauches, l'extinction est maximale quand les deux stimuli (ipsi- et controlatéral) sont présentés simultanément alors que la perception subjective de la temporalité est complètement perturbée. Le stimulus ipsilatéral est en effet perçu comme arrivant avant l'autre, sauf si le stimulus controlatéral est présenté plusieurs centaines de millisecondes avant le stimulus ipsilatéral. Ainsi, les stimuli controlatéraux à la lésion sont subjectivement perçus comme retardés à cause de l'extinction. Ce résultat est en accord avec des travaux montrant que l'attention focalisée accélère la vitesse de traitement à l'endroit où elle est appliquée. Il est donc clair qu'une lésion pariétale semblant affecter le traitement spatial a en réalité des conséquences beaucoup plus larges, affectant surtout la sphère des traitements conscients, qu'il s'agisse des traitements spatiaux ou temporels.

2.2.2 Apports de l'imagerie fonctionnelle

La revue de la littérature en neuropsychologie, bien entendue incomplète, ne montre pas directement que des mécanismes sériels sont impliqués dans le traitement visuel. Les tableaux cliniques sont très variés, mettant surtout en avant l'implication du cortex pariétal dans la perception consciente et la gestion des actes sensori-moteurs plutôt que dans l'analyse sérielle de la scène visuelle. La présente section explore des éléments additionnels de réponse fournis par certaines études en imagerie cérébrale chez le sujet sain.

Dans une expérience mettant en jeu la tomographie par émission de positons (TEP), les changements locaux dans le flux sanguin cérébral étaient mesurés tandis que des sujets humains réalisaient une tâche de recherche visuelle d'un élément simple (couleur ou mouvement) ou d'une conjonction d'éléments (couleur et mouvement) (Corbetta et al., 1995). La pente de recherche était nulle pour un seul élément, contrairement à la conjonction d'éléments. La recherche d'une conjonction, par rapport à la recherche d'un élément, entraînait une augmentation du débit sanguin au niveau du cortex pariétal postérieur, particulièrement dans l'hémisphère droit, un patron d'activation très similaire à celui obtenu au cours d'une tâche impliquant des

déplacements d'attention spatiale (Corbetta et al., 1993). Les auteurs conclurent que la recherche d'une conjonction implique l'inspection attentive sérielle des éléments d'une scène (Corbetta et al., 1995).

Cependant, cette interprétation fut remise en cause par des travaux ultérieurs en TEP et en IRMf (imagerie par résonance magnétique fonctionnelle, mesure les variations d'apport en oxygène dans une zone corticale entre deux tâches) montrant une activation du cortex pariétal postérieur au cours de tâches nécessitant un effort de traitement, mais sans déplacements de l'attention (Vandenberghe et al., 1997 ; Wojciulik & Kanwisher, 1999 ; voir aussi Culham & Kanwisher, 2001, pour la grande difficulté à comprendre le fonctionnement du cortex pariétal). En accord avec ces derniers résultats, il a été montré que la recherche d'une conjonction d'éléments ne permet pas de prédire une activation du cortex pariétal ; par contre, certaines zones du cortex pariétal présentent une activité proportionnelle à la difficulté de la tâche lors de recherches de conjonctions (Nobre et al., 2003). Une autre expérience a montré qu'à difficulté égale seule la recherche d'une conjonction met en œuvre le cortex pariétal droit (Shafritz et al., 2002), mais seulement lorsque les deux objets sont présentés simultanément et non pas successivement. Ainsi, le cortex pariétal interviendrait préférentiellement pour résoudre une ambiguïté spatiale, mais pas pour encoder des conjonctions d'éléments simples. Puisqu'une activité pariétale droite a également été associée à la sélection d'une cible parmi des distracteurs sans nécessité d'un liage perceptif (Marois et al., 2000), il est tout à fait possible que l'activité pariétale soit due à un processus de sélection tardive, quand plusieurs choix comportementaux sont possibles au même moment (Schadlen & Movshon, 1999). A noter que dans l'expérience de Marois et al. (2000), l'activité pariétale était aussi plus importante lorsque l'interférence des distracteurs était due à une proximité temporelle et pas seulement spatiale. Ceci renforce à nouveau l'idée que le cortex pariétal est important pour la perception consciente des objets et celle de leurs relations spatiales et temporelles, mais pas pour former des représentations de haut niveau.

En conclusion, les travaux en imagerie cérébrale ne soutiennent pas l'existence de mécanismes sériels à l'œuvre dans le système visuel. La formation de descriptions complexes des objets semble ne pas avoir besoin du cortex pariétal. Cependant il reste frappant de constater que les études qui démontrent clairement un traitement inconscient

dans la voie ventrale mettent typiquement en jeu un seul objet à la fois. Suivant les protocoles de masquage de Dehaene et al. (1998, 2001) montrant un traitement inconscient de chiffres et de mots isolés, serait-il possible de mettre en évidence un traitement de haut niveau pour deux chiffres ou deux mots présentés simultanément ? Probablement, mais jusqu'à quel niveau d'intégration ? L'étude des patients héminégligents laisse présager que la compétition s'effectue au niveau moteur, mais cela reste à prouver directement chez le sujet sain.

2.2.3 Apports de la TMS

La stimulation magnétique transcrânienne (TMS) consiste à appliquer à la surface du scalp des impulsions magnétiques qui perturbent la conduction des fibres blanches de la structure corticale sous-jacente. Certaines études en TMS ont fourni des résultats compatibles avec l'implication du cortex pariétal postérieur dans des tâches nécessitant la recherche d'une conjonction. Notamment, dans une expérience d'Ashbridge et al. (1997), tout comme dans l'expérience de Corbetta et al. (1995), les sujets réalisaient une tâche de recherche visuelle d'éléments simples (couleur ou orientation) ou d'une conjonction. A des temps différents à partir du début de la présentation du stimulus, la TMS était appliquée en un point de la région pariétale des sujets. Quand la stimulation avait lieu 100 ms après le début d'un stimulus contenant une cible ou 160 ms après le début d'un stimulus ne contenant pas de cible, l'application de la TMS sur le cortex pariétal postérieur entraînait une augmentation des TR dans la recherche d'une conjonction mais pas dans la recherche d'un élément simple. De plus, après un entraînement intensif des sujets ayant permis d'obtenir des pentes de recherche nulles pour les mêmes conjonctions d'éléments, la même stimulation TMS n'avait plus aucun effet (Walsh et al., 1998). Le même résultat a été rapporté pour la recherche d'une conjonction donnée d'éléments visuels associée à des pentes de recherche non-nulles après un entraînement intensif sur une autre conjonction d'éléments, suggérant un transfert d'apprentissage d'une tâche à l'autre (Walsh et al., 1999, expérience 2). Ashbridge et al. (1997) ont postulé que la TMS aurait perturbé la communication entre voies ventrale et dorsale. En particulier, la voie ventrale enverrait des signaux correspondant à la composition des différents items de la scène pour indiquer aux mécanismes d'attention spatiale dans le cortex pariétal les

positions probables de la cible. La TMS perturberait le transfert de ces signaux vers le cortex pariétal droit. Cette explication prédit qu'une augmentation du nombre d'items ainsi que la similarité entre cibles et distracteurs et l'hétérogénéité entre distracteurs devrait augmenter proportionnellement le temps nécessaire pour calculer la probabilité que chacun des items soit une cible et donc décaler le temps critique pour la TMS. Cependant, cette étude s'est bornée à une application de la TMS entre 0 et 200 ms après le début de l'apparition du stimulus et ceci seulement pour les essais comportant 8 items. L'emploi de tâches de recherche visuelle de difficulté croissante pourrait bien révéler un recrutement progressivement plus important du cortex pariétal, invalidant ainsi l'idée d'une différence qualitative entre recherches d'éléments simples et de conjonctions d'éléments comme cela a déjà été démontré en IRMf. D'ailleurs, le transfert d'apprentissage rapporté par Walsh et al. (1999, expérience 2) laisse supposer que la mise en jeu du cortex pariétal a plus à voir avec un apprentissage sensori-moteur qu'avec la formation de représentations de conjonctions d'éléments visuels. Si le cortex pariétal était impliqué dans la formation de représentations de conjonctions d'éléments par un balayage sériel de la scène, la perturbation par la TMS devrait pouvoir intervenir n'importe quand pendant l'inspection de la scène. Or, l'effet restreint de la TMS à un instant précis après l'apparition de la stimulation visuelle (Ashbridge et al., 1997) est incompatible avec l'idée que la recherche d'une conjonction nécessite un déplacement sériel de l'attention. De plus, une étude récente a mis en évidence que la détection d'un point isolé en vision périphérique est perturbée après une longue stimulation du cortex pariétal, alors qu'un point est typiquement considéré comme un élément pré-attentif (Hilgetag et al., 2001). Hilgetag et al. rapportent non seulement une diminution de la performance pour les stimuli controlatéraux à la stimulation TMS, mais également une augmentation de la performance pour les stimuli ipsilatéraux à celle-ci. Selon les auteurs, cette découverte s'accorde très bien avec un modèle de compétition inter-hémisphérique (voir aussi des données allant dans ce sens dans Walsh et al., 1999, expérience 1). Dans cette perspective, on peut imaginer que les représentations spatiales associées aux différents stimuli d'une scène visuelle pourraient entrer en compétition dans les cortex pariétaux afin d'être pris en compte au niveau comportemental et dans la sphère consciente (Duncan, 1998 ; Duncan et al., 1997) sans pour autant impliquer que le cortex

pariétal soit responsable de la structuration des représentations de ces stimuli. La similarité entre cibles et distracteurs dans une recherche de conjonction pourrait induire une compétition à un niveau supérieur entre les différentes réponses associées à chacun, le cortex pariétal étant justement impliqué dans la représentation d'associations sensori-motrices alternatives (Bunge et al., 2002).

En conclusion, des études en neuropsychologie, en imagerie fonctionnelle et en TMS mettent en avant une implication spécifique et critique du cortex pariétal postérieur, particulièrement de l'hémisphère droit, dans des tâches nécessitant la détection d'une cible définie par une conjonction d'éléments associée à une pente de recherche non nulle. Cependant, il n'y a pas de preuve irréfutable que cette implication du cortex pariétal soit reliée à son rôle dans le contrôle du déplacement sériel de l'attention spatiale. Un nombre de plus en plus important d'éléments tend plutôt à montrer que le cortex pariétal aurait une fonction très importante dans le contrôle sensori-moteur permettant d'expliquer de nombreux résultats.

2.2.4 Apports des potentiels évoqués

L'étude des potentiels évoqués (PE) enregistrés à la surface du scalp pendant que des sujets réalisent des tâches de recherche visuelle a permis de mieux comprendre la nature des mécanismes de sélection qui contraignent la perception des scènes naturelles. Les PE sont dérivés de la cartographie électrique (ou EEG) qui est en relation directe avec l'activité neuronale post-synaptique de vastes populations de neurones. Ils sont obtenus en moyennant l'EEG associé à la présentation d'un stimulus, la référence temporelle pour réaliser ce calcul étant le début de l'apparition des stimuli. Les PE étant une mesure des champs électromagnétiques associés au traitement d'un type de stimulus, ils ont une très bonne résolution temporelle, puisque ces champs correspondent à des déplacements de charges électriques. Par contre leur résolution spatiale est mauvaise dans la mesure où il est difficile d'estimer l'origine cérébrale des champs électriques enregistrés en surface.

Les deux questions fondamentales auxquelles se sont attachées à répondre les études en PE sont (1) le niveau de sélection des objets cibles (est-il précoce, i.e. perceptuel, ou au contraire tardif, post-perceptuel ?) et (2) les facteurs influençant la

nécessité de l'attention sélective à certains niveaux de traitement. Il apparaît que l'attention sélective opère à des niveaux multiples de traitement et que la présence d'une sélection attentionnelle à un niveau de traitement donné dépend de la présence d'une compétition entre stimuli à ce niveau, qui dépend à son tour de la nature des stimuli et de la tâche.

Une partie des travaux sur l'attention visuelle en PE s'est focalisée sur la sélection spatiale. Il a été montré que celle-ci a une influence très précoce sur le traitement visuel, intervenant probablement au niveau perceptuel plutôt que post-perceptuel, puisque l'attention spatiale module l'amplitude de la composante P1, dès 70-90 ms après l'apparition d'un stimulus (P1 est la première composante endogène évoquée par une stimulation visuelle, Hillyard, Vogel & Luck, 1998 ; Mangun, 1995). Des effets de cette nature ont été observés seulement dans des études mettant en jeu l'attention spatiale et n'ont pas été observés quand les stimuli attendus et non attendus sont présentés à la même position mais diffèrent selon d'autres dimensions comme la couleur (Hillyard & Anllo-Vento, 1998). La sélection par la couleur pourrait intervenir plus tard, vers 150 ms (Hillyard & Anllo-Vento, 1998). Si l'attention opérait après l'identification du stimulus, il n'y aurait aucune raison pour que la position du stimulus soit une dimension traitée différemment des autres dimensions. En revanche, un mécanisme spécial de sélection par la position est tout à fait en accord avec les représentations topographiquement organisées utilisées dans les étapes précoces et intermédiaires du traitement visuel. Des modulations de l'amplitude de la P1 sont observées quel que soit le statut du stimulus pour la tâche, c'est-à-dire pour les cibles et les non cibles, ainsi que pour des stimuli sans aucune importance pour la réalisation de la tâche (Luck, Fan & Hillyard, 1993). On peut supposer que des effets de l'attention opérant après l'identification du stimulus devraient être limités aux stimuli pertinents pour la tâche, parce qu'il n'y aurait aucune raison de porter son attention sur un stimulus déjà identifié comme non pertinent pour la tâche. Au contraire, tout effet attentionnel opérant avant l'identification doit nécessairement être insensible à l'identité du stimulus. Les générateurs de l'effet attentionnel sur la P1 ont été modélisés par des dipôles situés dans les aires corticales extra-striées de la voie ventrale (Heinze et al., 1994). De fortes similarités entre les effets attentionnels obtenus sur la P1 et les effets attentionnels observés en enregistrements unitaires chez le singe laissent

penser qu'ils pourraient influencer le traitement dans l'aire V4 (Luck, Chelazzi et al., 1997). Cette sélection précoce serait nécessaire dans les nombreux cas où une surcharge perceptuelle pèse sur le système, par exemple lorsque des réponses très rapides sont requises, lorsque les cibles sont difficiles à détecter, lorsque la cible est entourée de distracteurs ou en présentant les stimuli à des fréquences importantes comme dans les séquences RSVP.

Notons cependant qu'un effet précoce de l'attention spatiale n'implique pas nécessairement que la dimension spatiale d'un stimulus soit traitée systématiquement avant ses autres propriétés comme la forme et la couleur. Les effets précoces de l'attention spatiale reflètent plutôt le fait que les propriétés spatiales d'un objet sont disponibles plus précocement dans la voie ventrale, grâce aux petits CR de V1 par exemple, que les propriétés intégrées comme la forme. Ce résultat ne peut donc pas être tenu comme une preuve en faveur des modèles de sélection précoce.

Une autre série de travaux mettant en jeu la technique des PE a porté plus spécifiquement sur l'étude de la tâche de recherche visuelle. Il a été montré que l'attention focalisée est dirigée de manière réflexe vers la position d'une cible peu après le début d'un stimulus pop-out (Luck & Hillyard, 1995), en accord avec certaines données comportementales. Cependant, cette attraction de l'attention par un stimulus pop-out ne serait pas automatique (Luck & Hillyard, 1994a). En effet, quelle que soit la tâche, les sujets reçoivent toujours pour instruction de chercher un élément cible. Donc l'effet pop-out ne peut être considéré comme une preuve de capture attentionnelle automatique, « bottom-up ». Des mécanismes « top-down » sont toujours mis en jeu. En accord avec ce résultat électrophysiologique, un test direct de l'hypothèse de capture attentionnelle à permis de montrer qu'une propriété pop-out, comme une différence de couleur, si elle n'est pas associée de manière systématique avec la cible, n'attire pas plus l'attention que les autres stimuli qui n'ont pas d'effet « pop-out » (Yantis, 1998).

Dans une tâche de recherche visuelle d'une conjonction d'éléments, il a été montré que la réponse en PE est plus négative au niveau des électrodes occipitales temporales controlatérales à la cible par rapport aux électrodes ipsilatérales, environ 175 ms après l'apparition du stimulus (Luck, Girelli et al., 1997). Cette activité différentielle, maximale entre 200-300 ms après le début de la stimulation visuelle, est appelée N2pc

(pour négativité postérieure controlatérale apparaissant vers 200 ms). De nombreux éléments suggèrent un lien très fort entre la N2pc et la sélection spatiale attentionnelle (Luck & Hillyard, 1994a, 1994b). La N2pc semble refléter la sélection de la cible et la suppression progressive des distracteurs dans des tâches de recherche visuelle (Hopf et al., 2002a). La N2pc est controlatérale, ce qui est en accord avec le fait que les neurones de V4 répondent de manière quasi exclusive aux stimuli controlatéraux et au fait que les réponses des neurones de IT sont largement dominées par les stimuli controlatéraux. Plusieurs analyses de source suggèrent fortement que la N2pc est générée par un réseau d'aires visuelles corticales à distribution occipito-temporale ventrale, en accord avec l'idée qu'elle reflète en partie des mécanismes de compétition entre représentations visuelles dans les aires à haut niveau d'intégration de la voie ventrale (Hopf et al., 2000, 2002a). Cette origine ventrale de la N2pc semble doublée d'une participation des aires pariétales lorsque la cible visuelle est située à faible proximité d'éléments distracteurs, en accord avec la littérature revue plus haut sur la participation du cortex pariétal aux mécanismes de sélection attentionnelle (Hopf et al., 2000). Sa modulation semble dépendre des mêmes facteurs qui affectent les réponses des neurones enregistrés chez le singe (Luck, Girelli et al., 1997). La latence de 175 ms est la même que celle rapportée pour un effet attentionnel similaire au niveau d'une population de neurones de IT chez le singe (Chelazzi et al., 1993, 1998)[2]. Non seulement la latence de l'effet attentionnel est le même en PE et en enregistrements unitaires, mais l'effet N2pc est affecté par un ensemble de manipulations expérimentales de la même manière que les effets attentionnels observés au niveau unitaire par Chelazzi et ses collègues. La N2pc est plus large pour des tâches de discrimination de conjonctions que pour des tâches de détection d'éléments simples ; elle est plus large quand une cible est entourée de distracteurs proches ; enfin elle est aussi plus large quand un élément simple doit être localisé, par exemple par une saccade, que quand il doit simplement être détecté. Ainsi, que ce soit chez le primate humain ou chez le primate non humain, un rôle majeur de l'attention serait de résoudre le codage neuronal ambiguë qui a lieu quand plusieurs items sont présents au sein du CR d'un neurone. En accord avec ces travaux, Luck, Girelli et al.

[2] Cette similarité est troublante, notamment parce que la latence des neurones ou des actes moteurs des singes est souvent très inférieure à celle des humains (Fabre-Thorpe, Richard & Thorpe, 1998).

(1997) ont proposé une théorie de l'attention (« Ambiguity resolution theory ») qui postule que l'ambiguïté dans le codage neuronal peut être résolue par un mécanisme attentionnel qui limite le traitement à un seul objet a la fois. Un tel filtrage serait nécessaire seulement dans des conditions conduisant à un codage neuronal ambigu, il dépendrait donc du niveau d'intégration neuronale requis par la tâche et de la nature des stimuli.

Ces études en PE sur l'attention ont donc montré une place prépondérante du facteur spatial dans la sélection d'un objet cible au cours d'une tâche de recherche visuelle. Une étude complémentaire sur la N2pc a même suggéré que l'attention se déplacerait rapidement, par bonds successifs allant d'un objet cible potentiel à l'autre (Woodman & Luck, 1999).

Cependant, toutes ces études souffrent, tout comme les études comportementales rapportées plus haut sur la tâche de recherche visuelle, de l'usage quasi exclusif de stimuli plutôt simples et artificiels, très différents des objets que nous rencontrons dans la vie de tous les jours. Il est tout à fait envisageable que des capacités insoupçonnées de traitement parallèle d'objets complexes existent dans la voie ventrale. Des travaux complémentaires ont d'ores et déjà commencé à déterminer le mode de traitement de stimuli plus complexes.

Ainsi, une autre partie de la littérature en PE, s'intéressant elle au traitement de stimuli complexes, a montré que des mécanismes de discrimination d'objets isolés s'activent très tôt, dès 120-150 ms après l'apparition d'un stimulus (e.g. Bentin et al., 1996 ; Rossion et al., 2000 ; Schendan et al., 1998 ; Vogel & Luck, 2000), mettant en jeu de larges portions des aires corticales visuelles occipito-temporales ventrales (Hopf et al., 2002b). Ces latences de début de traitement sont remarquablement courtes. Ce qui est encore plus surprenant est qu'il ne faut pas plus de 150 ms après la présentation de photographies de scènes naturelles pour observer une différence entre les PE associés à la catégorisation d'images cibles contenant des animaux ou des véhicules de celles n'en contenant pas (Thorpe et al., 1996 ; VanRullen & Thorpe, 2001b). Ce traitement rapide semble mettre en jeu un réseau distribué d'aires corticales dans la voie ventrale (Fize et al., 2000). Une telle rapidité de traitement pourrait dépendre en grande partie de mécanismes essentiellement vers l'avant (« feedforward », Thorpe & Fabre-Thorpe,

2001 ; VanRullen & Thorpe, 2002). La robustesse de ces mécanismes permettrait d'expliquer la si grande rapidité des sujets humains à détecter des formes complexes dans les scènes naturelles. Ce qui est pertinent ici pour notre propos est que la catégorisation d'objets dans les scènes naturelles met déjà en jeu une certaine forme de parallélisme. En effet, dans les tâches décrites ci-dessus, les sujets n'ont aucun moyen de savoir par avance où va se trouver l'animal cible dans l'image. A l'appui de cette idée, la catégorisation rapide de scènes naturelles est toujours associée au niveau des PE à une activité différentielle précoce entre images cibles et images non-cibles vers 150 ms même avec des stimulations extra-fovéales (Fabre-Thorpe et al., 1998). Cette hypothèse d'un traitement en parallèle des scènes naturelles reste cependant à tester. Malgré les résultats rapportés par VanRullen et al. (sous presse) selon lesquels les scènes naturelles ne seraient pas traitées de manière parallèle mais plutôt de manière pré-attentive, il reste tout à fait possible que les PE puissent révéler un parallélisme plus massif que le comportement ne le laisse prévoir. Les articles 1 et 2 de cette thèse fournissent des éléments en faveur de cette hypothèse (Rousselet et al., 2002 ; Rousselet, Thorpe & Fabre-Thorpe, en préparation).

2.3 Codage neuronal dans la voie ventrale : au-delà des apparences

La littérature en neurosciences cognitives revue ci-dessus est contradictoire en apparence. En effet, quelle que soit la technique utilisée, les premières études réalisées ont en majorité conclu à l'implication de mécanismes sériels précoces dans le traitement des scènes naturelles, ceci reposant en grande partie sur la mise en jeu de ressources spatiales gérées par le cortex pariétal. En revanche, les études ultérieures insistent sur le caractère tardif de la sélection attentionnelle, la voie ventrale semblant capable de traitements bien plus sophistiqués que ne le laisse penser la description classique du système visuel. Pour comprendre cette opposition, la présente section a pour but d'aller au-delà de la description classique de la voie ventrale, source d'erreurs et de préjugés tenaces.

2.3.1 Les briques de base de la perception visuelle

Tous les modèles de la perception visuelle postulent l'existence d'unités fondamentales à partir desquelles des représentations complexes sont construites. De manière relativement implicite, la majorité des chercheurs en perception visuelle ne dénigreraient pas l'idée selon laquelle ces briques de base correspondent aux éléments relativement simples codés dans V1 et d'autres aires corticales supposées « hautement » spécialisées. C'est d'ailleurs le postulat de départ de la plupart des modèles sériels ou hybrides : tout est codé de manière fine mais locale dans V1, la combinaison de ces traits de base implique une intégration spatiale dans la voie ventrale dont les neurones codent des propriétés de plus en plus complexes mais en perdant de l'information spatiale, d'où le problème du liage perceptif et la nécessité de mécanismes sériels… Qu'en est-il réellement ?

Les cartes d'éléments de base, telles qu'elles sont définies par les études utilisant le paradigme de recherche visuelle ne correspondent pas aux cartes corticales telles qu'on les trouve dans V1. En effet, des travaux en psychophysique ont établi que les différences visuelles les plus fines discriminables entre deux stimuli (les « JND », ou just noticeable differences), probablement codées par V1, sont toujours plus fines que les différences permettant une recherche parallèle. Cela a déjà été souligné par plusieurs chercheurs du domaine (voir notamment Wolfe, 1998) mais cette idée semble encore tenace dans une partie de la communauté scientifique.

Les recherches dites parallèles d'éléments simples, sans nous révéler la nature des briques de base de la perception visuelle, pourraient entièrement s'expliquer par des mécanismes d'interactions centre/pourtour dans le système visuel. En effet, puisque les distracteurs forment très souvent une sorte de texture homogène, la cible apparaît alors comme un élément incongru pouvant être détecté par un simple codage du contraste local indépendamment du nombre de distracteurs. Ces mécanismes de codage du contraste local sont à l'œuvre dans l'ensemble du système visuel, dès la rétine, et ont été tout particulièrement étudiés dans V1 (Gilbert et al., 2000) et dans le cortex pariétal (Gottlieb, 2002).

Les modèles attentionnels classiques sont fondés en partie sur une mauvaise compréhension de l'organisation fonctionnelle du système visuel. Les modèles qui

stipulent l'existence d'une première phase d'extraction en parallèle d'éléments de base font référence à l'organisation fonctionnelle supposée modulaire et parallèle d'aires telles que V1 (e.g., Treisman, 1998a). Cependant, de nombreux travaux montrent que dès V1 l'organisation du système visuel n'est pas modulaire. D'une part, les neurones corticaux codent pour un ensemble de propriétés, ce sont donc des analyseurs multifonction plutôt que des détecteurs de traits. D'autre part, le regroupement des neurones codant les mêmes propriétés visuelles n'est pas une règle absolue mais plutôt une tendance (Bullier & Nowak, 1995 ; Schiller, 1997). Par exemple, il n'y a nulle part une aire strictement dédiée au traitement de la couleur ou du mouvement, bien que beaucoup de chercheurs associent la couleur à V4 et le mouvement à MT (Gegenfurtner & Kiper, 2003 ; Schiller, 1997). Partout les neurones codent pour plusieurs propriétés, même s'ils sont plus ou moins « spécialisés » dans un type de codage ou un autre. Les propriétés de forme et de couleur ne sont pas codées dans des « modules » distincts, mais interagissent très tôt dans le traitement visuel (Kubovy et al., 1999).

De plus, les neurones des aires corticales précoces semblent répondre à des combinaisons d'éléments ayant une configuration spatiale particulière. Par exemple, des neurones de V1 chez le chat répondent à deux lignes formant une croix, un « L » ou un « T » (Shevelev et al., 1995). De plus, des modulations complexes en provenance de l'extérieur du CR classique de neurones de V1 ont été mises en évidences, permettant sans doute de coder des jonctions et des coins (Sillito et al., 1995). Ceci montre de manière plus générale que les neurones de V1 ne codent pas simplement des barres ou des bords isolés. La structuration spatiale des formes visuelles commencent très tôt, sans faire appel à des mécanismes attentionnels.

Ces mécanismes de structuration de la forme sont très rapides, des objets complexes comme des visages pouvant être encodés par un flot massivement vers l'avant de décharges neuronales (Keysers et al., 2001 ; Rolls et al., 1999). Ainsi, ce qui est codé rapidement dans le système visuel n'étant pas nécessairement simple, il n'y a pas de raison de limiter le monde pré-attentif à une collection d'éléments bas niveau.

2.3.2 Compétition et codage spatial dans la voie ventrale

Il a été reproché au modèle de la compétition biaisée de ne pouvoir fonctionner que lorsqu'une cible est définie par avance, nécessitant l'intervention d'une carte de saillance pour diriger l'attention dans les autres cas (Itti & Koch, 2001). En effet, les modulations de l'attention dans la recherche d'objets sont typiquement mises en évidence lorsque l'animal qui réalise la tâche maintient en mémoire une description visuelle de la cible avant que l'ensemble des stimuli tests n'apparaissent à l'écran (Chelazzi et al., 1998). En l'absence de cette description par avance du stimulus, il est tout à fait possible que le modèle de la compétition biaisée soit inopérant et qu'il faille avoir recours à des mécanismes sériels d'exploration dépendant d'une carte de saillance (Itti & Koch, 2002). Cette carte de saillance définirait dans quel ordre les objets sont explorés en fonction de leur saillance intrinsèque, essentiellement le contraste figure/fond.

La nécessité de l'usage d'une carte de saillance est cependant fortement remise en cause (Einhäuser & König, 2003 ; Hayhoe et al., 2003 ; Torralba, 2003 ; VanRullen, sous presse). Néanmoins, cette remise en question du modèle de la compétition biaisée ne semble pas être fondée. En effet, comme cela a été mentionné plus haut (Figure 17C), la compétition entre représentations d'objets activées en parallèle intervient *après* un certain délai. La phase initiale de la réponse neuronale est identique dans les différentes conditions de stimulation. Au cours de cette première phase, les neurones répondent de la même manière que le stimulus efficace soit présenté seul ou parmi d'autres. Ce n'est que dans une deuxième phase que la réponse est affectée par la présence simultanée d'autres stimuli. Il faut noter que cette phase initiale, supposée représenter un encodage parallèle, est également présente lorsqu'aucun stimulus ne porte le statut de cible (Figure 17B). Ceci suggère qu'un traitement de plusieurs objets serait possible en parallèle même en l'absence d'intervention de mécanismes mnésiques à court terme, invalidant la remise en cause du modèle. Cependant, il faut noter que cette première phase semble plus courte en l'absence de cible que lorsque l'un des stimuli à été prédéfini comme cible. De plus, on peut se demander comment un tel encodage de plusieurs objets en parallèle est possible alors que la description fonctionnelle du système visuel suggère que cela est impossible. Les différentes sections qui suivent constituent un argumentaire détaillé montrant

comment la voie ventrale intègre l'information spatiale et pourrait traiter plusieurs objets en parallèle.

Codage spatial

Il est bon ici de rappeler ce qui apparaît souvent dans la littérature comme un lieu commun, souligné par exemple par Treisman (1996) : « Objects and locations appear to be separately coded in ventral and dorsal pathways, respectively, raising what may be the most basic binding problem : linking 'what' to 'where'. » Ce point de vue prend la forme d'un dogme dans les études sur le traitement visuel, à savoir que dans IT les représentations spatiales seraient perdues de telle sorte que IT ne pourrait pas représenter plus d'un objet à la fois sans ambiguïté (Reynolds & Desimone, 1999 ; Treisman, 1998a ; von der Malsburg, 1999 ; Wolfe & Cave, 1999). Cependant, une analyse attentive des données disponibles et de récentes expériences montrent qu'un codage spatial est présent tout le long de la voie ventrale.

Tout d'abord, il faut noter que la taille des CR de IT n'est pas si importante que ce que mentionnent de nombreux articles et qu'en réalité très peu d'études systématiques ont porté sur ce sujet. Récemment, Op de Beeck & Vogels (2000) ont rapporté chez le macaque des tailles de CR allant de 2.8° à 26°, avec une moyenne de 10.3° et un écart-type de 5° pour des stimuli isolés. On est donc loin des CR couvrant l'ensemble du champ visuel. En revanche, Rolls et al. (2003) ont rapporté que la taille moyenne des CR de 17 neurones de IT était de 78° d'angle visuel quand des stimuli apparaissaient sur un fond blanc, mais seulement de 22° sur un fond complexe « naturel ». Cette diminution de la taille des CR entre stimuli présentés sur fond de scènes naturelles et sur fond blanc ne s'accompagnait pas d'une diminution significative du taux de décharge des neurones ni de leur sélectivité (voir aussi Sheinberg & Logothetis, 2001). Dans les scènes naturelles, les neurones de IT présentent donc une forte diminution d'invariance à la translation.

De plus, les neurones présentant des propriétés d'invariance à la taille et à l'orientation des stimuli seraient une conséquence des grands CR supposés de IT. On s'attendrait donc à trouver de nombreux neurones présentant des réponses invariantes dans IT. Au lieu de cela, ils constituent en réalité une portion restreinte des neurones enregistrés dans IT et sont typiquement situés juste à côté des neurones présentant des

réponses fortement non invariantes (Logothetis & Sheinberg, 1996 ; Booth & Rolls, 1998). Ceci suggère que les réponses invariantes sont construites localement par des interactions entre neurones présentant des réponses sélectives pour certaines vues (Booth & Rolls, 1998). Ajoutons que les neurones de IT peuvent présenter une forte sensibilité à la position dans le CR, même lorsque celui-ci est de grande taille (DiCarlo & Maunsell, 2003 ; Kline et al., 2003 ; Op de Beeck & Vogels, 2000). Un codage rétinotopique grossier est ainsi préservé dans IT, constituant un argument supplémentaire en faveur de l'hypothèse d'un codage spatial dans IT beaucoup plus important que ne le suggère la description classique du système visuel.

En outre, la réponse à un objet est modulée par la présence d'un second objet et pas simplement supprimée après un certain délai comme le décrit par exemple Chelazzi et al. (1993, 1998). La force de cette modulation dépend de manière non linéaire de la distance entre les deux objets, de leur positions relatives par rapport au centre du CR et aussi de la forme du second objet (Missal et al., 1997, 1999 ; Rolls & Tovee, 1995). Ainsi, le fait que certains neurones de IT aient de grands CR et présentent une relative invariance à la position, n'exclut pas la possibilité d'un codage des relations entre 2 objets et potentiellement plus. D'autres études récentes ont montré que dans l'aire V4, qui projette directement sur IT, les neurones sont sensibles à des indices d'orientation 3D (Hinkle & Connor, 2002), à la distance de fixation (Dobbins et al., 1998 ; Rosenbluth & Allman, 2002) et à la disparité rétinienne (Watanabe et al. 2002). Si ces propriétés étaient également présentes dans IT, ce qui reste à vérifier, cela permettrait de cerner encore plus précisément l'étendue du codage spatial dans cette zone corticale.

Convergence ventrale-dorsale dans les cortex perirhinal et entorhinal

Une autre possibilité de codage spatial dans IT dépend des connexions réciproques qui existent entre IT et le cortex perirhinal et, par l'intermédiaire de ce dernier, entre IT et le cortex entorhinal. Les cortex perirhinal et entorhinal présentent tous deux des réponses sélectives à des objets complexes et à des scènes naturelles, ainsi que des propriétés mnésiques et associatives (Erickson et al., 2000 ; Murray & Richmond, 2001 ; Suzuki et al., 1997). Le cortex perirhinal reçoit de très nombreuses afférences de IT et projette massivement sur le cortex entorhinal. Ce dernier est réciproquement

connecté avec l'hippocampe et reçoit également des informations spatiales en provenance de la voie dorsale par l'intermédiaire de nombreuses connexions avec les aires parahippocampiques (Suzuki et Amaral, 1994). Ces dernières connexions pourraient expliquer pourquoi les neurones du cortex entorhinal sont capables d'encoder une information spatiale relativement fine sur la position d'un objet, bien que cette propriété pourrait aussi, de manière alternative ou complémentaire, être due à des projections du cortex préfrontal (Suzuki et al., 1997). Quoiqu'il en soit, étant donné la vitesse impressionnante de propagation des influx nerveux dans la voie dorsale (Bullier, 2003 ; Nowak & Bullier, 1997), il est tout à fait possible que même dans le cadre d'un modèle vers l'avant de traitement de l'information visuelle, les réponses des neurones de IT (même les toutes premières) soient modulées spatialement par des afférences provenant du cortex entorhinal via le cortex perirhinal. Une telle modulation spatiale pourrait par exemple faciliter la stabilisation des réponses de sous-populations de neurones de IT en autant de solutions locales correspondant à différents objets. Cette possibilité renforce l'hypothèse des MAX locaux développée plus loin.

Le code neuronal

Le code neuronal à propos de l'identité d'un stimulus peut être grossièrement réduit au nombre de potentiels d'action générés et à leur date d'arrivée par rapport au début de la stimulation, bien que nous allons voir qu'il existe d'autres possibilités de codage. Comme nous l'avons déjà vu plus haut, un paramètre de la réponse neuronale peut représenter plusieurs attributs physiques à la fois, (Mel & Fiser, 2000 ; Oram & Foldiak, 1996 ; Wallis & Rolls, 1997). Ainsi, une importante capacité des neurones de la voie ventrale, souvent sous estimée, réside dans l'incorporation dans leur réponse non seulement de la présence de 2 ou plusieurs attributs, mais aussi de leurs relations spatiales (Edelman & Intrator, 2000 ; Elliffe et al., 2002).

Concrètement, Elliffe et al. (2002), dans la lignée de modèles plus anciens de convergences synaptiques (e.g. Barlow, 1972), ont proposé un modèle dans lequel la sélectivité à des configurations spatiales particulières d'éléments, observée dès V1 (voir plus haut), serait préservée le long de la voie ventrale. Les neurones des niveaux supérieurs pourraient devenir sélectifs à des combinaisons spatialement ordonnées

d'éléments codés par des neurones des niveaux inférieurs. Ceci serait possible dans la mesure où à chaque niveau l'apprentissage synaptique est rendu sensible au contenu spatial de l'information afférente, de telle sorte qu'un neurone donné ne répond par exemple que lorsqu'un cercle vert apparaît au-dessus d'un carré bleu, et pas pour d'autres configurations des mêmes éléments. Ceci rendrait obsolète les problèmes de liage perceptif. Un tel mécanisme est tout à fait plausible dans la mesure où il met en œuvre des règles simples d'apprentissage synaptique non supervisé et exploite les régularités statistiques de notre environnement pour construire des représentations de plus en plus complexes. Il y a de bonnes raisons de penser qu'un tel mécanisme existe. En effet, de nombreuses études ont rapporté des réponses sélectives de neurones de IT à des arrangements spatiaux particuliers d'éléments, formant par exemple des visages (e.g. Perrett et al., 1982) et divers types d'objets (Logothetis & Sheinberg, 1996 ; Tanaka, 1996, 2003). Si le liage perceptif n'est plus un véritable problème, une information spatiale suffisante étant en réalité codée dans la voie ventrale, cela laisse beaucoup plus de possibilités pour un codage en parallèle de plusieurs objets.

Cette possibilité se trouve accrue par la nature du codage de l'information visuelle dans la voie ventrale. En effet, la perception de notre environnement semble dépendre d'un « coarse/sparse coding », ou « codage clairsemé », particulièrement bien adapté au traitement simultané de plusieurs objets (Földiák, 2002). Au niveau de la population de neurones, le sparse coding se traduit par le fait qu'à un instant donné, seule une toute petite partie des neurones est activée en réponse à un environnement naturel complexe (Földiák, 2002). Des enregistrements unitaires dans l'aire V1 de macaques ont montré que des stimuli « naturels » augmentent la « sparseness » (faible densité) et l'intensité des réponses neuronales. Ceci serait dû au fait que les stimuli présents dans notre environnement sont très redondants, occupant un ensemble restreint de l'espace des formes et au fait que l'information dans les scènes naturelles a une distribution très inhomogène et de relative faible densité. Le code neuronal pourrait être particulièrement bien adapté pour capturer ces aspects (Barlow, 1972 ; Simoncelli & Olshausen, 2001 ; Touryan & Dan, 2001 ; Reinagel, 2001). Ainsi, un couplage entre des décharges intenses et une activité globale faible augmenterait la quantité d'information par potentiel d'action et l'efficacité par neurone (Vinje & Gallant, 2000, 2002). Dans ce contexte, l'encodage

simultané de plusieurs stimuli serait possible si le nombre d'unités qu'ils activent n'est pas trop important et s'il n'y a pas trop de recouvrement entre leurs représentations, i.e. ils n'activent pas les mêmes unités de manière simultanée (Földiák, 2002). Ceci montre qu'un modèle complet de la perception des scènes naturelles ne peut pas se contenter de prendre en compte seulement les contraintes spatiales, il doit aussi incorporer la contrainte orthogonale d'un espace multi dimensionnel du codage des formes des objets (ce point est développé dans la section suivante).

Une autre possibilité de codage, complémentaire au « sparse coding », pourrait permettre de pallier certaines limitations dues aux recouvrements possibles entre deux représentations d'objets dans l'espace des formes. En effet, si la force de la réponse neuronale est influencée principalement par l'orientation, le degré d'occlusion et la taille du stimulus, la latence des réponses, en revanche, est fortement influencée par le contraste du stimulus (Gawne et al., 1996 ; Oram et al., 2002). Cette dernière pourrait donc servir de signal facilitant le liage perceptif (Gawne et al., 1996), en plus du mécanisme décrit par Elliffe et al. (2002). Les neurones ayant des débuts de réponse proches les uns des autres sont fortement susceptibles de véhiculer de l'information à propos du même objet (Gawne et al., 1996). Un tel mécanisme de reconnaissance d'objet pourrait être implémenté par le biais d'une compétition de type « winner-take-all » (« le gagnant prend tout ») dans le domaine temporel, les premiers neurones qui déchargent inhibant la réponse de ceux qui déchargent plus tard (Thorpe, 1990). Une stratégie similaire, faisant usage des mécanismes très rapides d'inhibition latérale existant dans le système visuel (Swadlow, 2003), pourrait servir à implémenter un codage par rang, selon lequel les neurones sont sensibles à l'ordre de décharge de leurs afférences (Thorpe et al., 2001b ; VanRullen et al., 1998). Ceci constitue une manière concrète d'implémenter une fonction MAX, qui, au niveau computationel, semble tout à fait capable de coder simultanément les représentations de plusieurs objets (Riesenhuber & Poggio, 1999, 2002 ; Rousselet, Thorpe & Fabre-Thorpe, 2003).

L'hypothèse des MAX locaux

Si le système visuel peut, semble t-il, prendre avantage de la dimension temporelle du signal neuronal pour pouvoir coder plusieurs objets simultanément, par

exemple par le biais d'une fonction MAX, cette stratégie semble néanmoins limitée. En effet, même dans un modèle vers l'avant implémentant une fonction MAX, le taux de fausses reconnaissances est relativement élevé lorsque deux objets apparaissent simultanément dans le champ visuel (VanRullen & Thorpe, 1999 ; mais voir Riesenhuber & Poggio, 1999).

Il reste cependant la possibilité, encore inexplorée, que le codage neuronal puisse prendre avantage de l'existence d'une organisation parcellisée de IT et du cortex perirhinal (Erickson et al., 2000 ; Kreiman et al., 2000 ; Tanaka, 1996, 2003 ; Wang et al., 2000). En effet, les zones corticales comprenant des neurones codant pour des propriétés complexes des objets ne forment pas un tout uniforme mais sont organisées en colonnes corticales, c'est-à-dire en groupes de neurones ayant des propriétés de codage relativement similaires mais variant progressivement à mesure que l'on déplace une électrode d'enregistrement dans une colonne. Cette organisation pourrait être la conséquence de la projection de l'espace à plusieurs dimensions des caractéristiques visuelles des objets de notre environnement sur l'espace cortical en deux dimensions de notre système visuel. Cette projection pourrait mettre en jeu des mécanismes auto-associatifs et compétitifs de cartographie corticale, minimisant ainsi les distances entre neurones fortement connectés (Rolls & Deco, 2002). Un tel espace parcellisé, dans lequel les interactions entre neurones, notamment compétitives, s'effectuent essentiellement au sein de chaque colonne corticale (Erickson et al., 2000 ; Wang et al., 2000), rend possible un codage par « prises de décisions » locales. Concrètement, il est tout à fait envisageable que plusieurs fonctions MAX locales puissent co-exister dans la mesure où chacune concerne des populations relativement distinctes dans l'espace des caractéristiques des objets. Cette possibilité est renforcée par l'existence d'un codage relativement peu dense des scènes naturelles (voir plus haut), ceci augmentant, au moins au niveau théorique, la capacité à coder plusieurs objets en même temps.

Le biais fovéal

L'importante capacité de la voie ventrale à encoder des stimuli de manière spatialement ordonnée remet en cause tous les modèles de traitement des scènes naturelles qui postulent que l'information spatiale est perdue au fur et à mesure de la

progression de l'influx nerveux vers les aires de plus haut niveau de la voie ventrale. Cependant, la portée de cette description moderne de la voie ventrale doit être relativisée par l'existence d'un très fort biais fovéal (Rolls & Tovee, 1995 ; Rolls et al., 2003). Lorsque deux objets apparaissent dans le champ visuel, la réponse des neurones de IT est fortement biaisée par le stimulus présent au niveau de la fovéa, puisqu'en général cette réponse véhicule beaucoup plus d'information à propos de cet objet qu'à propos des objets périphériques. Ceci semble être le cas pour les objets apparaissant sur un fond uniforme (Rolls & Tovee, 1995) ou sur un fond texturé naturel (Rolls et al., 2003). Ce biais pourrait s'expliquer par un modèle de IT incluant de forts poids synaptiques au niveau des afférences ayant des CR couvrant la fovéa, à cause de l'important facteur de magnification corticale pour cette région du champ visuel (Trappenberg, et al., 2002).

Ce biais fovéal pourrait fortement limiter la capacité du système visuel à traiter des scènes naturelles et semble constituer un élément en faveur des modèles sériels. Le problème est en réalité plus complexe. Au niveau comportemental, plusieurs études portant sur des tâches de recherche visuelle ont rapporté l'existence d'un tel biais fovéal (Carrasco et al., 1995 ; Wolfe et al., 1998). La découverte tardive de ce phénomène est attribuable à l'absence de contrôle des mouvements oculaires dans la quasi-totalité des expériences de recherche visuelle réalisées dans les années 80. Il apparaît clairement que des contraintes spatiales sont présentes même au niveau des représentations de haut niveau. Cette forte contrainte spatiale suggère que ce qui limite la perception claire et simultanée des objets dans notre champ visuel n'est pas la nécessité de recourir à des mécanismes sériels de liage perceptif, mais plutôt qu'à chaque instant, notre perception est fortement influencée par ce qui se trouve au niveau de la fovéa. Il semble ainsi y avoir une mise en cohérence entre le point de fixation et la zone où se porte par défaut l'attention focalisée (Wolfe et al., 1998). Suivant cette idée, il a aussi été suggéré que notre activité motrice en direction des objets de l'environnement pourrait être guidée par ce qui est présent à la fovéa (Rolls & Tovee, 1995 ; Rolls et al., 2003). Cette hypothèse s'accorde très bien avec une récente proposition selon laquelle la voie dorsale serait plutôt spécialisée dans les ajustements visuo-moteurs rapides et automatiques, alors que la voie ventrale permettrait l'initiation de corrections lentes et volontaires associées à l'identité des objets (Pisella et al., 2000).

Cependant, le biais fovéal ne doit pas être considéré comme absolu. Tout d'abord, la latence à laquelle ce biais se manifeste dans la réponse des neurones de IT n'a jamais été rapportée à ma connaissance. Les résultats décrits par Rolls et ses collaborateurs sont toujours sous la forme de décharges moyennes dans une fenêtre de temps relativement large. Or le décours temporel précis des décharges neuronales est un facteur crucial permettant d'apprécier la capacité du code neuronal à traiter les scènes naturelles en parallèle. De plus, même si le sujet n'est pas encore documenté chez le singe, il apparaît que la difficulté de la tâche pourrait déterminer l'échelle à laquelle les objets d'une scène peuvent être traités. Plus la tâche est compliquée, plus la portion de la scène qui peut être analysée en un temps donné est réduite (Nakayama, 1990). Ceci est en accord avec la proposition selon laquelle une grande partie du champ visuel peut être prise en compte à une faible résolution mais que l'augmentation de la résolution réduit la portion du champ visuel couverte par l'attention (« zoom lens model », Murphy & Eriksen, 1987). Contrairement aux modèles sériels de l'attention dans lesquels une seule zone de l'espace est prise en compte à la fois et où par conséquent les ressources attentionnelles sont distribuées dans un espace en deux dimensions, le modèle du « zoom lens » opère simultanément dans le champ visuel et dans l'espace des fréquences spatiales. En accord avec cette idée, Oliva & Schyns (1997) ont montré que le système visuel peut encoder simultanément deux stimuli superposés composés de fréquences spatiales différentes (basses et hautes) mais que l'attention ne peut se porter que sur une gamme de fréquences à la fois, rendant possible la reconnaissance de la scène représentée par ces fréquences spatiales, l'autre scène n'étant pas traitée jusqu'au stade de la reconnaissance. Selon le modèle développé par Oliva & Schyns (1997, voir aussi Schyns & Oliva, 1997 ; Schyns, 1998), l'attention opère suivant deux dimensions orthogonales, l'une étant la dimension spatiale mise en avant par la plupart des modèles attentionnels, l'autre étant une dimension centrée sur les propriétés diagnostiques des objets. Cette seconde dimension permettrait de présélectionner par exemple l'échelle spatiale qui est la plus prédictive d'un stimulus pour pouvoir le reconnaître. Cette hypothèse suggère l'existence de contraintes autres que spatiales dans le codage simultané de plusieurs objets. Il reste donc à évaluer l'influence de ces différents facteurs sur le comportement des neurones dans IT. Une expérience récente a par exemple mis en évidence que les zones corticales sensibles

à des visages représentent surtout le centre du champ visuel alors que d'autres aires plus sensibles à des maisons représentent des zones plus périphériques de celui-ci (Levy et al., 2001). Il reste à évaluer le comportement de ces aires corticales lorsque deux stimuli (comme deux visages, deux maisons ou bien un visage et une maison) sont présentés simultanément en différentes positions du champ visuel.

Liage perceptif par synchronisation des décharges neuronales

La synchronisation temporelle des réponses neuronales est un mécanisme souvent décrit comme indispensable au liage perceptif, il pourrait donc également contraindre les capacités de codage parallèle de la voie ventrale. Je ne m'attarderai pas sur le sujet car la synchronisation ne saurait constituer, par elle-même, le mécanisme permettant de former des représentations de haut niveau des objets. L'idée est relativement simple (e.g. Gray, 1999 ; Singer, 1999 ; Von der Malsburg, 1999) : si tous les neurones codant pour un objet ont leurs décharges synchronisées et que différents groupes de neurones entrent en synchronisation en même temps, alors plusieurs objets peuvent être codés de manière simultanée. Ainsi, l'activité neuronale associée à un objet étant isolée de celle associée aux autres objets de la scène visuelle, les problèmes d'interférences compétitives, nécessitant selon certains auteurs la mise en jeu de mécanismes sériels de liage perceptif, pourraient potentiellement disparaître. En outre, la synchronisation des réponses neuronales a récemment été impliquée dans des mécanismes de sélection attentionelle en l'absence de modulation du taux de décharge des neurones codant pour un objet cible (Fries et al., 2001). Plusieurs doutes subsistent pourtant quant à l'importance réelle de la synchronisation dans le codage neuronal des scènes naturelles. Au niveau expérimental, le lien entre synchronisation et formation de représentations complexes d'objets est plutôt spéculatif ; il y a aussi de sérieuses limites pratiques et théoriques à l'emploi de la synchronie neuronale pour coder des représentations d'objets (pour plus de détails voir Shadlen & Movshon, 1999). Par exemple, il ne semble pas que l'activité synchrone de neurones véhicule une information permettant la discrimination de plusieurs objets possibles (Panzeri et al., 1999). Il reste donc aux défenseurs de la synchronisation à démontrer qu'elle est porteuse d'une information supplémentaire non présente dans le taux de décharge des neurones. Un autre problème pour la théorie de la synchronisation

émerge d'une simple considération : le codage par synchronisation est potentiellement très puissant, beaucoup plus que le codage utilisé par le système visuel. En effet, von der Malsburg (1999) a démontré qu'un réseau neuronal synthétique à deux couches est capable d'effectuer de la reconnaissance d'objets en utilisant peu de neurones. Au contraire, le système visuel des primates occupe une grande partie de la surface corticale et a une organisation hiérarchique à nombreux niveaux. Il semble donc que la solution adoptée par les systèmes de vision biologiques soit plutôt celle où un grand nombre de neurones communiquent selon un code beaucoup plus simple mais contraint par une organisation hiérarchique et des notions de proximité entre neurones. Pour finir, le codage par synchronie ne permet pas, à lui seul, de spécifier les relations spatiales entre les parties d'un objet : si les parties A et B d'un objet sont associées par synchronisation, le code ne permet pas de savoir si A est au-dessus ou au-dessous de B par exemple. Il ne peut donc pas constituer le mécanisme par lequel le liage perceptif se fait ; par contre il pourrait fournir un mécanisme permettant de maintenir en mémoire de travail la cohérence de représentations de haut niveau une fois formées (Tallon-Baudry & Bertrand, 1999).

Avec un codage par synchronisation, ou dans les modèles sériels, une question reste toujours en suspend. Comment un neurone peut-il savoir que plusieurs stimuli sont présents dans son CR ? Si l'information nécessaire pour déterminer l'existence d'une configuration ambiguë de stimuli est présente, c'est qu'une information spatiale est disponible pour chaque objet et donc que le problème du liage perceptif est déjà au moins partiellement résolu.

3. Attention et conscience

Alors que certains chercheurs mettent en avant des observations expérimentales telles que les pentes de recherche visuelle non-nulles, les conjonctions illusoires, des activités pariétales…, comme autant de *preuves* de la mise en jeu de mécanismes sériels de traitements visuels, une revue détaillée de la littérature permet de montrer qu'il n'existe en réalité aucune preuve tangible en faveur de l'existence de mécanismes sériels. Un nombre croissant d'éléments à la fois théoriques et expérimentaux s'accumulent en faveur de l'hypothèse de la mise en jeu de mécanismes parallèles ayant des ressources

limitées. Il semble qu'une grande partie du débat repose sur un problème de méthodologie. Il paraît impossible de mettre en évidence un traitement inconscient par les seules données comportementales, notamment lors des tâches de recherche visuelle, ce qui implique la nécessaire mise en jeu de l'attention dans de telles tâches. Dès lors, il devient très difficile de faire la part des choses. Comment des variations de TR pourraient-elles être attribuées à des mécanismes de sélection précoces ou tardifs, dans la mesure où l'ensemble de la chaîne de traitement est mise en jeu dans la génération d'une réponse ? Pour conclure ce chapitre, nous allons succinctement tenter d'expliquer pourquoi des mécanismes parallèles peuvent donner lieu à des comportements sériels.

3.1 De l'attention à la conscience

Si l'attention n'est pas suffisante pour entraîner la perception consciente d'un stimulus (Lamme, 2003), il n'en reste pas moins qu'elle est nécessaire à cette perception consciente. Ainsi que le montrent les expériences de masquage (Keysers et al., 2001 ; Rolls et al., 1999), la décharge forte et prolongée de neurones semble être nécessaire pour qu'une perception consciente stable et durable se développe. Aussi, la décharge brève de neurones en réponse à une stimulation visuelle ne semble pas suffisante pour permettre d'engendrer un acte moteur, ni une perception consciente (Dehaene & Naccache, 2001). Etant donné que les neurones visuels présentent des décharges spontanées (i.e. en l'absence de stimulation extérieure), le recours à un signal neuronal fort pour déclencher une action pourrait permettre au système de faire la différence entre le bruit interne et un véritable signal extérieur.

Une erreur importante des modèles sériels du traitement visuel consiste à faire l'amalgame entre deux types de processus, d'une part les mécanismes de segmentation et de construction de la forme, d'autre part les mécanismes de sélection attentionnelle. Le fait qu'il puisse y avoir un traitement de la sémantique d'un objet de manière implicite implique qu'il s'agit bien de deux mécanismes différents. Il est donc important de distinguer traitements implicites et explicites dans les modèles du système visuel.

Au niveau implicite, l'hypothèse de la compétition biaisée (voir plus haut) suggère que des mécanismes neuronaux massivement parallèles sont sous jacents aux phénomènes

d'attention et de perception visuelle. Ici l'idée est simple : ce qui est sélectionné par l'attention contrôle le comportement et la sphère consciente.

Dans un modèle novateur de l'attention visuelle, Deco (Rolls & Deco, 2002) propose d'intégrer l'hypothèse de la compétition biaisée à l'idée selon laquelle V1 pourrait servir d'aire de référence à haute résolution (« high resolution buffer hypothesis »). Ceci permettrait à différentes zones corticales visuelles, de haut ou de bas niveau hiérarchique, d'effectuer des calculs nécessitant des détails fins de l'image et une grande précision spatiale (Bullier, 2001 ; Mumford, 1991). Les réponses précoces des neurones de V1 pourraient être considérées comme des filtres visuels. Par contre, leurs réponses plus tardives pourraient refléter un traitement interactif élaboré impliquant l'ensemble de la hiérarchie visuelle au travers des boucles vers l'avant/vers l'arrière entre les différentes aires corticales. Le principe de base de cette théorie est donc que différentes aires corticales pourraient interagir au travers de V1 dans le but d'effectuer des calculs particuliers demandés par la tâche.

Le modèle attentionnel de Deco montre que V1 peut fournir une représentation topologique appropriée permettant l'interaction des voies ventrale et dorsale dans l'organisation de l'attention. Un nouvel élément introduit dans le modèle de Deco est la possibilité d'une interaction à différents niveaux dans la hiérarchie entre les voies ventrale et dorsale (V1, V2 ou V4 par exemple), en fonction du degré de finesse requis pour la réalisation de la tâche visuelle en cours. Dans ce système, l'attention est une propriété dynamique émergente plutôt qu'un mécanisme indépendant. La dynamique du système, fonctionnant entièrement en parallèle, est telle que ses propriétés temporelles laissent penser à des modes de fonctionnement allant du « sériel » au « parallèle », selon les modèles classiques de l'attention. Les interactions compétitives au sein des aires corticales de la voie ventrale et de la voie dorsale et entre chacune de ces voies, permettent de rendre compte de l'ensemble des propriétés temporelles des tâches de recherche sérielle observées au niveau psychophysique chez le sujet sain comme chez le patient cérébrolésé, mais sans jamais avoir recours à un mécanisme explicite d'exploration sérielle de la scène ou à une carte de saillance perceptive.

Notons que le fait que des mécanismes en apparence sériels lors d'une recherche visuelle puissent s'expliquer par la mise en jeu d'un système à dynamique entièrement

parallèle n'implique pas que l'ensemble des processus attentionnels doivent recevoir la même explication. En effet, les modèles parallèles tels que celui de Deco s'appliquent aux déplacements implicites de l'attention, lorsqu'il n'y a pas de mouvements oculaires. Dans une situation réelle, les yeux bougent, par le biais de saccades, ce qui constitue bien un processus sériel par essence. Néanmoins, l'endroit où s'effectuera la prochaine saccade pourrait être dicté par des mécanismes parallèles de recherche implicite.

Nous pouvons désormais partiellement expliquer ce qui se passe lorsque, avec notre système visuel parallèle, nous recherchons sans succès un trousseau de clés posé sur le bureau et que nos yeux balayent cette scène de manière presque désordonnée. Ainsi, n'importe quel type de recherche visuelle nécessiterait la coopération de mécanismes compétitifs et parallèles et de mécanismes d'attention spatiale focalisée (Rolls & Deco, 2002 ; voir aussi Chelazzi, 1999). Ces mécanismes d'attention spatiale seraient recrutés pour focaliser les ressources de traitement sur un élément pouvant être une cible parmi l'ensemble des éléments qui sont entrés en compétition. Lors d'une recherche d'une cible qui 'pop out', l'accumulation de preuves en faveur de sa présence est très rapide et efficace, alors qu'elle est plus longue et moins efficace pour une cible définie par une conjonction. Dans le premier cas, l'attention focalisée serait rapidement dirigée vers la cible et pourrait être indispensable pour que la cible puisse être perçue et rapportée consciemment. Dans le cas d'une recherche de conjonction, l'attention spatiale servirait à amplifier le résultat du mécanisme compétitif parallèle, c'est-à-dire à augmenter la séparation dans la force de la représentation neuronale entre la cible et les non cibles, permettant ainsi à la cible de prendre le contrôle temporaire à la fois de la sphère consciente et de la sphère comportementale. Le cortex pariétal pourrait participer à cette opération en permettant un déplacement unique de l'attention vers la cible supposée puis en maintenant disponibles des ressources de traitement afin d'accomplir la tâche en cours. Ceci permettrait l'entrée en mémoire de travail de la cible et sa prise en compte au niveau comportemental (idée de Chelazzi, 1999). Le degré d'activation du cortex pariétal pourrait être proportionnel à la quantité d'analyses à effectuer sur l'objet sélectionné. Dans les cas extrêmes où cibles et non cibles sont très similaires ou lorsque les non cibles sont très différentes les unes des autres, l'accumulation de preuves au sein du réseau

compétitif peut être si lente qu'aucun signal clair n'est émis qui permette d'identifier et de localiser la cible. Chaque objet est alors inspecté à son tour.

Au final, la vitesse de traitement d'un système visuel parallèle et compétitif dépend de la sélectivité des neurones dans la voie ventrale au moment où l'on effectue la tâche mais aussi de la tâche en cours et de la disponibilité des informations nécessaires pour l'effectuer. Ce cadre de réflexion met en avant la flexibilité du système visuel.

L'importance du cortex préfrontal

Une autre source de contraintes expliquant la sélection sérielle de stimuli provient de certaines aires corticales frontales. Le modèle de la compétition biaisée (Desimone & Duncan, 1995 ; Rolls & Deco, 2002) postule que le cortex préfrontal biaise la compétition en favorisant les représentations correspondant à la cible (objet ou position spatiale) au sein d'un vaste réseau compétitif mettant en jeu les voies ventrale et dorsale. Il semble également que le cortex préfrontal n'ait pas simplement un rôle de contrôleur à distance mais soit aussi limité intrinsèquement dans ses capacités à décider à propos de plusieurs objets en même temps. Ces limites au niveau décisionnel pourraient fortement contraindre le degré de parallélisme du système visuel à un niveau explicite, celui qui correspond aux réponses comportementales. Le champ oculomoteur frontal (frontal eye field), par exemple, est directement impliqué dans les mécanismes de décision mis en jeu lors d'une tâche de recherche visuelle, allant de l'évaluation physique des stimuli jusqu'à la sélection de la saccade à effectuer pour amener les yeux sur la cible (Schall, 1997 ; Schall & Thompson, 1999). Il a récemment été suggéré que de tels mécanismes frontaux pourraient être à l'origine, par projection en retour, des effets attentionnels observés dans l'aire V4 (Moore & Armstrong, 2003).

De telles limitations au niveau décisionnel peuvent se comprendre facilement d'un point de vue moteur : nous avons deux yeux qui convergent vers un point de l'espace, il ne peut donc y avoir qu'une seule cible définie de manière explicite à la fois. Même en distribuant notre attention dans l'espace, nous n'avons que deux mains pour interagir avec les objets de l'environnement, et chaque fois qu'une tâche précise est entreprise elle requiert la définition spatiale importante seulement disponible au niveau

de la fovéa. Le réseau sensori-moteur de prise de décision se stabilise donc sur une seule solution à la fois dans la plupart des cas.

Cette idée trouve un écho dans l'étude des patients split-brain (à hémisphères séparés par section du corps calleux) réalisée par Luck et al. (1989, 1994). Chez ces patients, chaque hémisphère semble pouvoir travailler indépendamment de l'autre dans une tâche de recherche visuelle, ce qui se traduit par une capacité à trouver une cible deux fois plus importante lorsque les stimuli sont répartis de manière bilatérale que lorsqu'ils sont présents dans un seul hémichamp visuel à la fois. Ceci n'est pas le cas des sujets sains, chez qui la mise en commun des informations visuelles extraites par les deux hémisphères est source d'une forte interférence. Cette interférence pourrait avoir lieu dans le cortex visuel. Cependant, cette hypothèse est peu probable étant donné que des enregistrements unitaires chez le singe montrent que deux stimuli présentés dans deux hémichamps différents n'entrent pas en compétition (Chelazzi et al., 1998). L'hypothèse la plus probable est celle d'une compétition au niveau frontal qui nécessiterait un corps calleux intact. Récemment, Hines et al. (2002) ont montré que le corps calleux est nécessaire pour coordonner les déplacements de l'attention. Chaque hémisphère semble ainsi être doté de ses propres ressources attentionnelles dirigées vers l'hémichamp visuel controlatéral. Une part de l'important « goulet computationnel » souvent attribué à la voie ventrale pourrait donc se situer en réalité au niveau frontal. Les articles 1 et 2 de cette thèse présentent des données en faveur de cette hypothèse.

En conclusion, l'idée simple à retenir ici est que les mécanismes de prise de décision mis en jeu au niveau sensori-moteur et conduisant à des actions par essence sérielles, pourraient contraindre la compétition parallèle au niveau sensoriel et donner l'impression de mécanismes sériels.

3.2 De la conscience à l'attention

Nous venons de voir que ce que certains chercheurs attribuent à des mécanismes sériels de sélection sensorielle d'objets pourrait résulter de la mise en jeu de mécanismes décisionnels (éventuellement frontaux) à un niveau de traitement post-perceptuel permettant la sélection d'actes moteurs appropriés.

Cependant, l'aboutissement de ces processus de sélection n'est pas seulement de générer des commandes motrices mais aussi de permettre l'accès limité de certains stimuli à la sphère consciente. Une durée minimale d'activité neuronale liée au stimulus serait nécessaire pour permettre la mise en jeu de connexions longue distance dans le cerveau, reliant de manière dynamique et temporaire de nombreuses aires corticales, permettant ainsi à la conscience d'émerger (le sujet est largement détaillé dans Damasio, 1999 ; Edelman & Tononi, 2000 ; Dehaene & Naccache, 2001). Il faut envisager que ces mécanismes de la conscience, qui imposent au cerveau de nombreux rythmes endogènes, puissent contraindre à leur tour les mécanismes de décision au niveau sensoriel et au niveau sensori-moteur (idée proposée par Varela pendant la soutenance de thèse de Rufin VanRullen). En effet, la conscience implique l'intégration, pendant une certaine durée, de l'activité d'un vaste réseau d'aires à la fois perceptuelles et motrices. Il est donc concevable qu'un tel réseau ne puisse intégrer des informations, mêmes visuelles, qui excèderaient les capacités de décision motrices.

La problématique s'enrichit donc d'un niveau de complexité supérieur par cette apparente circularité. Tout ceci porte à croire que les capacités de codage en parallèle du système visuel sont sans doute largement sous-estimées et que les principales limites du système se trouvent à un niveau supérieur d'intégration, à la frontière entre le conscient et le non conscient.

Catégorisation rapide de scènes naturelles : travaux antérieurs de l'équipe

Avant de présenter et de discuter les expériences qui font l'objet de cette thèse, voici un résumé succinct des travaux réalisés antérieurement par l'équipe au sein de laquelle j'ai effectué cette thèse. Ces travaux sont discutés plus en détails en différents points du manuscrit. Ce résumé permet de situer le contexte dans lequel mes travaux de thèse se sont déroulés.

En 1996, une expérience princeps a montré que dans une tâche de catégorisation de scènes naturelles utilisant un protocole go/no-go les sujets sont à la fois précis et rapides pour détecter un animal dans une image flashée pendant seulement 20 ms (Thorpe et al., 1996). L'analyse des potentiels évoqués a révélé une activité différentielle entre essais cibles et essais distracteurs commençant dès 150 ms après l'apparition de l'image, suggérant qu'à cette latence assez d'information a été extraite par le système visuel pour commencer à discriminer les cibles des distracteurs. Cette activité différentielle aurait son origine dans les aires occipitales extrastriées (Fize et al., 2000, en révision).

Les travaux réalisés depuis 1996 ont permis de préciser les caractéristiques de ces mécanismes rapides de traitement. Cette tâche go/no-go de catégorisation :

1) peut être réalisée par le singe macaque (Fabre-Thorpe et al., 1998 ; Delorme et al., 2000) ;
2) ne nécessite pas d'informations de couleur (Delorme et al., 2000) ;
3) est très peu perturbée en vision extra fovéale (Fize et al., en révision) ;
4) peut être faite en périphérie lointaine avec des scores atteignant 61% de bonnes réponses à 71% d'excentricité (Thorpe et al., 2001a) ;
5) repose sur des mécanismes optimisés (Fabre-Thorpe et al., 2001) ;
6) n'est pas spécifique des animaux, mais s'applique à des catégories artificielles comme les moyens de transport (VanRullen & Thorpe, 2001a,b).
7) Finalement, l'activité différentielle enregistrée à 150 ms semble refléter un mécanisme de décision liée à la tâche, indépendant des différences physiques entre les images (VanRullen & Thorpe, 2001b).

Ce résumé n'est pas exhaustif mais fournit un point de départ pour comprendre la motivation des expériences réalisées durant cette thèse. Les diverses contraintes mise en évidence, notamment temporelles, ont permis de postuler que les mécanismes impliqués étaient majoritairement parallèles et essentiellement feed-forward. Dans ma thèse, j'ai tenté de cerner les limites de ce parallélisme et d'évaluer la spécificité de certains objets (notamment les visages humains) dans cette tâche de catégorisation rapide.

Article 1

Parallel processing in high-level categorization of natural images

Rousselet, G.A., Fabre-Thorpe, M. & Thorpe, S.J.

Nature Neuroscience 5, 629-630, 2002

Résultats comportementaux et électrophysiologiques de 20 sujets adultes dans une expérience visant à tester l'hypothèse d'un traitement parallèle des scènes naturelles.

L'article est suivi :

* des informations supplémentaires publiées en ligne sur le site de Nature Neuroscience ;
* d'analyses complémentaires non publiées ;
* de la reproduction d'un poster illustrant ce travail présenté à la conférence internationale de la Cognitive Neuroscience Society (CNS meeting) à San Francisco en 2002.

Ce travail a également fait l'objet :

* d'une présentation orale dans un symposium sur le traitement des scènes naturelles lors de la European Conference on Visual Perception (ECVP) en 2001 à Kusadasi en Turquie.
* d'un article en français publié dans un numéro spécial « découvertes » de *La Lettre du neurologue* en 2003.

Introduction

La revue de la littérature présentée au chapitre 1A suggère l'existence de mécanismes massivement parallèles dans la voie ventrale. Notamment, il a été montré que l'analyse des objets dans les scènes naturelles peut être extrêmement rapide (Thorpe et al., 1996 ; Fabre-Thorpe et al., 2001). Or, les scènes naturelles contiennent typiquement plusieurs objets et des fonds riches et texturés. Ceci suggère que les mécanismes mis en jeu dans le traitement rapide des scènes naturelles opèrent en parallèle dans le champ visuel. Une telle hypothèse est soutenue par les résultats de deux expériences montrant que la catégorisation d'une scène naturelle peut

s'effectuer en vision périphérique (Fize et al., en révision ; Thorpe et al., 2001). L'expérience décrite dans ce premier article avait pour objectif de tester directement l'hypothèse d'un traitement des informations en parallèle même lorsque des scènes naturelles différentes en termes d'échelle spatiale, de contenu sémantique, etc., étaient présentées simultanément. Elle consistait à comparer la performance comportementale et les potentiels évoqués chez des sujets humains réalisant une tâche de catégorisation (animal/non-animal) en présence d'une image ou de deux images de scènes naturelles flashées brièvement de part et d'autre d'un point de fixation sur le méridien horizontal. L'anatomie des voies visuelles permet la latéralisation des entrées visuelles, chacune des deux scènes naturelles étant prioritairement traitée par l'hémisphère controlatéral.

Résultats

Les résultats ont révélé trois points importants.

1) Les performances comportementales étaient entièrement compatibles avec un traitement parallèle de deux scènes naturelles. A) Les temps de réaction étaient identiques dans les conditions avec une et deux images. B) La légère baisse de précision observée dans la condition « 2 images » était expliquée par un modèle parallèle de traitement de l'information (présenté dans l'article) dans lequel chaque image est prise en charge par un hémisphère cérébral pour être amenée vers un nœud décisionnel unique (probablement au niveau du cortex frontal).

2) Au niveau électrophysiologique, l'activité différentielle entre les essais cibles et distracteurs, qui nous sert d'index de la vitesse de traitement dans cette tâche, apparaissait à la même latence qu'il y ait une ou deux images, soit 150 ms pour les effets les plus précoces. Ceci était vrai qu'il s'agisse de l'activité enregistrée en regard des électrodes postérieures ou des électrodes frontales.

3) Cependant, les activités occipitales et frontales étaient asymétriques. La latence d'apparition de l'activité frontale était plus tardive et surtout, son amplitude était plus importante dans la condition 1 image que dans celle à 2 images, alors que l'activité occipitale était identique dans les deux cas.

Discussion

Les résultats ont été interprétés dans le cadre d'un modèle attentionnel à sélection tardive : chaque hémisphère pourrait traiter une scène naturelle indépendamment de l'autre, les résultats de leurs analyses étant combinés tardivement afin de prendre une décision motrice. Cette expérience

confirme aussi des travaux antérieurs ayant suggéré que chaque hémisphère pourrait constituer un stock de ressources de traitement indépendantes (Friedman & Campbell Polson, 1981 ; Luck et al., 1989, 1994 ; Sereno & Kosslyn, 1991). Elle est aussi en accord avec des résultats chez le singe montrant que deux objets présentés chacun dans un des hémichamps visuels n'entrent que très peu ou pas du tout en compétition parce que les champs récepteurs des neurones de IT sont essentiellement controlatéraux (Chelazzi et al., 1998). Ces données chez le singe suggèrent aussi que le traitement en parallèle de scènes naturelles pourrait être limité au cas particulier d'une image par hémisphère. C'est ce qu'ont notamment suggéré VanRullen et al. (sous presse), dans un article où ils montrent qu'au niveau comportemental, les scènes naturelles seraient traitées de manière sérielle, puisque le temps de réaction des sujets augmentait avec le nombre de scènes à traiter. Il était donc nécessaire de poursuivre ces travaux sur le parallélisme. Une deuxième expérience est ainsi décrite dans l'article 2 qui avait pour objectif de mieux caractériser les capacités de traitement en parallèle du système visuel lorsqu'il doit faire face à 1, 2 ou 4 images présentées dans des quadrants.

Parallel processing in high-level categorization of natural images

Guillaume A. Rousselet, Michèle Fabre-Thorpe and Simon J. Thorpe

Centre de Recherche Cerveau and Cognition (UMR 5549, CNRS-UPS), Faculté de Médecine de Rangueil, 133 route de Narbonne, 31062 Toulouse, France

Correspondence should be addressed to G.A.R.(guillau@cerco.ups-tlse.fr)

Published online: 28 May 2002, doi:10.1038/nn866

© 2002 Nature Publishing Group http://neurosci.nature.com

Models of visual processing often include an initial parallel stage that is restricted to relatively low-level features, whereas activation of higher-level object descriptions is generally assumed to require attention[1–4]. Here we report that even high-level object representations can be accessed in parallel: in a rapid animal versus non-animal categorization task, both behavioral and electrophysiological data show that human subjects were as fast at responding to two simultaneously presented natural images as they were to a single one. The implication is that even complex natural images can be processed in parallel without the need for sequential focal attention.

High-order representations, up to the semantic level, can be accessed very rapidly from brief picture presentations[5,6]. Event-related potential (ERP) experiments show that complex processing of natural scenes is achieved 150 ms after stimulus onset[7]. Thus, when humans are asked to decide whether a briefly presented photograph contains an animal, the ERPs in response to targets and distractors diverge sharply from 150 ms. There is evidence that these differences reflect a real visual decision rather than physical differences between stimulus categories[8]. The scenes used in such experiments typically contain several objects, suggesting that there is at least some degree of parallelism in the underlying processing. To explore this issue, we analyzed whether processing speed is affected when subjects are asked to process two pictures simultaneously.

Twenty subjects (mean age, 32.5 ± 10.9) performed a modified version of the animal versus non-animal go/no-go task used in previous studies[7,8] (see **Supplementary Fig. 1** and **Supplementary Methods** online). In 20 blocks of 96 trials, single brief presentations (20 ms) of one image appearing 3.6° to the left or right of a central fixation point were randomly mixed with the same number of dual presentations in which two images were flashed simultaneously at the same eccentricities. In both conditions, an animal target was presented on half of the trials. Target location (left versus right hemifield) was equiprobable.

Notably, subjects were able to process dual and single presentations at the same speed (**Fig. 1a**). This is shown by both the median reaction times (RTs, 390 versus 391 ms, respectively) and by the latencies of the earliest responses which were equal or shorter with two images than with one image (means of 255 versus 260 ms, respectively; see **Supplementary Table 1** online).

Subjects tended to be more accurate in the one-image condition (90.4%) than with dual images (86.7%). This accuracy decrease was predicted by a simple parallel model of processing (**Fig. 1b**) in which each of two simultaneously presented images is processed by a separate and independent mechanism, and both mechanisms eventually converge on a single output system (see **Supplementary Methods**). Further support for a parallel processing model comes from the tight fit between the experimental and the predicted cumulative performance accuracy (*d'*) curves (**Fig. 1b**).

The similarity in processing speed between the two conditions was confirmed by electrophysiological data (**Fig. 2**). Associated ERPs were averaged off-line for each condition and difference waves were obtained by subtracting the ERP for correct distractor trials from the ERP for correct target trials. Differential activation, probably generated within high-order extrastriate visual areas[9], was clearly seen at both occipito-temporal and frontal sites (see **Supplementary Fig. 2** online). There was no effect of image condition on the onset of the differential occipital activity. Target and distractor signals diverged sharply around 140–150 ms after stimulus onset with an enhanced occipital negativity on target trials. This differential occipital activity became significant at similar latencies in both

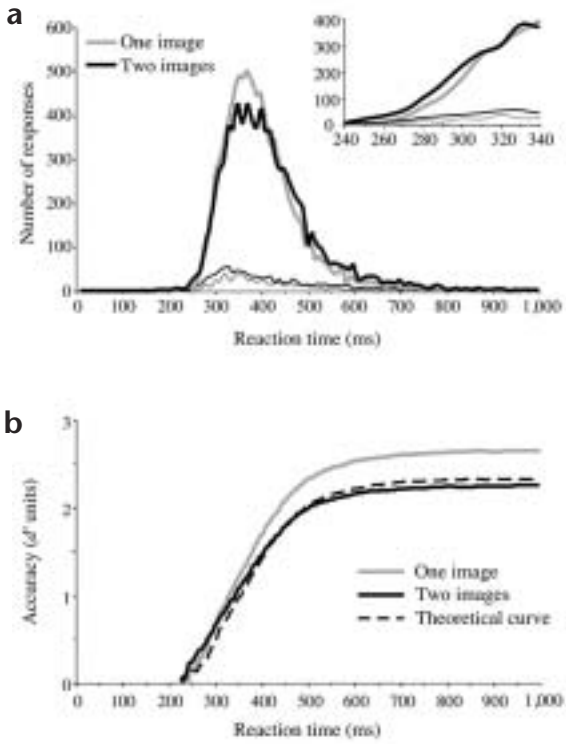


Fig. 1. Behavioral results. (**a**) Reaction time distributions. Number of responses are expressed over time, with time bins of 5 ms. Correct responses or 'hits' (thick top curves) are shown for the one target alone (gray) or for the target flanked by distractor (black). False alarms (thin bottom curves) are shown for the one distractor alone (gray) or for the two distractors (black). (**b**) Performance time course functions and predictions of a parallel model of processing. Average performance accuracy (in *d'* units) is plotted as a function of processing time (in ms) for one image (gray curve) and for two images (black curve). The dynamic *d'* was calculated from the cumulative number of hits and false alarms at each successive 10 ms time step. The predicted curve from the model was calculated using the probabilities of hits and false alarms calculated from the experimental data in the one-image condition. A global fall in accuracy from 90.4% in the one-image condition to 87.7% in the two-image condition was predicted by our model (see **Supplementary Methods**). The experimental procedures were authorized by the local ethical committee (CCPPRB No. 9614003) and all subjects gave informed consent to participate.

© 2002 Nature Publishing Group http://neurosci.nature.com

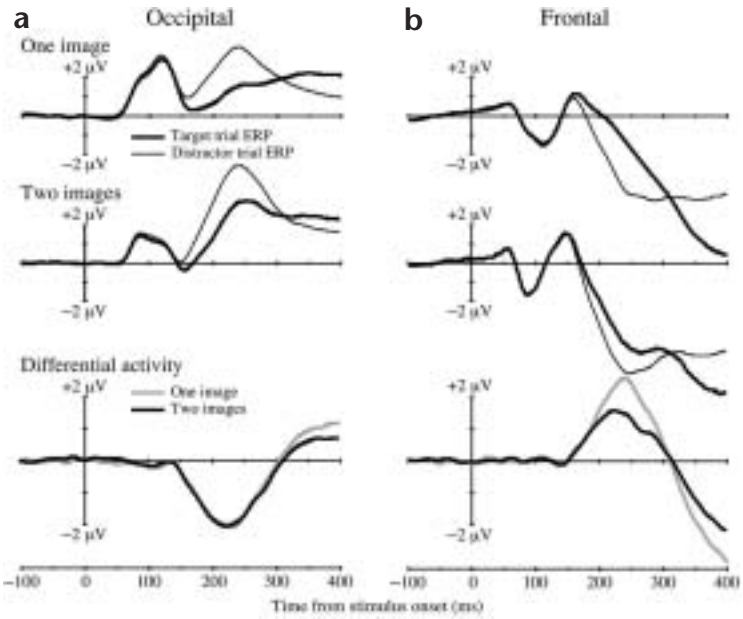


Fig. 2. Grand average ERPs and associated differential activities. Grand average ERPs are plotted for correct target trials (thick line) and for correct distractor trials (thin line). Results are shown for contralateral occipital electrodes (**a**) and all frontal electrodes (**b**), for the one-image (top panel) and the two-image (middle panel) conditions. Bottom, differential activity between the one- (gray) and two-image (black) conditions.

conditions (152 ms with one image versus 150 ms with two images, $P < 0.0005$) and then developed at the same rate, with the same slope and amplitude in both conditions.

Differential activity was also seen at frontal sites (**Fig. 2b**), starting at around 160–170 ms in both conditions and becoming significant ($P < 0.0005$) at about the same latency: 173 ms (one image) and 175 ms (two images). At 190 ms after stimulus onset, the differential activity recorded in the one-image condition began to diverge from that of the two-image condition, developing with a steeper slope and finally reaching a higher amplitude.

These behavioral and electrophysiological results provide strong evidence that processing speed is unchanged between the one- and two-image conditions. Furthermore, the slight accuracy impairment (<4%) with two images can be explained using a very simple model in which the two images are processed by separate mechanisms that pool their outputs. The brief image presentations and initial lateralization of visual inputs to the contralateral striate visual cortex indicate that each hemisphere could work in parallel on a different visual scene. This interpretation is strengthened by the high lateralization of the differential occipital activity.

The RT distributions (**Fig. 1a**) show that the number of 'go' responses in the two-image condition, although initially similar to that seen in the one-image condition, was considerably lower around the mean RTs. This effect might be explained by some form of competitive process occurring in the two-image condition. Given the strong similarity between the occipito-temporal differential activity in the two conditions (**Fig. 2a**), it seems unlikely that this competition affects the initial visual processing. Competition is more likely to occur later on at the point of 'sensorimotor decision'[10]. Evidence for a late competitive process at frontal sites comes from the late divergence seen between the one- and two-image conditions after 190 ms. High-level representations in occipito-temporal visual areas would be activated independently in each hemisphere. At frontal sites, by contrast, when integration of the outputs of the two cerebral hemispheres is needed for decision-making, competition could result from frontal processes related either to category-specific decision-making[11] or to response inhibition on no-go trials[12].

These physiological results directly support psychological models in which the competitive bottleneck is situated at a high level of integration[13,14]. Further support comes from behavioral findings that show that this animal versus non-animal categorization task can be done simultaneously with another attentionally demanding task[15]. We have no evidence, however, to suggest that the ability to process more than one image at the same time extends beyond the specific case of two images presented simultaneously in the two hemifields. Further experiments will be required to explicitly test whether the system can process other stimulus arrangements in parallel, such as two images presented within the same hemifield or four images presented simultaneously.

Taken together, our data show that high-level object categorization of natural scenes can be done in parallel very rapidly and without the need for sequential focal attention. Whereas classic models of allocation of attentional resources consider 'early' vision as being early in complexity and restrict low-level vision to the lower part of the cortical hierarchy (namely V1 and V2), early vision might more appropriately be considered as processing that is early in time.

Note: Supplementary information is available on the Nature Neuroscience website.

Acknowledgments

This work was supported by the Cognitique program (COG35 and 35b). Financial support was provided to G.A.R. by a PhD grant from the French Government.

Competing interests statement

The authors declare that they have no competing financial interests.

RECEIVED 24 JANUARY; ACCEPTED 29 APRIL 2002

1. Treisman, A. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **353**, 1295–1306 (1998).
2. Wolfe, J. M. *Vis. Res.* **34**, 1187–1195 (1994).
3. Kinchla, R. A. *Annu. Rev. Psychol.* **43**, 711–742 (1992).
4. McElree, B. & Carrasco, M. *J. Exp. Psychol. Hum. Percept. Perform.* **25**, 1517–1539 (1999).
5. Potter, M. C. *J. Exp. Psychol. [Hum. Learn.]* **2**, 509–522 (1976).
6. Biederman, I. *Science* **177**, 77–80 (1972).
7. Thorpe, S., Fize, D. & Marlot, C. *Nature* **381**, 520–522 (1996).
8. VanRullen, R. & Thorpe, S. J. *J. Cogn. Neurosci.* **13**, 454–461 (2001).
9. Fize, D. *et al. Neuroimage* **11**, 634–643 (2000).
10. Schall, J. D. *Nat. Rev. Neurosci.* **2**, 33–42 (2001).
11. Freedman, D. J., Riesenhuber, M., Poggio, T. & Miller, E. K. *Science* **291**, 312–316 (2001).
12. Sasaki, K., Gemba, H., Nambu, A. & Matsuzaki, R. *Neurosci. Res.* **18**, 249–252 (1993).
13. Duncan, J. *Psychol. Rev.* **87**, 272–300 (1980).
14. Chun, M. M. & Potter, M. C. *J. Exp. Psychol. Hum. Percept. Perform.* **21**, 109–127 (1995).
15. Li, F. F., VanRullen, R., Koch, C. & Perona, P. *Proc. Natl. Acad. Sci. USA* (in press).

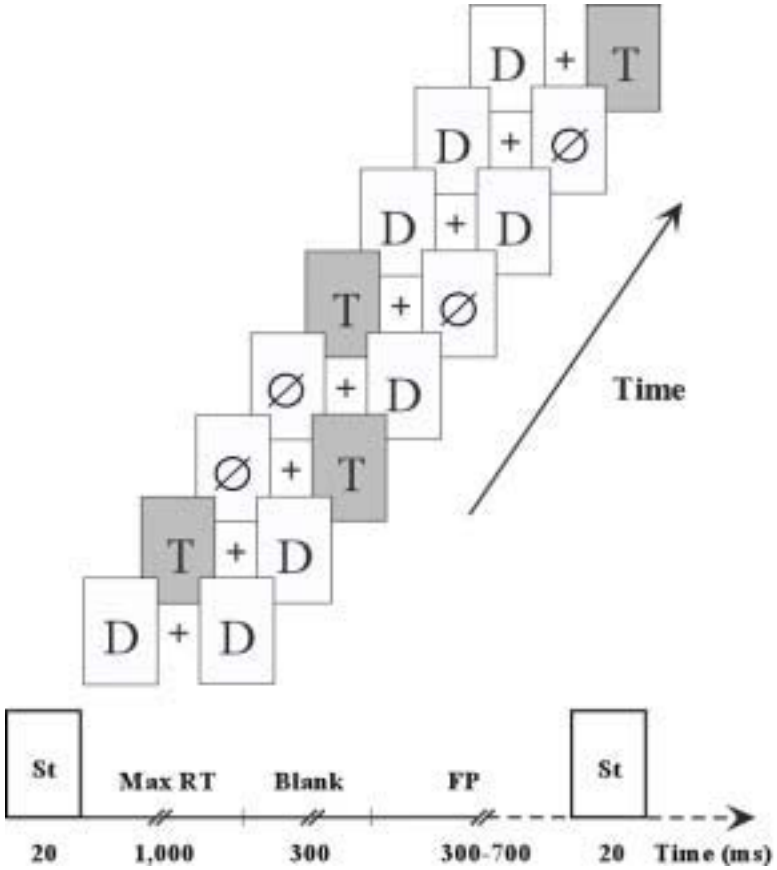
Supplementary Methods

Subjects. 20 volunteers (age range 22-54, mean 32, 10 males and 10 females) participated in this study. All subjects had normal or corrected to normal vision.

Stimuli. We used photographs of natural scenes taken from a large commercial CD-ROM library (Corel Stock Photo Library) as stimuli. From this data bank, 1920 distractors and 960 targets were selected. Each was seen by every subject but randomly distributed across all series with all conditions counterbalanced to avoid any bias. Vertical photographs (384 by 256 pixels, sustaining about 8.4° by 5.6° of visual angle) were chosen to be as varied as possible. Animals included mammals, birds, fish, insects and reptiles... There was no a priori information on the size, position, or number of targets in any particular photograph. There was also a very wide range of distractor images that included outdoor or indoor scenes, natural landscapes (mountains, fields, forests, beaches...) and street scenes, pictures of food, fruits, vegetables or plants, buildings, tools or other man-made objects...

Task and set-up. Subjects were sat in a dimly lit room at 120 cm from a computer screen (resolution: 800 x 600, vertical refresh rate: 75 Hz) piloted from a PC computer. To start a block of trials, they had to place their finger on a response pad for one second. A trial was organized as follows (see supplementary Fig. 1). A stimulus (one or two photographs) was presented for two frames, i.e. 20 ms. Participants had to raise their finger as quickly and as accurately as possible (go response) each time an animal was present. Responses were detected using infrared diodes. Subjects had 1000 ms to respond, after which delay their response was considered as a no-go response. This maximum response time delay was followed by a 300 ms black screen, then by a 300-700 ms fixation point (0.1° of visual angle), resulting in a random 1600-2000 ms intertrial interval. When the photographs contained no animal, subjects had to keep their finger on the pad for at least 1000 ms (no-go response). An experimental session included 20 blocks of 96 trials in which target and distractor trials were equally likely. To prevent learning, each image was only seen once by each subject. Half of the stimuli contained one picture, the other half two pictures. When one picture was presented, it could appear on the left or the right of the fixation point (centered at 3.6° eccentricity) and be either a target (T) or a distractor (D) with no image (∅) on the other side. When two pictures were presented, there were either two distractors or a target and a distractor. Each block contained

12 trials for each of the 6 conditions: DT, TD, ∅T, T∅, ∅D, D∅; the last condition: DD was twice as common. The design was counterbalanced so that overall, each of the images was presented the same number of times on the left and on the right, and the same number of times in the one and the two image conditions.



Supplementary Fig. 1. Task and set-up. St, stimulus; Max RT, maximum reaction time delay; FP, fixation point. See Supplementary Methods for details.

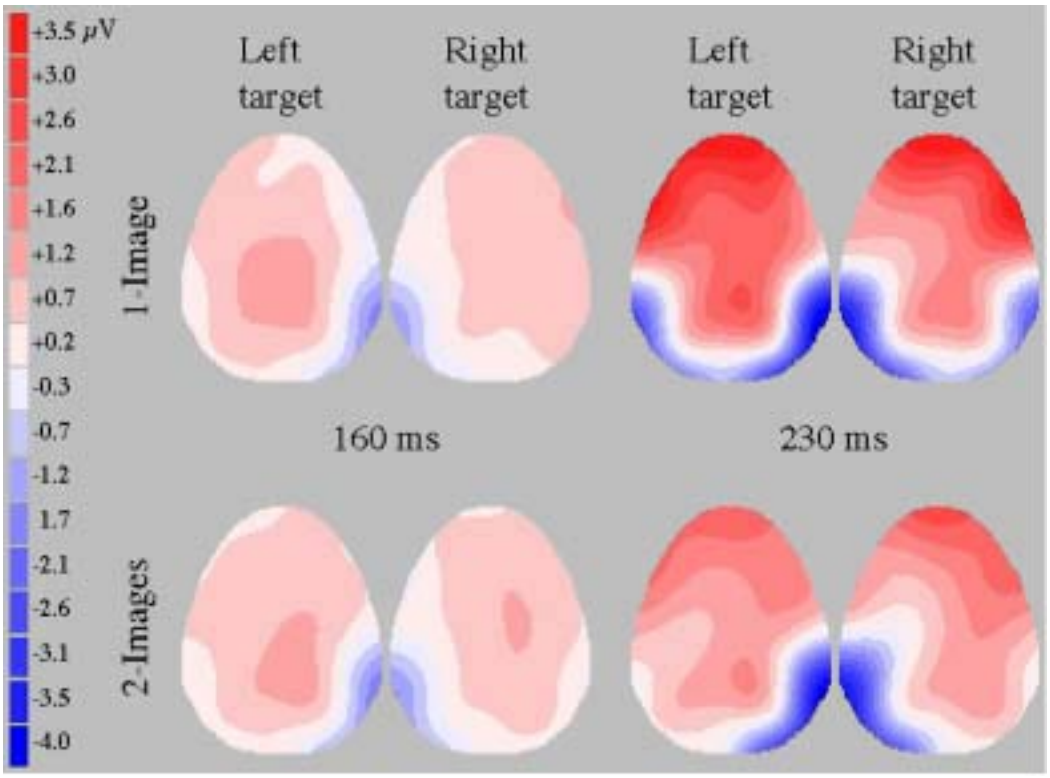
Predictions from a parallel model of processing.

In the two image condition, a simple model of parallel processing postulates that each of the two simultaneously presented images is processed by a separate and independent mechanism working with the level of performance reached in the one image condition and that the two outputs are then pooled together. When a single image is processed, the probabilities of hits and false alarms can be calculated from the experimental data: p (Hit) = 0.88 and p (FA) = 0.07. Thus, a correct response on a distractor trial (no-goDD) is only obtained when both distractors are correctly ignored: no-goDD = $(1\text{-p (FA)})^2$; expected

value: $(1\text{-}0.07)^2 = 0.865$. For target trials, a correct response (goTD) is produced either by a hit in response to the target or by a false alarm to the simultaneously presented distractor: goTD = 1 – (1 – p (Hit)) x (1 – p (FA)); expected value: 1 – (1-0.88) x (1-0.07) = 0.888. As target and distractor trials are equiprobable, the overall probability of correct responses if both images are processed in parallel should be = (NoGoDD + GoTD) / 2 (expected value: (0.865+0.888) / 2 = 0.877. A global fall of accuracy from 90.4% in the one image condition to 87.7% in the two image condition is predicted by this simple parallel model. The expected result (87.7%) is very close to the observed value (86.7%). Moreover, the same model can also be applied to evaluate performance over time at each successive 10 ms time step to generate the expected *d′* curve shown in **Fig. 1b**. The *d′* was calculated from the formula $d' = z_n - z_s$, where z_n is chosen such that the area of the normal distribution above that value is equal to the false-alarm rate, and where z_s is chosen to match the hit rate. The very close fit between the observed and predicted *d′* curves for the two image results again supports the hypothesis that a parallel processing model can explain our experimental data.

EEG analysis. Electrical activity was recorded from 32 electrodes mounted in an elastic cap in accordance with the 10-20 system (Oxford Instruments) with the addition of extra occipital electrodes and using a Synamps amplifier system (Neuroscan Inc.). The ground electrode was placed along the midline, ahead of Fz. Impedances were systematically kept below 5 kΩ. Signals were digitized at a sampling rate of 1000 Hz (corresponding to a sample bin of 1 ms) and low-pass filtered at 100 Hz with a notch at 50 Hz. Potentials were on-line referenced on electrode Cz and re-referenced off-line by subtracting the average of all signals from each signal. Baseline correction was performed using the 100 ms of pre-stimulus activity. Two artifact rejections were applied over the [-100 ms; +400 ms] time period: on frontal electrodes with a criterion of [-50; +50 μV] to reject trials with eye movements, and on parietal electrodes with a criterion of [-30; +30 μV] to remove trials with excessive activity in the alpha range. Only correct trials were averaged. ERPs were low-pass filtered at 40 Hz before analysis. Four differential activities were computed by subtracting evoked potentials for distractors from evoked potentials for targets: DT - DD, TD - DD, ∅T - ∅D, T∅ - D∅. Analysis concentrated on four groups of electrodes: occipital (left: T5, O1, O1', CB1, CB1'; right: T6, O2, O2', CB2, CB2') and frontal (left: FP1, F3, F7; right: FP2, F4, F8) where the differential activity was clearest.

Latency of the differential activity. To evaluate the start of a differential activity effect, paired *t*-tests (19 degrees of freedom) were performed at the $p < 0.0005$ level to compare the responses on target and distractor trials. The presence of 15 successive significant *t* values was used to index a differential effect (for similar strategies see Rugg, M., Doyle, M. & Wells, T. *J. Cogn. Neurosci.* **7**, 209-227, 1995; Thorpe S.J., Fize, D., Marlot, C., *Nature* **381**, 520-522, 1996).



Supplementary Fig. 2. Cartography of the differential activity. Interpolated maps (frontal lobe at the top) of the differential activity (DA) are shown for the one image (top) and the two image (bottom) conditions at two different latencies: 160 ms (left) and 230 ms (right) after stimulus onset. In the two image condition, the target is always flanked by a distractor in the opposite hemifield. For each condition and each latency, the left (right) map corresponds to the DA obtained with a target in the left (right) hemifield. At 160 ms, the negative going occipital differential activity is strongly lateralized over contralateral electrode sites in both experimental conditions, while the positive going frontal differential activity is more widely distributed. Close to the peak of the differential activity (230 ms), the occipito-temporal differential activity remains clearly lateralized over contralateral electrode sites in the two image condition. The positive going activity peaked over all frontal electrodes, with higher amplitude in the one image condition compared to the two image condition.

		1-Image	2-Images	Statistical test	*p*
Behavior	**Reaction Time (ms)**				
	Mean	410 ± 10	413 ± 11	$F(1, 19) = 2.9$	.11
	Median	391 ± 10	390 ± 11	$F(1, 19) = .05$	.83
	Minimal processing time (ms)				
	Mean on Individual data	330 ± 7	330 ± 9	$F(1, 19) = .00$	1.0
	Mean on Overall data	260	255	$\chi 2$, df = 1	< .001
	Accuracy (%)				
	Overall	90.4 ± .6	86.7 ± .7	$F(1, 19) = 96.9$	< .0001
	Targets	87.8 ± 1.5	83.7 ± 1.7	$F(1,19) = 45.4$	< .0001
	Distractors	93.1 ± .7	89.7 ± 1.1	$F(1,19) = 63.4$	< .0001
Differential Activity	**Latency (ms)**				
	Occipital	152	150	paired t-test, df = 19	< .0005
	Frontal	173	175	paired t-test, df = 19	< .0005
	Peak Amplitude (μV)				
	Occipital	-3.30 ± .15	-3.37 ± .19	$F(1,19) = .46$	.51
	Frontal	4.31 ± .27	3.0 ± .27	$F(1,19) = 32.5$	< .0001
	Peak Latency (ms)				
	Occipital	227 ± 4.8	230 ± 4.5	$F(1,19) = .51$	.48
	Frontal	249 ± 4.9	241 ± 7.2	$F(1,19) = 2.4$	.14

Supplementary Table 1. Behavioral, electrophysiological data and statistical analysis. For each parameter, values are given with their standard error on the mean for the one image and the two image conditions. The two right columns indicate the statistical test used and the probability reached (*p*). ANOVAs (*F* are given) were used to compare the data obtained in the two different conditions. The minimal -behavioral-processing time is defined as the latency at which correct go-responses start to significantly outnumber incorrect go-responses in the RT histogram. For the differential activity, the peak amplitude and the peak latency data were entered into ANOVAs with number of pictures as a within-subject factor. Only the peak of the highest amplitude was taken into account for each region and for each subject.

Parallel processing in high-level categorization of natural images: non published complementary analyses

1. Detailed behavioral statistics

1.1 Left versus right image presentations

Left and right target images were processed with the same overall accuracy (89% in both cases). In the 1-image condition, the same pattern applied to left and right distractor images (93% in both cases). Erroneous responses made toward left and right distractors were triggered at approximately the same latency (respectively 433 ms vs. 439 ms, *n.s.*). However, there was a small but significant RT advantage for right targets over left targets (mean RT: left image = 416 ms, right image = 407 ms, $p < 0.0001$; median RT: left image = 394, right image = 387 ms, $p < 0.0001$). This effect is barely visible in the RT histogram below (Figure 1).

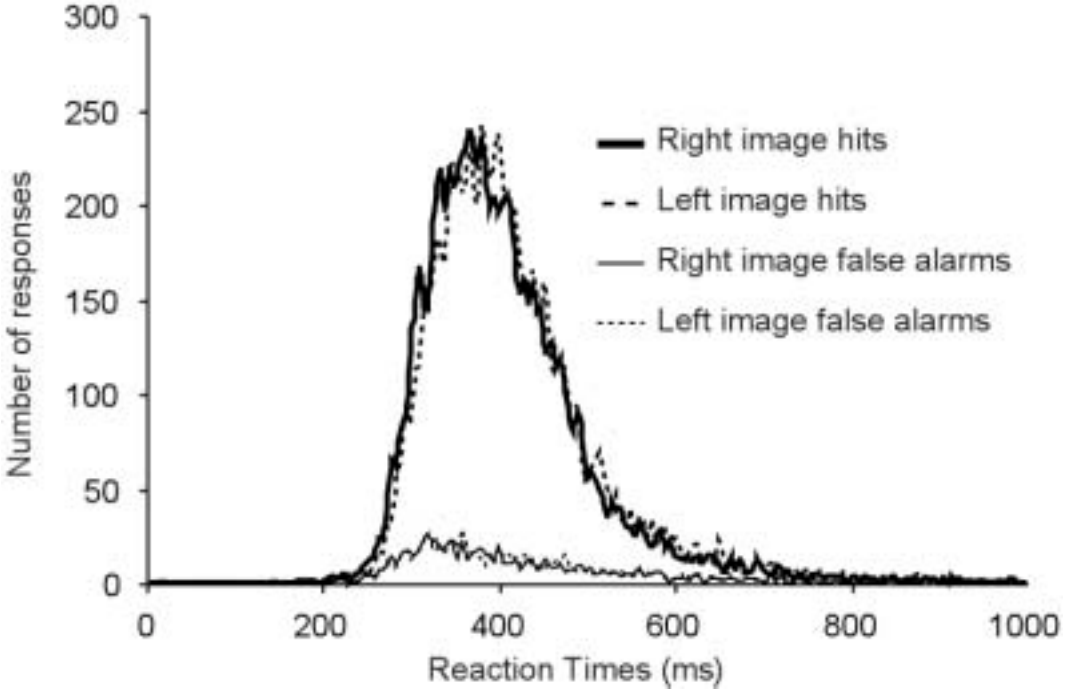


Figure 1. Distributions of reaction times associated with responses toward left and right images. Data from the 1-image and the 2-image conditions have been collapsed.

This advantage was not present in the early part of the RT distribution as indicated by similar minimum RT (left image = 332 ms, right image = 329 ms, *n.s.*). It might be explained by a stimulus-response compatibility effect as most subjects were right handed. This is confirmed by the fact that a small RT effect in favor of right targets was also seen while subjects performed a simple image detection task (see below the control behavioral experiment).

1.2 Targets versus distractors

Accuracy

Distractors were better categorized than targets (respectively 91 % vs. 86 %, p = .017), be it in the 1- or the 2-image condition (1-image: 93% vs. 88%; 2-images: 90% vs. 84%). The accuracy drop in the 2-image condition compared to the 1-image condition affected distractors and targets to the same extend (main effect of the image factor: 1-Image = 90 % vs. 2-Images = 87 %, p < .0001, no interactions between status and image factors).

Mean reaction times

Correct responses toward targets were triggered faster than erroneous responses toward distractors (respectively 411 ms vs. 435 ms, p = 0.002). This was true with one and two images (1-image: 410 ms vs. 433 ms; 2-images: 413 ms vs. 437 ms).

2. Control behavioral experiment

This control experiment was designed to find an explanation to the slightly shorter but non significantly different minimal RT observed at the population level in the 2-image condition compared to the 1-image condition (255 ms vs. 260 ms). This difference can be observed as a small shift toward shorter latencies of the 2-image RT distribution compared to the 1-image RT distribution in Figure 1b of the original paper. This shift was hypothesized to be due to higher stimulus energy in the 2-image condition compared to the 1-image condition. Twenty subjects (9 of whom had participated in the main study) were tested on a simple detection task of 1 or 2 pictures. They had to respond each time an image appeared on the screen, irrespective of the nature of the image. They performed four series of 96 trials. The stimulation design was the same as in the main experiment. On average, they tended to respond 10 ms faster in the 2-image condition than in the 1-image condition (202 ms vs. 211 ms, respectively, $F = 49.338$, $p < 0.0001$), a result that could explain the leftward RT distribution shift seen in the main experiment with the 2-image condition (Figure 2). Right images tended to elicit slightly shorter RT than left images (respectively, 209 ms vs. 213 ms, p = .012). This result strengthens the idea that the small RT laterality effect found in the main experiment could find its origin in non task-specific constrains.

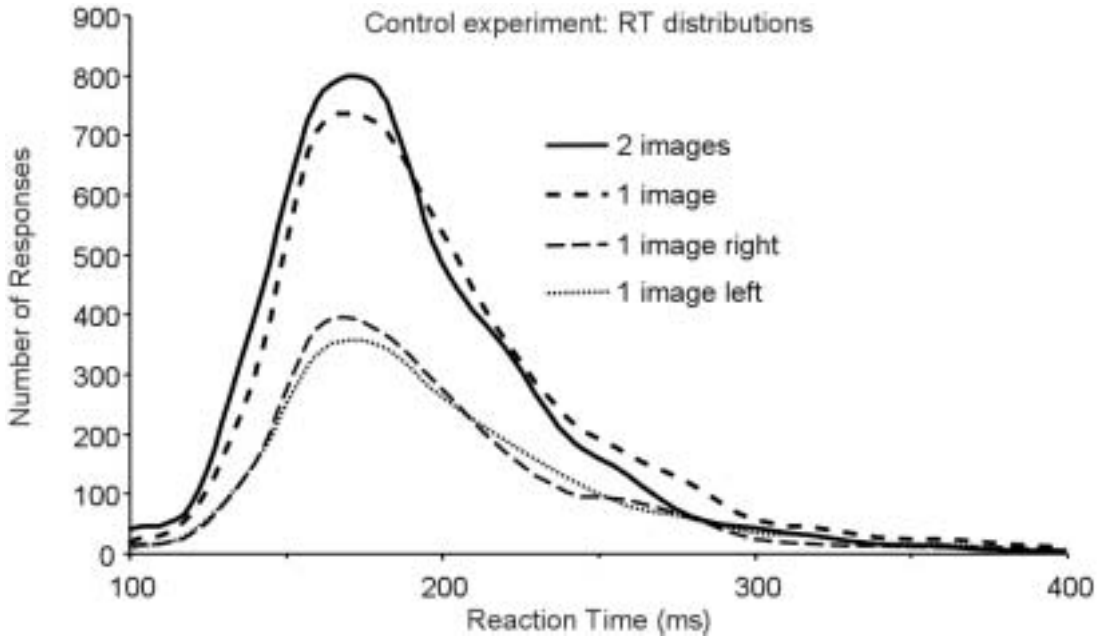


Figure 2. Reaction time distributions in the control experiment.

3. ERP peak analysis – data from experiment 1

3.1 Overall data

On the graph below (Figure 3, see also Figure 5) are plotted occipito-temporal ERP components recorded on the lateral sites T5 & T6. Two components are clearly visible: a first component characterized by a positive deflection peaking around 100 ms (P1) and a second component characterized by a negative deflection peaking around 150-170 ms (N1). Differential activity between correct target ERP and correct distractor ERP starts to emerge at around 150 ms over contralateral occipital electrodes in both image conditions (1 or 2 images). Over ipsilateral sites, the differential activity appears at about 200 ms when 2 images were presented, while no early ERP components (P1 N1) were present when only 1 image was presented. Note the large differences in ERP component amplitudes between the two image conditions, starting as early as 50 ms after stimulus appearance over ipsilateral sites. The sections below provide a statistical analysis of the latencies and amplitudes of P1 and N1 in target present trials. These two components were measured at the posterior sites where they were largest, namely T5-T6, O1'-O2', CB1-CB2. For each component and each hemisphere, the latency was measured from the electrode at which the largest signal was recorded. The amplitude was then measured at that latency for the remaining electrodes, following the ERP guideline (Picton et al., 2000).

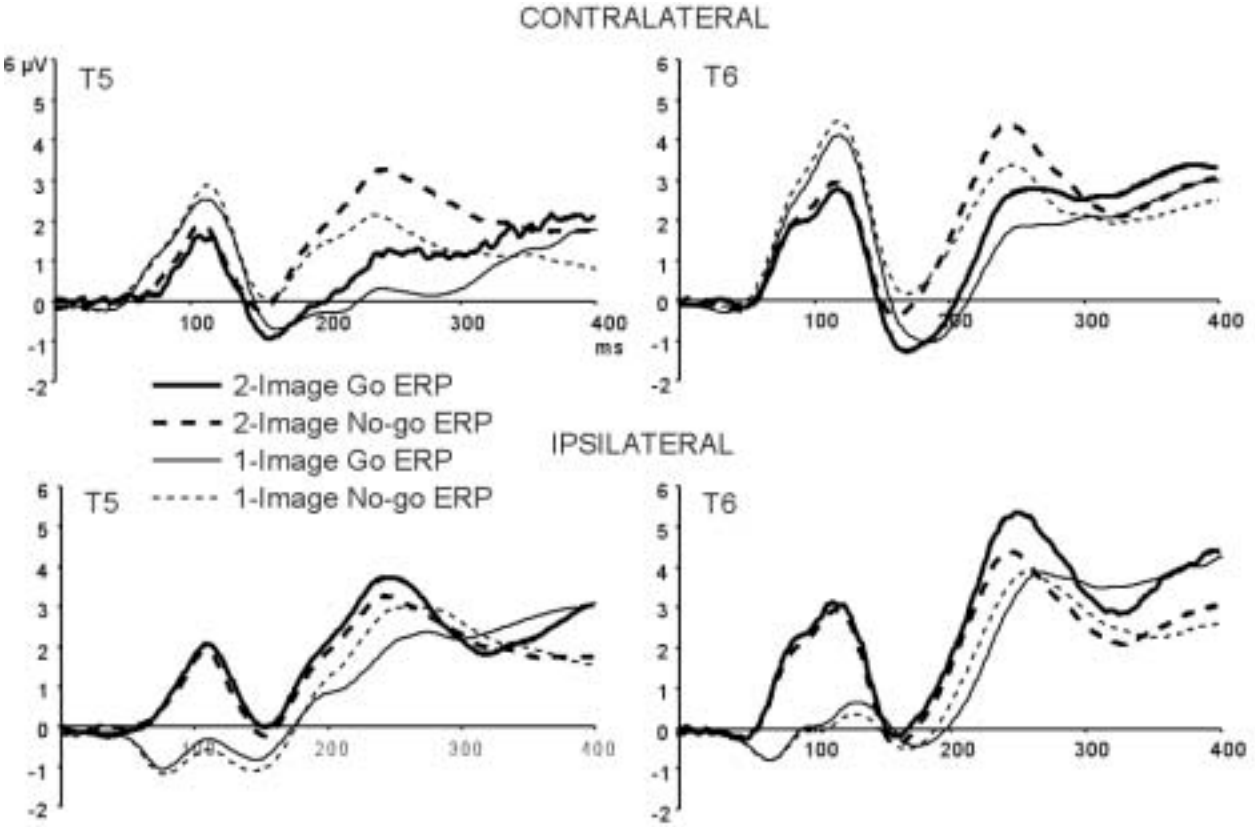


Figure 3. ERP components recorded over temporal electrodes ipsi- and contralateral to the target in the two image conditions. These target potentials are compared to distractor potentials. In the 2-image conditions, the same potentials are used for comparison in the ipsi- and contralateral graphs.

P1 Latency

P1 latency was much shorter in the 2-image compared to the 1-image (respectively 105 ms and 124 ms, $p < 0.0001$; contralateral sites alone: 106 ms and 115 ms, $p = 0.01$), probably reflecting the stronger energy in trials with 2 images. Left and right targets were associated with identical P1 latencies (115 ms and 114 ms respectively). Latencies recorded on the left (114 ms) and right (116 ms) hemispheres did not differ either. Finally, in the 1-image condition, P1 latency was considerably shorter for sites contralateral to the target (115 ms) compared to those ipsilateral to the target (133 ms) as shown by an interaction between the image, hemifield and hemisphere factors ($p < 0.0001$).

P1 amplitude

P1 amplitude did not differ significantly between conditions (1 image = 3.9 µV, 2 images = 4.1 µV), hemifields (left target = 4.1 µV, right target = 4.0 µV) and hemispheres (left = 3.8 µV, right = 4.3 µV). Although, when contralateral sites were tested separately, the signal in the 1-image condition was significantly larger than that in the 2-image conditions (respectively, 5.3 µV and 4.0 µV, $p < 0.0001$). Finally, in the 1-image condition, P1 amplitude was larger

for sites contralateral to the target (5.3 μV) compared to those ipsilateral to the target (2.6 μV) as shown by an interaction between the image, hemifield and hemisphere factors ($p < 0.0001$).

N1 latency

Like P1, the N1 component peaked earlier in the 2-image condition compared to the 1-image condition (respectively, 150 ms and 178 ms, $p < 0.0001$; contralateral sites alone: 152 ms and 175 ms, $p < 0.0001$). With ipsi- contralateral sites considered altogether, there was no difference between left (164 ms) and right (163 ms) targets. However, when contralateral sites were analyzed separately, there was a significant bias in favor of right targets/left hemisphere (159 ms) compared to left targets/right hemisphere (168 ms, $p = 0.007$).

N1 amplitude

N1 was larger after the presentation of 2 images than 1 image (2 images = -2.7 μV, 1 images = -1.9 μV, $p = 0.03$) which was also true with contralateral sites tested separately (2 images = -3.0 μV, 1 image = -2.0 μV, $p = 0.02$). This effect is opposite to the one found at the level of the P1. It would be interesting to disentangle this point, but it is hard to draw any conclusion at the moment. N1 peak amplitude, be it in the case of ipsi- and contralateral signals, was not affected by hemifield and hemisphere factors (left target = -2.3 μV, right target = -2.2 μV; left hemisphere = -2.1 μV, right hemisphere = -2.5 μV).

Overall, the stimulation with 2 images add a strong impact on the P1 and N1 ERP components. From these analyses it was not possible to tease apart the low level effects due to the addition of 1 image, from the task related ones. Differential activities were thought to largely isolate task related factors, eliminating large low level differences between trials with 1 and 2 images.

4. Differential activity peak data analysis

Large differential activities were recorded at various electrodes sites with occipital, frontal and parietal topographies (Figure 5).

4.1 Occipital differential activity

Here we consider only the activity recorded over electrodes contralateral to the targets, given that the differential activity was highly lateralized.

Latency

The peak latency of the differential activity did not vary significantly with 1 image (227 ms) compared to 2 images (230 ms), or with left targets (228 ms) compared to right targets (228 ms) (Figure 4).

Amplitude

The peak amplitude of the differential activity did not vary significantly with 1 image (-3.3 μV) compared to 2 images (-3.4 μV), or with left targets (-3.4 μV) compared to right targets (-3.3 μV) (Figure 4).

4.2 Frontal differential activity

Latency

The frontal differential activity peaked at about the same latency in the 1-image (249 ms) and the 2-image (241 ms) conditions. Left and right targets were associated with peak of differential activities that did not differed significantly from one another (left = 246 ms, right = 244 ms). The peak latency of the frontal differential activity was on average shorter in the left than the right hemisphere (left = 242 ms, right = 248 ms, p = 0.01).

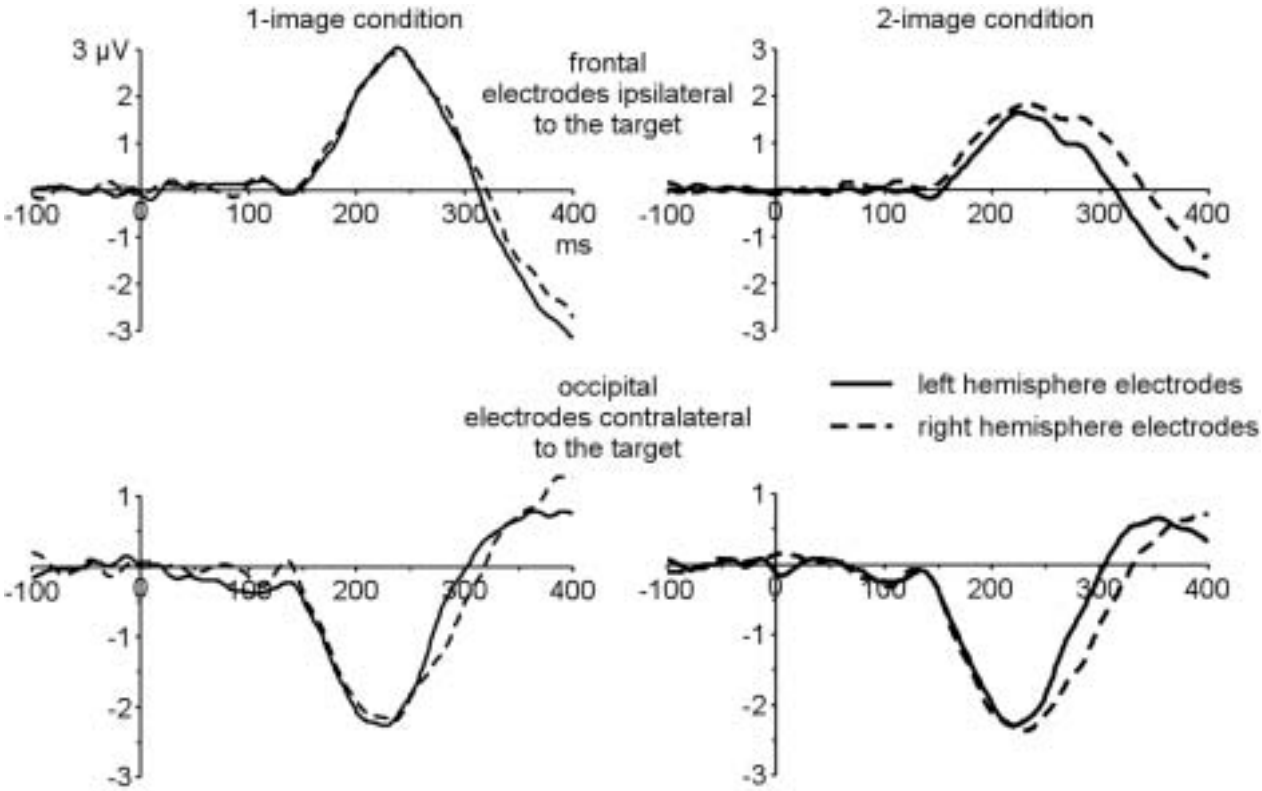


Figure 4. Occipital and frontal differential activities recorded respectively over electrodes contralateral and ipsilateral to the targets. For each image condition, signal recorded over left and right hemispheres are compared. The occipital signal results from the average of signals recorded over electrodes T5-O1'-CB1 (left hemisphere) and T6-O2'-CB2 (right hemisphere). The frontal signal results from the average of signals recorded over electrodes FP1-F3-F7 (left hemisphere) and FP2-F4-F8 (right hemisphere).

Amplitude

The frontal differential activity was larger with 1 image (4.3 μV) than with 2 images (3.0 μV, $p < 0.0001$). It was not significantly different for left than right targets (left = 3.6 μV, right = 3.7 μV) and for left compared to right hemisphere (left = 3.6 μV, right = 3.8 μV). The frontal activity was significantly stronger over ipsi- than contralateral electrodes (hemifield per hemisphere interaction, $p = 0.003$). When ipsilateral sites were tested separately, the same effects were found for latency and amplitude, except that the shorter latency found for left hemisphere electrodes was not significant.

A direct comparison of the differential activities recorded over contralateral occipital and ipsilateral frontal electrodes showed the occipital activity peaked earlier (occipital = 228 ms, frontal = 244 ms, $p < 0.0001$) but with a lower amplitude than the frontal activity (occipital = 3.3 μV, frontal = 3.9 μV, $p = 0.03$). This difference strengthens the idea according to which, even if the signal recording over frontal electrodes reflects in large part occipito-temporal activity, part of it actually originates from other sources (possibly frontal) than the signal recorded over occipital electrodes.

4.3 Parietal differential activity

Finally, results from a large parietal differential activity are reported. It was measured at midline electrodes Pz and Pz' but it is only illustrated at Pz' where it was the largest (Figure 5). Given its late onset, it might be related to motor/late decisional mechanisms. This component was analyzed as a peak although a more appropriate analysis would be to measure its amplitude in different time windows. Such analyses were applied to some of the data reported in the second article of the thesis.

Latency

The parietal differential activity peaked at about the same latency in the two image conditions (1-image = 421 ms, 2-images = 424 ms). There was no difference between target hemifield positions (left = 429 ms, right = 424 ms).

Amplitude

The parietal activity was significantly larger in the 1-image condition (4.7 μV) than in the 2-image condition (3.9 μV, $p < 0.0001$). Target hemifield positions did not significantly affect its amplitude (left = 4.3 μV, right = 4.2 μV).

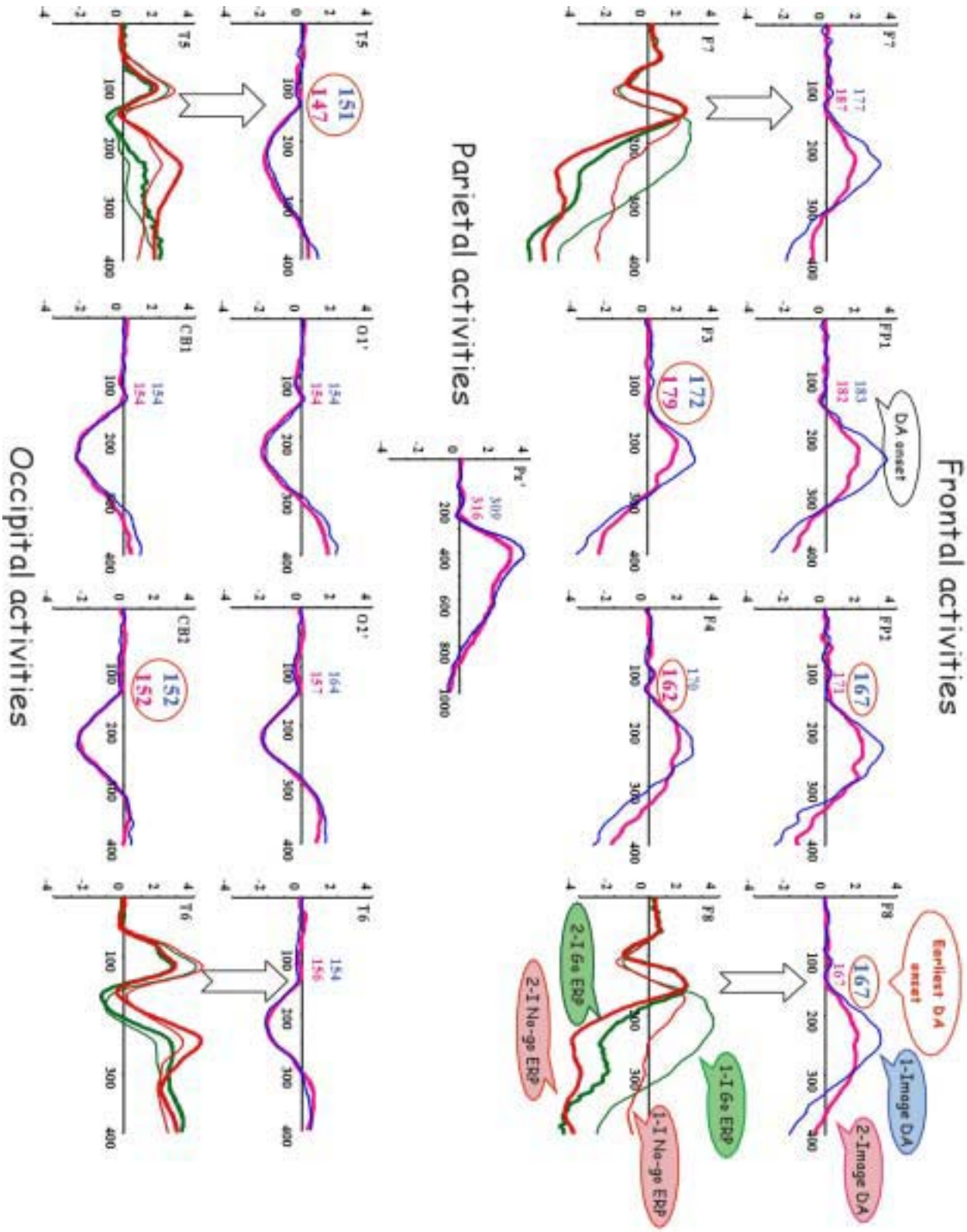


Figure 5. ERP and differential activities recorded at occipital, frontal and parietal electrodes in the 1-image and the 2-image conditions. ERP and differential activities are expressed in microvolt as a function of time in millisecond Potentials are plotted schematically according to the electrode positions on the scalp.

Two natural images can be processed as fast as one in a superordinate visual categorization task

Guillaume A. Rousselet, Michele Fabre-Thorpe, Simon J. Thorpe

CERCO, UMR 5549 CNRS-UPS, Toulouse, France

guillaud@cerco.ups-tlse.fr, mft@cerco.ups-tlse.fr, thorpe@cerco.ups-tlse.fr

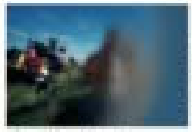
INTRODUCTION: two visions of vision

The world through a CCD camera
(thanks to Laurent Itti for the picture)

The classic view:

- Visual system = clever camera that takes deductive decisions from objective information.
- Parallel processing of low-level features => serial binding process.
- 1 high-level representation at a time.

Early vision = early on the complexity scale



The world through a retina

The modern view:

- Visual system = guessing machine that takes heuristic decisions by betting on what is out there based on little evidence but a lot of experience.
- Several high-level representations at a time that constrain low-level processes.

Early vision = early on the time scale

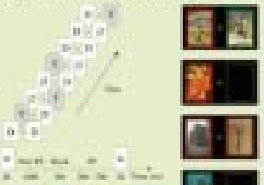
EXPERIMENT

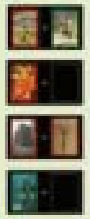
The visual brain works very fast: first frontal activations around 80 ms (Baron et al., 2000; Foxe & Simpson, 2002). An object discrimination process starts around 100 ms (Vogel & Luck, 2000) that can lead in 150 ms to the categorization of natural images (Thorpe et al. 1996).

Question = Does this rapid processing of natural images apply in parallel across the visual field?

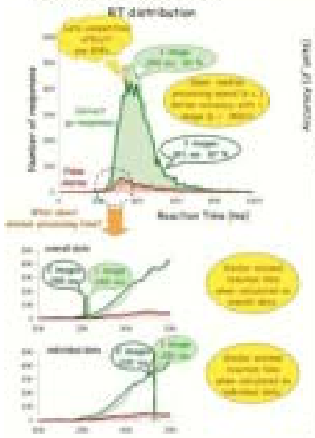
Setup:

- 20 subjects, 10 women, 10 men, mean age 32
- go / no-go categorization task: target = animal
  - target (30%): release button (< 1 sec), go response
  - non-target (70%): keep pressing button (1 sec), no go response
- 1 or 2 vertical pictures (5.6° / 10.4°) centered at 9.6° of eccentricity to the left or (and) right of the fixation point
- 7 conditions: 1_, 1R, _1, 2V, 2L, _2, 2H (x2)
- 20 series of 96 trials
- 5 longer and brief presentations (30 ms)
- No a priori knowledge of number, position and species of animals
- 2000 photographs of natural scenes (Corel library) => 1920 distractors and 960 targets
- 8000 microdipoles: 32 channels recorded in a cap, bandpassed 30 to 100 Hz, digitalized 3000 Hz, epochs: 100 ms to 1000 ms, impedances < 5 kΩ, average reference





BEHAVIORAL RESULTS

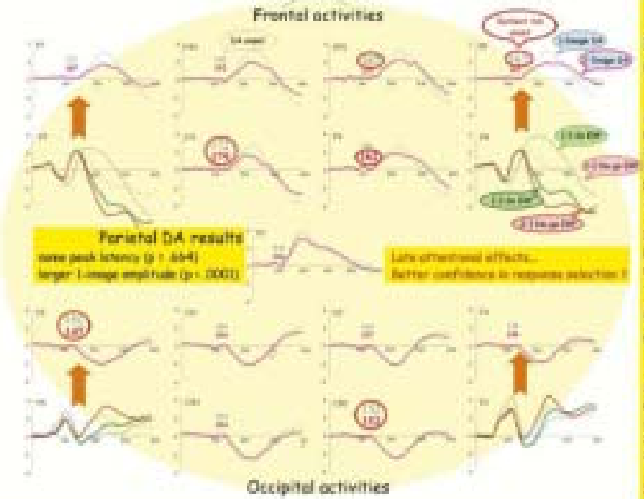


CONCLUSION

Similar time course when processing either 1 or 2 natural images. The small accuracy impairment can be explained by a very simple model where the 2 images are processed by separate mechanisms that pool their outputs together (thanks to Rufin VanRullen for the idea).

ERP RESULTS

Differential activities (DA) = correct target ERPs minus correct non-target ERPs = index of visual decision-making (VanRullen & Thorpe, 2001).
DA onset: 15 successive significant paired t-test values, 19 df, $p < .0005$.
DA peak amplitude and peak latency data: ANOVAs.



MAIN RESULTS

Frontal DA results
(ipsilateral to targets)

- Same onset latency (see figure).
- Same slope in early phase.
- Same peak latency ($p = .123$)
- Larger amplitude with 1 image ($p < .003$)

Identical processing speed.
Late sensori-motor competition.

Occipital DA results
(contralateral to targets)

- Same onset latency (see figure).
- Same slope.
- Same peak latency ($p = .482$)
- Same amplitude ($p = .506$)

Identical processing speed.

DISCUSSION

Processing speed is unchanged between the 1-Image and the 2-Image conditions. The slight accuracy impairment with 2 images can be explained by a very simple model. Brief image presentation => lateralization of visual inputs to the contralateral striate visual cortex => each hemisphere works in parallel on a different visual scene, as shown by the pattern of occipital differential activity. Frontal results and RT histograms suggest a late competition hypothesis. High-level representations in the visual occipito-temporal areas would be activated independently in each hemisphere. At frontal sites, when integration of the outputs of the 2 cerebral hemispheres is needed to make a unique response, the decision mechanism would be biased in favor of a response by a target or against a response by a non-target.

Early vision might be considered as early in time with no a priori limits about complexity.

Early vision => abstract representations => hypothesis generation
=> confrontation with low-level information

Chelazzi et al., 1998
2 (or more) objects in one hemisphere

Keysers et al., 2001
"pipeline" processing, 1 image per level of the visual system

& the present study...

What are the limits of parallelism???

Article 2

Processing of one, two or four natural scenes in humans: the limits of parallelism.

Rousselet, G.A., Thorpe, S.J. & Fabre-Thorpe, M.

(article sous presse)

Résultats comportementaux et électrophysiologiques de 16 sujets adultes dans une expérience visant à tester les limites du traitement en parallèle des scènes naturelles.

Introduction

Cette expérience avait pour objectif d'évaluer si les capacités de traitement en parallèle rapportées dans l'article 1 (dans lequel les images sont présentées autour du méridien horizontal et chacune dans un hémichamp) constituent un cas exceptionnel, chaque image étant prise en charge par un hémisphère, ou bien si ce résultat peut être généralisé à d'autres situations.

Dans cette expérience, les sujets réalisaient toujours une tâche de catégorisation go/no-go animal/non-animal. Cette fois-ci, entre 1 et 4 images étaient présentées brièvement dans des quadrants (Figure 1). Dans le cas de 1 image, elle pouvait apparaître dans 1 des 4 quadrants. Dans le cas de 2 images, elles pouvaient apparaître soit de part et d'autre du méridien vertical, dans l'hémichamp inférieur ou dans l'hémichamp supérieur, soit de chaque côté du méridien horizontal dans le même hémichamp. La première situation (proche de celle utilisée dans la première expérience) permettait de comparer le parallélisme au sein des deux hémichamps inférieur et supérieur – condition inter-hémisphérique. La seconde situation permettait d'aborder le parallélisme intra-hémisphérique et de comparer le comportement de chacun des deux hémisphères dans une telle situation. La condition 4 images constituait le test le plus difficile, mettant en jeu une compétition à la fois inter- et intra-hémisphérique.

Il faut souligner que cette expérience n'est pas directement comparable aux expériences classiques de recherche visuelle. En effet, il ne s'agissait pas de répondre pour un animal parmi un nombre d'objets distracteurs variant entre 0 et 3, mais pour un animal dans une scène naturelle

parmi un nombre variable de scènes non cibles. Le comportement différent des neurones de la voie ventrale en réponse à des scènes naturelles par rapport à des objets isolés, comme cela a été montré dans le chapitre 1A, justifie cette approche.

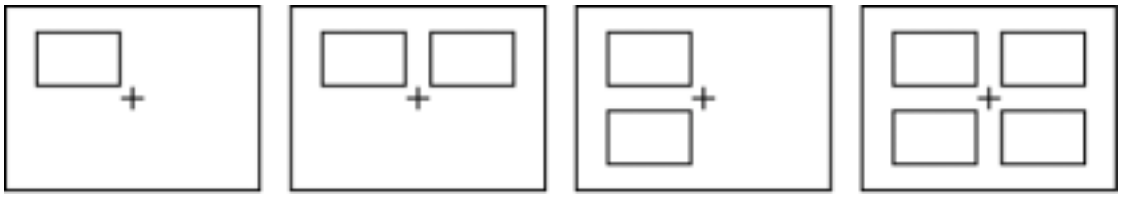


Figure 1. Représentation schématique des 4 conditions expérimentales. De gauche à droite : 1 image ; 2 images inter-hémisphériques ; 2 images intra-hémisphériques ; 4 images. La position des images était contrebalancée entre sujets.

Résultats

Dans cette expérience les différentes conditions expérimentales à analyser étaient nombreuses, donnant lieu à un nombre considérable de résultats. Certains résultats ont mis en évidence la nécessité d'effectuer des expériences contrôles permettant de lever l'ambiguïté sur les interprétations possibles, ce qui explique que cet article soit encore en préparation. Voici succinctement les plus importants.

1) Dans la condition une image, le traitement d'une scène naturelle dans un quadrant était beaucoup moins efficace que dans la première expérience, i.e. dans le cas d'une image présentée à gauche ou à droite du point de fixation, chevauchant l'axe horizontal.

2) Le temps de réaction augmentait et le pourcentage de bonnes réponses diminuait avec le nombre d'images à traiter. Ces données comportementales n'étaient pas entièrement compatibles avec un traitement en parallèle de 2 ou de 4 scènes naturelles, bien que cela soit le cas chez certains sujets.

3) Les deux conditions avec 2 images étaient associées avec des résultats comportementaux très similaires, indiquant que la source majeure d'interférence n'était pas d'origine intra hémisphérique.

4) Les données électrophysiologiques présentaient à nouveau une dissociation entre les activités différentielles occipitale et frontale. L'activité occipitale indiquait un traitement parallèle de 2 images, mais pas de 4 images. En revanche, l'activité frontale distinguait entre la condition 1 image d'une part et les autres conditions d'autre part, ces dernières différant très peu entre elles.

5) Enfin, des analyses complémentaires de l'activité différentielle occipitale suggèrent qu'elle pourrait refléter une sélection attentionnelle tardive, impliquant peut-être des ressources spatiales.

Discussion

La faible performance des sujets dans la condition 1 image par rapport aux données rapportées précédemment dans d'autres articles de l'équipe n'est pas complètement élucidée (Fize et al., en révision ; Rousselet et al., 2002 ; Thorpe et al., 2001a). Deux interprétations sont possible. Dans la première, les scènes naturelles seraient traitées moins efficacement quand elles sont présentées dans des quadrants plutôt que latéralement de part et d'autre du méridien horizontal. Dans la seconde, les séries d'essais présentant au hasard des stimuli à 1, 2 ou 4 images, on peut imaginer que le sujet, pour ne pas réaliser trop de réponses incorrectes dans les essais difficiles, ait en quelque sorte « seuillé » sa prise de décision à un niveau plus élevé que lorsqu'il se trouve uniquement confronté à des stimuli à 1 et 2 images. Une expérience contrôle suggère que la première hypothèse est la plus probable. Ce sujet mérite de plus amples investigations.

Contrairement à la première expérience sur le parallélisme, cette deuxième étude a révélé un effet perturbateur important du nombre de scènes à traiter au niveau comportemental. Ceci était vrai même dans les conditions où les 2 images étaient présentées chacune dans un hémichamp visuel différent, des conditions expérimentales très proches de la condition 2 images de la première expérience. Cet effet renforce l'idée d'un traitement différent dans les quadrants. De manière générale, la diminution de la performance comportementale avec le nombre d'images est en accord avec les données rapportées par VanRullen et al. (sous presse) montrant un traitement sériel des scènes naturelles. Selon eux, l'existence de neurones sélectifs pour une catégorie d'objets dans IT, par exemple des animaux, permet de les détecter sans attention focalisée en vision périphérique alors qu'une autre tâche est réalisée en vision centrale, comme cela a été montré par Li et al. (2002). Ils seraient donc traités de manière préattentive. Par contre, si plusieurs objets stimulent en même temps le champ récepteur d'un de ces neurones, une compétition s'établirait quand même, ralentissant la détection de l'objet cible. Dans ce cas le traitement serait sériel. Cependant, le chapitre 1A montre à quel point il est difficile d'établir un lien direct entre le comportement et les mécanismes neuronaux sous-jacents.

Les données électrophysiologiques soutiennent plus l'idée d'un traitement parallèle dans les aires visuelles postérieures, la compétition entre les scènes naturelles ayant surtout lieu au niveau

frontal. Une telle dissociation entre activités différentielles occipitales et frontales a été rapportée dans plusieurs études en potentiels évoqués. Dans différentes tâches de sélection attentionnelle, les principaux effets rapportés étaient des déflexions négatives enregistrées en regard des électrodes postérieures et des positivités frontales (Bass et al., 2002 ; Hillyard & Anllo-Vento, 1998 ; Karayanidis & Michie, 1996, 1997 ; Luck et al., 1997b ; Smid et al. 1999). Ces composantes ont reçu différentes appellations : SN (selection negativity), OSN (occipital SN), N2, N2pc, N2b, etc., pour les activités occipitales ; FSP (frontal selection positivity), SP, P2a, pour les activités frontales. Si bien souvent l'interprétation de ces deux composantes reste spéculative, certains ont montré, par des analyses de sources, qu'une grande partie du signal frontal pourrait s'expliquer par des dipôles postérieurs (Anllo-Vento et al., 1998 ; Kenemans et al., 2002), notamment dans la tâche animal/non-animal avec une image (Fize et al., en révision). D'autres ont cependant suggéré que les activités enregistrées en regard des électrodes frontales pourraient constituer une activité d'origine frontale en relation avec l'évaluation des stimuli dans le cadre de la tâche et le choix d'une réponse comportementale appropriée (Hopf & Mangun, 2000 ; Lange et al., 1998 ; Potts et al., 1996, 2001, 2002). L'origine de ce signal reste cependant à déterminer ; des structures comme le cortex cingulaire antérieur (Lange et al., 1998) et le cortex orbito frontal (Potts et al., 2001) ont été proposées. Le plus probable est cependant une situation mixte, dans laquelle les potentiels frontaux reflètent à la fois l'intervention de dipôles postérieurs et l'intervention additionnelle de sources frontales. Les résultats obtenus dans les deux expériences sur le parallélisme pourraient donc être interprétés selon un modèle de sélection tardive, la plus grande partie de la compétition étant d'origine décisionnelle. Cette hypothèse sera testée prochainement dans une expérience consistant à répliquer l'étude avec 2 images en utilisant un bonnet à 256 électrodes afin de réaliser une analyse de source contrainte par l'IRM structurale des sujets.

Si la plus grande partie de l'interférence entre les images est d'origine frontale, il reste cependant à expliquer pourquoi il y avait une baisse d'amplitude de l'activité différentielle occipitale dans la condition 4 images par rapport aux autres conditions. Selon une première interprétation, cet effet pourrait refléter une forte compétition dans la voie ventrale due aux larges champs récepteurs de IT qui intègreraient l'information de l'ensemble du champ visuel. Cette interprétation n'est cependant pas compatible avec certaines données présentées au chapitre 1A. Sur la base d'analyses complémentaires décrites dans l'article et en m'appuyant sur certains éléments de la

littérature, j'ai proposé une interprétation alternative selon laquelle l'effet occipital dans la condition 4 images pourrait être dû à une interaction rapide entre des aires décisionnelles frontales et des aires visuelles occipitales. Cette hypothèse est compatible avec des travaux ayant montré des activités frontales dès 80 ms après l'apparition d'un stimulus, suggérant que des boucles en retour se mettent en place précocement dans le système visuel (Foxe & Simpson, 2002). Elle est aussi soutenue par la mise en évidence de modulations d'origine frontale des potentiels évoqués occipitaux dès 125 ms (Barcelo et al. 2000). Finalement, j'ai proposé qu'une partie de l'activité différentielle enregistrée en regard des électrodes occipitales pourrait refléter un mécanisme tardif de sélection spatiale d'un objet. Cependant, des expériences complémentaires seront nécessaires pour étayer cette hypothèse. Par exemple, il est possible que cet effet occipital reflète le processus descendant de seuillage au niveau des aires visuelles évoqué plus haut pour expliquer les données comportementales.

Pour conclure, je voudrais revenir sur le protocole expérimental mis en œuvre dans ces deux expériences sur le parallélisme. On pourrait penser qu'une condition encore plus écologique consisterait à mettre en œuvre une seule scène naturelle par essai, mais dont la taille serait variable ainsi que le nombre d'objets non cibles. Cette approche est intéressante mais beaucoup plus difficile à mettre en œuvre. De plus, traiter l'ensemble d'une scène naturelle dans laquelle tout est a priori « congruent », au moins en ce qui concerne les échelles de taille et les relations sémantiques entre objets et contexte, est totalement différent de traiter 4 scènes qui peuvent être présentées à différentes échelles et qui font chacune appel (le plus souvent) à une représentation sémantique différente…

Processing of one, two or four natural scenes in humans: the limits of parallelism

Guillaume A. Rousselet, Simon J. Thorpe & Michèle Fabre-Thorpe

Abstract

The visual processing of objects in natural scenes is fast and efficient, as indexed by behavioral and ERP data (Thorpe, Fize & Marlot, Nature 381 (1996) 520). The results from a recent experiment suggested that such fast routines work in parallel across the visual field when subjects were presented with two natural scenes simultaneously (Rousselet, Fabre-Thorpe & Thorpe, Nature Neurosci. 5 (2002) 629). In the present experiment, the visual system was driven to its limits by presenting one, two or four scenes simultaneously. Behavior and ERP reveal a clear cost in processing an increasing number of scenes. However, a parallel-late selection model can still account for the results. This model is developed and discussed with reference to behavioral, single-unit and ERP data.

1. Introduction

Our ability to face complex and unpredictable environments seems to rely on several mechanisms involving fast and parallel visual routines. The very rapid categorization of briefly presented pictures (Biederman, 1972; Potter, 1976) could well rely on neurons in higher order areas such as infero-temporal cortex (IT) that have been shown to fire selectively and very rapidly to a wide range of complex stimuli, both in monkeys (Gross, Bender & Rocha-Miranda, 1969; Perrett, Rolls & Caan, 1982; Tanaka, 1996; Vogels, 1999) and in humans (Allison, Puce, Spencer & McCarthy, 1999; Kreiman, Koch & Fried, 2000). Other evidence comes from studies using event-related potentials (ERP) that indicate that the discrimination of isolated stimuli might start at about 120-150 ms after stimulus onset (Jeffreys, 1996; Schendan, Ganis & Kutas, 1998; Rossion, Gauthier, Tarr *et al.*, 2000; Vogel & Luck, 2000). Whether such fast mechanisms can apply in parallel across the visual field is still very controversial. Indeed, many models of visual processing assume that after an early parallel encoding of object features, there is a computational bottleneck that prevents more than one set of features at a time from forming 'high-level' object representations (Treisman, 1998; Wolfe, 1998). However, single-unit recordings in monkeys have shown that when more than one object is present within the receptive field of IT neurons, they can all affect the response of the cell, although there is evidence that the different stimuli compete for control of the neuronal response (Olson, 2001; Chelazzi, Duncan, Miller & Desimone, 1998). Such phenomena have also been seen in the context of natural scenes (Sheinberg & Logothetis, 2001). Nevertheless, it is clear that relatively little is known about the degree of parallelism in the processing of natural objects and even less in the context of natural scenes where several objects are typically present simultaneously. In previous studies, we showed that in 150 ms the human brain has accumulated enough information to start to categorize a natural scene as containing or not an animal (Thorpe, Fize & Marlot, 1996; Fabre-Thorpe, Delorme, Marlot & Thorpe, 2001) or a non-biological category such as a means of transport (VanRullen & Thorpe, 2001). This rapid categorization of natural scenes already suggests some degree of parallel processing in the visual system. In a recent experiment, we tried to challenge the visual parallelism by requiring from human subjects the simultaneous processing of two different natural scenes. The results strengthen the idea that such fast routines

can work in parallel across the visual field (Rousselet, Fabre-Thorpe & Thorpe, 2002). Compared with a single scene condition, human subjects were shown to be just as fast at detecting animals when they had to process two different natural scenes flashed for 20 ms on each side of a central fixation point, with the images centered on the horizontal meridian. Over contralateral occipital electrodes, ERPs on correct go trials diverged from ERPs on correct no-go trials at 150 ms in both conditions. This pattern of differential activity suggests that some high-level object properties are accessed in parallel during natural scene processing. But the brief presentation of images in the left or the right hemifield had induced an initial lateralization of the visual inputs to the contralateral hemisphere so that each hemisphere could independently process one of the two different scenes. However, it is not known whether this result would extend to a situation in which the two images would be presented in the same hemifield, hence entering into an intra-hemisphere competition.

In the present experiment the visual system was driven to its limits by tackling both inter- and intra-hemispheric parallel processing. The task required subjects to respond as fast and accurately as possible each time there was an animal in a briefly (26 ms) presented visual display. This display contained one, two or four photographs of natural scenes appearing simultaneously centered at 4.9° from a central fixation point, in 1 to 4 of the quadrants. When two pictures were presented, they could appear in the upper or the lower visual field or in the left or the right visual hemifield.

In addition, some aspects of the paradigm used in the present experiment might give us better insights as to the actual mechanisms generating the differential activity. This point will be addressed in the last part of the electrophysiology result section.

Here we present the behavioral performance and electrical scalp surface recordings from 16 human subjects.

2. Experimental Procedures

2.1. Subjects

The 16 adult volunteers in this study (7 women, all right-handed, 9 men, 2 left-handed, mean age 28 ranging from 21 to 50) gave their informed written consent. All subjects had normal or corrected to normal vision.

2.2. Stimuli

We used photographs of natural scenes taken from a large commercial CD-ROM library (Corel Stock Photo Libraries). From this data bank, 3360 distracters and 960 targets were selected. . Horizontal photographs (384 by 256 pixels, sustaining about 7.8° by 5.2° of visual angle) were chosen to be as varied as possible. Animals included mammals, birds, fish, insects and reptiles... There was no a priori information on the size, position, or number of targets in any particular photograph. There was also a very wide range of distracter images that included outdoor or indoor scenes, natural landscapes (mountains, fields, forests, beaches...) and street scenes, pictures of food, fruits, vegetables or plants, buildings, tools or other man-made objects...

2.3. Task and set-up

Subjects sat in a dimly lit room at 100 cm from a computer screen (resolution: 1024 x 768, vertical refresh rate: 75 Hz) piloted from a PC computer. Stimulus presentation and behavioral response recording was achieved using the Presentation software application (http://nbs.neuro-bs.com/). To start a block of trials, subjects had to place their finger on a response pad for one second. Stimulus displays were composed of 1, 2 or 4 images that appeared centered at the corners of an imaginary rectangle. The distance between the center of each image and the central fixation point was 3.7° vertically and 4.9° horizontally. In the 1-image condition, the photograph could appear in four possible locations and be either a target (T) or a distractor (D). In the 2-image conditions, the scenes could appear either (1) on each side of the vertical meridian, whether in the upper or the lower visual field (inter-hemifield condition), or (2) above and below the horizontal meridian, whether in the left or the right hemifield (intra-hemifield condition). In all cases the 2 images were either two distractors or a target and a distractor. The 4-image condition involved the presentation of either one target among three distractors on target trials or four distractors otherwise.

A trial was organized as follows: a 300-600 ms fixation point (about 0.1° of visual angle) appeared in the middle of the screen after which a display composed of one, two or four photographs was presented for two frames, i.e. 26 ms. Participants had to raise their finger as quickly and as accurately as possible (go response) each time an animal was present. Responses were detected using infrared diodes. Subjects had 1000 ms to respond, after which their response was considered as a no-go response. This maximum response time delay was followed by a 300 ms black screen before the reappearance of the 300-600 ms fixation point, resulting in a random 1600-2200 ms inter-trial interval. When the photographs contained no animal, subjects had to keep their finger on the pad for at least 1000 ms (no-go response). An experimental session consisted of 20 blocks of 96 trials in which target and distractor trials were equally likely in each condition. To prevent learning, each image was seen only once by each subject. On any of the four target trials (1-image, 2-image inter, 2-image intra and 4-image trials), the target image could be shown in each of the four quadrants, resulting in a total of 16 target conditions. On distractor trials, the position of a given distractor was not considered in the analysis, thus resulting in 9 distractor conditions (4 conditions in the 1-image trials in which the distractor could appear in each of the quadrants, 2 conditions in each of the 2-image trials in which both distractors could appear in one - upper/lower or left/right- hemifields and 1 condition in the 4-image trials). The design was counterbalanced so that overall, each image was presented the same number of times in all the different conditions. Subjects were given two training series before the test session. Training images were not used in the test session. Task effects on behavioral measurements were assessed by ANOVA with a Greenhouse-Geisser correction for non-sphericity. Post-hoc analyses were performed using paired *t*-tests with a Bonferroni correction or Wilcoxon tests.

2.4. EEG analysis

Electric cortical activity was recorded from 32 tin electrodes mounted in an elastic cap in accordance with the 10-20 system (Oxford Instruments) with the addition of extra occipital electrodes and using a Synamps amplifier system (Neuroscan Inc.). The ground electrode was placed along the midline, ahead of Fz. Impedances were systematically kept below 5 kΩ. Signals were digitized at a sampling rate of 1000 Hz (corresponding to a

sample bin of 1 ms) and low-pass filtered at 100 Hz. Potentials were on-line referenced to electrode Cz and averaged-referenced off-line. Baseline correction was performed using the 100 ms of pre-stimulus activity. Two artifact rejections were applied over the [-100 ms; +400 ms] time period: on frontal electrodes with a criterion of [-80; +80 μV] to reject trials with eye movements, and on parietal electrodes with a criterion of [-40; +40 μV] to remove trials with excessive activity in the alpha range. Only correct trials were averaged except when specified in the text. ERP components were further low-pass filtered at 40 Hz before analysis. Sixteen differential activities (one for each target condition) were computed by subtracting evoked potentials for distractor trials from evoked potentials for target trials. Analysis concentrated on three groups of electrodes where distinct differential activities were clearly identified: occipital (left: P3', T5, O1', O1, CB1, CB1'; right: P4', T6, O2', O2, CB2, CB2'), frontal (left: FP1, F3, F7; right: FP2, F4, F8) and parietal (P3, P3', Pz, Pz', P4, P4') electrodes. These groups of electrodes were selected based on our previous experiment on parallel processing (occipital and frontal electrodes: Rousselet et al., 2002; parietal electrodes: Rousselet, Thorpe & Fabre-Thorpe, unpublished data). Among the occipital electrodes, O1 & O2 pertain to the 10-20 system. The additional occipital electrodes have the following spherical coordinates (theta/phi): O1' = -92/54, O2' = 92/-54, CB1 = -115/54, CB2 = 115/-54, CB1' = -115/72, CB2' = 115/-72, P3' = -74/61, P4' = 74/-61. Note that CB1-CB2, O1'-O2' and P3'-P4' are part of the 10-10 system where they appear respectively as PO9-PO10, PO7-PO8 and PO3-PO4. The parietal electrode Pz' is also referred as POz in the 10-10 system.

Electrophysiological measurements were entered in omnibus ANOVAs with five within-subject factors: image (4 levels), upper/lower visual field, left/right visual field, left/right hemisphere electrodes (occipital and frontal differential activities only), and electrodes (occipital and parietal = 6 levels, frontal = 3 levels). A Greenhouse-Geisser correction for non-sphericity was applied when necessary. Post-hoc analyses were performed using paired *t*-tests with a Bonferroni correction.

The onset at which the ERP amplitude on target trials significantly diverged from that on distractor trials was evaluated by performing paired *t*-tests (15 d.f., $p<0.05$) at each time bin, i.e. every ms with a 1000 Hz sample rate. In our previous experiment paired *t*-tests were considered significant only when $p<0.0001$ (Rousselet et al., 2002). In the present study, because of the large number of different conditions, the signal to noise ratio was lower justifying the use of a threshold at $p<0.05$ to index differential activity latencies. However, in order not to underestimate those latencies, we fixed an arbitrary threshold of 20 successive significant *t* values to index a differential effect rather than the 15 steps previously used. Thus, a given differential activity latency reported in this paper is the time at which the two conditions start to differ for at least 20 ms.

3. Behavior

In this behavioral section, we will first focus on the comparison between the conditions with 1, 2 and 4 images. The comparison between the intra- and inter-hemifield 2-image conditions will be the subject of a second part. Finally, laterality effects will be approached.

3.1.Comparing the 1-, 2- and 4-image conditions

Despite the very challenging nature of the task, a good level of performance was reached (Table 1, Fig.1). Mean accuracy decreased as a function of the number of images to process ($F(1.6, 23.3)=167$, $p<0.0001$) (Fig.1 & 2). A higher accuracy was reached in the 1-image condition (80.7%) than in the 2- and 4-image conditions, the 2-image condition (74.7%) being in turn associated with better performance compared to the 4-image condition (67.6%) (Wilcoxon tests, all $z<-3.5$, all $p<0.0001$). Across conditions, accuracy was better for distractors (no-go responses: 78.8%) than targets (go responses: 70.0%), reflecting a common bias of human subjects ($F(1, 15)=11.7$, $p=0.004$). However, this was not true for the 4-image condition in which accuracy on no-go responses was not different from accuracy on go responses (interaction between go/no-go and image factors, $F(1.4, 21.1)=18.0$, $p<0.0001$). The number of simultaneously presented photographs affected both the proportion of correct go responses ($F(1.5, 22.6)=9.1$, $p<0.003$) and the proportion of correct no-go responses on distractors ($F(1.3, 19.5)=175$, $p<0.0001$). Accuracy on targets was significantly better with 1 image (74.1%) than with 2 images (68.2%) and 4 images (67.4%) (Wilcoxon tests, both $z<-2.5$, both $p<0.02$). But the comparison between the 2- and the 4-image conditions failed to reach statistical significance. The accuracy on distractors was also better in the 1-image condition (87.4%) than in the 2- (81.2%) or 4- (67.7%) image conditions, the two last conditions being also significantly different from one another (Wilcoxon tests, all $z<-3.4$, all $p<0.001$).

Behavior	**1 image**	**2 images**	**4 images**
accuracy (%)			
mean accuracy	80.7 (1.1)	74.7 (0.9)	67.6 (0.9)
model predictions	n.a.	76.9 (1.1)	70.8 (1.3)
correct go	74.1 (2.2)	68.2 (1.8)	67.4 (1.4)
correct nogo	87.4 (1.3)	81.2 (1.8)	67.7 (2.0)
RT (ms)			
mean	477 (11)	493 (11)	504 (12)
median	457 (11)	469 (11)	475 (12)
min RT (10 ms bins)	310	320	350

Table 1. Summary of behavioral results: 1 vs. 2 vs. 4 images. Data shown here have been pooled over quadrants for clarity. Standard error is indicated in brackets. A simple parallel model of processing was used to estimate from the 1-image results the accuracy reduction due to the addition of distractor images (second row). This model (Rousselet, Fabre-Thorpe & Thorpe, 2002) postulates that each of the two simultaneously presented images is processed by a separate and independent mechanism whose accuracy is adjusted to the one reached in the 1-image condition; the two outputs are then pooled together. In the 2-image condition, a correct no-go response on a distractor trial with two different distractors (no-goDD) is only obtained when both distractors are correctly ignored: no-goDD = $(1\text{-p (FA)})^2$. For target trials, in which a target is simultaneously presented with a distractor, a correct go response (goTD) is produced either by a hit in response to the target or by a false alarm to the simultaneously presented distractor: goTD = 1 – (1 – p (Hit)) x (1 – p (FA)). As target and distractor trials are equiprobable, the overall probability of correct responses if both images are processed in parallel should be: (no-goDD + goTD) / 2 = $((1\text{-p(FA)})^2+1\text{-(1-p(Hit))x(1-p(FA))})/2$. The same logic was applied to predict the results with 4 images, with no-goDDDD = $(1\text{-p(FA)})^4$ and goTDDD = 1 – (1 – p (Hit)) x $(1 – \text{p (FA)})^3$. Thus, the probability of correct responses with 4 images is: $((1\text{-p(FA)})^4+1\text{-(1-p(Hit))x(1-p(FA))}^3)/2$.

We assessed whether the general drop in accuracy due to the increasing number of pictures to process was accounted for by a simple parallel model of processing (as in Rousselet et al., 2002). In this model, each of the simultaneously presented images is processed by a separate and independent mechanism, each mechanism converging on a single output system (see Table 1 caption for details). In our task, a prediction for the accuracy

in the 2-image condition can be made on the basis of hit rate and false alarm rate obtained in the 1-image condition. A prediction was computed for each subject. The expected average result (76.9%) is very close to the value observed in the 2-image condition (74.7%). However, this difference between the model and the actual data was significant, showing that subjects performed on average 2% worse than expected by our very simple model of parallel processing (Wilcoxon test: z=-2.4; p<0.02). In the 4-image condition, the prediction from the 1-image results also tended to be more optimistic (70.8%) than the observed results (67.6%) (Wilcoxon test: z=-2.8; p<0.005). Overall, this simple parallel model gives a relatively good account of the observed data but tended - on average - to overestimate performance significantly. However, it has to be noted that this was not true for every subject, given that out of 16 subjects, 5 subjects in the 2-image condition and 4 subjects in the 4-image condition performed better than expected by the model.

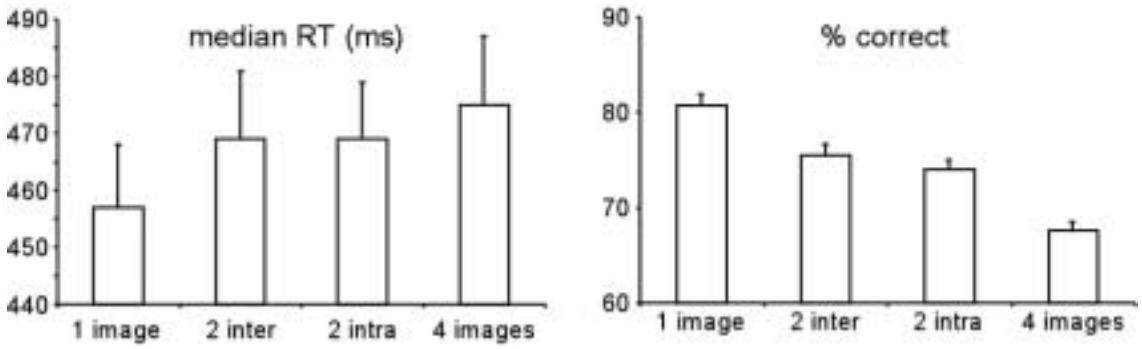


Fig. 1. Median reaction times and mean accuracy with the associated standard errors in the 1-image, 2-image intra- and interhemifield and 4-image conditions.

Mean and median reaction times increased with the number of images presented, (respectively F(1.3, 19.7)=17.5, p<0.0001; F(1.3, 19.5)=10.7, p=0.002). All comparisons for the mean RT values (respectively 477, 493 and 504 ms for the 1-, 2-, and 4-image condition) were significant (paired t-tests: all p<0.02). This pattern was also true for median RT values (respectively 457, 469 and 475 ms) (p<0.01) but the difference between the 2- and the 4-image conditions was not significant. To evaluate how fast subjects can perform the task, we also used as an index the minimal processing time defined as the latency of the bin at which correct go-responses started to significantly outnumber incorrect go-responses in the RT histogram (Fig.2; $\chi 2$ tests on cumulated data at each 10 ms time bin, p<0.01). The minimal processing time needed to correctly respond with 1-image was 310 ms (Table 1, Fig.2). The temporal cost induced by the addition of one distractor image was 10 ms but increased to 40 ms when four images had to be processed simultaneously.

3.2. The 2-image condition: comparing inter- vs. intra-hemifield competition

In the previous section, the three main conditions in this experiment were compared. Clear evidence was found for a strong competition when four images were presented in the visual field, but some competition was also present with only two simultaneously presented images. However, the global results of the 2-image condition average two different cases. In the first one the two images are presented in different (left and right) hemifields and can be processed independently by each hemisphere, whereas in the second case the two images

are presented in the same hemifield and have to be processed by the same hemisphere. Comparing these two cases was important to address the issue of the level of interference between two competing images in our task. If this competition mainly took place within hemispheres, we expected better performances in the inter- than in the intra-hemifield condition. If competition mainly took place at a higher level of integration, for example at a decision stage in frontal areas, then no difference between the two conditions would be expected.

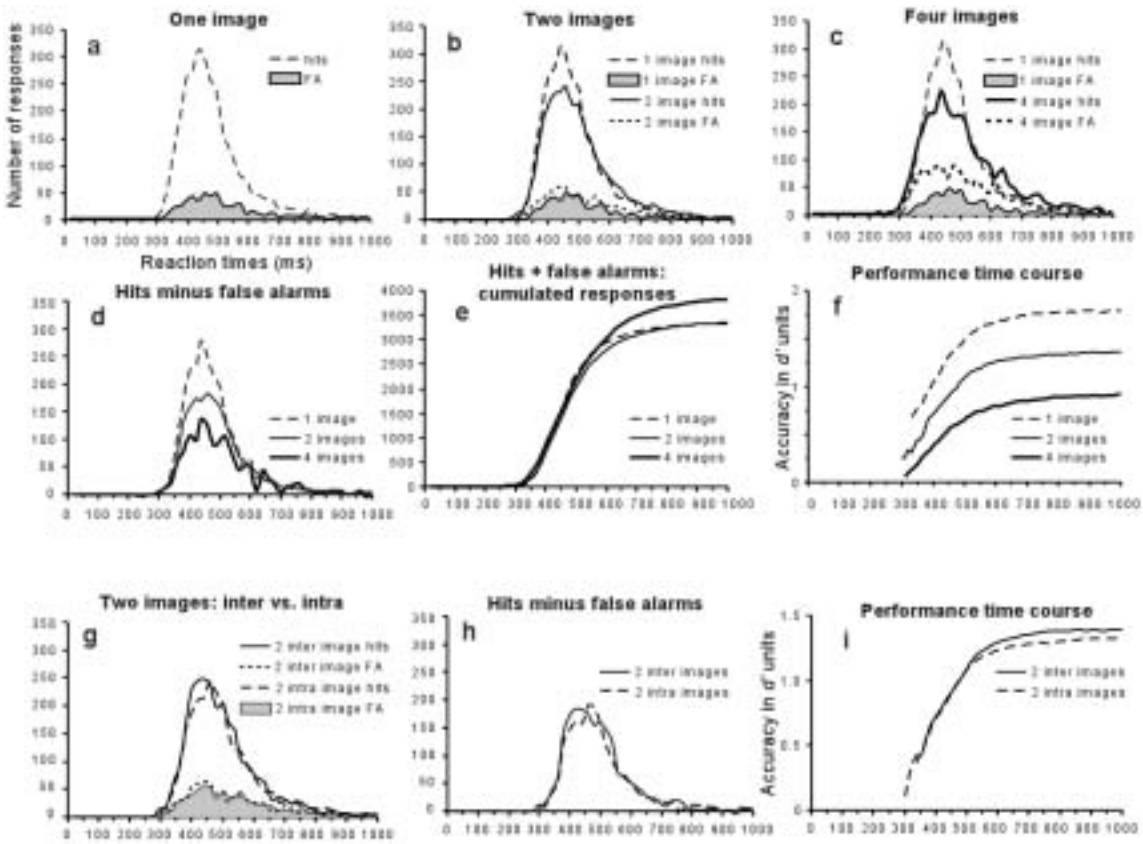


Fig. 2. Reaction time distributions and performance time course. The top two rows compare the behavioral results associated with the 1-, 2- and 4-image conditions. The 2-image curves result from the averaging of the intra and inter-hemifield 2-image conditions. These two conditions are directly compared in the bottom row. In all RT distributions (a, b, c, d, g, and h), the number of correct (hits) and incorrect go-responses (false alarms: FA) are expressed over time, with time bins of 20 ms. As targets and distractors were equally likely in the task, the difference between hits and FA allowed a careful examination of how accuracy varies over time. The RT distributions obtained in the 1-image condition and shown in (a) are also plotted in b and c to allow better comparison with the 2- and 4-image conditions. In panel **d)**, FA have been subtracted from hits to allow direct comparisons of mean accuracy over time. The cumulated response curves in panel (e) illustrate that subjects tended to produce more go-responses in the 4-image condition. Performance was also analyzed over time using a dynamic *d'* calculated from the cumulative number of hits and FA at each successive 20 ms time bin. Plateau values correspond to the *d'* values calculated on global results and were affected by the number of images to process. Comparing the intra- and inter-hemifield 2-image conditions in the bottom row shows that RT distributions were very similar and that the accuracy reached a somewhat higher level when the two images were presented in different hemifields rather than in the same (left or right) hemifield (**i**).

No reliable difference was found in the global mean accuracy (Fig.1, Table 2) obtained in the inter- and intra-hemifield 2-image conditions (75.5 vs. 74.0%). However when considering separately the accuracy on distractors and targets we found that although the proportion of correct no-go responses on distractors was similar (80.4 vs. 81.9%), go accuracy was significantly higher when the two images were presented in separate

hemifields (70.5 vs. 66.0% respectively for inter- and intra-hemifield conditions; interaction between inter/intra and go/no-go factors, $F(1, 15)=7.3$, $p<0.02$). However, this effect did not concern the earliest responses triggered in the two conditions (RT<500ms, Fig.2i). Processing speed was remarkably similar between the two conditions: mean, median and minimal RT were virtually identical and did not present any reliable differences. Similarly, the RT distributions had very close profiles (Fig.2g,h).

Behavior		**2 inter images**	**2 intra images**
mean accuracy (%)		75.5 (1.1)	74.0 (0.9)
	model predictions	76.9 (1.1)	76.9 (1.1)
	correct go	70.5 (2.1)	66.0 (1.9)
	correct no-go	80.4 (1.8)	81.9 (1.8)
RT (ms)	mean	491 (12)	495 (10)
	median	469 (12)	468 (10)
minimal RT (10 ms bins)		320	320

Table 2. Summary of the 2 image behavioral results: 2 inter- vs. 2 intra-hemifield image conditions. Data shown here have been collapsed over quadrants for clarity. Standard error is indicated in brackets.

3.3. Hemifield comparisons: left vs. right & upper vs. lower

Although it was not the main purpose of this experiment, it was interesting to examine possible bias between the different hemifields. Accuracy on go responses did not present any reliable effect for the left/right and upper/lower comparisons. On the other hand, processing speed presented some reliable effects. Median RT was slightly shorter in response to targets that appeared in the right visual field (466 ms) compared to those appearing in the left visual field (469 ms) ($F(1, 15)=6.0$, $p<0.03$), a difference that was not present at the level of mean RT (right=491ms, left=492ms, *n.s.*). The difference between lower and upper visual field targets was more pronounced. Targets presented in the upper visual field were processed significantly faster than targets presented in the lower visual field (mean RT: upper = 488 ms, lower = 497 ms, median RT: upper = 463 ms, lower = 472 ms). Both differences were significant using an ANOVA with images (4 levels), upper/lower and left/right visual field within-subject factors (respectively, $F(1, 15)=9.8$, $p<0.007$; $F(1, 13)=11.3$, $p<0.004$). This effect was not seen for the earliest responses as minimal RT was the same in both cases (349 ms) and did not interact with other factors.

3.4. Behavior: discussion

In the present study we investigated the capacity of the human visual system to categorize natural scenes at a superordinate level using a very challenging task in which one, two or four images were briefly and simultaneously flashed in different quadrants of the visual field. The main question we wanted to address was whether there was any evidence for parallel processing in such a demanding task.

First of all, it is worth noting that the task used in the present experiment was much more challenging than the one used in our previous report (Rousselet et al., 2002). Monitoring up to four quadrant images instead of one or two images presented along the horizontal meridian had a dramatic impact on subjects' performances: in the 1-image condition, accuracy decreased by 10%, median reaction time (RT) increased by 66 ms and minimal RT

increased by 50 ms compared to the former experiment. This discrepancy might be due to an increase in spatial uncertainty in the present experiment because the target could appear in one of four locations instead of one between two in the former experiment. However, such explanation is very unlikely in the light of the results from a previous experiment testing large eccentricities (Thorpe, Gegenfurtner, Fabre-Thorpe & Bülthoff, 2001). In that experiment, subjects had to perform the same go/no-go animal/non-animal task used here, but with photographs of natural scenes centered at up to 70° from the fixation point along the horizontal meridian. Despite the very high spatial uncertainty, subjects still responded correctly on 90% of the trials at 13° of eccentricity. Thus, a more likely explanation for the drop of accuracy in the 1-image condition between the two experiments relies on a change of decisional strategy. This bias might have been introduced by the random presentation of 1-, 2- and 4-image conditions in the same series.

To test this hypothesis, 16 additional subjects who had not participated in the original experiment were tested in two control behavioral studies (8 subjects per study). Different subjects were tested in these two experiments because not enough images were available to avoid stimulus repetition. In the first experiment, subjects were only presented with the 1-image condition and the scene could appear in one of four quadrants. In the second experiment, they were only presented with the 4-image condition. The experimental conditions were identical to those in the original experiment except that subjects performed 15 series of 96 trials in the 1-image experiment and 8 series of 96 trials in the 4-image experiment. Results show that even when subjects were tested with a constant image set size, their performance was not better than when all conditions were mixed together.

In the 1-image experiment the 8 subjects (3 women, 5 men, mean age 25.6 ranging from 21 to 31, 3 left handed) scored 75.8% on average (individual range [69.9-81.6]), which is below the 80.7% obtained with the 16 original subjects. Accuracy was 84.0% [71.7-97.8] on distractors, 67.6% [42.1-84.2] on targets. Mean RT was only 10 ms slower compared to the original data, reaching 487 ms [380-591].

In the 4-image experiment the 8 subjects (3 women, 5 men, mean age 24.1 [21-29], 2 left handed) performed almost exactly like the 16 original subjects. Mean accuracy was 67.3% [58.6-72.1], reaching 69.6% [50.5-85.7] on distractors and 65.0% [50.8-74.2] on targets. Mean RT was 503 ms [401-643].

These data collected with two control experiments suggest that the results obtained in the present study cannot be explained by response biases due to the random alternation of all the different experimental conditions within the same series. It also suggests that human observers are worse at categorizing animals in natural scenes in visual quadrants than along the horizontal meridian. This issue clearly deserves further investigation.

We now turn to the effects of processing an increasing number of pictures in this task. Compared to the 1-image condition, the addition of a second image in the opposite hemifield (inter-hemifield condition) decreased accuracy and increased RT significantly. In keeping with the idea that even a parallel model of visual processing with unlimited capacities predicts impairments because of the presence of distractors (Kinchla, 1992; Palmer, 1998), the decrease in accuracy was in large part accounted for by a very simple model of parallel processing. However, the increase in RT obtained in the present study contradicts our previous null effect on RT when the images were presented along the horizontal meridian (Rousselet et al., 2002). This might be taken as evidence that the search was not truly parallel in this task, at least not in the sense of Treisman (1998), where parallelism is defined by flat visual search functions. It is important to note here that the animal categorization task used in the present experiment cannot be directly compared to classic visual search tasks. Indeed, searching for animals in one natural scene already relies on some form of parallelism. Hence processing 4 natural scenes is

considerably more challenging than searching for one target object among three distractor objects. Despite this qualitative difference in the nature of the stimuli, VanRullen, Reddy and Koch (in press) showed recently with stimulus set sizes ranging from 1 to 16 natural scenes, that the animal/non-animal categorization task used in our experiment cannot be performed in parallel, i.e. RT increased with the number of distractors, mirroring the effects found with simple forms. Surprisingly, the animal task can be performed without focal attention, in the periphery, while subjects are concentrated on a demanding central task (Li, VanRullen, Koch & Perona, 2002). The puzzling contrast between the experimental results found in the visual search task and in the dual-task paradigm has led VanRullen et al. (in press) to hypothesize that natural scenes are processed pre-attentively but not in parallel. While the term "parallel" refers to low-level segmentation mechanisms that can extract odd objects from an array, the term "pre-attentive" implies the existence of neurons in the ventral pathway coding for high-level representations of objects in natural scenes that can be activated without focused attention. These representations would be built through our interaction with the world. Such high-level object "filters" might allow the detection of objects in the dual task used by Li et al. However, these representations would not be immune to local competition in the ventral pathway, which typically occurs in natural scenes as a consequence of the presence of multiple objects (Chelazzi et al., 1998). Thus, a possible explanation for the 'parallel' processing presented in our previous report might be due to the fact that neurons responding to complex objects in the occipito-temporal areas are strongly biased toward contralateral stimuli, receiving virtually no interference from ipsilateral stimuli (Chelazzi et al., 1998). Consequently, more competition was expected in the 2-image intra-hemifield condition.

Presenting the distractor in the same hemifield as the target had the same consequences as those reported in the inter-hemifield condition, except that the capacity to detect targets decreased. This effect could indeed reflect intra-hemisphere competition. Alternatively, as we have suggested previously, this competition might rather take place at a higher level of integration, for example in frontal areas (Rousselet et al., 2002). Teasing these two hypotheses apart is rather difficult on the basis of behavioral data. The next section provides electrophysiological evidence that favors the second alternative. But already, the 4-image results are providing cues. The fact that there was a further drop in performance with 4 images compared to the 2-image condition seems to fit with the classic view that IT neuronal receptive fields cover the entire visual field, so that 3 distractor images would normally be expected to increase the competitive effects on the visual processing of the target image. Paradoxically, in the next section we develop an opposite argument: because IT receptive fields have recently been reported not to cover the entire visual field (Op De Beeck & Vogels, 2000; Rolls, Aggelopoulos & Zheng, 2003) and appear typically biased toward the contralateral hemifield (Chelazzi et al., 1998), the performance drop between the 2-intra image and the 4-image conditions might be due to a competition taking place at a higher level of integration, possibly in prefrontal cortex.

4. Electrophysiology

Differential activities were computed by subtracting correct no-go trial ERPs from correct go trial ERPs. In the go/no-go paradigm used here, this technique has been shown to allow access to task related effects independently of non-controlled low-level differences (VanRullen & Thorpe, 2001; Macé, Rousselet, Sternberg et al., 2002) and without the need to make assumptions about putative links between ERP components and

underlying sources (Makeig, Westerfield, Jung et al., 2002). In previous experiments, the onset latency of the differential activity as proved to be a good indicator of processing speed in categorization tasks (Delorme, Rousselet, Macé & Fabre-Thorpe, in press; Fabre-Thorpe et al., 2001; Rousselet et al., 2002). In addition, the amplitude of the differential activity has been found to increase with subject accuracy, somehow reflecting the quality of processing (Fabre-Thorpe et al., 2001; Rousselet et al., 2002; Thorpe, Bacon, Rousselet et al., 2002). Two components, one occipito-temporal and one frontal, were isolated. Their topography was identical to the one presented in our previous report (Rousselet et al., 2002). They are analyzed in the two next sections.

4.1. Effect of processing an increased number of images on occipital ERP

The left row of Fig.3 shows the event-related potentials recorded over occipital electrodes. Independently of image status (target or distractor), there was a strong effect of image condition on the amplitude of the overall electrophysiological signal. This effect was certainly due to the large physical differences between experimental conditions. Differential activities were thus used to get access to task related effects independently of these physical differences. Occipital differential activities were almost superimposed in the 1-image condition and the two 2-image conditions. With four images, the differential activity tended to have a later onset and its amplitude was clearly reduced compared to the three former conditions. Paradoxically, the onset latency of the differential activity was significantly longer with one image (175 ms) than with two images presented in both the inter- (155 ms) or intra-hemifield (164 ms) conditions. The longest onset latency was found in the 4-image condition (190 ms). This result is at odds with our previous findings showing either no differences in differential activity onset as a function of behavioral RT (Thorpe et al., 1996) or an earlier onset associated with shorter RT (Delorme, Rousselet, Macé & Fabre-Thorpe, in press). This result might be due to a higher variability in the electrophysiological data in this experiment compared to the previous ones probably because of task difficulty. Therefore, we used several other measurements to assess the task effects on visual processing. First, we analyzed the latency and the amplitude of the peak of the differential activity. As in our previous results (Rousselet et al., 2002; Fize, Fabre-Thorpe, Richard et al., in revision), the occipital differential activity was strongly biased toward sites contralateral to the target (as shown by an interaction between the laterality and the hemisphere factors, F=30.8, p<0.0001), thus the analysis concentrated exclusively on contralateral posterior electrodes. Regardless of the 1-, 2- or 4-image conditions, the differential activity reached its peak at the same latency, around 250 ms (Figure 3). However, its amplitude tended to decrease with task difficulty and thus with error rate (F=5.4, p=0.008). The peak amplitude in the 4-image condition was significantly lower than in each of the three other conditions (all p<0.03). However, peak amplitude in these three other conditions did not differ from one another. Post-hoc comparisons performed separately on each posterior electrode also failed to reveal differences between these conditions. Mean amplitude between 200 and 250 ms post-stimulus presented the same pattern, with two occipital sites at which there was a significant effect of the number of images (CB1-CB2 and CB1'-CB2', respectively F=5.3, p=0.007; F=7.3, p=0.002), the amplitudes associated with the processing of the three conditions with 1 or 2 images being higher than the one associated with the processing of 4 images (paired t-test, all p<0.03). No mean amplitude differences were found in the 150-200 ms interval. Thus it appeared that one or two images, whether presented in the same or different

hemifields, were processed to the same extent in posterior visual areas. It is only in the four-image condition that target processing suffered significantly from the competition induced by the distractors.

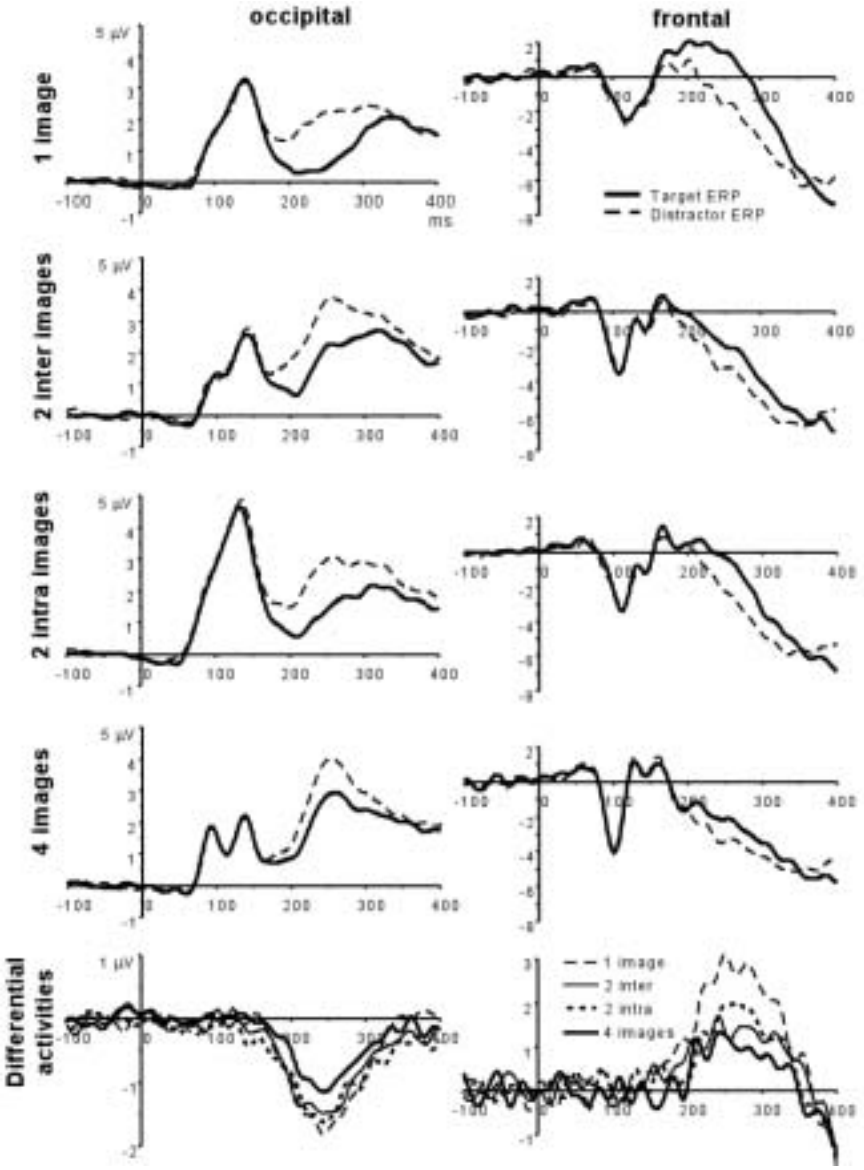


Fig. 3. ERPs and differential activities. ERPs on correct target (plain line) and distractor (dotted line) trials are shown for each condition of presentation at occipital sites contralateral to the target and frontal sites ipsilateral to the target. Occipital signals were characterized by an initial positive deflection followed by a negative going potential. This negativity was increased on target trials, giving rise to a differential activity between target and distractor trials in the 150-350 ms time window. Frontal signals presented the reverse pattern. Data are shown by pooling signals over quadrant and hemisphere dimensions from contralateral electrodes CB1-CB2 (occipital) and FP1-FP2 (frontal).

4.2. Effect of processing an increased number of images on frontal ERP

Frontal differential activity was higher over sites ipsilateral to the presentation (F=5.5, p=0.034) and therefore analysis concentrated on ipsilateral anterior electrodes. This was expected given recent evidence showing that the signal recorded over frontal electrodes in tasks requiring the categorization of a central image

can be explained in large part by dipoles situated in the ventral pathway (Delorme et al., in press). Thus, with one image, the frontal activity seems to partially mirror with a reverse polarity, the contralateral activity recorded over occipital electrodes. However, we found recently a dissociation between these two activities, frontal electrodes capturing in addition signals related to late stimulus evaluation (Rousselet et al., 2002; see also Hopf & Mangun, 2000; Lange, Wijers, Mulder & Mulder, 1998; Potts, Liotti, Tucker & Posner, 1996; Potts & Tucker, 2001). In the present experiment, focusing on ipsilateral electrodes was thus intended to highlight differences between occipital and frontal signals.

Frontal differential activity onsets were found to present a pattern similar to the pattern found at the occipital level. The shortest differential activity in the 1-image condition appeared at 183 ms, which was before 2-intra hemifield images (194 ms) and 4-images (203) but after 2-inter hemifield images (174 ms).

As in the case of the occipital differential activity, the frontal peak latency (around 260 ms, Fig.3) was not significantly different between the four conditions. Its amplitude was affected by the image factor (F=4.9, p=0.017) but with a different pattern from the one found at occipital sites. The largest amplitude was found in the 1-image condition (Fig.3, bottom right). It diverged significantly from that obtained in the intra-hemifield 2-image condition (p=0.016) and the 4-image condition (p=0.024) but not from the amplitude in the inter-hemifield 2-image condition. Those three last conditions did not differ from on another. When performed on each frontal electrode, post-hoc analysis showed the same effects (although the comparison between 1-image and 2-image inter-hemifield activities just failed to reach statistical significance, p=0.051 on electrodes FP1-FP2). From 250 to 300 ms post-stimulus, the differential activity mean amplitude reached in the 1 image condition surpassed the mean amplitude in the 3 other conditions at sites FP1-FP2 (all p<0.025).

4.3. ERP Results: discussion

In this experiment, ERPs were used to get a better insight into the mechanisms involved in the simultaneous processing of two and four photographs of natural scenes. In particular, the mean amplitude of the differential activity in the different experimental conditions provided a measure of processing as a function of time before behavioral responses were triggered.

One of the main outcomes of this analysis was that the pattern of differential activities in the 2-image inter-hemifield condition replicated the dichotomy found previously between occipital and frontal sites (Rousselet et al., 2002). With a single scene, the differential activity recorded over frontal sites has been shown to mirror in large part the occipital differential activity (Delorme et al., in press; Fize et al., in revision; see also Anllo-Vento, Luck & Hillyard, 1998; Kenemans, Lijffijt, Camfferman & Verbaten, 2002). However, when two different scenes presented in a different hemifield have to be processed simultaneously, an additional mechanism is reflected in the frontal differential activity (Rousselet et al., 2002). Specifically, we found evidence in our previous study and in the present one that the amplitude of the frontal differential effect was reduced when two images were presented despite there being no effect whatsoever on the peak latency and the peak amplitude of the differential effect at occipito-temporal sites. This result is in keeping with the behavioral literature (Friedman & Campbell Polson, 1981; Sereno & Kosslyn, 1991) as well as studies in patients (Luck, Hillyard, Mangun & Gazzaniga, 1994) showing that each hemisphere might act as an independent resource limited visual processor. In our experiments using photographs of natural scenes, it seems that the occipital areas in each hemisphere were

working independently. This was probably due to the brief and lateralized image presentations, each hemisphere being first stimulated by the contralateral image. Furthermore, single-unit recordings in monkeys suggest that the interference between two stimuli presented in two different hemifields does not appear to take place in the ventral pathway, despite the existence of trans-callosal connections (Chelazzi et al., 1998). We thus reiterate our initial conclusion that much of the interference in the 2-image inter-hemifield condition might arise at the level of prefrontal cortex; a proposition compatible with a two-stage competitive model of visual processing (Chun & Potter, 1995). Indeed, the present results seem to fit with the idea of a bottleneck located close to the response output stage. The fact that the differential neuronal activity at frontal electrodes was higher on average for one-image trials than for 2-image inter-hemifield trials suggest the existence of a competition taking place at an integration stage at which object representations processed in occipito-temporal areas would compete to gain control of the unique response output, which is a kind of mechanism well documented in prefrontal cortex (Bichot, Chenchall Rao & Schall, 2001; Freedman, Riesenhuber, Poggio & Miller, 2001; Rao, Rainer & Miller, 1997; Schall & Thompson, 1999; Tanji & Hoshi, 2001). In keeping with this idea, previous studies have suggested that ERP signals recorded over frontal electrodes might reflect a frontal activity related to stimuli evaluation and behavioral response choice (Hopf & Mangun, 2000; Lange et al., 1998; Potts et al., 1996; Potts & Tucker, 2001).

Another major outcome of this experiment was that adding a second image in the same hemifield as the target (intra-hemifield condition) had virtually the same consequences as those reported in the inter-hemifield condition. It was only at the behavioral level that a main difference between the two 2-image conditions appeared, the capacity to detect targets slightly decreasing in the intra-hemifield condition relatively to the inter-hemifield condition, as already mentioned above. This could be the hallmark of intra-hemisphere competition, as suggested by data showing that competition takes place mainly between stimuli presented in the contralateral visual hemifield (Chelazzi et al., 1998; Hopf, Luck, Girelli et al., 2000). However, we found again that the major source of interference might happen in the prefrontal cortex rather than in the ventral pathway. So it does appear that two images might be processed concurrently in the same hemisphere without much interference. However, it might be that our ERP recordings were not able to capture an occipital effect between the inter- and intra-hemifield conditions because of insufficient signal to noise ratio. A difference between these two conditions could also be expected at the level of the frontal differential activity, where neurons involved for instance in visual-motor decisional mechanisms seem to preserve retinotopic (Moore & Armstrong, 2003; Schall & Thompson, 1999) or hemisphere preferences (Barcelo, Suwazono & Knight, 2000). Further experiments, perhaps with many more experimental trials to improve the signal to noise ratio, will be necessary to capture these putative signal differences in frontal and occipital activity.

It was only in the 4-image condition that a significant effect was found on the occipital differential activity. The clear impact seen on its amplitude with four simultaneously presented images compared to the 2-image conditions might indicate that the competition in one hemisphere integrated information from both hemifields due to the large receptive fields of IT neurons or to competition involving trans-callosal connections. However, there are two reasons why such conclusion cannot be drawn from this result. First, contrary to popular belief, IT neuronal receptive fields do not typically cover the entire visual field and can instead be rather small (Op De Beeck & Vogels, 2000); they may even be particularly small in size in response to objects in the context of natural scenes as opposed to blank backgrounds (Rolls et al., 2003). Second, there is evidence that ipsilateral

stimuli do not enter into competition with contralateral stimuli because IT neurons are strongly driven by contralateral stimuli (Chelazzi et al., 1998). This is also evident in patients suffering from visual extinction after frontal and/or parietal lesions (Driver & Vuilleumier, 2001). When two objects are presented to the two hemifields of these patients, the one contralateral to the side of the lesion tends not to be perceived consciously because it enters in competition with the other one. However, there is evidence that this competition is not taking place (or only partially) in the ventral pathway because the extinguished object seems to be recognized implicitly on the basis of neuronal responses in occipito-temporal areas classically responding to objects (Driver & Vuilleumier, 2001). Hence, as an alternative to the idea that the competition observed with 4 images is taking place in IT, we suggest that the effect of the 4-image condition on occipital activity might be due to feedback from prefrontal cortex that integrates evidence from the two hemispheres to make a category related decision (Freedman et al., 2001; Rainer, Asaad & Miller, 1998). This hypothesis fits with the existence of first frontal activations as early as 80 ms after stimulus onset, suggesting that feedback loops have enough time to take place very early during visual processing (Foxe & Simpson, 2002). It also fits with the recent demonstration of frontal modulations at 125 ms on occipital ERPs (Barcelo et al., 2000).

The 4-image condition being so challenging probably led to a very slow and less efficient accumulation of evidence, explaining the later onset and lower amplitude of the differential activity in this condition compared to the 1- and 2-image conditions. However, we do not want to overstate this conclusion, which is only plausible if one assumes that the differential activity recorded over occipito-temporal electrodes reflects at least to some extent the direct involvement of high-level object mechanisms implemented in the ventral pathway. In a similar vein, complementary experiments with high-density electrode recordings will be necessary to isolate precisely the origin of the interference when human subjects are presented with several images simultaneously. Until careful source analyses are performed in this kind of task, the present conclusions are only speculative, but provide a realistic account of the data obtained so far.

5. Some insights into the origin of the differential activity

Although it is not yet possible to draw definitive conclusions regarding the patterns of differential activities recorded in the present task, our experimental protocol can provide some evidence about where in the visual system the target-distractor interference occurred and what is the origin of the differential activity.

5.1. Differential activity on non-correct trials

In this experiment, because of task difficulty, a sufficient number of incorrect responses were available to evaluate the cerebral activity associated with false alarms and missed targets. This analysis was thought to provide a better understanding of the relationship between differential activity amplitude and decision mechanisms underlying response selection. The differential activity presented so far in this paper was calculated by subtracting the mean ERP associated with correct no-go distractor trials from ERP recorded on correct go target trials. Using the ERP on correct no-go distractor trials as a reference, we determined the differential responses produced by incorrect go trials (false alarms) and by incorrect no-go trials (target misses). To this aim, the signal associated with the correct no-go distractor trials was subtracted separately from the signal associated

with each of the two incorrect trials. This produced a false alarm differential activity and a target miss differential activity respectively. The mean amplitudes of the correct, target miss and false alarm differential activities were determined for occipital, frontal and parietal electrodes with time windows of 50 ms (Fig. 4). There was no evidence for a differential activity associated with missed targets, while a clear differential activity was seen with false alarms. The false alarm differential activity was conspicuous and shared the same time course as the classical differential activity but with a smaller amplitude (see Fig. 4 for details and statistical results).

In summary, the cerebral activity elicited by target-images in which the subjects did not detect the target did not diverge from those induced by distractor images. Inversely, when a subject responded incorrectly to a distractor image as if it contained an animal (false alarm), the cerebral activity diverged from other distractor images as in the case of target-images.

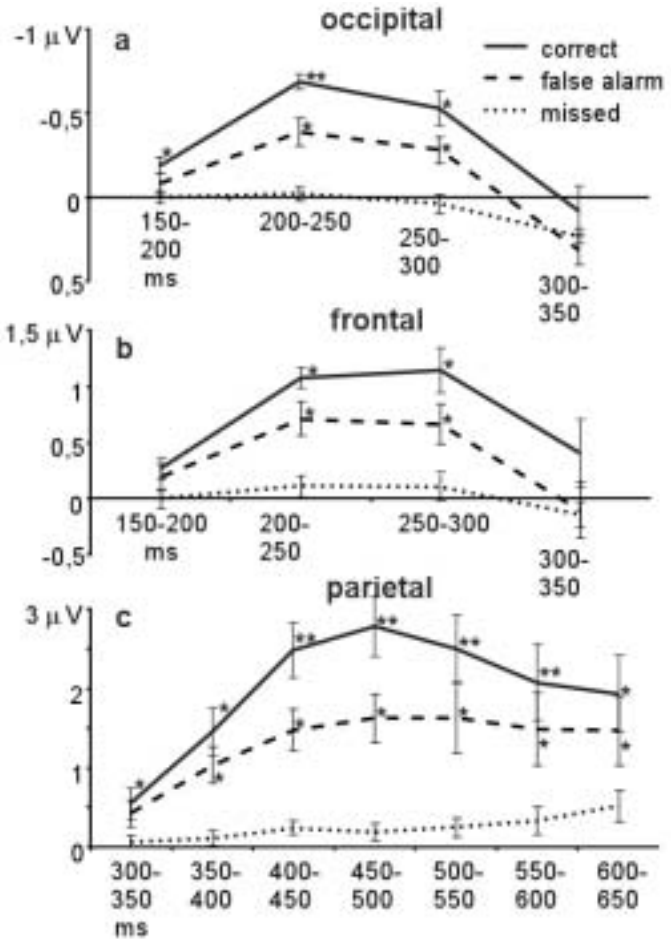


Fig. 4. Mean amplitude of the differential activity to false alarms and missed targets. Note that the differential activity was negative at occipital sites and positive at frontal and parietal sites. Data with associated standard errors are pooled across the image and the quadrant factors using 50 ms time windows. Data were first entered in omnibus ANOVAs with 6 within-subject factors: correct/false alarm/missed, image, upper/lower visual field, left/right visual field, left/right hemisphere electrodes (occipital and frontal differential activities only), electrodes. Small stars indicate significant post-hoc two by two comparisons (all $p<0.05$). A single star over the classical or the false alarm differential activity mean amplitude indicates that it differed significantly from the differential activity to missed targets. Two stars over the classical differential activity amplitude indicate that it differed significantly from the amplitude reached in the false alarm and the target missed conditions. At those

occipital sites where accuracy effects were larger (CB1-CB2 & CB1'-CB2', 150-200 ms: F=4.3, p=0.008; 200-350 ms: all F>12.0, all p<0.0001) post-hoc analysis revealed that a higher amplitude was also associated with correct trials compared to FA trials in the 150-200 ms time window (CB1'-CB2', p=0.046) and in the 250-300 ms time window (CB1-CB2, p=0.013; CB1'-CB2', p=0.014). At those frontal sites where accuracy effects were stronger (FP1-FP2, F=4.7, p=0.006) this effect was nearly significant in the 250-300 ms time window (p=0.052). Note that parietal effects are plotted on a different time scale and that this delayed effect probably reflects motor response generation.

5.2. Upper versus lower visual fields

Finally, as the upper and lower hemifields do not have the same cerebral representation (Tootell, Hadjikhani, Mendola et al., 1998), the data were analyzed separately for each hemifield and compared. When an image containing an animal was presented in the upper or lower visual field, it defined respectively upper and lower target trials. Results were entered in an ANOVA with five within-subject factors: image (4 levels), upper/lower visual field, left/right visual field, left/right hemisphere electrodes, electrodes (6 levels). The data presented below have been collapsed over the image, left/right visual field and hemisphere dimensions for simplicity. This was possible given that these three factors did not interact with the upper/lower factor.

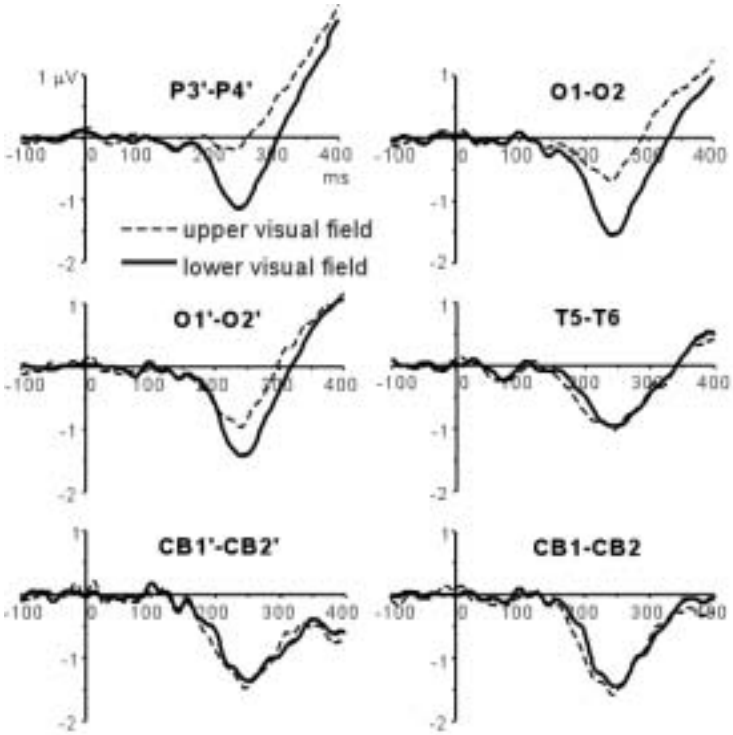


Fig. 5. Upper vs. lower occipital differential activities. Occipital differential activity is shown at each of the six posterior electrode pairs where it was recorded: CB1-CB2 and CB1'-CB2', O1-O2, O1'-O2' and T5-T6 and P3'-P4'. Data are pooled across image, left/right field and hemisphere factors.

The occipital differential activity onsets were virtually identical in upper (169 ms) and lower (173 ms) visual fields, but the peak latency presented a reliable advance when the targets appeared in the upper visual field compared to the lower visual field, consistent with the 9 ms behavioral effect (F=16.7, p=0.001, upper = 240 ms, lower = 253 ms). On the other hand, the peak amplitude of the differential activity was higher when targets appeared in the lower visual field (F=5.8, p=0.029). Although this effect seems to contradict the behavioral

results, it differed significantly depending on the electrodes (F=36.2, p<0.0001). As depicted in Fig. 5, there were no significant differences between upper and lower signals at the temporal T5-T6 sites and the more posterior CB1'-CB2' and CB1-CB2 sites. At sites medial to T5-T6 and anterior to CB1-CB2 and CB1'-CB2', differential activity was significantly higher for lower visual field targets when compared to upper visual field ones (P3'-P4': F=22.5, p<0.0001; O1'-O2': F=9.0, p=0.009; O1-O2: F=34.0, p<0.0001). A study of the mean amplitude of the differential activity between 150 and 350 ms with time windows of 50 ms revealed that the interaction between position and electrode factors was already present in the first time window (150-200 ms: F=10.0, p<0.0001) and in all the following ones (all F>9.6, p<0.001). At frontal sites, these effects were more difficult to assess because of the lower electrode coverage. Differential activity peak amplitude presented a borderline interaction between the hemifield and the electrode factors (F=3.7, p=0.057). This effect was significant between 200-350 when analysis was performed on mean amplitude in 50 ms time windows (all F>3.7, all p<0.05). This was due to a higher amplitude of the differential activity with upper hemifield targets at the most anterior sites (FP1-FP2) and a higher differential activity amplitude with lower targets at more lateral and dorsal frontal sites (respectively F7-F8 and F3-F4). However, post-hoc analysis on each electrode failed to reach significance.

5.3. Origin of the differential activity: discussion

When the original study on differential ERP effects related to animal categorization was published, the cause of the differential activity was unclear (Thorpe et al., 1996). The differential activation could reflect the activity of neural mechanisms selectively responding to animals. Alternatively, it could reflect inhibitory mechanisms specific to no-go trials. Indeed, some of the results, and in particular the fact that there was no correlation between the onset latency of the differential effect and behavioral reaction time, as recently confirmed by Johnson & Olshausen (2003), were consistent with such a hypothesis (Thorpe et al., 1996). This activity might also be related to the decision that an animal is present, a decision being made in the ventral pathway, in cortical areas such as V4 and IT, or at a higher level of integration, like in the prefrontal cortex where explicit categorization is thought to take place.

In the present experiment, we were able to test more directly these various hypotheses. The absence of differential activity for missed targets and the presence of a reliable differential activity effect associated with false alarms is consistent with the hypothesis that this activity could reflect the activation of neurons tuned to animals or animal features. It is reasonable to imagine that once a sufficient number of these neurons are recruited by the visual stimulation, their activity triggers a behavioral response, whether the target was really there or not. Although this conclusion might provide us with a simple account of the origin of the differential activity, an additional argument suggests that it might not be related directly to the activation of populations of "animal detectors". Indeed, we have argued above that the pattern of occipital differential activity in the 4-image condition speaks rather in favor of the involvement of feedback from prefrontal cortex to the ventral pathway in generating such activity. According to this stance, the differential activity would reflect late stages in the target selection process. Given the very indirect way by which this conclusion is reached, we do not want to make a strong case of it. Further experiments are strongly needed to strengthen or falsify this hypothesis. One piece of evidence that strengthens the idea that the occipital differential effects result from feedback related phenomena

comes from recent data using a choice saccade task in which two images were presented simultaneously to the left and right of the fixation point and the subjects were required to make an eye movement to the side that contained an animal. Remarkably, the fastest behavioral responses occurred between 130 and 150 ms, that is to say, before the onset of the main differential ERP effect at occipital sites (Kirchner, Bacon-Macé & Thorpe, 2003).

The analysis of the upper versus lower bias in the present results also provides evidence that favors a late account of the differential activity. The fact that the distribution and amplitude of the differential activity depended on the retinal location of the images might be taken as evidence that the structures involved in its generation are themselves retinotopically organized. In contrast with this point of view, if the differential activity would reflect directly the activation of a unitary decision mechanism, there would be no reason for it to show differences depending on where the target is located. Luck, Girelli, McDermott & Ford (1997a) made a similar deduction about the N2pc, an ERP component registered over posterior electrodes contralateral to the target in a visual search task. They found that the N2pc was larger for lower compared to upper visual field targets, in agreement with the hypothesis that this activity was generated in a human area V4 homologous to the monkey area V4. In monkeys, V4 is organized so that most of the lower visual field is represented dorsally, while the upper visual field is represented ventrally (Gattass, Sousa & Gross, 1988). If we make the plausible assumption that this organization is preserved in humans, neuronal activity originating in V4 would be more easily recorded by posterior electrodes following the presentation of lower visual field targets, because these electrodes would be situated closer to the putative dorsal representation of the lower visual field. Given that we found the same pattern of results in the present experiment, it seems unlikely that the effects reported here are produced in areas homologous to monkey IT cortex. Indeed, while neuronal responses in IT preserve some retinotopic information (DiCarlo & Maunsell, 2003; Kline, Amador-Garza, McAdams et al., 2003), the population of IT neurons as a whole does not present a bias such as the one found in V4 in its anatomical organization. The idea of the involvement of an area like V4 in generating the differential activity is further strengthened by the lack of interference between two images presented in the same hemifield, as discussed previously. Furthermore, we have already argued that the decrease of the occipital differential activity amplitude in the 4-image condition is not likely to reflect the involvement of IT cortex in the generation of the differential activity.

If we now suppose that intermediate level areas are involved, equivalent to V4 for example, there are various options. One is that neurons at this level in the visual system are already capable of showing category specificity at the moment they start firing. But the relatively late (150 ms) latency for the start of this activation seems rather too long for feed-forward V4 activation. An alternative would be to suppose that the differential activation of intermediate level structures could result from the activation of back-projections from structures such as IT and possibly prefrontal cortex. One reason for such reactivation might be to form a more detailed visual representation of the selected object. We must also leave open the possibility that the categorization of animals in natural scenes, as indexed by the differential activity, does not rely on “high-level” representations, but rather on features of intermediate complexity that might be more diagnostic for this kind of task (Rousselet, Macé & Fabre-Thorpe, 2003; Ullman, Vidal-Naquet & Sali, 2002). This would leave more time for interactions in the ventral pathway to occur. Alternatively, or in addition, the differential activity might reflect the spatial selection of a target based on its component features. This spatial selection might require interactions between prefrontal cortex and the ventral pathway (e.g. Barcelo et al., 2000; Gehring & Knight, 2002; Moore &

Armstrong, 2003). This proposal is very similar to the one made by Luck and colleagues using visual search of relatively low-level properties (Hopf et al., 2000; Luck et al., 1997a) and follows the lines of evidence showing that visual discrimination might rely on spatial selection before a response can be produced (Chelazzi, 1999).

However, it is not clear for the moment whether the differential activity reported in studies from our group can be directly compared to the N2pc reported by Luck and colleagues. The N2pc typically has an onset at about 180 ms post-stimulus and is proportionally larger for increasingly difficult searches, for example when distractors share more and more features with the target, and is absent for simple search tasks (Luck & Hillyard, 1994). It is larger for conjunction targets than for single-feature pop-out targets (Luck et al., 1997a). It is also larger for a target and a distractor placed close together in one hemifield than when a target and a distractor are in different hemifields (Luck et al., 1997a) and appears to reflect the attenuation of distractor interference (Hopf, Vogel, Woodman et al., 2002). Together with the finding of a larger N2pc when subjects are required to foveate the target (Luck et al., 1997a), this suggests a close link between this occipital modulation and spatial attention. The generators of this component seem to be in lateral occipito-temporal regions, with an additional contribution from posterior parietal cortex when the task is particularly challenging (Hopf et al., 2000). This pattern of results directly links the N2pc component to single-unit attention effects observed in areas IT and V4 of the macaque (Chelazzi et al., 1998; Chelazzi, Miller, Duncan & Desimone, 2001; Luck, Chelazzi, Hillyard & Desimone, 1997b) as discussed in Luck et al. (1997a).

The problem is that we never manipulated all these factors in our own experiments with natural scenes. Moreover, the amplitude of the differential activity was almost systematically larger with higher accuracy. In fact, most of the time the amplitude of the differential activity appears to be inversely proportional to the difficulty of the task as stated in the introduction and as shown in the present experiment. New experiments are required to further understand the relationship between the differential activity and the N2pc, but it would be surprising if the two components reflected totally different mechanisms.

Alternatively, the differential activity reported here might be more related to the occipito-temporal N1 component, which shares a similar latency and distribution to the differential activity effects seen in our task. Interestingly, the mechanism reflected by the N1 seems to be a discrimination process (Mangun & Hillyard, 1991; Vogel & Luck, 2000) that takes place within the focus of spatial attention (Luck, Fan & Hillyard, 1993; Luck, Hillyard, Mouloua et al., 1994; Luck & Hillyard, 1995). But contrary to the finding that the amplitude of the N1 discrimination effect is unaffected by the difficulty of the task (Vogel & Luck, 2000) we found in the present experiment and repeatedly in previous experiments from our group that the amplitude of the differential activity is modulated by the task difficulty (Delorme et al., in press; Fabre-Thorpe et al., 2001; Thorpe et al., 2002).

For the moment it is not clear how the differential activity reported here and in previous experiments from our lab is related to the N2pc component and N1 discrimination effects. But at the same time it is striking to see how all these studies converge on the conclusion that the discrimination of a target stimulus, be it in a visual search task or a foveal discrimination task, systematically relies on some sort of spatial selection. A similar conclusion has been reached from the studies of patients suffering from hemineglect following a parietal lesion. Despite an intact ventral pathway ipsilateral to the lesion that allows object processing up to the semantic level, these patients are not conscious of and cannot act upon stimuli controlateral to the lesion (Driver & Vuilleumier, 2001). It has been suggested that even if objects are processed in parallel in the ventral pathway,

the parietal cortex, probably in relation with the frontal cortex, might trigger a final shift of spatial attentional resources toward the potential target in order to make an explicit judgment about it (Chelazzi, 1999). In difficult conditions, when signal to noise ratio associated with the target can be relatively low, this shift of spatial resources is also thought to amplify the outcome of the target selection by a parallel competitive mechanism in the ventral pathway (Chelazzi, 1999). Thus, whether the ventral pathway works in parallel or not, there may well be a serial stage in visual processing that is needed to explicitly select a stimulus representation. This might be what is reflected by the differential activity reported here.

If this hypothesis is true, it would provide stronger temporal constraints for models of visual object processing than was assumed previously. Specifically, feedforward models of the ventral visual system might be able to account for the initiation of the spatial selection of targets in natural scenes in 150 ms. This view is also consistent with several innovative models of visual processing in which high level units interact very rapidly with low level units in order to refine and/or select object representations, possibly by an interplay between ventral and dorsal visual pathways (Bullier, 2001; Deco, Pollatos & Zihl, 2002; de Kamps & van der Velde, 2001).

6. Summary

This experiment investigated the limits of parallelism in a task requiring human subjects to detect animals in one, two or four photographs of natural scenes presented briefly and simultaneously in different quadrants. At the behavioral level, accuracy decreased and reaction times increased with the number of images to process. Thus, animals did not “pop-out” from natural scenes. However, a simple parallel model of visual processing provided a relatively good fit of the accuracy data obtained. At the electrophysiological level, the subtraction of distractor ERPs from animal ERPs revealed a differential activity whose amplitude seemed to be related to behavioral accuracy, whereas no correlation could be found between its latency and the behavioral reaction times. Occipital differential activities suggested a parallel processing of two natural scenes, whether they were presented in different hemifields or in the same hemifield. Both behavioral and electrophysiological data suggested that the main interference in this task was not due to intra-hemisphere competition. Furthermore, based on the literature reporting single-unit and neuropsychological data, the drop in behavioral performance and in amplitude of the occipital differential activity observed in the 4-image condition was interpreted as being due to feedback from prefrontal cortex. This hypothesis is strengthened by the finding that the major source of interference was found at the level of frontal electrodes, and not occipital electrodes, a dissociation taken as evidence for a late selection account of the behavioral data. More generally, additional analyses suggested that the occipital differential activity reflects late stages in the target selection process, involving feedback from higher-level areas on retinotopically organized areas such as V4. During the rapid categorization of objects in natural scenes, the occipital differential activity could reflect a final shift of attentional resources within the ventral pathway towards a potential target.

Acknowledgements

The work was supported by the ACI Cognitique grants n°COG35 & 35b. Financial support was provided to Guillaume A. Rousselet by a Ph.D. grant from the French government. The authors declare that they have no

competing financial interests. We thank Nadège M. Bacon-Macé for her invaluable help in making counterbalanced experimental series and programming image presentation. We also thank Oliver Joubert for his very valuable help running the experimental sessions in the two control behavioral experiments. The manuscript was improved by many discussions with Rufin VanRullen.

References

Allison, T., Puce, A., Spencer, D. D., & McCarthy, G. (1999). Electrophysiological studies of human face perception. I: Potentials generated in occipitotemporal cortex by face and non-face stimuli. *Cerebral Cortex, 9*(5), 415-430.

Anllo-Vento, L., Luck, S. J., & Hillyard, S. A. (1998). Spatio-temporal dynamics of attention to color: evidence from human electrophysiology. *Human Brain Mapping, 6*, 216-238.

Barcelo, F., Suwazono, S., & Knight, R. T. (2000). Prefrontal modulation of visual processing in humans. *Nature Neuroscience, 3*(4), 399-403.

Bichot, N. P., Chenchal Rao, S., & Schall, J. D. (2001). Continuous processing in macaque frontal cortex during visual search. *Neuropsychologia, 39*(9), 972-982.

Biederman, I. (1972). Perceiving real-world scenes. *Science, 177*(43), 77-80.

Bullier, J. (2001). Integrated model of visual processing. *Brain Research Reviews, 36*(2-3), 96-107.

Chelazzi, L. (1999). Serial attention mechanisms in visual search: a critical look at the evidence. *Psychological Research, 62*(2-3), 195-219.

Chelazzi, L., Duncan, J., Miller, E. K., & Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *Journal of Neurophysiology, 80*(6), 2918-2940.

Chelazzi, L., Miller, E. K., Duncan, J., & Desimone, R. (2001). Responses of neurons in macaque area V4 during memory-guided visual search. *Cerebral Cortex, 11*(8), 761-772.

Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance, 21*(1), 109-127.

de Kamps, M., & van der Velde, F. (2001). Using a recurrent network to bind form, color and position into a unified percept. *Neurocomputing, 38-40*, 523-528.

Deco, G., Pollatos, O., & Zihl, J. (2002). The time course of selective visual attention: theory and experiments. *Vision Research, 42*, 2925-2945.

Delorme, A., Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (in press). Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Cognitive Brain Research*.

DiCarlo, J. J., & Maunsell, J. H. R. (2003). Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *Journal of Neurophysiology, 89*, 3264–3278.

Driver, J., & Vuilleumier, P. (2001). Perceptual awareness and its loss in unilateral neglect and extinction. *Cognition, 79*(1-2), 39-88.

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience, 13*(2), 171-180.

Fize, D., Fabre-Thorpe, M., Richard, G., Doyon, B., & Thorpe, S. (in revision). Foveal vision is not necessary

for rapid categorisation of natural images: a behavioural and ERP study.

Foxe, J. J., & Simpson, G. V. (2002). Flow of activation from V1 to frontal cortex in humans. A framework for defining "early" visual processing. *Experimental Brain Research, 142*(1), 139-150.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science, 291*(5502), 312-316.

Friedman, A., & Campell Polson, M. (1981). Hemispheres as independent resource systems: limited-capacity processing and cerebral specialization. *Journal of Experimental Psychology: Human Perception and Performance, 7*(5), 1031-1058.

Gattass, R., Sousa, A. P., & Gross, C. G. (1988). Visuotopic organization and extent of V3 and V4 of the macaque. *Journal of Neuroscience, 8*(6), 1831-1845.

Gehring, W. J., & Knight, R. T. (2002). Lateral prefrontal damage affects processing selection but not attention switching. *Cognitive Brain Research, 13*(2), 267-279.

Gross, C. G., Bender, D. B., & Rocha-Miranda, C. E. (1969). Visual receptive fields of neurons in inferotemporal cortex of the monkey. *Science, 166*(910), 1303-1306.

Hopf, J.-M., & Mangun, G. R. (2000). Shifting visual attention in space: an electrophysiological analysis using high spatial resolution mapping. *Clinical Neurophysiology, 111*, 1241-1257.

Hopf, J.-M., Vogel, E., Woodman, G., Heinze, H.-J., & Luck, S. J. (2002). Localizing Visual Discrimination Processes in Time and Space. *Journal of Neurophysiology, 88*, 2088–2095.

Hopf, J. M., Luck, S. J., Girelli, M., Hagner, T., Mangun, G. R., Scheich, H., & Heinze, H. J. (2000). Neural sources of focused attention in visual search. *Cerebral Cortex, 10*(12), 1233-1241.

Jeffreys, D. (1996). Evoked potential studies of face and object processing. *Visual Cognition, 3*, 1-38.

Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. *Journal of Vision*, *3*, 499-512, http://journalofvision.org/3/7/4/, doi:10.1167/3.7.4.

Kenemans, J. L., Lijffijt, M., Camfferman, G., & Verbaten, M. N. (2002). Split-second sequential selective activation in human secondary visual cortex. *Journal of Cognitive Neuroscience, 14*(1), 48-61.

Kinchla, R. A. (1992). Attention. *Annual Review of Psychology, 43*, 711-742.

Kirchner, H., Bacon, N., & Thorpe, S. J. (2003). In which of two scenes is the animal? Ultra-rapid visual processing demonstrated with saccadic eye movements. *Perception (suppl.), 32*, 170 (abstract).

Kline, K., Amador-Garza, S., McAdams, C., Maunsell, J., & Sereno, A. (2003). Spatial and eye position modulation of neuronal activity in anterior inferior temporal and perirhinal cortices. *Journal of Cognitive Neuroscience (suppl.), E293*, 188 (abstract).

Kreiman, G., Koch, C., & Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience, 3*(9), 946-953.

Lange, J. J., Wijers, A. A., Mulder, L. J., & Mulder, G. (1998). Color selection and location selection in ERPs: differences, similarities and 'neural specificity'. *Biological Psychology, 48*(2), 153-182.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences of the United States of America, 99*(14), 9596-9601.

Luck, S. J., Chelazzi, L., Hillyard, S. A., & Desimone, R. (1997b). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of Neurophysiology, 77*(1), 24-42.

Luck, S. J., Fan, S., & Hillyard, S. A. (1993). Attention-related modulation of sensory-evoked brain activity in a visual search task. *Journal of Cognitive Neuroscience, 5*, 188-195.

Luck, S. J., Girelli, M., McDermott, M. T., & Ford, M. A. (1997a). Bridging the gap between monkey neurophysiology and human perception: an ambiguity resolution theory of visual selective attention. *Cognitive Psychology, 33*(1), 64-87.

Luck, S. J., & Hillyard, S. A. (1994). Spatial filtering during visual search: evidence from human electrophysiology. *Journal of Experimental Psychology: Human Perception and Performance, 20*(5), 1000-1014.

Luck, S. J., & Hillyard, S. A. (1995). The role of attention in feature detection and conjunction discrimination: an electrophysiological analysis. *International Journal of Neuroscience, 80*(1-4), 281-297.

Luck, S. J., Hillyard, S. A., Mangun, G. R., & Gazzaniga, M. S. (1994). Independent attentional scanning in the separated hemispheres of split-brain patients. *Journal of Cognitive Neuroscience, 6*(1), 84-91.

Luck, S. J., Hillyard, S. A., Mouloua, M., Woldorff, M. G., Clark, V. P., & Hawkins, H. L. (1994). Effects of spatial cuing on luminance detectability: psychophysical and electrophysiological evidence for early selection. *Journal of Experimental Psychology: Human Perception and Performance, 20*(4), 887-904.

Macé, M. J.-M., Rousselet, G. A., Sternberg, C. R., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Very early ERP effects in rapid visual categorization of natural scenes: distinguishing the role of low-level visual properties and task requirements. *Perception (suppl.), 31*, 132 (abstract).

Makeig, S., Westerfield, M., Jung, T. P., Enghoff, S., Townsend, J., Courchesne, E., & Sejnowski, T. J. (2002). Dynamic brain sources of visual evoked responses. *Science, 295*(5555), 690-694.

Mangun, G. R., & Hillyard, S. A. (1991). Modulations of sensory-evoked brain potentials indicate changes in perceptual processing during visual-spatial priming. *Journal of Experimental Psychology: Human Perception and Performance, 17*(4), 1057-1074.

Moore, T., & Armstrong, K. M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature, 421*(6921), 370-373.

Olson, C. R. (2001). Object-based vision and attention in primates. *Current Opinion in Neurobiology, 11*(2), 171-179.

Op De Beeck, H., & Vogels, R. (2000). Spatial sensitivity of macaque inferior temporal neurons. *Journal of Comparative Neurology, 426*(4), 505-518.

Palmer, J. (1998). Attentional effects in visual search: relating search accuracy and search time. In R. D. Wright (Ed.), *Visual attention* (Vol. 8, pp. 348-388). Oxford (UK): Oxford University Press.

Perrett, D. I., Rolls, E. T., & Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Experimental Brain Research, 47*(3), 329-342.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 509-522.

Potts, G. F., Liotti, M., Tucker, D. M., & Posner, M. I. (1996). Frontal and inferior temporal cortical activity in visual target detection: Evidence from high spatially sampled event-related potentials. *Brain Topography, 9*, 3-14.

Potts, G. F., & Tucker, D. M. (2001). Frontal evaluation and posterior representation in target detection. *Cognitive Brain Research, 11*, 147-156.

Rainer, G., Asaad, W. F., & Miller, E. K. (1998). Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature, 393*(6685), 577-579.

Rao, S. C., Rainer, G., & Miller, E. K. (1997). Integration of what and where in the primate prefrontal cortex. *Science, 276*(5313), 821-824.

Rolls, E. T., Aggelopoulos, N. C., & Zheng, F. (2003). The receptive fields of inferior temporal cortex neurons in natural scenes. *Journal of Neuroscience, 23*(1), 339-348.

Rossion, B., Gauthier, I., Tarr, M. J., Despland, P., Bruyer, R., Linotte, S., & Crommelinck, M. (2000). The N170 occipito-temporal component is delayed and enhanced to inverted faces but not to inverted objects: an electrophysiological account of face-specific processes in the human brain. *Neuroreport, 11*(1), 69-74.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience, 5*(7), 629-630.

Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *Journal of Vision, 3*(6), 440-455.

Schall, J. D., & Thompson, K. G. (1999). Neural selection and control of visually guided eye movements. *Annual Review of Neuroscience, 22*, 241-259.

Schendan, H. E., Ganis, G., & Kutas, M. (1998). Neurophysiological evidence for visual perceptual categorization of words and faces within 150 ms. *Psychophysiology, 35*(3), 240-251.

Sereno, A. B., & Kosslyn, S. M. (1991). Discrimination within and between hemifields: a new constraint on theories of attention. *Neuropsychologia, 29*(7), 659-675.

Sheinberg, D. L., & Logothetis, N. K. (2001). Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. *Journal of Neuroscience, 21*(4), 1340-1350.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience, 19*, 109-139.

Tanji, J., & Hoshi, E. (2001). Behavioral planning in the prefrontal cortex. *Current Opinion in Neurobiology, 11*(2), 164-170.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520-522.

Thorpe, S. J., Bacon, N., Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2002). Rapid categorisation of natural scenes: feedforward vs feedback contribution evaluated by backward masking. *Perception (suppl.), 31*, 150 (abstract).

Thorpe, S. J., Gegenfurtner, K. R., Fabre-Thorpe, M., & Bülthoff, H. H. (2001). Detection of animals in natural images using far peripheral vision. *European Journal of Neuroscience, 14*(5), 869-876.

Tootell, R. B. H., Hadjikhani, N. K., Mendola, J. D., Marrett, S., & Dale, A. M. (1998). From retinotopy to recognition: fMRI in human visual cortex. *Trends in Cognitive Sciences, 2*(5), 174-183.

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London series B: Biological Sciences, 353*(1373), 1295-1306.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5*(7), 682-687.

VanRullen, R., Reddy, L., & Koch, C. (in press). Visual search and dual-tasks reveal two distinct attentional resources. *Journal of Cognitive Neuroscience*.

VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: from early perception to decision-making. *Journal of Cognitive Neuroscience, 13*(4), 454-461.

Vogel, E. K., & Luck, S. J. (2000). The visual N1 component as an index of a discrimination process. *Psychophysiology, 37*(2), 190-203.

Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study. *European Journal of Neuroscience, 11*(4), 1239-1255.

Wolfe, J. M. (1998). Visual search. In H. Pashler (Ed.), *Attention* (pp. 13-73). Hove (UK): Psychology Press Ltd.

Partie B : Catégorisation visuelle des scènes naturelles : le contexte et son influence sur la perception des objets

Introduction

S'il pouvait parler, l'ours blanc des deux photographies de la page précédente poserait certainement cette question : « A votre avis, suis-je plus facile à catégoriser dans un contexte compatible que dans un contexte non compatible ? ». Cette question a été à la base d'un travail sur la catégorisation du contexte global d'une scène naturelle et ce chapitre propose une synthèse des éléments disponibles actuellement sur ce sujet. Nous verrons tout d'abord que la question de l'influence du contexte sur la catégorisation des objets est loin d'être résolue. Une partie du problème réside dans notre manque de connaissance des mécanismes de la catégorisation du contexte lui-même, contrairement au savoir accumulé sur le traitement des objets. Cet aspect sera abordé dans une seconde partie.

1. De l'influence du contexte sur la catégorisation des objets

"In research on high-level scene perception, the concept of *scene* is typically defined (though often implicitly) as a semantically coherent (and often nameable) view of a real-world environment comprising background elements and multiple discrete objects arranged in a spatially licensed manner. Background elements are taken to be larger-scale, immovable surfaces and structures, such as ground, walls, floors, and mountains, whereas objects are smaller-scale discrete entities that are manipulable (e.g. can be moved) within the scene" (Henderson & Hollingworth, 1999, p. 244).

Selon cette définition, les scènes naturelles sont plus qu'une simple collection d'objets. Cependant, la majeure partie des recherches sur les scènes naturelles est dédiée au traitement des objets dans les scènes, laissant de côté la question du traitement des scènes elles-mêmes, dans leur ensemble. Cette question est importante dans la mesure où les informations grossières d'une scène, notamment sa catégorie sémantique et sa structure spatiale, pourraient servir à intégrer de manière harmonieuse l'information visuelle acquise d'une saccade à la suivante. Ce mécanisme serait particulièrement important dans le cadre des modèles de la perception visuelle qui soutiennent que très peu d'information est retenue entre deux saccades (Hochberg, 1986 ; O'Regan, 1992; Rensink, 2000, 2002; Wolfe, 1999). Par exemple, Hochberg (1986) a proposé que l'intégration d'une série de vues partielles d'une séquence vidéo serait réalisée grâce à

l'usage d'un schéma mental servant à guider l'interprétation et l'intégration de vues successives du monde visuel. Ce schéma pourrait être considéré comme une représentation abstraite de la structure spatiale de la scène dont les points de repères caractéristiques ne préserveraient pas les détails sensoriels. L'existence d'un schéma mental des scènes naturelles est suggéré par l'effet dit « d'extension des bordures » (« boundary extension effect », Intraub & Richardson, 1989 ; Intraub, 1999) : la mémoire que nous avons d'une scène visuelle semble contenir des informations qui n'étaient pas présentes dans une image mais étaient fortement suggérées par le contexte.

Cependant, les modèles « amnésiques » du système visuel sont remis en cause par les travaux récents présentés au chapitre 1A (Henderson & Hollingworth, 2003 ; Hollingworth, 2003 ; Hollingworth & Henderson, 2002 ; Simons et al., 2002).

Plus généralement, le contexte visuel semble pouvoir guider le déploiement de l'attention vers des cibles potentielles dans des tâches de recherche visuelle (Chun, 2000). Une question plus spécifique consiste à déterminer si le contexte fourni par une scène a une influence directe sur la catégorisation des objets qui s'y trouvent (une importante synthèse sur ce sujet est disponible dans Henderson, 1992 ; Henderson & Hollingworth, 1999 ; Hollingworth & Henderson, 1998). Différents modèles ont été proposés qui stipulent que le contexte influence soit les mécanismes de structuration de la forme, soit le stade de la reconnaissance proprement dite, ou bien encore plus tardivement, les étapes post-perceptives de prise de décision, une position notamment défendue par Henderson & Hollingworth (1999).

Le problème principal de ces modèles est qu'ils sont discutés dans un cadre théorique complètement dépassé, celui du modèle purement vers l'avant et cloisonné en étapes de traitement bien délimitées décrit par Marr (1982, voir aussi Biederman, 1987). Des modèles plus récents du système visuel suggèrent au contraire une architecture dynamique dans laquelle les activations neuronales en provenance de la rétine sont comparées très rapidement avec des hypothèses générées à différents niveaux de la hiérarchie visuelle, probablement au travers d'interactions incessantes entre les connexions vers l'avant et les connexions vers l'arrière (Bullier, 2001 ; Rao & Ballard, 1999 ; Ullman, 1995). Dans ce contexte, il est très peu probable que l'interprétation visuelle d'un objet ne soit pas contrainte, au moins en partie, par les informations

véhiculées par le contexte dans lequel il apparaît. Hollingworth & Henderson (1998, 1999) rapportent une absence d'effet de contexte qui pourrait provenir de l'usage d'images flashées contenant à la fois un objet cible et un contexte. Cette approche n'est pas des plus écologiques dans la mesure ou dans notre vie de tous les jours le contexte est stable pendant un long moment. Par exemple, pendant que je tape cette thèse devant mon ordinateur, mon environnement immédiat reste le même, ce qui me permet de générer des attentes sur les objets possibles lorsque je balaye la pièce des yeux. Même en l'absence d'un souvenir précis des objets qui m'entourent, la présence d'articles et de livres est hautement prédictible, par contre la présence d'une grosse pépite d'or l'est beaucoup moins. Mais cela veut-il dire qu'un livre sera perçu plus rapidement que la pépite d'or, si elle était là ? Dans une étude récente en potentiels évoqués, Ganis & Kutas (2003) ont utilisé un protocole dans lequel une scène naturelle apparaît pendant 300 ms, permettant en principe aux sujets de générer des attentes sur les objets associés à son contexte. La même scène était présentée immédiatement après pendant 300 ms, mais cette fois-ci un objet congruent ou non congruent était ajouté (par exemple une scène avec des joueurs de football, l'objet critique étant soit un ballon, soit un rouleau de papier toilette). Malgré ce protocole plus écologique, la latence de l'activité différentielle entre la condition congruente et la condition non congruente était au minimum de 300 ms, bien au-delà des latences rapportées pour la catégorisation d'objets dans les scènes naturelles (Thorpe et al., 1996). On pourrait aussi interpréter le résultat de Ganis & Kutas comme marquant une très forte influence du contexte : le sujet s'attend tellement à voir un ballon qu'il ne réalise que très tardivement qu'il s'agit d'un rouleau de papier toilette. Une autre explication de ce résultat contraire aux prédictions des modèles interactifs pourrait résider dans le fait que le traitement visuel est tellement rapide dans des conditions optimales où les objets sont bien contrastés (c'était le cas dans l'expérience de Ganis & Kutas) que l'influence du contexte devient marginal. Une prédiction qui reste à tester est que le contexte pourrait avoir une influence maximale dans des conditions pauvres de perception, par exemple la nuit ou par temps de brouillard. Cette hypothèse pourrait être testée en répliquant l'expérience de Ganis & Kutas mais en variant le contraste des objets congruents et non congruents. Il est également possible qu'il faille présenter la scène contextuelle plus longtemps afin que son sens soit extrait et que des prédictions soient

générées. Pour approfondir cette question nous allons maintenant aborder l'étude des mécanismes de catégorisation des scènes.

2. Mécanismes de la catégorisation du contexte

La compréhension des interactions entre le sens d'une scène et la représentation des objets qui s'y trouvent dépend en grande partie de notre compréhension du traitement des scènes naturelles. Cette seconde partie propose une synthèse des avancées récentes dans ce domaine, qui comporte d'une part des travaux d'imagerie cérébrale et d'autre part des travaux de psychologie expérimentale et de modélisation.

2.1 Bases anatomiques de la catégorisation des scènes naturelles

La catégorisation des scènes naturelles ne fait l'objet de travaux en neurosciences cognitives que depuis très peu de temps. Des études récentes suggèrent qu'un ensemble d'aires corticales serait impliqué dans différents aspects du traitement des scènes. De manière générale, le cortex parahippocampique et le cortex pariétal semblent préférentiellement activés pendant la perception d'une scène naturelle par rapport par exemple à la perception d'un visage (Nakamura et al., 2000 ; Sato et al., 1999). De manière plus spécifique, la région corticale parahippocampique serait impliquée dans le traitement de la structure spatiale de la scène (Epstein & Kanwisher, 1998 ; Epstein et al., 1999, 2001, 2003), l'apprentissage et la reconnaissance de bâtiments et de paysages (Maguire et al., 2001). Elle permettrait également, en conjonction avec le cortex retrosplenial, le traitement des informations contextuelles à la fois spatiales et non spatiales (Bar & Aminoff 2003). Enfin, le sulcus lingual droit a été impliqué dans la perception des bâtiments (Aguirre et al., 1998).

L'existence d'un réseau largement distribué pour le traitement des différents aspects des scènes naturelles est confirmée par la neuropsychologie. Divers troubles de la perception, de la reconnaissance ou de la mémorisation d'éléments propres aux scènes elles-mêmes, indépendants des agnosies pour des objets, ont été décrits suite à des lésions du cortex pariétal, du gyrus lingual et du cortex parahippocampique (Aguirre & D'Esposito, 1999 ; Epstein et al., 2001 ; Farah, 1990 ; Mendez & Cherrier, 2003).

2.2 Bases fonctionnelles de la catégorisation des scènes naturelles

Il existe donc un réseau largement distribué d'aires corticales permettant divers aspects du traitement des scènes naturelles. S'il est pour l'instant difficile d'envisager comment chaque partie de ce réseau fonctionne, des données comportementales ainsi que des travaux de modélisation permettent de mieux cerner les contraintes qui pèsent sur le système dans son ensemble. Un traitement très rapide des scènes semble ainsi possible, certains suggérant qu'il serait assez rapide pour influencer directement les mécanismes de traitement des objets.

Une telle proposition peut paraître surprenante dans la mesure où selon le point de vue dominant dans la littérature, la catégorisation des scènes est conçue comme l'aboutissement d'une série de traitements effectués dans la voie ventrale (voir Chapitre 1A). Selon cette logique, les objets seraient systématiquement catégorisés avant les scènes.

Cependant, un point de vue alternatif émerge de plusieurs études ayant montré que la catégorisation d'une scène peut s'effectuer très efficacement même lorsque des dessins ou des photographies sont très brièvement présentées à des sujets humains (Biederman, 1972 ; Biederman et al., 1973, 1982 ; Intraub, 1999 ; Oliva & Schyns, 1997, 2000 ; Potter, 1975, 1976 ; Potter & Levy, 1969 ; Schyns & Oliva, 1994). Le fait que la catégorisation d'une scène soit toujours possible dans des conditions très brèves de présentation est souvent considéré comme une preuve que les mécanismes sous-jacents sont également rapides. Selon cette logique, il a été suggéré que la catégorisation des scènes pourrait être réalisée de manière concomitante ou même avant la catégorisation des objets. En effet, certaines théories postulent que contrairement au point de vue dominant, les scènes pourraient être identifiées à partir d'indices visuels qui leurs seraient propres. Par exemple, Biederman (1988) a proposé que son modèle structural de la reconnaissance d'objet fondé sur l'usage de « géons » (primitives tridimensionnelles) pourrait être étendu à la reconnaissance des scènes. Des primitives avec une échelle spatiale plus importante que celles servant à la reconnaissance d'objet pourraient ainsi représenter des informations visuelles spécifiques des scènes indépendamment des informations sur les objets. Cette proposition n'a pas encore été testée de manière empirique mais pose déjà problème au niveau théorique. D'une part, les modèles par

description structurale ne sont pas plausibles biologiquement (Rolls & Deco, 2002). D'autre part, comme le font remarquer Henderson & Hollingworth (1999), les scènes ne sont probablement pas représentées comme de gros objets étant donné le manque de contraintes fortes sur leur organisation structurale. Il reste néanmoins possible que la structure spatiale d'une scène, même a un niveau grossier, puisse permettre son identification (Sanocki & Epstein, 1997). La représentation de la structure spatiale d'une scène "may include information about the extent and location of the ground plane and other reference objects and surfaces, as well as size and distance relations within the scene" (Sanocki & Epstein, 1997, p.378). Une telle information pourrait servir à catégoriser une scène. Cette hypothèse a été confirmée par Schyns & Oliva (1994 ; Oliva & Schyns, 2000) qui ont montré que le sens général d'une scène pouvait être extrait à partir de photographies dont seules les basses fréquences spatiales sont préservées ou bien encore des taches de couleurs ayant une configuration spatiale particulière. Le point très intéressant est que ces mêmes photographies ne contenaient pas assez d'information pour pouvoir catégoriser les objets se trouvant dans la scène. Ces expériences renforcent donc l'idée selon laquelle la catégorisation des objets et des scènes dépendrait d'indices visuels distincts. Finalement, des données issues de travaux de modélisation suggèrent que des « filtres » visuels spécifiques des scènes et dérivés de la combinaison d'un ensemble restreint de filtres bas niveau pourraient être suffisants pour discriminer la plupart des types de scènes constituant notre environnement (Oliva & Torralba, 2001 ; Torralba & Oliva, 2003).

Ce dictionnaire restreint de propriétés physiques spécifiques des scènes rend tout à fait plausible la possibilité d'une identification des scènes naturelles tout aussi rapide, ou même plus rapide que celle des objets. Dans ce dernier cas, l'identification du contexte pourrait même guider l'identification de l'objet. Cette hypothèse prend toute sa crédibilité lorsque l'on prend en compte la nature adaptative, en terme de stratégies, dont fait preuve le système visuel. Un exemple d'une telle approche est fournie par le cadre de réflexion de la reconnaissance diagnostique (« diagnostic recognition framework ») développé par Schyns (1998 ; Schyns & Oliva, 1997). Dans ce cadre théorique, la performance dans une tâche de reconnaissance est déterminée par l'interaction entre plusieurs facteurs dont les principaux sont les demandes de la tâche, c'est-à-dire

l'information nécessaire à la réalisation de la tâche, et la disponibilité perceptuelle de cette information. L'expertise pour une catégorie d'objets serait un autre facteur important influençant la performance, probablement en permettant aux experts d'accéder à des détails diagnostiques très précis auxquels les sujets naïfs ne seraient pas sensibles (Rossion & Gauthier, 2002 ; Schyns, 1998). Le système visuel pourrait ajuster dynamiquement ses propres stratégies en fonction de la tâche et des informations disponibles pour la réaliser afin de capturer le plus efficacement possible les propriétés diagnostiques des cibles de la tâche. Comme le soulignent Schyns & Murphy (1994, cité dans Schyns, 1998) : "if a fragment of a stimulus categorizes objects (distinguishes members from non-members), the fragment is instantiated as a unit in the representational code of object concepts" (Schyns, 1998, p.155). Autrement dit, cela veut dire qu'une catégorisation de haut niveau ne dépend pas forcément de représentations de hauts niveaux si des représentations de plus bas niveau permettent de catégoriser correctement un stimulus dans une tâche donnée (voir aussi Ullman et al., 2002).

Un modèle fonctionnel récent du traitement des scènes naturelles fournit également des éléments en faveur d'un traitement rapide du sens d'une scène (Rensink, 2000, 2002). Dans son architecture triadique ("triadic architecture"), Rensink inclut un système attentionnel de traitement des objets et un système non attentionnel à capacités limitées, dédié au traitement du sens général de la scène (sa catégorie) et de sa structure spatiale. Parce que ce système maintient en mémoire de travail des informations à propos d'aspects stables de la scène, il pourrait servir à guider l'attention et les yeux vers les objets d'intérêt. Dans cette architecture fonctionnelle, le sens d'une scène serait extrait essentiellement sans attention, par l'intégration d'un ensemble restreint de propriétés de bas niveau. On peut cependant se demander si un tel système non attentionnel devrait théoriquement être plus rapide que le système attentionnel des objets. Rensink fonde son argumentation sur le travail de Oliva & Schyns (1997) portant sur la catégorisation de photographies hybrides de scènes visuelles. Dans cette expérience, chaque photographie était composée de la superposition de deux scènes, l'une étant exprimée en terme de basses fréquences spatiales, l'autre en terme de hautes fréquences spatiales. Certains sujets devaient catégoriser les scènes en utilisant par exemple seulement les basses fréquences spatiales, sans savoir que les hautes fréquences spatiales véhiculaient une

information sur une autre scène superposée. Un des résultats importants de cette étude montre qu'alors même que les sujets rapportant la catégorie d'une scène à une fréquence spatiale donnée n'étaient pas conscients de la présence simultanée d'une autre scène dans l'image, leur système visuel était capable de traiter implicitement la scène non perçue. Mais contrairement à l'interprétation de Rensink, ce résultat ne montre pas que l'extraction du sens d'une scène peut être réalisé sans attention au sens strict. Il montre plutôt que cette extraction peut avoir lieu sans attention *focalisée*, ce qui est très différent (voir Chapitre 1A). A l'appui de cette argumentation, on peut citer le travail de Naccache et al. (2002), qui ont démontré que l'existence d'effets d'amorçage sémantique dans des conditions de fort masquage rétrograde (Dehaene et al., 1998, 2001), qui pourrait traduire la mise en jeu de traitements automatiques non attentionnels, dépend en réalité de la focalisation de l'attention sur la fenêtre temporelle dans laquelle le stimulus masqué apparaît. Il est donc fort probable que si un train d'images hybrides était présenté à des sujets, un traitement implicite de certaines fréquences spatiales n'aurait lieu que pour les images sur lesquelles l'attention temporelle est portée. En revanche, même s'il apparaît que l'idée d'un traitement automatique du sens des scènes n'est pas très décisif dans l'argumentation sur leur vitesse de traitement, les résultats présentés plus haut stimulent la poursuite des investigations dans ce domaine.

Au niveau empirique, il faut cependant noter que les seuls indices en faveur d'un traitement rapide des scènes naturelles découle d'expériences ayant eu recours à des temps très brefs de présentation. Malheureusement, des présentations visuelles brèves, particulièrement dans le cas des séquences RSVP (Rapid Sequential Visual Presentations ; pour des revues voir Intraub, 1999 ; Potter, 1999) fournissent un débit de traitement et non pas une durée absolue de traitement visuel.

Le troisième article de cette thèse décrit deux expériences dont le but était d'évaluer plus précisément le temps nécessaire pour extraire le sens général d'une scène (Rousselet & Fabre-Thorpe, soumis). La méthodologie employée est la même qui a été utilisée pour évaluer le degré de parallélisme dans le système visuel (articles 1 et 2). Il s'agissait d'un paradigme go/no-go dans lequel les images sont flashées brievement, un tel protocole étant particulièrement bien adapté pour évaluer les contraintes temporelles qui pèsent sur le traitement visuel.

Article 3

How long to get to the “gist” of real-world natural scenes?

Rousselet, G.A. & Fabre-Thorpe, M.

(article soumis à *Visual Cognition*)

Résultats comportementaux de 48 sujets adultes dans deux expériences visant à tester la vitesse de catégorisation du contexte global de scènes naturelles présentées en couleur et en noir et blanc.

L’article est suivi de la reproduction d’un poster illustrant une partie de ce travail présenté à la conférence internationale de la Cognitive Neuroscience Society en 2003 à New York.

Introduction

L’une des questions importantes lorsque l’on étudie la catégorisation d’un objet dans une scène naturelle concerne l’aide que peut apporter l’arrière-plan de la scène dans le traitement de l’objet lui-même. L’idée d’une reconnaissance plus rapide d’un objet présenté dans un contexte congruent (par rapport à un contexte non congruent) va dans le sens d’une interaction favorable des traitements simultanés de l’objet et du contexte de la scène. Une telle hypothèse formalise ainsi le parallélisme du traitement des informations visuelles. Cependant, la question de la rapidité à laquelle notre système visuel accède à la reconnaissance d’une scène dans sa globalité est une question incontournable pour étayer cette hypothèse. Un des points importants soulevé dans le chapitre 1B est le manque de connaissances portant sur la catégorisation du contexte, c'est-à-dire du sens général d’une scène naturelle. Cette étude avait pour objectif d’évaluer le temps de traitement du sens global d’une scène en utilisant la méthodologie go/no-go mise en œuvre précédemment pour étudier le traitement des objets en parallèle. Les sujets réalisaient tour à tour 4 tâches de catégorisation de photographies de scènes naturelles flashées brièvement. Il y avait 2 catégories dites « naturelles », des scènes de mer et des scènes de montagne, et 2 catégories dites « artificielles », des scènes urbaines d’extérieures et des scènes d’intérieur. Dans une première expérience toutes les photographies étaient en couleur. La seconde expérience comportait 50% d’images en couleur et 50% en noir et blanc afin d’évaluer l’importance des indices de couleur dans cette tâche.

Résultats

1) Les sujets humains étaient à la fois très précis et rapides pour extraire le sens général d'une scène, confirmant les résultats d'expériences antérieures. Les scènes naturelles étaient plus rapidement catégorisées que les scènes dites artificielles.

2) La couleur avait une influence très faible sur la performance et apparaissait comme un facteur facultatif pour la catégorisation du contexte. Ce facteur pouvait prendre plus d'importance pour certaines des catégories, comme la mer. Dans ce cas, la couleur était extraite rapidement et pouvait accélérer le traitement des scènes naturelles appartenant à cette catégorie.

3) La comparaison des résultats acquis dans la première des deux expériences avec ceux obtenus dans une tâche de catégorisation animal/non-animal suggère que les objets pourraient en moyenne être traités plus rapidement que le contexte. Le chevauchement des histogrammes de temps de réaction montre également qu'une interaction figure/fond est tout à fait possible notamment lorsque les réponses sont déclenchées avec des TR moyens et lents.

Discussion

Des travaux antérieurs avaient déjà montré que la catégorisation du contexte pouvait être effectuée très rapidement. Dans la présente étude, l'emploi d'une tâche go/no-go, d'un nombre important d'essais et l'analyse des distributions de temps de réaction permet de fournir des contraintes un peu plus fortes que précédemment sur les mécanismes sous-jacents. Le traitement plus rapide des scènes naturelles par rapport aux scènes artificielles pourrait refléter un plus grand recouvrement entre les propriétés physiques des scènes urbaines d'extérieur et des scènes d'intérieur par rapport aux scènes de mer et de montagne, nécessitant ainsi l'intégration de plus d'éléments pour permettre de les distinguer.

De plus, il apparaît que la couleur pourrait, dans certains cas, être disponible précocement et être utilisée même dans le cas des réponses les plus précoces. Ce résultat est en accord avec une étude récente chez le singe montrant que dans IT les neurones aux réponses les plus précoces à des objets complexes peuvent être sensibles aux informations de couleur (Edwards et al. 2003). Cependant, la couleur n'est pas indispensable pour réaliser la catégorisation rapide du contexte, comme cela avait déjà été montré pour les objets auparavant dans des conditions expérimentales très semblables (Delorme et al., 2000). Ce manque d'impact de la couleur sur la catégorisation

rapide ne doit pas être considéré comme absolu. En effet, l'importance de ce facteur dépend de sa diagnosticité pour la tâche, c'est-à-dire que la couleur peut être un facteur primordial pour la catégorisation visuelle si elle permet de discriminer une catégorie parmi d'autres (Oliva & Schyns, 2000). L'importance des effets descendants pour « prédisposer » le système visuel au traitement d'une information pertinente rendrait ce traitement extrêmement rapide (voir plus loin).

Enfin, la catégorisation d'objets tels que des animaux semble plus rapide que la catégorisation du contexte. Ce résultat n'implique cependant pas que les objets soient systématiquement traités plus rapidement que le contexte, de telle sorte que ce dernier ne pourrait avoir d'effet sur le traitement des premiers. Tout d'abord, les distributions des temps de réaction dans les deux tâches montrent un large recouvrement, suggérant qu'une grande partie des images pourrait être traitée à la même vitesse dans les deux tâches. De plus, il faut considérer que dans des conditions écologiques, le contexte ne devrait pas changer abruptement d'un instant à l'autre, et devrait donc permettre de générer des attentes quant aux objets susceptibles d'être rencontrés dans un environnement donné. Finalement, les temps de traitement rapportés pour le contexte et les objets ne sont pas absolus. Ils dépendent de la prédictibilité des cibles.

A) Le contexte pourrait être traité beaucoup plus rapidement dans certaines conditions où cibles et distracteurs sont très différents les uns des autres. L'article 4 de cette thèse présente des données allant dans ce sens : quand une scène naturelle est apprise et doit être reconnue parmi d'autres scènes, maximisant ainsi les informations descendantes qui définissent la cible, le temps de traitement peut être significativement réduit par rapport au temps nécessaire à la réalisation d'une tâche de catégorisation animal/non-animal (Delorme et al., accepté pour publication). Ainsi, une prochaine expérience testera la vitesse de catégorisation du contexte pour deux catégories seulement : d'une part divers types d'environnements naturels, et d'autre part des environnements artificiels, urbanisés. Les sujets devront répondre « naturel » dans une tâche et « artificiel/urbain » dans une autre tâche. La distinction relativement nette entre ces deux grandes catégories d'environnement devrait permettre d'optimiser les informations descendantes permettant de pré-activer certains groupes de neurones dans la voie ventrale et ainsi de diminuer le temps de traitement du contexte qui pourrait dès lors être beaucoup plus proche de celui des objets.

B) La vitesse de catégorisation des objets n'est pas non plus fixe. Des expériences récentes menées dans notre équipe ont en effet montré que la catégorisation d'oiseaux ou de chiens (catégories subordonnées) est associée à des temps de réaction moyens environ 40-70 ms plus lents que ceux obtenus dans une tâche de catégorisation d'animaux (Macé, Thorpe & Fabre-Thorpe, en préparation). Le temps de réaction augmente et la précision diminue de manière proportionnelle au nombre d'images non-cibles contenant d'autres types d'animaux. Ainsi, quand on change la diagnosticité des cibles, on change la stratégie du sujet. Le système visuel est un système dynamique qui s'adapte en permanence en fonction de la tâche. L'augmentation des temps de réaction dans la catégorisation des oiseaux ou des chiens indique aussi que le niveau de représentation atteint dans la tâche animal/non-animal au moment ou le sujet prend une décision n'est sans doute pas aussi complexe qu'on le pense, ces représentations pourraient au contraire rester très rudimentaires. D'autres données, décrites dans la discussion des articles 4 et 5 soutiennent cette hypothèse. Il est donc tout à fait possible que le contexte puisse influencer très rapidement la catégorisation des objets en fonction de la tâche que les sujets réalisent.

Finalement, une prochaine expérience devrait tester directement l'influence du contexte sur la catégorisation animal/non-animal. Cette expérience mettra en jeu des contextes congruents et non congruents avec les cibles. J'ai déjà commencé à mettre en place une base de données d'images test. Pour cela, j'ai découpé plus d'une centaine d'animaux dans des photographies de scènes naturelles. Chaque animal a ensuite été placé au même endroit et à la même échelle dans deux scènes différentes, l'une congruente avec sa présence, l'autre non. Le critère de congruence retenu est pour l'instant relativement grossier, des animaux sauvages étant placé dans des environnements naturels ou bien dans des environnements urbains. Les images de l'ours blanc qui illustrent le chapitre 1B en sont un exemple. Cependant, différents facteurs compliquent considérablement la mise en œuvre de cette expérience, notamment le respect rigoureux du critère de congruence pour l'ensemble des images réalisées, ainsi que des problèmes de contrastes locaux qui peuvent varier considérablement d'un contexte à l'autre. D'autre part, des expériences préliminaires seront nécessaires pour définir un protocole adapté à la mise en évidence d'effets de contexte. Il est tout à fait envisageable que le contexte n'ait pas d'influence pour des objets nets et bien contrastés. Par contre, il pourrait avoir un très fort effet si les objets sont flous ou très bruités, mimant par exemple des conditions de brouillard.

How long to get to the "gist" of real-world natural scenes?

Guillaume A. Rousselet & Michèle Fabre-Thorpe

Abstract

This study aimed at assessing the processing time of a natural scene in a fast categorization task of its context or "gist". In experiment 1, human subjects performed 4 go/no-go categorization tasks in succession with color pictures of real-world scenes belonging to 2 natural categories: 'sea' and 'mountain', and 2 man-made categories: 'indoor' and 'urban'. Experiment 2 used color and grey-level scenes in the same tasks to assess the role of color cues on performance. Pictures were flashed for 26 ms. Both experiments showed that the gist of real-world scenes can be extracted with high accuracy (>90%), short median RT (400-460 ms) and early responses triggered with latencies as short as 260-300 ms. Natural scenes were processed faster than man-made scenes, and color information did not appear as a crucial feature to perform the tasks studied here. The processing speed is compared for scene and object categorization tasks.

INTRODUCTION

Natural scenes are more than a simple collection of objects. However, much of the research on scene processing has been devoted to the understanding of object processing in scenes, letting on the side the question of how we process the whole scene itself.

This issue is important given that we do not only process objects but we also analyze the context in which they appear. Global coarse information about a scene (mainly its category, or *gist*, and its spatial structure, or *layout*) is relatively crucial in memory-free models of scene perception in which little information is integrated across saccades. According to this idea, perception is constructed by integrating abstract scene representations with volatile object representations available at the locus of attention (O'Regan, 1992; Rensink, 2000, 2002; Wolfe, 1999; but see Henderson & Hollingworth, 2003; Hollingworth, 2003; Hollingworth & Henderson, 2002; Simons, Chabris, Schnur & Levin, 2002). More generally, visual context has been shown to guide attention toward potential target objects (for a review see Chun, 2000). The role played by the background of a scene in object identification is still controversial (see reviews in Henderson, 1992; Henderson & Hollingworth, 1999), with evidence in favor or against such a view (see among many others Biederman, Mezzanotte & Rabinowitz, 1982; Boyce, Pollatsek & Rayner, 1989; but see Hollingworth & Henderson, 1998, 1999; Ganis & Kutas, 2003). But what is known on the global processing of a scene is far limited compared to the knowledge accumulated about object processing. In particular, for scene context to influence object identification, one fundamental constraint is the speed at which scene context can be extracted.

What do we know exactly about scene processing? Only recently this topic has received more attention from cognitive neuroscience researchers, revealing a set of cortical areas involved in different aspects of scene processing, like the parahippocampal and parietal cortices (Nakamura, Kawashima, Sato et al., 2000; Sato, Nakamura, Nakamura et al., 2000). More specifically, the parahippocampal area has been attributed different functions like processing of the spatial layout of the scene (Epstein & Kanwisher, 1998; Epstein, Graham & Downing, 2003), learning and

recognition of buildings and landscapes (Maguire, Frith & Cipolotti, 2001). It is also thought to mediate, in conjunction with the retrosplenial cortex, both spatial and nonspatial contextual processing (Bar & Aminoff 2003). Finally, the right lingual sulcus has been implicated in the perception of buildings (Aguirre, Zarahn & D'Esposito, 1998). This distributed system, largely separated from the object system, might lead to hypothesize that scenes are processed very efficiently, maybe fast enough to be able to influence object processing.

Scene categorization is often regarded as the ultimate representation generated along the ventral pathway in which a scene would be reconstructed progressively by integrating local contrasts (Marr, 1982). Following such a view, objects would be processed almost systematically before the scene (Biederman, 1987; Riesenhuber & Poggio, 2000).

Alternatively, many studies have suggested that scene categorization can be performed very efficiently from very brief visual presentations (Biederman, Mezzanotte & Rabinowitz, 1982; Intraub, 1997; Oliva & Schyns, 1997, 2000; Potter, 1975, 1976; Potter & Levy, 1969; Schyns & Oliva, 1994). The fact that scene categorization is possible with very brief presentations is often taken as an evidence for fast underlying mechanisms. Thus, scene categorization could be performed simultaneously or even precede object identification. Indeed, although some theories have suggested that scene categorization might result from the identification of the component objects (e.g. Friedman, 1979), other theories have emphasized that scenes might also be identified from scene-specific cues. For instance, Biederman (1988) suggested that his original structural model of object recognition using "geons" (3D primitives) might be extended to scene recognition. He proposed that primitives with a larger spatial scale than those used to represent objects could represent scene specific information independently of object information. Although this proposal has not been tested empirically, Henderson & Hollingworth (1999) have suggested that, given the lack of strong constraints on their structural organization, scenes are not likely to be represented as large objects. However, it remains possible that the spatial organization of a scene (even at a coarse level) might mediate its identification. Sanocki & Epstein (1997, p.378) suggested that the representations of the spatial layout "may include information about the extent and location of the ground plane and other reference objects and surfaces, as well as size and distance relations within the scene". This information about the spatial structure of the scene might be used to extract its meaning. Indeed, the gist of a scene can be extracted from low spatial frequency versions of photographs preserving coarse spatial layouts, or spatially arranged color blobs, but in which information to categorize component objects was not preserved (Schyns & Oliva, 1994; Oliva & Schyns, 2000). This strengthens the idea that object and scene categorization might be mediated by distinct visual cues. Finally, computational evidence suggests that scene-based visual filters derived from the combination of a restricted set of low-level filters might be sufficient to perform most of the visual discrimination needed to categorize the context of our environment (Oliva & Torralba, 2001; Torralba & Oliva, 2003).

Given this reduce dictionary of scene-based physical properties and the known capacity of the visual system to dynamically adjust its strategies as a function of task constraints to pick the most adequate image features (Schyns, 1998), it is plausible that scenes might be identified very fast, probably as fast or even faster than objects. So far, the evidence for fast scene processing comes from experiments using brief visual presentations (see above). Unfortunately, brief visual presentations and particularly RSVP sequences (Rapid Sequential Visual Presentations; e.g. Intraub, 1997; Potter, 1975, 1976; Potter & Levy, 1969) provide a rate of visual processing rather than an

absolute evaluation of the visual processing time. Furthermore, experiments like those performed by Oliva & Schyns (1997, 2000) used vocal responses or involved a matching task between a written name and a visual scene with a two-choice response. Using 16 different categories in the matching task, they showed that the averaged mean reaction times (RT) were largely distributed from 476 ms (*city* category) to 631 ms (*valley* category).

This mean RT can be compared to the 400 ms mean RT that has been commonly found for object categorization in various studies from our group. These studies employed a go/no-go animal categorization task in which human subjects were required to respond as fast and as accurately as possible each time a natural photograph, that was flashed for the first time and for only 20-40 ms, contained an animal (Delorme, Richard & Fabre-Thorpe, 2000; Fabre-Thorpe, Delorme, Marlot & Thorpe, 2001; Thorpe, Marlot & Fize, 1996). This finding has been extended to other object categories such as means of transport (VanRullen & Thorpe, 2001), human faces and animal faces (Rousselet, Macé & Fabre-Thorpe, 2003), although food objects might take up to 30 ms longer (Delorme, Richard & Fabre-Thorpe, 2000). Compared with Schyns and Oliva's studies, objects might thus be identified before extraction of scene context. But the difference in processing time might also originate in the motor response required in both tasks as our go/no-go task relies only on a single motor output, whereas their matching task requires a choice of response.

In this study, we have assessed the time course of the categorization of the scene context, or its “gist”, with the same go/no-go visual categorization task previously used to study object categorization. We selected 4 categories of color pictures, two of them were natural, ‘sea’ and ‘mountain’, and the two others were man-made, ‘indoor and urban’ scenes. In a first experiment, subjects were asked to perform four go/no-go categorization tasks, one per category. A second experiment was designed to assess the effect of color on gist processing speed.

EXPERIMENT 1

The aim was to provide an estimate of the temporal constraints in the visual processing of the gist of a natural scene, for 4 scene categories representative of our environment (sea, mountain, urban and indoor). These categories were relatively coarsely defined in order to present subjects with pictures as varied as possible (see Stimuli and Figure 1).

Method

Participants. Twenty-four adults (12 women and 12 men, mean age 30, ranging from 19 to 51, 3 of them left handed), volunteered in this study and gave their informed written consent. All participants had normal or corrected to normal vision.

Stimuli. We used 24-bit (16 millions of colors) photographs of natural scenes (768 by 512 pixels, sustaining a visual angle of about 15.6° x 10.5°) taken from a large commercial CD-ROM library (Corel Stock Photo Libraries). From this data bank, we selected 384 images for each of the four environmental categories. For each category, half of them were horizontal photographs, the other half were vertical photographs. They were all chosen to be as varied as

possible, representing the four types of scenes from a large range of viewpoints and perspectives (Figure 1). Each image was seen only once by a given subject to prevent learning.

Sea pictures were composed of various coast scenes (including beach scenes, cliff scenes, or showing various rocks, icebergs…) as well as "full sea" pictures with boats, sailboards, and surfboards. In all cases the sea was largely visible on the pictures. The mountain category contained pictures that showed large mountain backgrounds at different distances in all seasons as well as various photographs taken from the point of view of mountain hikers. Urban pictures were almost exclusively taken from the point of view of someone walking in towns ranging from small villages to large cities. Photographs depicted streets, buildings, houses, public squares, etc., from many places around the world. Indoor scenes were photographs taken from inside various man-made constructions like houses, apartments, churches, museums, and stores…

There was no overlap in the pictures from the four target categories: sea scenes did not contain harbors or mountains in the background; mountain scenes did not contain villages; street scenes did not include streets constructed along the sea, etc.

Procedure. Image presentation and behavioral response measurement were carried out using the software Presentation (NeuroBehavioral Systems, http://nbs.neuro-bs.com/). Subjects sat in a dimly lit room at 100 cm from a computer screen (horizontal resolution = 1024 pixels, vertical resolution = 768 pixels, vertical refresh rate: 75 Hz) piloted by a PC computer. To start a block of trials, they had to place their finger on a infra-red response pad for one second. A trial was organized as following: a fixation cross (0.1° of visual angle) appeared for 300-900 ms and was immediately followed by the stimulus presented for two frames, i.e. about 26 ms, in the middle of the screen. These brief presentations prevented any exploratory eye movements. Participants had to lift their finger as quickly and as accurately as possible (go response) each time a target scene was presented and to withhold their response (no-go response) when the photograph did not belong to the target category. Responses were detected using infrared diodes. Subjects had 1000 ms to respond; longer reaction times were considered as no-go responses. This maximum response time delay was followed by a 300 ms black screen, before the fixation point was presented again for a variable duration, resulting in a random 1600-2200 ms inter-trial interval.

Subjects were tested in two experimental sessions on two different days. A given picture category was the target category for 4 consecutive series. In each session they performed two categorization tasks for a total of 8 blocks of 96 trials with target and non-target trials being equiprobable. This led to a total of 1536 trials per subject. The order in which the subjects performed the four category tasks was counterbalanced across subjects. In a given task, the 48 non-target images belonged equally to the 3 other environmental categories. Thus, when performing the sea categorization task, a 96 trial series contained 48 target sea pictures, 16 non-target mountain scenes, 16 non-target urban scenes and 16 non-target indoor scenes. To avoid any bias, the design was organized so that across subjects each of the 384 pictures of a given category was seen 12 times as a target and 12 times as a distractor. Furthermore, when seen as a distractor, each image appeared the same number of times in the three different categorization tasks. Subjects had one training block of 48 images before starting the 4 series of a given categorization task. Training pictures were not used during testing.

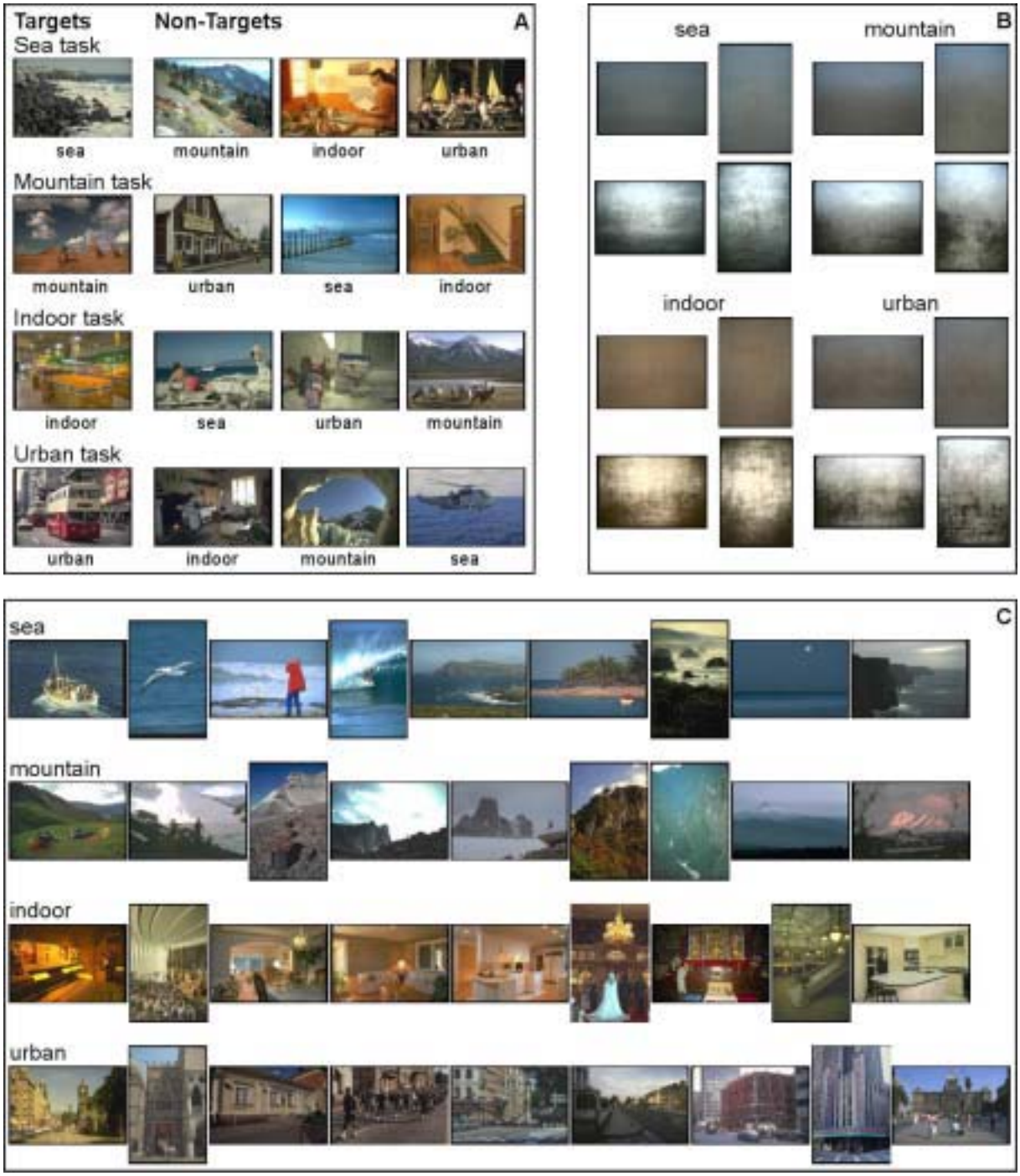


Figure 1. Tasks and stimuli in experiment 1. A) Tasks: while performing one of the four scene categorization tasks (sea, mountain, indoor, urban), non-targets belonged equally to the three other categories. Note the variety of stimuli used in this experiment. B) Pixel by pixel average picture for each stimulus category presented separately for horizontal and vertical images. For each category, the top two images represent the raw average pictures and the two bottom images are the equalized versions obtained using the “equalize” function in Photoshop 5.5. For each color channel and the luminance channel, the function attributes a "black" value to the darkest pixel and a "white" value to the brightest one. It then redistributes regularly the intermediate pixel values of the distribution between these two extremes. C) Examples of pictures used in experiment 1. For each category, the 9 target pictures associated with the fastest reaction times of each subject are presented.

Performance was evaluated by determining the percentage of correct trials and the latency (computed between stimuli onset and finger lift) at which subjects triggered their finger movement response. When repeated measures ANOVA were used, a Greenhouse-Geisser correction for non-sphericity was applied.

Results

Performance in the four tasks was evaluated by analyzing separately accuracy and reaction times (RT). A summary of the results can be found in Table 1.

Accuracy. Subjects performed remarkably well in the four tasks. Mean accuracy was 96.2% correct in the sea categorization task; 95.6% in the mountain task; 95.5% in the indoor task; and 95.1% in the urban task. A one-way analysis on ranks was performed on these data and showed no significant effect (Friedman test: $\chi 2$(3 df)=6.8, n.s.d.). We then analyzed separately go and no-go responses.

In all 4 tasks subjects were better at responding on target scenes (97,4% correct) than at withholding their response on non-target-trials (93,8% correct). Subjects were really good at detecting targets, correct go response reached 98.1% in the sea task; 97.5% in the mountain task; 97.0% in the indoor task; and 97.1% in the urban task. These results were not homogenous (Friedman test: $\chi 2$(3 df)=10.4, p=0.016), subjects scored better with the sea targets. Planned post-hoc Wilcoxon tests showed that this higher accuracy with sea scenes reached significance when compared to indoor and mountain targets (both Z<-2.3, both p<0.02). All other comparisons failed to reach significance.

	SEA	MOUTAIN	INDOOR	URBAN
accuracy (%)				
mean	96.2 (1.9) [91.1-99.7]	95.6 (2.6) [87.8-98.4]	95.5 (2.6) [88.0-98.4]	95.1 (2.7) [88.0-99.2]
correct go	98.1 (2.4) [88.5-100]	97.5 (1.8) [92.2-99.5]	97.0 (2.6) [91.1-100]	97.1 (3.1) [86.0-100]
correct no-go	94.5 (2.7) [88.5-99.5]	93.6 (4.5) [81.8-98.4]	93.9 (3.7) [84.9-99.0]	93.4 (4.8) [80.7-99.0]
correct no-go related	85.9 (7.9)	83.5 (10)	84.3 (9.7)	83.8 (13)
correct no-go unrelated	98.8 (1.2)	98.7 (2.0)	98.7 (1.2)	98.2 (1.7)
d'	3.9 (0.6) [2.7-5.7]	3.7 (0.4) [2.4-4.3]	3.7 (0.6) [2.4-4.6]	3.6 (0.6) [2.5-5.3]
RT (ms)				
mean	422 (37) [348-494]	444 (46) [358-535]	466 (50) [373-555]	482 (45) [403-569]
median	405 (37) [332-477]	425 (45) [334-499]	448 (47) [359-523]	463 (42) [384-538]
minimal RT (ms)				
overall data	260	290	300	300
individual data	331 (28) [280-370]	346 (35) [290-420]	363 (32) [310-420]	372 (36) [310-460]

Table 1. Experiment 1: summary of results. Correct no-go related (or unrelated) accuracy refers to correctly categorized distractor images that belonged (or did not belong) to the same high-level category (natural vs. man-made scenes) as the target images. Standard deviation is indicated in brackets. Range of individual responses is indicated in square brackets [min-max].

Correct no-go responses reached 94.5% in the sea task; 93.6% in the mountain task; 93.9% in the indoor task; and 93.4% in the street task. As sea and mountain scenes both belonged to natural categories whereas indoor

and urban scenes belonged to man-made categories. This means that, in each categorization task, one third of the distractors had a very strong relationship with the target category.

Thus, the performance on distractors was studied separately depending on whether distractors were "related" or "unrelated" to the target category.

Data were analyzed using repeated measures ANOVA with category (4 levels) and related/unrelated (2 levels) as within-subject factors. The analysis showed that subjects performed equally well at ignoring distractors regardless of the task (category factor: $F(2.6,23)=1.0$, n.s.d.). However, correct no-go responses were strongly modulated by the categorical relationship between distractors and targets. Accuracy was significantly worse with distractors that belonged to the related category (84.4%) compared to the two others (98.6%) (related/unrelated effect: $F(1,23)=82.5$, $p<0.0001$). This was true for all categories (no interaction with the category factor; further confirmed by separate Wilcoxon tests on each category, all $Z<-4.1$, all $p<0.0001$).

Reaction times. Although the analysis of accuracy did not reveal major differences between tasks, speed of processing measured by mean and median reaction times (RT) differed strongly between the four categorization tasks (Friedman tests: both $\chi 2$(3 df)>42, both $p<0.0001$). Mean and median RT were respectively 423/405 ms in the sea task; 444/425 ms in the mountain task; 466/448 ms in the indoor task; and 482/463 ms in the urban task. All two by two comparisons on mean and median RT were significant (Wilcoxon tests: all $Z<-2.6$, all $p<0.01$) except the comparisons between the urban task and the indoor task (Figure 2A). Thus the four tasks were ranked according to processing speed as follows: (1) sea, (2) mountain, (3) indoor = urban.

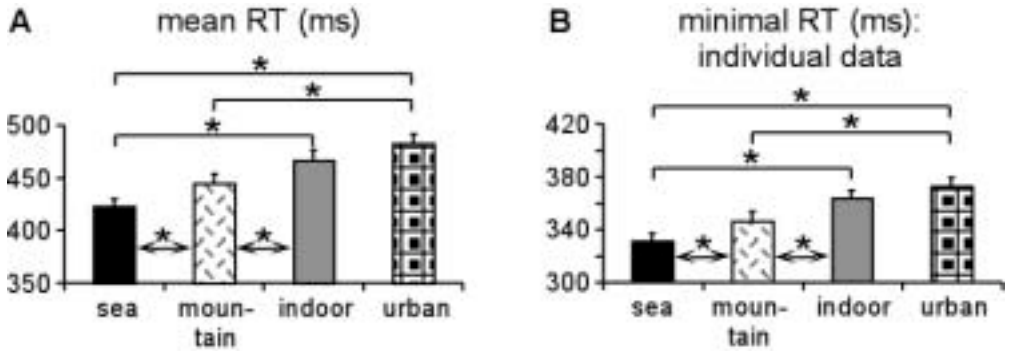


Figure 2. Mean and minimal reaction time with associated standard errors obtained for each of the four scene categorization tasks in experiment 1. Asterisks indicate statistically significant differences (see text for details). In B, asterisks correspond to Wilcoxon tests where all $Z<-2.4$ and all $p<0.02$.

These differences in processing speed can be observed in the RT distributions of Figure 3. Speed of processing was thus faster for the sea context and was also less variable (Figure 3, A, B, C & D) as shown by a narrower RT distribution in the sea task compared to the mountain task (standard deviation: sea = 37 ms, mountain = 46 ms, indoor = 50 ms, urban = 45 ms). Moreover this faster processing speed for sea pictures and, to a lesser extent, for mountain pictures, could be seen even on the fastest responses triggered by the subjects. Thus, a complete shift of the RT distributions towards shorter latencies could be seen for sea and mountain pictures (Figure 3E). Expressing

performance as cumulative *d'* curves as a function of time revealed that discriminative information was available earlier in the sea task than in the three other tasks and accumulated faster to reach a higher value (Figure 3F).

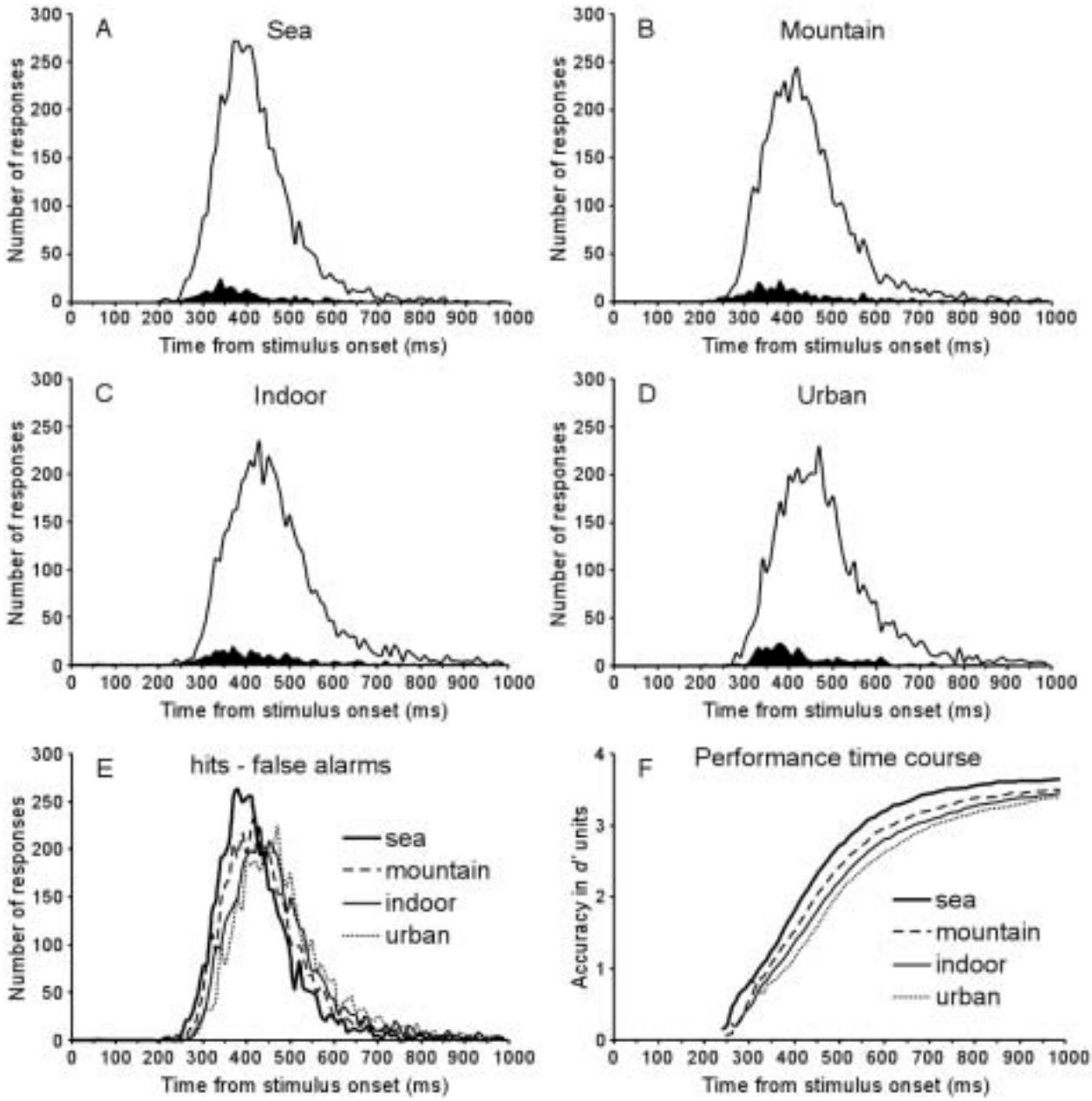


Figure 3. Time course of visual processing in experiment 1. From A to D, RT distributions on correct (upper curve) and incorrect go-responses (black histogram) are presented with the number of responses expressed over time, with 10 ms time bins. In E, false alarm distributions have been subtracted from hit distributions to allow a more direct comparison of the four tasks. In F, average performance accuracy (in *d'* units) is plotted as a function of processing time with 10 ms time bins. Cumulative numbers of responses were used. The *d'* was calculated from the formula $d' = z_n - z_s$, where z_n is chosen such that the area of the normal distribution above that value is equal to the false-alarm rate, and where z_s is chosen to match the hit rate. Note that the *d'* calculated here is not presumed to represent the actual distributions of signal and noise that underlie performance in the response time task. By taking into account the hit and FA rates in a single value at each time point, this time course of performance gives an estimation of the

processing dynamics for the entire subject population. The plateau values correspond to the *d′* calculated from the overall accuracy results.

We then assessed more directly whether this average processing speed ranking of the four tasks was also true for the earliest responses that could be triggered. As target and distractor trials were equally likely, a random behavior should equalize hits and false alarm; hence the minimal behavioral processing time was determined by the latency at which correct go-responses started to significantly outnumber incorrect go-responses ($\chi 2$(1 df), $p<0.001$) using a non-cumulated RT histogram with 10 ms time bins. Such early correct go-responses cannot be considered as anticipations. The analyses were performed both on the overall data (pulling together all trials from all subjects), and for each subject separately. With the overall data set, the minimal processing time was 260 ms in the sea task, 290 ms in the mountain task and 300 ms in both the indoor and the street task. Individual data (computed using cumulated RT histograms with 10 ms time bins) confirmed this tendency with a mean individual minimal processing speed of 331, 346, 363 and 372 ms respectively for sea, mountain, indoor and outdoor target photographs (Figure 2B).

In conclusion, natural environments could be classified faster than man-made environments and this was true for the whole range of responses produced, from the earliest to the latest responses. Furthermore, among natural environments, sea scenes presented a clear processing speed advantage over mountain scenes. Although the accuracy performance was not very different between the four tasks, the rate of information processing was higher in the natural scenes (especially the sea scenes) compared to the man-made scenes.

Comparison with an object categorization task. Results presented above show that the gist of a natural scene flashed for 26 ms can be extracted both very efficiently and very quickly. But how fast is that processing compared to the categorization of objects in natural scenes? Previous studies from our laboratory have extensively assessed the performance of human subjects with the same go/no-go categorization task using animal as target category. A recent study (Rousselet, Macé & Fabre-Thorpe, 2003) is particularly adequate to compare the present human performance on global scene categorization with animal categorization because the same number of subjects were tested (n=24) with the same number of trial per category, an identical set-up, the same image data bank, the same number of trial per category, and the same behavioral procedures (subjects had to alternate between the animal categorization task and a human face categorization task). A similar level of accuracy was also reached in the animal task (96.3%, n.s.d.) but the speed of processing was faster than with scenes. This faster processing was seen when using the median RT which was significantly shorter in the animal task (371 ms) than in any of the four scene categorization tasks used here (two by two comparisons using Mann-Whitney tests: all $U<-2.8$, all $p<0.005$). The fastest discriminative responses were found at the same latency than in the sea categorization task (thus earlier than for any other scene context, Mann-Whitney tests: all $U<-2.6$, all $p<0.01$), an effect that was true for both the overall data set (260 ms in both tasks) and the individual data. On the other hand, performance accuracy increased more rapidly in the animal task than in the sea task. This can be seen on the RT distribution and even more clearly when accuracy performance is expressed in function of time by a d' curve (Figure 4).

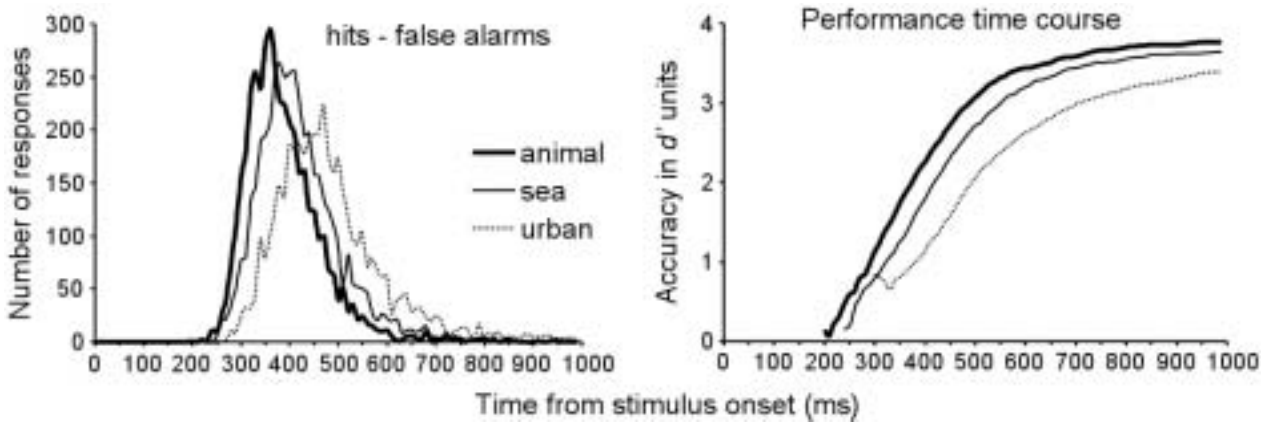


Figure 4. Comparison between scene and object categorization. Top panel: the RT distribution obtained in a preceding study (Rousselet, Macé & Fabre-Thorpe, 2003) on an object categorization task (target = animal) is compared to the RT distributions obtained with the fastest (sea task) and the slowest (urban task) gist categorization tasks (false alarm distributions have been subtracted from hit distributions to allow a more direct comparison). Bottom panel: the *d'* curves show that signal discrimination started earlier in the animal task compared to the sea task, a processing speed advantage that was maintained over the entire range of response latencies. For more details see text and caption of figure 3.

Discussion

This first experiment confirmed that the general meaning of a visual scene can be extracted both very rapidly and highly efficiently with only brief visual presentations (e.g. Biederman, 1972; Potter, 1975, 1976). It also sets a minimal and an average processing time, respectively in the range of 260-300 ms and 400-460 ms, to extract this meaning from natural photographs. These figures are not very different from those obtained for object categorization in a variety of studies performed in our group. However a difference clearly emerged between natural scenes (see, mountain) and man-made scenes (indoor and urban), natural scenes being categorized faster.

Among many properties that might be used to categorize natural scenes, the most obvious one, and probably the easiest to test, is color (e.g. Torralba & Oliva, 2003). For instance, it is very plausible that color was used as a diagnostic cue to categorize sea scenes. The influence of color would be maximal in the sea task, because of large blue textured surfaces in the lower part of the picture and might explained the faster speed of processing in this task. The importance of color cues depends on whether they constitute diagnostic features of the target category (Oliva & Schyns, 2000; Tanaka & Presnell, 1999). Color contrasts may also be used to improve image segmentation, accelerating image analysis (Gegenfurtner, 2003). For example, the importance of color has already been demonstrated in a recognition test using natural scenes (Gegenfurtner & Rieger, 2000). Indeed, such diagnostic cue could allow the pre-setting of specific groups of 'diagnostic' neurons and speed up the processing of the expected visual information (Delorme, Rousselet, Macé & Fabre-Thorpe, in press).

The use of color cues in the rapid categorization of objects has been investigated and it was unexpectedly shown that removing chromatic information had little effect on average accuracy and speed of processing as well as on minimal processing speed (Delorme et al, 2000). But color might not be as an efficient diagnostic cue in animal categorization than in global scene categorization. Experiment 2 was thus designed to test the effect of removing color cues on the categorization of the gist of natural scenes.

EXPERIMENT 2

In this second experiment, we wanted to assess more specifically how the removing of color information from natural scenes would impair performance and whether color could be considered as one of the cue mediating the fast processing of natural scenes revealed in the first experiment. Thus, we tested another group of subjects using the same paradigm and the same set of images employed in experiment 1. The difference between the two experiments was that half of the images were presented in color and half in black and white (BW pictures - 256 grey levels). Color and BW images were mixed at random in the series of stimulation in order to prevent subjects from relying on different strategies when categorizing color and BW pictures and to allow a more direct comparison of human performance in these two conditions.

Method

Participants. Twenty-four adults (12 women and 12 men, mean age 30, ranging from 20 to 52, 3 of them left handed) volunteered in this study and gave their informed written consent. All participants had normal or corrected to normal vision. None of them had been tested in the first experiment.

Stimuli. The same set of natural scenes photographs used in experiment 1 served as stimuli in experiment 2. For each 24-bit (16 millions of colors) photograph, an 8-bit version (256 grey levels) was generated using Photoshop 5.5.

Procedure. The design of experiment 2 was identical to the one of experiment 1 except on two points. First, subjects were tested in a single session. Second, subjects were presented with 50% color and 50% BW photographs. Thus, each experimental condition was subdivided into two color conditions. The design was counterbalanced so that across subjects each image was seen the same number of times in color and in black and white.

Results

In experiment 2, the mean accuracy was not significantly different from the one reached in the first experiment (95.1% and 95.6% correct respectively; between-subject analysis on ranks, Mann-Whitney test: U=281, Z=-0.4, n.s.d.). There was also a non-reliable tendency in favor of fastest responses in the first experiment compared to the second (mean RT were 454 ms and 476 ms respectively, U=216, Z=-1.5, n.s.d.). As in experiment 1, we analyzed separately accuracy and reaction time data. A summary of the results can be found in Table 2.

Accuracy. Global accuracy was very good in experiment 2. Data were entered in a repeated measure ANOVA with category (4 levels) and color (2 levels) as within-subject factors. This analysis showed that the levels of accuracy reached with color and BW images were not significantly different (color = 95.1%; BW = 94.7%; $F(1,23) = 3.8$, n.s.d.). This was true for all four categories of natural scenes as there was no significant interaction between category and color factors.

	SEA		MOUTAIN		INDOOR		URBAN	
	color	**bw**	**color**	**bw**	**color**	**bw**	**color**	**bw**
accuracy (%)								
mean	95.1 (2.6) [89-99]	95.6 (2.8) [86-99]	95.1 (3.4) [83-99]	94.5 (2.8) [84-98]	95.2 (4.3) [77-98]	94.5 (3.5) [82-99]	95.0 (3.5) [84-99]	94.2 (3.1) [87-98]
correct go	95.0 (3.2) [86-100]	95.6 (3.4) [88-100]	95.1 (5.2) [74-100]	95.1 (4.5) [78-100]	96.1 (7.1) [64-100]	95.1 (5.3) [75-100]	95.7 (4.8) [81-100]	95.1 (4.5) [82-100]
correct no-go	95.2 (3.8) [85-99]	95.5 (3.4) [85-100]	95.1 (3.3) [86-100]	93.9 (3.2) [88-99]	94.4 (3.7) [86-99]	93.9 (4.0) [86-100]	94.4 (4.6) [80-100]	93.2 (4.7) [83-100]
correct no-go related	87.2 (10.8)	88.8 (9.2)	87.1 (8.8)	83.7 (8.1)	86.2 (8.7)	84.1 (10.3)	88.2 (7.4)	82.9 (11.3)
correct no-go unrelated	99.2 (1.5)	98.8 (1.5)	99.0 (1.9)	99.0 (1.6)	98.5 (1.9)	98.8 (2.0)	97.5 (3.8)	98.4 (1.9)
d'	3.5 (0.6) [2.4-5.2]	3.6 (0.6) [2.2-4.5]	3.5 (0.6) [2.1-4.9]	3.5 (0.6) [2.0-4.9]	3.7 (0.6) [1.7-4.3]	3.5 (0.8) [1.9-5.0]	3.6 (0.6) [2.0-5.2]	3.4 (0.6) [2.3-4.6]
RT (ms)								
mean	443 (56) [300-566]	461 (60) [323-586]	471 (58) [359-603]	462 (63) [332-595]	493 (70) [344-640]	503 (65) [352-632]	498 (55) [347-599]	499 (58) [360-636]
median	429 (56) [288-549]	444 (61) [310-589]	452 (59) [336-572]	443 (61) [320-572]	475 (70) [314-619]	485 (67) [324-625]	479 (57) [318-580]	478 (61) [322-615]
minimal RT (ms)								
overall data	290	300	310	310	320	340	330	310
Individual data	388 (45) [280-490]	398 (50) [280-520]	401 (50) [300-500]	400 (54) [290-510]	425 (64) [300-600]	434 (55) [320-560]	429 (48) [310-520]	432 (53) [310-580]

Table 2. Experiment 2: summary of results. For each condition, the results are indicated separately for color pictures (color) and for grey level pictures (bw). For other details see table 1 caption.

Contrary to experiment 1, accuracy on go and no-go responses did not differ significantly from one another (go = 95.4%; no-go = 94.5%; $F(1,23)=1.2$, n.s.d.). Separate analysis showed that the removal of color cues had no significant effect on either go responses (color = 95.5%; BW = 95.2%; $F(1,23)=0.7$, n.s.d) and no-go responses (color = 94.8%, BW = 94.1%; $F(1,23)=3.3$, n.s.d.). In addition, there was no difference in accuracy between the four tasks for both go responses ($F(2.2,51)=0.2$, n.s.d.) and no-go responses ($F(2.4,55)=2.2$, n.s.d.). There was no significant interaction between color and category factors.

Like previously found in experiment 1, no-go responses were made more frequently toward distractors that belonged to the same higher level category as the targets (natural versus man-made). In other words, subjects proved much better at categorizing distractors unrelated (98.6%) than related (86.0%) to the target category ($F(1,23)=92.6$, $p<0.0001$). The only effect induced by the removal of color cue was seen on related distractors in the urban task: indoor pictures were correctly ignored with a higher accuracy when presented in color (88.2%) than in BW (82.9%) ($Z=2.3$, $p=0.02$).

Reaction times. An ANOVA analysis showed that reaction times were affected both by the category of the target scene (category effect on both median and mean RT both $F>26$, $p<0.0001$) and by the availability of color cues (mean RT: $F(1,23)=9.1$, $p=0.006$; median RT: $F(1,23)=5.2$, $p=0.03$) so that speed of performance was analyzed separately on color and BW pictures for each categorization task (Figure 5).

However the results were not consistent from one category to another. Whereas sea and indoor pictures were categorized faster in color than in BW (about 15 ms and 10 ms faster respectively; Wilcoxon tests: both $Z<-3.2$, both $p<0.001$, for both mean and median RT), urban scenes showed no effect of color cues removal (mean and median

RT: both Z>-0.7, both *n.s.d.*), and, surprisingly, mountain pictures showed a tendency to be categorized with a slower speed in color (9 ms slower) than in BW. This tendency reached significance for mean RT (Z=-2.2, p=0.03) but not for median RT which presented only a borderline effect (Z=-2.0, p=0.05).

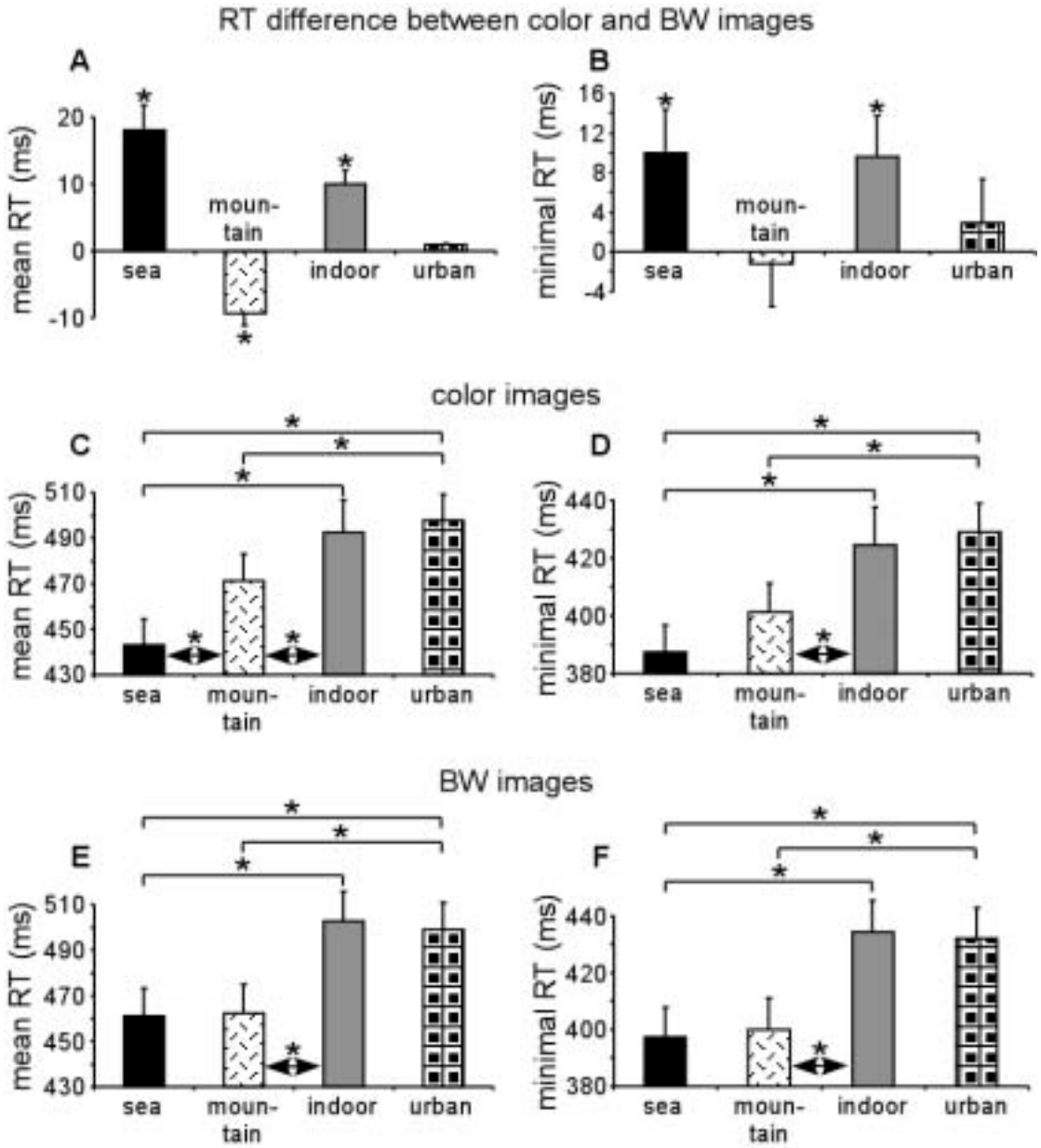


Figure 5. Speed of processing in experiment 2: mean RT (left column) and minimal RT (right column). The top two graphs illustrate the color processing speed advantage by subtracting, for each of the four categorization tasks, the value obtained with color images from the value obtained with BW images. Mean reaction times for each categorization task and associated standard errors are shown for color images (middle panel) and for BW images (bottom panel). An asterisk shows statistically significant between-category effects.

Regarding differences in processing speed between the four categories, the results obtained with color images tested separately showed the robustness of the results obtained in experiment 1. Indeed, as in this first

experiment, all two by two comparisons were significant for both mean and median RT (all Z<-2.5, all p<0.02) except between indoor and urban scenes. Thus the two man-made categories were categorized at about the same speed.

When BW pictures were analyzed separately, the same pattern of results appeared again for both mean and median RT (all Z<-3.5, all p<0.0001) with the exception that the two natural categories (mountain and sea pictures) were categorized at the same speed.

The very limited effect on performance speed linked to the removal of color information when extracting the gist of natural scenes can be seen in RT distributions (Figure 6). Color and BW RT distributions were virtually superimposed in the case of urban pictures and the amplitude of the shift towards shorter RT (for sea and indoor scenes) or towards longer RT (for mountain scenes) was indeed very restricted. The time course of performance (Figure 6, insets) again shows how small the effect of removing color was on subjects' capacity to discriminate between targets and distractors. Cumulated *d'* curves were virtually superimposed from the earliest responses to the plateau, with very similar slope, indicating that information accumulated at a similar speed for color and BW pictures. Regarding the small differences in plateau value, two by two one-way analyses on ranks revealed only one significant difference, namely that signal detection was slightly higher with color images compared to BW images in the urban task (color = 3.4; BW = 3.3; Z=-2.1, p=0.03; other comparisons were not significant).

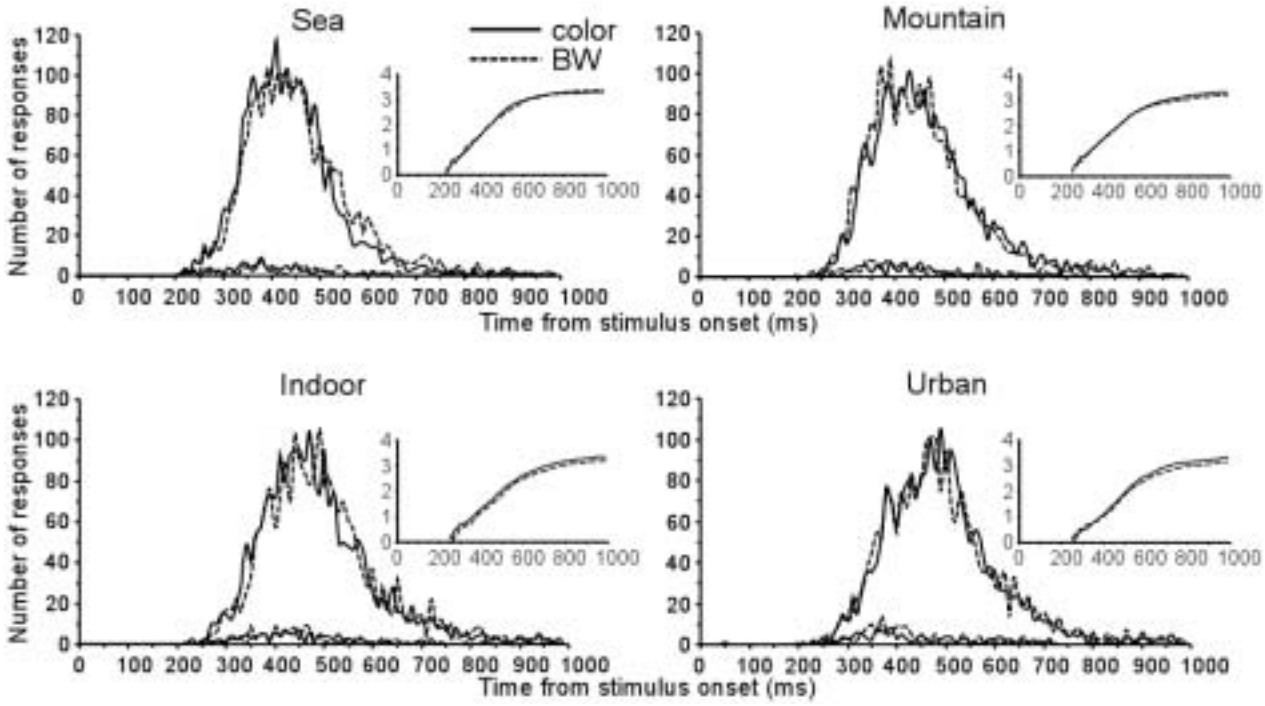


Figure 6. Reaction time distributions for correct hits (top two traces) and for false alarms (bottom two traces) for the four scene categorization tasks in experiment 2. The graphs compare for each target category the reaction time distributions associated with color (black trace) and BW images (grey trace). Each time, the insets show the *d'* computed from the cumulative number of responses in the RT histograms in both conditions. For details see caption of figure 3.

Discussion

Two main points have been stressed by the results obtained in these two studies. First, human subjects are both very accurate and very fast at categorizing the gist of real-world scenes with a processing speed advantage in favor of natural scenes compared to man-made scenes. Second, the removal of color cues has a very restricted effect on the performance of scene context categorization. Moreover, comparing the present results with those obtained for object categorization using the same task and the same set-up provides information on the possible interactions between the processing of objects and of the context in which they are presented.

Scene categorization is accurate and fast, but faster for natural scenes

Experiment 2 confirmed that subjects could extract very rapidly the global context of a natural scene. It also strengthened the finding that natural scenes can be processed faster than man-made scenes, and confirmed a slight processing speed advantage of sea images over mountain images.

Overall, experiments 1 and 2 showed that human subjects can differentiate complex scene categories with median RT ranging from about 400 to 460 ms but with early responses observed at latencies that can be as short as 260-300 ms. This processing time is remarkably short compared to RT observed when subjects are simply asked to detect the appearance of a natural scene on the screen. Indeed, human subjects are able to detect that a scene has appeared, whatever its category, with a mean RT of about 230 ms (VanRullen & Thorpe, 2001) or even shorter (Rousselet, unpublished data: 20 subjects, mean RT=211 ms). Thus, using a mean RT of 230 ms in such detection task as a reference, the additional cost to realize a complex gist categorization task is on average 170-230 ms, but can be as low as 30-70 ms for early responses. These strong temporal constraints are in favor of models of visual processing relying essentially on coarse, feedforward and massively parallel mechanisms to achieve scene recognition (Oliva & Torralba, 2001; Torralba & Oliva, 2003; VanRullen & Thorpe, 2002).

Both experiments also revealed a clear processing speed advantage in favor of natural scenes compared to man-made scenes. This advantage was not mediated by color cues, given that the same was true with BW pictures. Aside from color, several combinations of low-level scene-based properties could potentially explain how subjects classified pictures in these two experiments (Oliva & Torralba, 2001; Torralba & Oliva, 2003). But the bias toward natural scenes might find its explanation in a stronger variability in the physical properties of man-made environments compared to the natural environments used in these experiments. Indeed, although we used very various natural scenes, they were still limited to sea and mountain pictures. On the other hand, indoor and outdoor man-made scenes covered almost the whole collection of imaginable man-made scenes. This resulted in a more constrained dictionary of physical properties for natural compared to man-made scenes, which gave the opportunity to more specifically pre-set groups of 'diagnostic' neurons responding to target properties in the natural scene images. The stronger variability in man-made scenes compared to natural scenes can be appreciated from the average scenes and the examples in Figure 1 (B & C respectively). This in turn might allow faster decision making, on the basis of limited need for evidence, in the case of natural scenes compared to man-made scenes. Future experiments will be needed to evaluate more carefully how gist processing speed is affected by target diagnosticity, for instance by varying the physical similarities among targets and between target and distractor scenes.

Influence of color cues in scene categorization

The second experiment provided evidence about the role played by color information in task performance for the four scene contexts studied here. In the first place, it is surprising to note that color had virtually no effect on subjects' capacity to detect targets. Thus, color is not necessary to categorize real-world pictures in this fast go/no-go paradigm. Concerning the speed of processing of urban scenes, the lack of effect of removing the color cues is in agreement with one previous report from Oliva & Schyns (2000). However, in the present study we found that color helped the subjects at correctly ignoring non-target indoor pictures in the urban task (man-made category-related distractors) as they scored better when such scenes were presented in color than in BW. Color seems to play a role in the processing of indoor scenes as subjects were also faster to make a decision on color than on BW indoor target pictures. However, this effect is weak (only 10 ms), and it might have been strengthen by our protocol as it was not found by Oliva & Schyns (2000). The two man-made scene categories used in this experiment might share many low-level properties, forcing subjects to rely on distinctive features to discriminate between those two categories. One of these features might have been the diagnostic yellowish/brownish color present in many indoor scenes because of artificial lightings (see the color version of Figure 1B online). Note that this color advantage in the indoor categorization task could even be seen on the earliest responses produced. Thus, color removal might have a small effect on response speed in this task, but could be used very early during the course of visual processing, in keeping with results from Gegenfurtner & Rieger (2000).

Processing speed was also slightly faster (15 ms on average) when color was available in the sea task, which again affected the earliest discriminative behavioral responses. This is not surprising, given that the blue of the sea was largely predictive of the category. However, color alone does not appear to be able to explain the good performance in this task. First, BW sea pictures were categorized with a good efficiency. Second, the blue feature was not a specific attribute of the sea category; it was present in large proportions in the skies of the mountain task and some urban pictures. Thus, it seems that the use of a single strategy relying on the diagnostic use of blue was not sufficient to perform the task. But subjects could also rely on a slightly more sophisticated strategy based on the detection of a blue surface situated in the lower part of the scene (Delorme et al., in press).

The use of blue textured surfaces as a cue in the sea task might also explain the unexpected negative effect of color in the mountain task in which large surfaces of blue sky were often present and could induce ambiguity as a common feature with some of the distractors. Anyway, color effects were again relatively weak in this task (about 9 ms) and, contrary to the indoor and the sea task, did not affect the earliest responses, but essentially behavioral responses triggered with average RT larger than the median RT (>450ms).

The magnitude of the color effects reported here is relatively small compared to the average 50 ms advantage in naming RT reported for color pictures over BW pictures by Oliva & Schyns (2000). The category that is the more related to our *sea* category was their *coastline* category which presented a color effect of about 50 ms. This effect is much larger than the one reported here (about 15 ms). A very plausible interpretation for this discrepancy relies on the use of two different tasks, a go/no-go task in the present study and a naming task in Oliva & Schyns (2000). It is possible that the necessity to name pictures, which was also associated with an additional 200-300 ms RT, forced subjects to rely on different representations, more sensitive to the color factor. Another more plausible explanations stems in the fact that in Oliva & Schyns (2000) subjects had to name a picture belonging to 8

possible categories on every trial, while in our experiment subjects had a unique target for 384 consecutive trials, a protocol used to allow subjects to respond as fast as they could. This, in addition with the intermixed appearance of BW and color pictures, might have biased subjects' strategy towards the use of non-color properties like spatial frequencies, textures, depth of field, and other properties that have been shown to constitute valid cues for scene categorization (Torralba & Oliva, 2003).

Overall, these results are compatible with a previous report from Delorme, Richard & Fabre-Thorpe (2000), showing that in fast object categorization tasks color has only a limited effect (food categorization task) if any at all (animal categorization task). We thus conclude that although color can be extracted very rapidly during the course of visual processing and can be used to improve performance when it is diagnostic of the target scene category, it is not a crucial feature for processing speed in a fast go/no-go categorization task of real-world scenes.

Object categorization versus scene categorization

In experiment 1, we performed a tentative comparison between object and scene categorization performances. This analysis revealed that although subjects were fast in categorizing a whole scene, this fast processing of scene gist was actually on average 30-90 ms slower than the processing of objects in natural scenes, like animals, but also like humans, human and animal faces and means of transport (Delorme, Richard & Fabre-Thorpe, 2000; Rousselet, Macé & Fabre-Thorpe, 2003; VanRullen & Thorpe, 2001). This object advantage might very plausibly find its origin in the weaker structural constraints found in natural scenes. Indeed, the same gist can be assigned to scenes with relatively different low-level features and spatial arrangements. It is probably this relatively loosely defined structure of the scenes compared to component objects (like animals, vehicles, faces…) that can explain their slower processing speed. However, this does not mean that scene categorization relies on higher-level representations than object categorization. First, the fact that subjects did systematically more errors on distractors that belonged to the same higher-level category as the target of the task ('natural' vs. 'man-made' categories) might be taken as an evidence for the use of relatively low level cues in these tasks. Second, as we have argued recently (Rousselet, Macé & Fabre-Thorpe, 2003), the rapid categorization of objects like faces and animals in natural scenes might depend on coarsely defined features of intermediate complexity rather than on high-level complete descriptions (see also Ullman, Vidal-Naquet & Sali, 2002). Thus, the slower processing of scenes compared to objects might find its explanation in the need to integrate in parallel a larger conjunction of relatively low level features in order to reach a decision level in the processing of a natural context. The rapid categorization of objects in natural scenes would rely on the conjunction of a more limited number of features than the categorization of gist. Hence, because a given object category like animals has a more predictable physical description, neurons coding for target objects in the ventral pathway would benefit from a finer task-related top-down pre-setting in the animal task compared to the scene tasks. In addition, the difference in processing speed between objects and scenes could reflect the limitations of our visual system to process natural scenes in parallel. Indeed, although we recently demonstrated that two scenes can be processed in parallel (Rousselet, Fabre-Thorpe & Thorpe, 2002), we have now evidence that this capacity is limited to certain conditions (Rousselet, Thorpe & Fabre-Thorpe, in preparation; VanRullen, Redy & Koch, in press). Constraints on parallel processing would be even stronger on gist categorization because it requires integrating a large collection of low-level features to make a decision.

However, these results should not be considered as an argument in favor of models postulating that there is no early interaction between scene and object representations (e.g. Henderson & Hollingworth, 1999). Indeed, in everyday situations, the gist of a scene is not changing abruptly from one image to the next but is much stable over time, probably allowing predictive hypotheses about possible objects to build up, in keeping with modern interactive frameworks (Bullier, 2001; Rao & Ballard, 1999; Ullman, 1995). But even here, using very short presentations, a considerable overlap was observed between the RT distributions of object and scene categorization. This overlap shows that the processing of an object might benefit from the simultaneous processing of the context in which it appears. The activation of congruent populations of neurons would probably allow a faster identification of a cow in a field than in a church. This issue clearly deserves further investigations.

CONCLUSIONS

Confirming earlier studies that have used brief visual presentations, data from the two experiments reported here showed that the gist of real-world scenes could be extracted with a high accuracy and with short RT in a fast go/no-go visual categorization task. Furthermore, it was shown that natural scenes used in these experiments were processed faster than man-made scenes, probably because features of natural scenes might be more diagnostic than those of man-made scenes, allowing a stronger top-down presetting. In addition, we showed that if color information could be used very early to process more efficiently some specific scene categories, it does not appear as the most crucial aspect of the real-world scenes used by human subjects to perform the fast go/no-go categorization tasks studied here.

Acknowledgments
This work was supported by the CNRS and the Cognitique grant n°IC2. Financial support was provided to G.A. Rousselet by a Ph.D. grant from the French government. We thank Anne-Sophie Paroissien & Olivier Joubert for their very valuable help running the experimental sessions in experiments 1 and 2 respectively.

REFERENCES

Aguirre, G. K., Zarahn, E., & D'Esposito, M. (1998). An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. *Neuron, 21*(2), 373-383.
Bar, M., & Aminoff, E. (2003). Cortical analysis of visual context. *Neuron, 38*(2), 347-358.
Biederman, I. (1972). Perceiving real-world scenes. *Science, 177*(43), 77-80.
Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review, 94*(2), 115-147.
Biederman, I. (1988). Aspects and extensions of a theory of human image understanding. In Z. W. Pylyshyn (Ed.), *Computational processes in human vision: an interdisciplinary perspective* (pp. 370-428). Norwood (N.J.): Ablex.
Biederman, I., Glass, A. L., & Stacy, E. W., Jr. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology, 97*(1), 22-27.
Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*(2), 143-177.
Boyce, S. J., Pollatsek, A., & Rayner, K. (1989). Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 556-566.
Bullier, J. (2001). Integrated model of visual processing. *Brain Research Reviews, 36*(2-3), 96-107.
Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences, 4*(5), 170-178.
Delorme, A., Richard, G., & Fabre-Thorpe, M. (2000). Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans. *Vision Research, 40*(16), 2187-2200.

Delorme, A., Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (in press). Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Cognitive Brain Research*.
Epstein, R., Graham, K. S., & Downing, P. E. (2003). Viewpoint-specific scene representations in human parahippocampal cortex. *Neuron, 37*(5), 865-876.
Epstein, R., Harris, A., Stanley, D., & Kanwisher, N. (1999). The parahippocampal place area: recognition, navigation, or encoding? *Neuron, 23*(1), 115-125.
Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience, 13*(2), 171-180.
Friedman, A. (1979). Framing pictures: the role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General, 108*(3), 316-355.
Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Cognitive Brain Research, 16*(2), 123-144.
Gegenfurtner, K. R. (2003). Cortical mechanisms of colour vision. *Nature Reviews Neuroscience, 4*(7), 563 -572.
Gegenfurtner, K. R., & Rieger, J. (2000). Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology, 10*(13), 805-808.
Henderson, J. M. (1992). Object identification in context: the visual processing of natural scenes. *Canadian Journal of Psychology, 46*(3), 319-341.
Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50*, 243-271.
Henderson, J. M., & Hollingworth, A. (2003). Eye movements and visual memory: detecting changes to saccade targets in scenes. *Perception & Psychophysics, 65*(1), 58-71.
Hollingworth, A. (2003). Failures of retrieval and comparison constrain change detection in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance, 29*(2), 388-403.
Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General, 127*(4), 398-415.
Hollingworth, A., & Henderson, J. M. (1999). Object identification is isolated from scene semantic constraint: evidence from object type and token discrimination. *Acta Psychologica, 102*(2-3), 319-343.
Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance, 28*(1), 113-136.
Humphreys, G. W. (1998). Neural representation of objects in space: a dual coding account. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 353*(1373), 1341-1351.
Maguire, E. A., Frith, C. D., & Cipolotti, L. (2001). Distinct neural systems for the encoding and recognition of topography and faces. *Neuroimage, 13*(4), 743-750.
Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.
Nakamura, K., Kawashima, R., Sato, N., Nakamura, A., Sugiura, M., Kato, T., Hatano, K., Ito, K., Fukuda, H., Schormann, T., & Zilles, K. (2000). Functional delineation of the human occipito-temporal areas related to face and scene processing. A PET study. *Brain, 123*(9), 1903-1912.
O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: the world as an outside memory. *Canadian Journal of Psychology, 46*(3), 461-488.
Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology, 34*(1), 72-107.
Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology, 41*(2), 176-210.
Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*(3), 145–175.
Potter, M. C. (1975). Meaning in visual search. *Science, 187*(4180), 965-966.
Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2*(5), 509-522.
Potter, M. C. (1999). Understanding sentences and scenes: The role of conceptual short-term memories. In V. Coltheart (Ed.), *Fleeting memories* (pp. 13-46). Cambridge, Massachusetts: MIT Press.
Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology, 81*(1), 10-15.
Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79-87.
Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition, 7*(1/2/3), 17-42.
Rensink, R. A. (2002). Change detection. *Annual Review of Psychology, 53*, 245-277.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience, 5*(7), 629-630.
Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *Journal of Vision, 3*(6), 440-455.
Rousselet, G. A., Thorpe, S. J., & Fabre-Thorpe, M. (in preparation). Processing of one, two or four natural scenes in humans: the limits of parallelism.
Sanocki, T., & Epstein, W. (1997). Priming spatial layout of scenes. *Psychological Science, 8*(5), 374-378.
Sato, N., Nakamura, K., Nakamura, A., Sugiura, M., Ito, K., Fukuda, H., & Kawashima, R. (1999). Different time course between scene processing and face processing: a MEG study. *Neuroreport, 10*(17), 3633-3637.
Schyns, P. G. (1998). Diagnostic recognition: task constraints, object information, and their interactions. In M. J. Tarr & H. H. Bülthoff (Eds.), *Object recognition in man, monkey, and machine* (pp. 147-179). Amsterdam: Elsevier Science Publishers.
Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time and spatial scale dependant scene recognition. *Psychological Science, 5*, 195-200.
Schyns, P. G., & Oliva, A. (1997). Flexible, diagnosticity-driven, rather than fixed, perceptually determined scale selection in scene and face recognition. *Perception, 26*(8), 1027-1038.
Simons, D. J., Chabris, C. F., Schnur, T., & Levin, D. T. (2002). Evidence for preserved representations in change blindness. *Consciousness and Cognition, 11*(1), 78-97.
Smid, H. G., Jakob, A., & Heinze, H. J. (1997). The organization of multidimensional selection on the basis of color and shape: an event-related brain potential study. *Perception & Psychophysics, 59*(5), 693-713.
Syrkin, G., & Gur, M. (1997). Colour and luminance interact to improve pattern recognition. *Perception, 26*(2), 127-140.
Tanaka, J. W., & Presnell, L. M. (1999). Color diagnosticity in object recognition. *Perception & Psychophysics, 61*(6), 1140-1153.
Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520-522.
Torralba, A. (2003). Contextual priming for object detection. *International Journal of Computer Vision, 53*(2), 153-167.
Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: computation in neural systems, 14*, 391-412.
Ullman, S. (1995). Sequence seeking and counter streams: a computational model for bidirectional information flow in the visual cortex. *Cerebral Cortex, 5*(1), 1-11.
Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5*(7), 682-687.
VanRullen, R., Reddy, L., & Koch, C. (in press). Visual search and dual-tasks reveal two distinct attentional resources. *Journal of Cognitive Neuroscience*.
VanRullen, R., & Thorpe, S. J. (2001). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception, 30*(6), 655-668.
VanRullen, R., & Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vision Research, 42*(23), 2593-2615.
Wolfe, J. M. (1999). Inattentional amnesia. In V. Coltheart (Ed.), *Fleeting memories* (pp. 71-94). Cambridge, Massachusetts: MIT Press.

Processing speed of natural scenes: categorization of the global context

Guillaume A. Rousselet & Michèle Fabre-Thorpe
guillaume.rousselet@cerco.ups-tlse.fr michele.fabre-thorpe@cerco.ups-tlse.fr
Centre de Recherche Cerveau & Cognition (CERCO), UMR 5549 CNRS-UPS, 133 Route de Narbonne, 31062 Toulouse, FRANCE

Introduction

visual natural scene = discrete objects + background elements + spatial layout = global scene context or "gist"

objects are processed very efficiently and very fast in natural scenes
Thorpe et al. (1996)
Fabre-Thorpe et al. (1998)
Van Rullen & Thorpe (2001)

Very few studies on the processing of the global scene context

different cortical representations for objects and scenes
Humphreys (1998)
Aguirre et al. (1998)
Epstein et al. (1998, 2000)
Saito et al. (1995)
Nakamura et al. 2000

scene identification from very brief image presentations
Potter (1975, 1976)
Biederman et al. (1973, 1982)
Schyns & Oliva (1994)
Oliva & Schyns (1997)

types of information used for scene identification

classic view = high-level representations

alternative view = low- or mid-level representations

scene-based "gecore"
Biederman (1986)
diagnostic spatial information
Schyns & Oliva (1994)
Oliva & Schyns (1997)
diagnostic colors and colored blobs
Oliva & Schyns (2000)
medium-level spatial properties
Oliva & Torralba (2001)

The global meaning of a scene might be processed concurrently with the identification of objects, not after

Problem: rapid presentations give a rate not a speed of processing

Question: what is the precise processing speed of the global scene context?

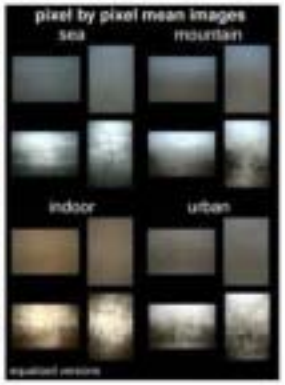
Categorization of the global scene context:

4 scene categories = 4 tasks

(superordinate level) natural scenes / man-made scenes
(basic level) sea / mountain / indoor / urban

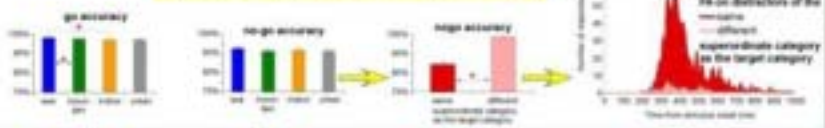
Methodology: go/no-go task

- 24 subjects 1536 trials per subject
- go responses on targets (release response key >1sec)
- no-go responses on non-targets (keep pressing response key >1sec)
- Unique & brief presentations (23 ms)
- ISI = 1600-2200 ms
- 4 blocks: 4 x 4 consecutive series = subjects were trained on 192 a series before each task
- task order counterbalanced over subjects
- 1 series = 96 trials (48 target trials + 48 non-target trials)
- Non-target pictures = 16 pictures of the 3 other target categories
- Each category: 384 images (192 horizontal + 192 vertical images)
- Images: 573 x 768 pixels = 10.5° x 15.8°



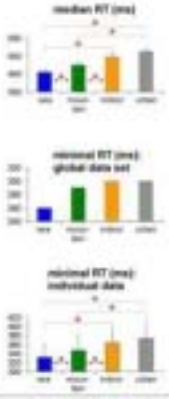
Accuracy

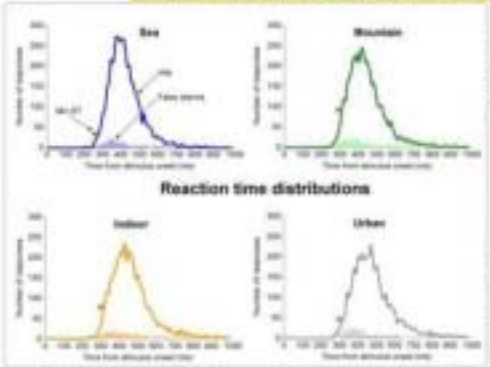
High-level of accuracy in every task.
Slightly better categorization of sea targets (color diagnosticity?)
More errors on distractors belonging to the same superordinate category.

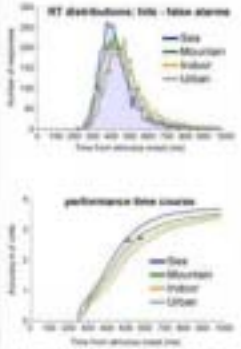


Processing time course

Processing speed is faster for natural scenes vs. man-made scenes.
This bias affects the entire range of responses produced, from the earliest to the latest responses.
Hyp: natural scenes are more homogeneous than man-made scenes and/or contain more diagnostic features (coarse spatial frequencies, colors...) thus enhancing their predictability?



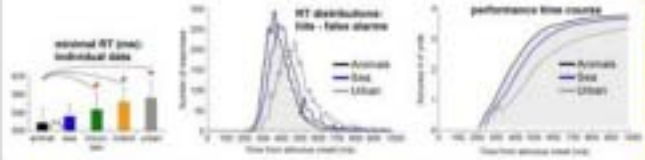




Between-subject comparison with an object categorization task

go/no-go animal task

data from a previous experiment
(Rousselet, Macé & Fabre-Thorpe, submitted)
*same number of trials
*same number of subjects
*same experimental design (2 tasks)
*pictures from the same collection



Earliest responses in the sea task might be triggered at about the same latency as in the animal task but are less numerous.

The global RT distribution is delayed towards longer latencies in the sea vs. animal task.

Processing speed advantage for objects vs. context?

Conclusions

1. Fast and accurate categorization of the global scene context.
2. Faster processing of natural vs. man-made scenes: differences in spatial and/or color cues?
3. Pattern of false alarms => use of relatively low-level cues in these tasks.
4. Scene context categorization is as fast or slightly slower than an object categorization task.
hypothesis: a) it takes time to integrate a large set of low-level features to categorize the context;
b) objects like animals are categorized using a more limited set of relatively low-level features.
5. Suggested experiments: 1) replicate the present experiment with 50% color / 50% BW pictures to test color hypothesis; 2) replicate the present experiment with e.g. indoor scenes as targets and sea and mountain scenes as distractors to test the hypothesis of a lack of diagnosticity / animal task.

Article 4

Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes

Delorme, A., Rousselet, G.A., Macé, M.J.-M. & Fabre-Thorpe, M.

(*Cognitive Brain Research*, sous presse)

Résultats comportementaux et électrophysiologiques de 14 sujets adultes dans une expérience comparant la vitesse de traitement d'une scène naturelle dans la tâche de catégorisation go/no-go habituelle et dans une tâche de reconnaissance dans laquelle la cible est entièrement prédictible.

Introduction

L'article 4 de cette thèse avait pour objectif de mieux cerner l'influence réciproque des facteurs ascendants et descendants dans le traitement des scènes naturelles. Les sujets alternaient entre une tâche de catégorisation animal/non-animal et une tâche de reconnaissance dans laquelle pour chaque série une image différente était apprise et constituait l'unique cible présentée 50 fois parmi 50 images non cibles. Cette tâche de reconnaissance a été conçue pour maximiser autant que possible les influences descendantes sur le traitement des scènes naturelles.

Résultats

Cette expérience a révélé deux résultats principaux.

1) Par rapport à la tâche de catégorisation, l'économie temporelle pour réaliser la tâche de reconnaissance était relativement faible, seulement 40 ms au niveau des réponses comportementales précoces et 30 ms en prenant en compte le début de l'activité différentielle dans les deux tâches.

2) Une analyse de source de l'activité différentielle suggère que dans les deux tâches la « décision visuelle » serait prise par les mêmes aires corticales, localisées grossièrement dans les régions postérieures et ventrales du cerveau.

Discussion

Si la précision de l'analyse de source réalisée sur ces données est limitée par l'emploi d'un bonnet à 32 électrodes, la co-localisation des dipôles expliquant la majeure partie de l'activité différentielle dans les deux tâches est à l'origine d'une hypothèse de travail très intéressante : quelque soit la difficulté de la tâche, la discrimination entre cibles et distracteurs (la 'décision' visuelle) pourrait être prise en charge par la même zone corticale. Selon cette hypothèse, l'activité différentielle résulterait de l'interaction entre informations montantes et descendantes dans certaines zones critiques de la voie ventrale. De plus amples investigations seraient nécessaires pour étayer cette hypothèse, par exemple en combinant potentiels évoqués et IRM fonctionnelle.

Le résultat le plus important de cette expérience est sans doute le délai relativement faible entre le temps nécessaire pour répondre sur une image particulière apprise et répétée 50 fois et la tâche de catégorisation animal/non-animal. En effet, la tâche de reconnaissance était conçue pour être une des plus simples possibles afin de fournir des contraintes temporelles aussi proche que possible d'une valeur plancher. La tâche était effectivement très simple et l'analyse des erreurs suggère que les sujets ont utilisé des indices relativement bas niveau pour la réaliser, tels que des patchs de couleur ou certaines orientations dans une zone particulière de la scène. Etant donnée la simplicité de la tâche de reconnaissance, il peut donc paraître surprenant que la tâche de catégorisation nécessite seulement 30 à 40 ms supplémentaires pour être réalisée. Cette faible différence suggère qu'il n'y a pas nécessairement besoin de mettre en jeu des représentations très détaillées pour réaliser la tâche animal/non-animal.

Si on replace ce résultat dans le cadre de la discussion de l'article 3, il renforce l'idée selon laquelle les contraintes temporelles qui pèsent sur la catégorisation du contexte et des objets dépendent fortement de la diagnosticité des cibles. D'une part, l'augmentation de la diagnosticité des catégories de contextes cibles pourrait donc être associée à une réduction significative de leur temps de traitement. D'autre part, la nécessité d'effectuer des catégorisations plus détaillées ralentirait parfois considérablement l'analyse des objets (Macé, Thorpe & Fabre-Thorpe, en préparation). Selon les cas, les objets pourraient ainsi être analysés beaucoup plus lentement ou beaucoup plus rapidement que le contexte.

On peut aussi replacer le résultat de l'article 4 dans le cadre du traitement en parallèle des animaux dans les scènes naturelles. En effet, si les résultats des articles 1 et 2 suggèrent une grande part de parallélisme, il reste tout à fait envisageable que celui-ci soit limité à des tâches

mettant en jeu une catégorisation super-ordonnée des objets. Il est envisageable que les résultats soient différents si la tâche des sujets été d'effectuer une catégorisation oiseau/non-oiseau ou chien/non-chien, qui semble requérir des représentations plus fines que la catégorisation animal/non-animal (Macé, Thorpe & Fabre-Thorpe, en préparation). A l'inverse, le parallélisme aurait pu être beaucoup plus évident si la tâche avait été de détecter une image apprise au préalable, comme c'était le cas dans la tâche de reconnaissance de l'article 4.

Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes

Delorme Arnaud, Rousselet Guillaume A., Macé Marc J.-M.
& Fabre-Thorpe Michèle

Abstract

The influence of task requirements on the fast visual processing of natural scenes was studied in 14 human subjects performing in alternation an "animal" categorization task and a single-photograph recognition task. Target photographs were randomly mixed with non-target images and flashed for only 20 ms. Subjects had to respond to targets within 1 s. Processing time for image-recognition was 30-40 ms shorter than for the categorization task, both for the fastest behavioral responses and for the latency at which event related potentials evoked by target and non-target stimuli started to diverge. The faster processing in image-recognition is shown to be due to the use of low-level cues, but source analysis produced evidence that, regardless of the task, the dipoles accounting for the differential activity had the same localization and orientation in the occipito-temporal cortex. We suggest that both tasks involve the same visual pathway and the same decisional brain area but because of the total predictability of the target in image-recognition, the first wave of bottom-up feed-forward information is speeded up by top down influences that might originate in the prefrontal cortex and preset lower levels of the visual pathway to the known target features.

1. INTRODUCTION

Spotting a specific object among others is an every day task that appears trivial but raises a number of questions concerning the underlying visual processing. In visual search tasks, subjects are asked to look for a pre-specified target embedded in distractor arrays. Typically, for low-level features, ERP studies suggest that a visual decision can be made in about 150 ms [1,21,34]. This latency increases when targets are defined by a conjunction of characteristics such as form and color [18], although pop out has been reported for some specific conjunction of low-level features [7,21,28,38]. Surprisingly, 150 ms has also been reported to be the minimal processing time to differentiate between different classes of natural images. Using a superordinate categorization task in which human subjects had to respond when a natural image that they had never seen before contained an animal, Thorpe et al. [36] showed that visual evoked potentials recorded on correct target trials differed sharply from those recorded on correct distractor trials at about 150 ms after stimulus onset. This differential brain activity has been found at the same latency with non-biological relevant categories of objects such as "means of transport" and has been shown to be related to "visual decision making" rather than physical differences between photographs belonging to different categories [40]. This speed of processing could well be seen for any well-learned object-category [32]. In such categorization tasks, very different objects have to be grouped together (i.e. a snake and a flock of sheep) and performance cannot rely on the analysis of a single low-level cue or even on a single conjunction of low-level cues. When considering this very short delay together with the anatomy and physiology of the visual system, it was argued that such severe temporal time constraints imply that the underlying processing probably relies on feed-forward mechanisms during a first wave of visual information [35,36].

It thus seems that high-level search tasks such as looking for an animal in a natural scene might be performed as fast as the simplest pop-out search tasks. To explain speed of processing in visual search tasks, emphasis had been put on the target saliency, and on the number of diagnostic stimulus features [33]. However, increasing stimulus diagnosticity in the animal categorization task of natural images by using highly familiar photographs failed to induce a decrease of the minimal processing time: subjects could categorize novel images as fast as images on which they had been extensively trained [8].

Thus, the fast visual processing mode that underlies rapid-categorization cannot be speeded up when top-down pre-setting of the visual system is optimized with experience. However, it is a difficult experimental issue to determine the relative importance of bottom-up and top-down processes. To investigate further how top-down knowledge related to task requirements could influence the visual analysis of natural images, we tested human subjects in a task in which they were assigned a given photograph as target and had to detect this single target-photograph among a variety of different non-target stimuli. Being fully briefed about the target should allow subjects to maximize the use of top-down influences and to rely only on a limited number of low-level cues specific to the target-image.

In the present experiment, we studied the fast processing of natural images in human subjects performing in alternation the superordinate "animal / non-animal" categorization task and the single-photograph recognition task. Along with behavioral performance, analysis involved associated ERPs and localization of brain sources to investigate the neural dynamics of early information processing. Since both tasks used the same natural images as stimuli and required the same motor response, any processing differences should be related to task requirements.

2. METHODS

Stimuli

All stimuli used in the two tasks were photographs of natural scenes (Corel CD-ROM library). In each group, images were chosen to be as varied as possible (Figure 1). Subjects were tested on blocks of 100 stimuli including 50 % targets and 50 % distractors. In the categorization task 1000 photographs were used (50 % distractors and 50 % targets) and each of them was seen only once by each subject. The target-photographs included pictures of mammals, birds, fish, arthropods, and reptiles. There was no a priori information about the size, position or number of targets in the photograph. There was also a wide range of non-target images, with outdoor and indoor scenes, natural landscapes or city scenes, pictures of food, fruits, vegetables, trees and flowers....

In the recognition task, as in the categorization task, targets and non-targets were equiprobable in each block of 100 images so that the target-photograph assigned to a given block was seen 50 times among 50 varied non-target photographs that did not contain an animal. Each of the 14 subjects was tested with 15 targets (a total of 210 targets) and the same 750 non-target stimuli. In the 210 photographs used as targets, 140 (10 images per subject) contained an animal and were thus similar to the target photographs used in the categorization task. They had been categorized by human subjects in a previous study [8] and were known to offer different levels of difficulty. The remaining 70 (5 images per subject) did not contain any animal and were thus homogenous with the non-targets used in both tasks.

Task and protocol

Fourteen human subjects (7 women and 7 men, mean age 26 ranging from 22 to 46), with normal or corrected to normal vision volunteered for this study. Participants sat in a dimly lit room at 110 cm from a color computer screen piloted from a PC computer. They were required to start a block of 100 images by pressing a touch-sensitive button. A small fixation point (< .1° of visual angle) appeared in the middle of the black screen. Then, an 8-bit color vertical photograph (256 pixels wide by 384 pixels high which roughly correspond to 4.5° X 6.5° of visual angle) was flashed for 20 ms using a programmable graphic board (VSG 2.1, Cambridge Research Systems).

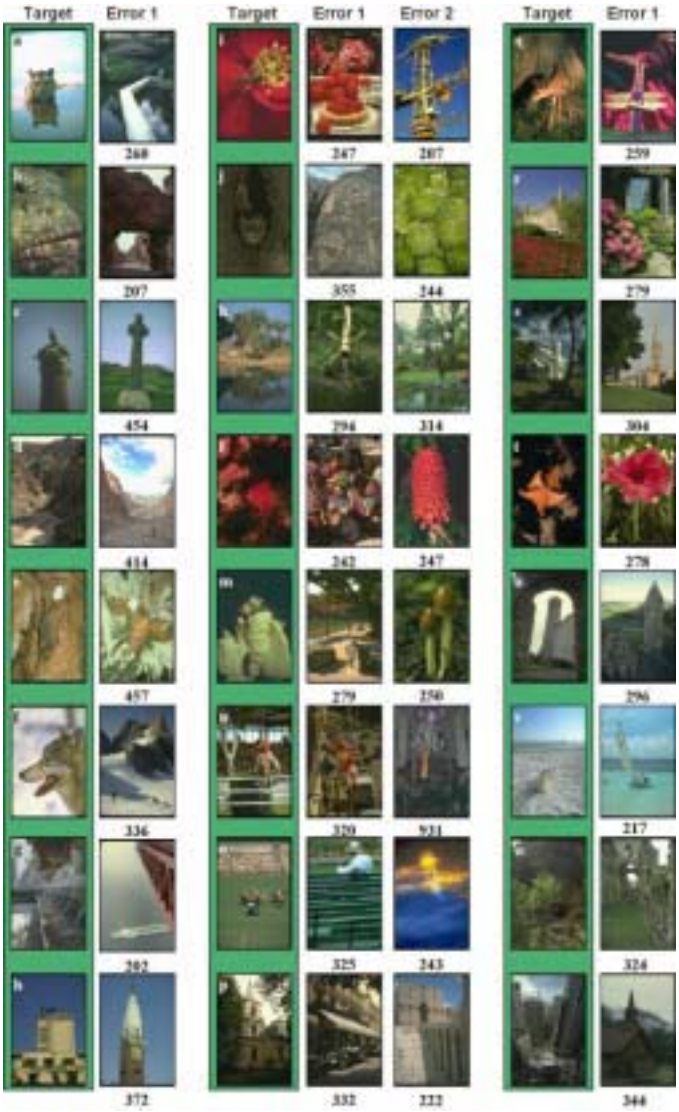


Figure 1. Targets and associated errors in the recognition task. Target-images used in the recognition task are illustrated on a green background. The figures show the high variety of the animal images used in the 10 testing blocks (images a, b, c, e, f, i, j, k, l, m, n, o, q, v) in which animals are sometimes hardly visible (e, i, j, v) and the non-animal images used in the 5 control blocks (images d, g, h, p, r, s, t, u, w, x). On the right of each target-image is shown the non-target photograph(s) that induced a false alarm. Errors can clearly be related to global orientation (a, c, d, g, h…), color (e, i, j, l…), color patches in specific locations (n, t…), object identity or semantics (p, s, x…), spatial layout of the scene (b, e, f, k, n, v….) or any combination. The figures below each error indicate the reaction time of the incorrect go response. Similar natural images were used in the categorization task.

The short presentation time prevented any exploratory eye movement. The stimulus onset asynchrony (i.e. time between the onset of one image and the onset of the next image in a series) was random between 1800 ms and 2200 ms.

Subjects had to give a go/no-go response: releasing the button as quickly and accurately as possible when they saw a target-image but keeping their finger(s) on the button on non-target trials. They were given a maximum of 1000 ms to respond, after which delay any response was considered as a no-go response.

On two different days, subjects were tested on 10 categorization blocks and 10 recognition blocks, alternating between the two tasks within a session while their associated EEG was recorded. In the animal categorization task, subjects had to respond whenever the picture contained an animal. In the target-image recognition task, a given animal image was assigned as the target for the following block of 100 images. The 5 image-recognition control blocks using images that did not contain an animal were inserted at regular intervals.

For the image-recognition task, each testing block was preceded by a learning phase during which the subject was presented with the target-photograph which was both repeatedly flashed for 20 ms (similar to the testing conditions) and presented for 1000 ms to allow ocular exploration (3*5 flashes intermixed with 2 long -1000 ms-presentations). Participants were instructed to carefully inspect and memorize the target-image in order to respond to it in the following sequence of images as fast and as precisely as possible. The testing block started immediately after the learning phase.

Evoked-Potential Recording and Analysis

Electric brain potentials were recorded from 32 electrodes mounted on an elastic cap (Electro-cap International Inc). Data acquisition was made at 1000 Hz using a SynAmps recording system (Neuroscan Inc.) coupled with a PC computer. The analog low-pass filter was set at 500 Hz and the default SynAmps analog 50-Hz notch filter was used. Impedances were kept below 5 kOhms. Potentials were recorded with respect to common reference Cz, then average re-referenced. Potentials on each trial were baseline corrected using the signal during the 100 ms that preceded the onset of the stimulus. Trials were checked for artifacts and discarded using a [-50; +50 μV] criterion over the interval [-100; +400 ms] at frontal electrodes for eye movements and a [-30; +30 μV] criterion on the period [-100; +100 ms] at parietal electrodes to discard alpha brain waves. Only correct trials were considered for ERP averages. The waveforms were low-pass filtered at 35 Hz for use in graphics. Inter-subject two-tailed statistical *t*-tests (13 degrees of freedom) were performed on unfiltered ERPs for each electrode to evaluate the latency at which target ERPs diverged from non-target ERPs. This differential activity onset was defined as the time from which 15 consecutive values were statistically different to compensate for multiple comparisons. We computed significance for all electrodes but focused on two groups: frontal electrodes (10-20 system nomenclature: Fz, FP1, FP2, F3, F4, F7, F8) and occipital electrodes (10-20 system nomenclature: O1 & O2 with the addition of Oz, I, O1', O2', PO9, PO10, PO9', PO10') where the differential activity reached the highest amplitude. The additional occipital electrodes have the following spherical coordinates (theta/phi): Oz = 92/-90, I = 115/-90, O1’ = -92/54, O2’ = 92/-54, PO9 = -115/54, PO10 = 115/-54, PO9’ = -115/72, PO10’ = 115/-72.

Source localization

The source analysis was performed using a 4-shell ellipsoidal model and using BESA (Brain Electrical Source Analysis, version 99). Because of temporal muscle contraction, the two most temporal electrodes were too noisy and were discarded from the analysis. All other electrodes were used to localize the equivalent dipoles. Grand-average waveforms were low-pass filtered at 35 Hz before analysis. Pairs of dipoles were placed in a central position,

given a spatial symmetry constraint, then fitted in location and orientation for a particular time window (simplex algorithm).

3. RESULTS

The aim of this study was to compare the visual processing of a natural image when the task requirements called for the representation of a high-level object category such as "animal" or when it could be performed using short-term memory of low-level cue(s). Behavior and ERPs were recorded and analyzed in all subjects.

Behavioral results: recognition vs. categorization

The analysis of behavioral performance included accuracy, speed of response and a study of the non-target images that incorrectly induced a go-response.

Accuracy. Although extremely good in both tasks (93.1 % correct in the categorization task; 98.7 % correct in the recognition task) accuracy was significantly better in the recognition task (two-tailed $\chi 2$: $df = 1$, $p < .0001$), an effect that was found to be significant at $p < .05$ for each individual subject. An accuracy bias was found in both tasks, but whereas this bias was in favor of correct no-go responses in the categorization task it was in favor of correct go responses in the recognition task. Thus, subjects were slightly better at ignoring distractors than responding to animal-targets in the categorization task (93.9 % vs. 92.4 %; two-tailed $\chi 2$: $df = 1$, $p < .0001$) whereas they were more accurate at detecting the target-image in the recognition task than at ignoring non-target images (99.7 % vs. 97.5 %; two-tailed $\chi 2$: $df = 1$, $p < .0001$). This result provides an argument for the use of different strategies in the 2 tasks that will be discussed later.

Reaction time (RT). As illustrated in Figure 2, reaction times were significantly faster for the recognition task (median RT: 337 ms) than for the categorization task (median RT: 400 ms; two-tailed Mann Whitney U test: $p < .0001$). For individual subjects this difference was always significant ($p < .01$).

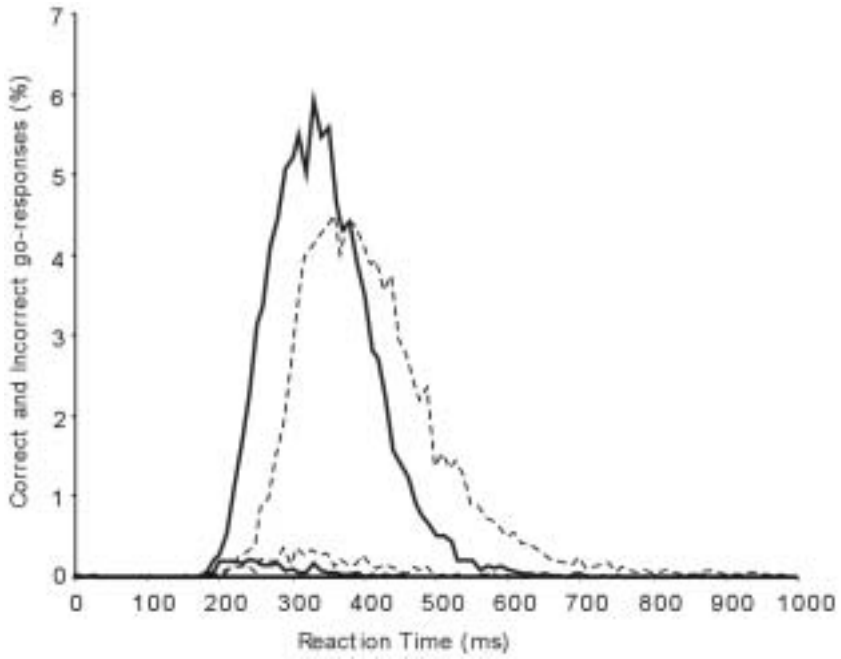


Figure 2. Overall reaction time distribution of go-responses in both the animal categorization task (black traces) and the recognition task (doted lines). The top two traces are for correct go responses towards targets, the bottom two traces are for false alarms induced by non-target stimuli.

Processing speed can be measured using median RT or mean RT, but these values do not reflect all aspects of processing speed. One very useful value is the minimal processing time needed to complete the tasks. The average slower speed in the categorization task could be due to some difficult photographs that need longer processing time [8]. Thus, although the average processing time could be shorter in the recognition task, the minimal processing time might be similar in both tasks. As in our experimental protocol targets and non-targets were equiprobable in both tasks, we defined the minimal processing time (Figure 2) as the first time bin for which correct hits to targets started to significantly outnumber false alarms to non-targets. Responses triggered with shorter latency but with no bias towards correct go-responses were presumably anticipations initiated before stimulus processing was completed. Using 10-ms time bins, this "minimal processing time" was found significant at 220 ms (two-tailed $\chi 2$: $df = 1$, $p < .0001$) in the recognition task and at 260 ms in the categorization task (two-tailed $\chi 2$: $df = 1$, $p = .0007$). The minimal processing time to reach decision was thus shortened by about 40 ms in the recognition task relatively to the categorization task. However, this shortening of RT latencies can be seen in Figure 2 as a shift of the entire RT distribution of the recognition task toward shorter latencies, from the earliest to the latest behavioral responses.

Control set. The results obtained in the recognition task with the control sets (that used non-animal target pictures) show again the better accuracy and the shorter processing time associated with tasks that only require image recognition (Figure 3). Subjects scored 98.3 % correct, with a median RT for correct go-responses at 348 ms. These scores are slightly below the performance level observed when the one-image target contained an animal (respectively 98.7 % and 337 ms), a result that could be due to higher similarities with the distractors, but the minimal processing time was found at exactly the same latency (220 ms) in both cases ($p < .0001$, $\chi 2$ test evaluated over every 10 ms time bin).

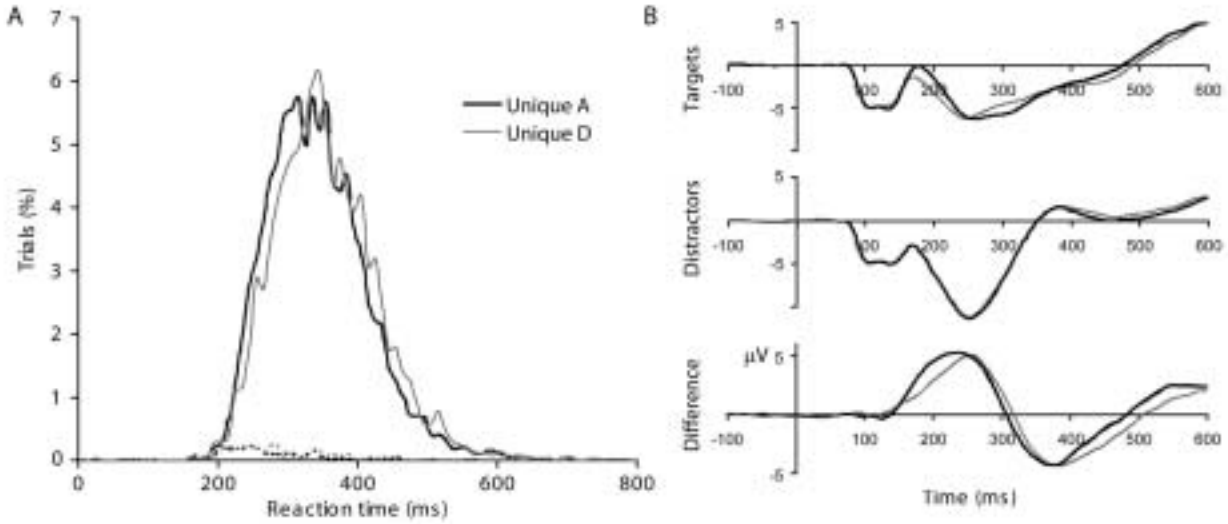


Figure 3. Overall results from the 14 subjects on the two different target-photograph sets in the recognition task. A, histogram of reaction time for the condition where pictures containing animals had to be recognized (Unique A) and for the condition where the target pictures did not contain animals (Unique Non-A). B, the differences between the frontal (FZ) ERPs recorded on correct targets trials (upper curves) and on non-targets trials (middle curves) are plotted (lower curves). In A and B: data are plotted in black for the animal set and in gray for the non-animal set.

Errors. A question that needs to be raised concerns the kind of errors that are produced in both tasks. In the categorization task, false alarms on distractors were slightly less common than target misses, and so far it has rarely been possible to objectively determine the reasons for these errors. In contrast, the errors produced in the recognition task were often seen with non-target images that share some obvious low-level properties with the memorized target image. These features (Figure 1) appear to be related to coarse orientation of objects, prevailing color, patches of color(s) in a given location, context or object identity, spatial layout or complexity of the scene… When performing the recognition task, subjects were thus relying on low-level visual cue(s) that could differ from one memorized target to another.

Event-related potentials

ERPs were considered separately for correct target and correct distractor trials (Figure 4). Using both individual data and grand average ERPs, the differential brain activity between the two types of trial was assessed in the two tasks by subtracting the average ERP on correct distractor trials from the average ERP on correct target trials. It is commonly assumed that the averaged electrical responses recorded from the scalp result from stimulus-evoked brain events and that the amplitude and latency of the various components of this evoked response reflect the most relevant features of the brain processing dynamics. Recently it has been shown [23] that these deflections might be generated by partial stimulus-induced phase resetting of multiple electroencephalographic processes. However, by using the difference between the two ERPs, no assumption is made about the relevance of the different ERP components, since the question that is addressed concerns only the differences in the cerebral processing of targets and distractors. The onset latency of this differential activity –that might correspond to the minimal visual processing time to differentiate a target from a distractor- was assessed using a two-tailed paired *t*-test performed for each 1 ms time bin and for each electrode (see Methods).

As reported in previous studies using this categorization task, a positive differential activity, was clearly seen on frontal electrodes [8,36]. On occipital sites, a mirror differential activity of inverse polarity was observed [10]. The results are illustrated on Figure 4 and show that ERPs to targets and non-targets superimposed very well until about 170 ms at which point they diverged abruptly (two-tailed paired t-test: $df = 13$, $p < .02$; occipital: 169 ms; frontal: 179 ms).

In the recognition task, the ERPs on correct target trials were computed separately for the two different sets of target-images (animal and control non-animal sets) and for their associated non-target images (Figure 3B). The grand average ERPs computed on all the non-targets superimposed perfectly (Figure 3B, middle traces) showing that there was no bias in the high variety of distractors used with the two different target sets. On the other hand, ERPs averaged separately on correct trials for the two target sets showed some differences (Figure 3B, upper traces). The onset latency of the differential ERPs (Figure 3B, lower traces) was found at 135 ms in the animal picture recognition task (two tailed paired t-test : $df = 13$, $p < .02$; occipital : 135 ms; frontal 148 ms), a latency virtually identical to the one found in the non-animal picture recognition task (two tailed paired t-test : $df = 13$, $p < .02$; occipital : 134 ms; frontal : 145 ms). Although the onsets were similar for these two sets of recognition targets, they diverged shortly after, the amplitude of the differential ERP increasing with a steeper slope with animal pictures targets. However, in the two sets of target-images, the computed differential activities reached similar amplitudes (on

FZ electrode, animal pictures: 5.5 µV; non-animal pictures: 5.1 µV); but, the peak amplitude was observed earlier with animal images (233 ms) than with the set of non-animal images (255 ms). These differences at the ERP level might reflect the higher diagnosticity of animal images among non-animal images compared to the recognition of non-animal images among similar pictures.

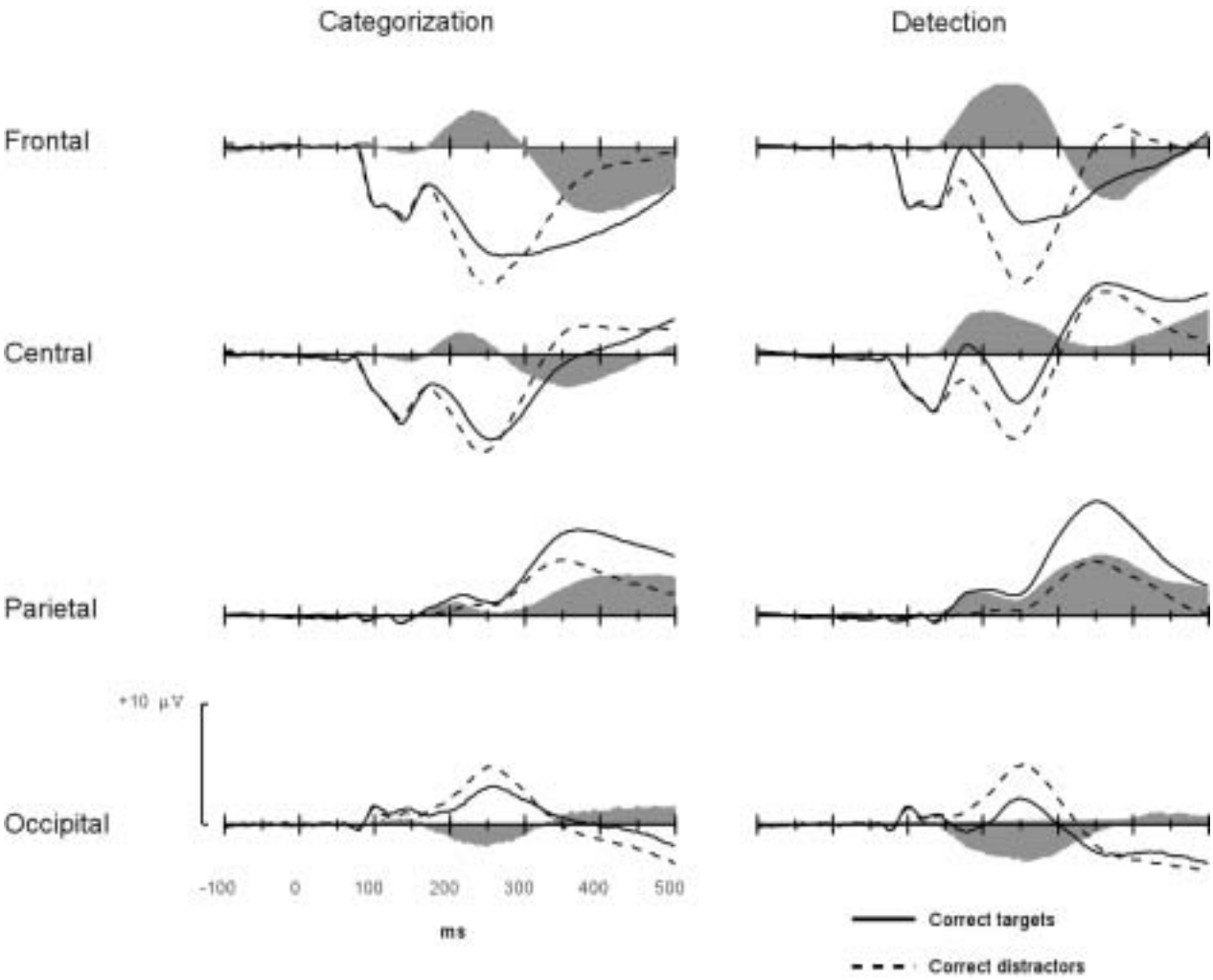


Figure 4. Grand average differential ERP activity. Average ERPs for all subjects in the categorization task (left column) and in the recognition task (right column) at different scalp locations: frontal, central, parietal and occipital sites corresponding respectively to the midline electrodes Fz, Cz, Pz and Oz. Average ERP on correct target trials (black line), average ERP on correct distractor trials (dashed lines), differential activity between correct target and distractor trials (shaded area). Note that the latency of the differential activity is always shorter in the recognition task.

Thus, in the picture recognition task, a clear differential activity was also observed at all sites but its onset was seen around 140 ms, much earlier than in the categorization task regardless of whether the images contained an animal or not. Consistent with this result, the difference between the two tasks also reached significance at about 140 ms (two-tailed paired t-test: $df = 13$, $p < .02$; occipital 141 ms; frontal 158 ms). Thus differential activity between target and non-target trials developed much earlier and reached a much higher amplitude in the recognition task than

in the categorization task (5.3 μV vs. 2.9 μV for electrode Fz). Moreover, the peak of amplitude was observed at similar latencies in both tasks when pictures contained an animal (animal categorization: 234 ms, image-recognition: 235 ms).

In both tasks the differential ERP between animal-target and non-target ERPs also showed an early small deflection that reached significance at about the same latency in the categorization task (two-tailed paired t-test: $df = 13$, $p < .02$; first occipital electrode: 98 ms; first frontal electrode 120 ms) and in the recognition task (two tailed paired t-test: $df = 13$, $p < .05$; occipital: 100ms; frontal: 112 ms). This small deflection does not appear with non-animal target images in the recognition task (Figure 3B, lower traces) and might thus be linked to statistical differences in physical properties of different subsets of images as documented recently [40].

Source localization and activation dynamics

For both tasks we used an ellipsoidal source model in the software BESA to analyze the dipole source localization of the differential ERP waveforms and the time course of their activities (Figure 5). Despite the strong constraints imposed on the model (large time window of 80 ms and only 2 dipoles that were required to be symmetrically positioned), residual variance was kept under 4 % for both tasks (residual variance: 3.9 % in the categorization task and 2.2 % in the recognition task), as already found in other studies using the categorization task [10]. Models using shorter and different time windows produced dipole localization that could not be distinguished from those illustrated in Figure 5. Thus, most of the difference between ERPs to target and non-targets can be explained by a single bilaterally activated brain area located ventrally and laterally in the occipital lobe, in a region that probably corresponds to extra-striate visual cortex. The localization and orientation of the dipoles were similar for the two tasks, the most obvious difference between the observed scalp signals being the time-course of the differential activity which started earlier in the recognition task.

In the recognition task, the two sets of images were analyzed separately and were found to be associated with non-distinguishable dipoles that accounted in both cases for about 98% of the differential ERP waveforms. The only difference was seen in the temporal dynamics of activation of both pairs of dipoles that were associated with a stronger activity increase from 150 ms onwards with the set of animal targets, reaching earlier its maximal amplitude.

4. DISCUSSION

The results of the present study show that the processing time of natural scenes by the human visual system depends on task instructions. When subjects are required to recognize a given target-image, they can rely on a variety of low-level cues, a hypothesis supported by the high similarity between the target and the non-target scenes that induced response errors. Consequently, the subjects were faster and more accurate in this natural scene recognition task than when they categorized the same type of natural images on the basis of the presence of an animal, a task that presumably requires access to more abstract representations. The results also provide some evidence that regardless of the visual analysis required in either task, the perceptual decision is made in the same brain structure and the visual information probably processed along the same visual pathway.

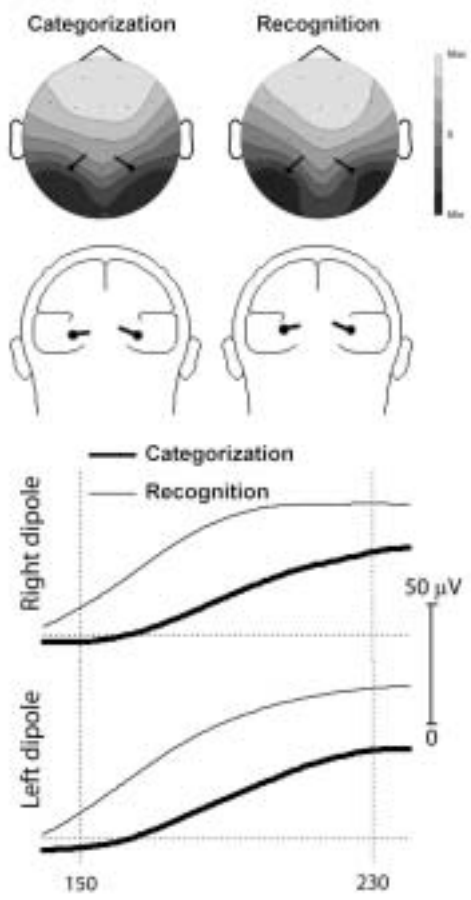


Figure 5. Cartography of the differential activity between the ERP waveforms of target and non-target data trials and localization of the electrical sources that accounted for this difference. For both tasks, the categorization task and the recognition task, a bilateral source accounted for more than 96 % of the differential ERP waveforms. Top: Gray-level scalp maps illustrate the averaged differential potential at 230 ms. Superimposed on these maps, the localization of the sources was virtually the same in both tasks. The location of the dipoles is also shown on frontal views. Bottom, the temporal dynamics of the left and right electrical source show that activation starts earlier and reaches a higher amplitude in the recognition task than in the categorization task.

The visual processing required for recognizing a given target-image is done in a delay that is about 30-40 ms shorter than the visual analysis required for detecting an animal in the same image. This delay is observed for both the latency at which the earliest behavioral responses are produced and the onset latency of the differential cerebral activity (used as an index of the perceptual decision). It increases to 60 ms when considering the median reaction time, reflecting the fact that the variation in response latencies is larger in the animal categorization task than in the image-recognition task (Figure 2) because of a larger difficulty range in the categorization task.

One could argue that the main difference between the two tasks is due to a novel vs. familiarity effect. Whereas the categorization task is exclusively performed with previously unseen images (trial unique presentations), the target-image recognition task involves the repetitive visual processing of a recently memorized photograph (i.e. "familiar") among non-target images that have never been seen before (i.e. "novel"). Indeed, it has been shown using event-related fMRI, that the activity of brain areas that are thought to be involved in scene categorization (extrastriate visual cortex, inferotemporal cortex and prefrontal cortex) is modulated by stimulus repetition in subjects performing a rapid classification of pictures [4]. However, in the "animal" categorization task, we have recently shown that extensive experience with a given set of natural scenes did not result in faster behavioral responses than with completely novel images nor reduce the latency of the differential ERPs [8]. In agreement with other ERP studies using words, faces and other visual stimuli [12,22,31,39], familiarity effects were not seen until about 300-360 ms post-stimulus and thus could not account for speeding up the visual processing in the recognition task used here.

Various interpretations could account for our results. As target-image recognition task relies on detection of low-level cues, one possibility is that the faster analysis could simply result from the by-pass of higher processing stages that would only be necessary to reach a decision in the superordinate "animal" categorization task, when access to abstract representations is specifically required. In the recognition task, the perceptual decision could be made in brain structures considered as lower in the hierarchy of visual processing but in which low-level features would be already fully analyzed and accessible. Decisions could be made in area V4 or even in the primary visual cortex V1 as suggested by Barbur et al. [2]. Alternatively we would like to argue that visual information is analyzed along the same brain pathway [16] but that the higher target predictability in the image-recognition task allows faster processing of the pertinent cues using top-down connections to preset neuronal assemblies at various levels of the visual pathway.

The main result supporting this alternative view is the location of the dipoles accounting for 96 % and more of the differential activity recorded in both tasks. Even though the 32-recording-site set-up and the ability of the BESA software to specify accurately the "absolute" location of the brain activity may be questioned, the fact that, regardless of the task, the dipoles were found at very similar positions and orientations in the brain appears difficult to explain if the underlying brain areas were not the same. In both tasks, the perceptual decision could therefore involve the same cerebral structures, most probably the occipito-temporal visual areas involved in object recognition. The location has been confirmed using the same categorization task with an event-related fMRI study [9], and found to be close to areas such as the fusiform gyrus involved in the recognition of various stimuli such as faces, objects or animals [5,14,20]. In correlation with the differential activity that develops 30-40 ms earlier in the target-image task, the main difference between our two tasks was found in the temporal dynamics and amplitude of the dipole activation (Figure 5) that developed earlier and reached higher amplitude in the image-recognition task.

In preceding studies using the animal categorization task we have already argued that the short latency at which the scalp differential activity starts to develop imposes such a high temporal constraint that the perceptual decision presumably relies essentially on feed-forward processing [8,35,36]. We postulated that information from the retina had to reach the primary visual cortex, area V1 (via the thalamus), and was subject to further processing in areas V2 and V4 before reaching the high-level brain areas involved in object recognition. These various processing steps are likely to be just as essential in the target-image recognition. Thus the most likely interpretation still relies on a faster visual processing of these images because of total target predictability.

In both tasks, speed of bottom-up processing would depend upon the tuning of neuronal populations along the visual pathways and thus on stimulus diagnosticity. Such bias has been shown for spatial frequencies [29], suggesting that a given scene might be flexibly encoded and perceived at the scale that optimizes information for the on-going task. Automatic target priming has been shown for color and spatial position in pop-out tasks [24,25] and has been attributed to temporary representations that could be updated on the basis of task demands. Saccade latency can be shortened by 30 ms and more, an effect linked to diagnosticity since it builds up with target color repetition [26]. In our tasks, we would expect top-down influences to bias bottom-up visual processing more heavily and more precisely in the recognition task than in the categorization task. The recognition of a target scene might be achieved using a carefully chosen low-level feature or a simple combination of characteristics (a blob of a given color or orientation for example). Compatibility would be maximal in this task because every target-image would activate all

preset neuronal populations. Moreover, as the specific location of this feature in the image is also known, focalized spatial attention could be allocated at the exact location of the screen where the cue is going to appear when the target is flashed; a view that is supported by our analysis of the images that induced false alarms. In contrast, in the categorization task, the subject needs to process evenly the whole natural scene: the location of the target-animal in the photograph is unknown and although many features (an eye, a paw, a tail, a beak, a wing…) are diagnostic of the presence of an animal, none of them is necessary to classify an image as a target. Thus the presetting of the visual system cannot be as highly specific as in the recognition task and could not rely on the same features. Indeed, whereas color appears as an important diagnostic feature in the image-recognition task, we have shown that the fast responses in the "animal" categorization task do not rely on color cues [6]. A strong modulation of color processing could be due to top-down influences from high-level predictions about color-specific features [19].

Among the brain structures that might heavily influence the visual pathway through descending connections depending on behavioral requirements is the prefrontal cortex [3,27]. In a categorization task, the firing of prefrontal neurons reflects category membership rather than simple processing of the physical characteristics of the stimuli [11]. In the target-image recognition task, the activity in the frontal cortex is probably very similar to that recorded in a delayed matching to sample task with elevated activity during delay periods [13,15]. Moreover, prefrontal neurons can also convey information about both the physical characteristics of a stimulus and its location [30], a combination of cues used in the target-image recognition task. Thus, in the target-image recognition task, prefrontal activity could very precisely modulate the neuronal activity along the visual pathway [17] to optimize, for each memorized target, the processing of the selected pertinent cues.

Whereas total predictability speeds up visual processing, we showed using a control set of target images that presetting does not have the same strength for all natural scenes. Scenes with animals were, on average, recognized faster than scenes without animals. Certainly some features might be more salient in animal photographs presented among non-animal photographs, whereas the control set of non-animal images presented among other non-animal pictures could lack this diagnostic advantage. Another possible explanation may lie in the performance, in alternation, of the animal categorization task and the image recognition. Subject might have difficulty in inhibiting totally the presetting of neuronal populations tuned to animal features.

Another point that needs stressing is the fact that, in our preceding studies, the onset of the differential activity was found at about 150 ms for the categorization whereas in the present study it was found about 20-30 ms later. Image size or presentation cannot account for this increased onset latency. On the other hand, this difference could be explained by the switching between two different tasks that required different presettings of the visual system as it has also been seen in another experimental protocol using two different interleaved tasks (manuscript in preparation). It might be that, had we used a blocked procedure in which subject would have completed all the testing series of one task before completing the second task we would have ended with even shorter differential activities.

Regardless of the task, we suggest that natural images are processed along the same visual circuit and that a perceptual decision is made in the same brain area but that the processing speed of bottom-up information is highly dependent upon the subject expectancy and the strength of top-down influences. However, we evaluated the temporal cost of the higher-level visual computations needed to perform the superordinate "animal" categorization task at

about 30-40 ms. This temporal cost appears low when considering the discrepancy in task requirements. The answer might be in the level of complexity of the most informative features for classification. Fast super-ordinate categorization might rely on diagnostic features of intermediate complexity [37], accessible with coarse visual information rather than on fully integrated high-level object representations.

Acknowledgments: This work was supported by the CNRS and fellowships from the French government. Experimental procedures with human subjects were authorized by the local ethical committee (CCPPRB No. 9614003).

References

[1] Anllo-Vento, L., Luck, S.J. and Hillyard, S.A., Spatio-temporal dynamics of attention to color: evidence from human electrophysiology, *Hum Brain Mapp*, 6 (1998) 216-38.
[2] Barbur, J.L., Wolf, J. and Lennie, P., Visual processing levels revealed by response latencies to changes in different visual attributes, *Proc R Soc Lond B Biol Sci*, 265 (1998) 2321-5.
[3] Barcelo, F., Suwazono, S. and Knight, R.T., Prefrontal modulation of visual processing in humans, *Nat Neurosci*, 3 (2000) 399-403.
[4] Buckner, R.L., Goodman, J., Burock, M., Rotte, M., Koutstaal, W., Schacter, D., Rosen, B. and Dale, A.M., Functional-anatomic correlates of object priming in humans revealed by rapid presentation event-related fMRI, *Neuron*, 20 (1998) 285-96.
[5] Chao, L.L., Haxby, J.V. and Martin, A., Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects, *Nat Neurosci*, 2 (1999) 913-9.
[6] Delorme, A., Richard, G. and Fabre-Thorpe, M., Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans, *Vision Res*, 40 (2000) 2187-200.
[7] Enns, J.T. and Rensink, R.A., Influence of scene-based properties on visual search, *Science*, 247 (1990) 721-3.
[8] Fabre-Thorpe, M., Delorme, A., Marlot, C. and Thorpe, S.J., A limit to the speed of processing in Ultra-Rapid Visual Categorization of novel natural scenes, *J Cog Neurosci*, 13 (2001) 171-180.
[9] Fize, D., Boulanouar, K., Chatel, Y., Ranjeva, J.P., Fabre-Thorpe, M. and Thorpe, S.J., Brain areas involved in rapid categorization of natural images: an event-related fMRI study, *Neuroimage*, 11 (2000) 634-43.
[10] Fize, D., Fabre-Thorpe, M., Richard, G., Doyon, B. and Thorpe, S.J., Rapid categorisation of foveal and extrafoveal natural images: Associated ERPs and effect of lateralisation, *(Submitted*).
[11] Freedman, D.J., Riesenhuber, M., Poggio, T. and Miller, E.K., Categorical representation of visual stimuli in the primate prefrontal cortex, *Science*, 291 (2001) 312-6.
[12] Friedman, D., Cognitive event-related potential components during continuous recognition memory for pictures, *Psychophysiol*, 27 (1990) 136-48.
[13] Fuster, J.M. and Alexander, G.E., Neuron activity related to short-term memory, *Science*, 173 (1971) 652-4.
[14] Gauthier, I., Skudlarski, P., Gore, J.C. and Anderson, A.W., Expertise for cars and birds recruits brain areas involved in face recognition, *Nat Neurosci*, 3 (2000) 191-7.
[15] Goldman-Rakic, P.S., Cellular basis of working memory, *Neuron*, 14 (1995) 477-85.
[16] Grill-Spector, K. and Kanwisher, N., Different recognition tasks activate a common set of object processing areas in the human brain, *Soc Neurosci Abstr* (2000) 686.6.
[17] Hasegawa, I. and Miyashita, Y., Categorizing the world: expert neurons look into key features, *Nat Neurosci*, 5 (2002) 90-1.
[18] Hillyard, S.A. and Anllo-Vento, L., Event-related brain potentials in the study of visual selective attention, *Proc Natl Acad Sci U S A*, 95 (1998) 781-7.
[19] Hopf, J.M., Vogel, E., Woodman, G., Heinze, H.J. and Luck, S.J., Localizing visual discrimination processes in time and space, *J Neurophysiol*, 88 (2002) 2088-95.
[20] Kanwisher, N., McDermott, J. and Chun, M.M., The fusiforme face area: a module in human extrastriate cortex specialized for face perception, *J Neurosci*, 17 (1997) 4302-11.
[21] Karayanidis, F. and Michie, P.T., Evidence of visual processing negativity with attention to orientation and color in central space, *Electroencephalogr Clin Neurophysiol*, 103 (1997) 282-97.
[22] Liu, T. and Cooper, L.A., The influence of task requirements on priming in object decision and matching, *Mem Cognit*, 29 (2001) 874-82.

[23] Makeig, S., Westerfield, M., Jung, T.P., Enghoff, S., Townsend, J., Courchesne, E. and Sejnowski, T.J., Dynamic brain sources of visual evoked responses, *Science*, 295 (2002) 690-4.
[24] Maljkovic, V. and Nakayama, K., Priming of pop-out: I. Role of features, *Mem Cognit*, 22 (1994) 657-72.
[25] Maljkovic, V. and Nakayama, K., Priming of pop-out: II. The role of position, *Percept Psychophys*, 58 (1996) 977-91.
[26] McPeek, R.M., Maljkovic, V. and Nakayama, K., Saccades require focal attention and are facilitated by a short-term memory system, *Vision Res*, 39 (1999) 1555-66.
[27] Miller, E.K. and Cohen, J.D., An integrative theory of prefrontal cortex function, *Annu Rev Neurosci*, 24 (2001) 167-202.
[28] Nakayama, K. and Silverman, G.H., Serial and parallel processing of visual feature conjunctions, *Nature*, 320 (1986) 264-5.
[29] Oliva, A. and Schyns, P.G., Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli, *Cognit Psychol*, 34 (1997) 72-107.
[30] Rainer, G., Asaad, W.F. and Miller, E.K., Memory fields of neurons in the primate prefrontal cortex, *Proc Natl Acad Sci U S A*, 95 (1998) 15008-13.
[31] Rugg, M.D., Soardi, M. and Doyle, M.C., Modulation of event-related potentials by the repetition of drawings of novel objects, *Brain Res Cogn Brain Res*, 3 (1995) 17-24.
[32] Schendan, H.E., Ganis, G. and Kutas, M., Neurophysiological evidence for visual perceptual categorization of words and faces within 150 ms, *Psychophysiol*, 35 (1998) 240-51.
[33] Schyns, P.G., Diagnostic recognition: task constraints, object information, and their interactions. In M.J. Tarr and H.H. Bülthoff (Eds.), *Object recognition in man, monkey, and machine*, Elsevier Science Publishers, Amsterdam, 1998, pp. 147-179.
[34] Sugita, Y., Electrophysiological correlates of visual search asymmetry in humans, *Neuroreport*, 6 (1995) 1693-6.
[35] Thorpe, S.J. and Fabre-Thorpe, M., Seeking categories in the brain, *Science*, 291 (2001) 260-3.
[36] Thorpe, S.J., Fize, D. and Marlot, C., Speed of processing in the human visual system, *Nature*, 381 (1996) 520-2.
[37] Ullman, S., Vidal-Naquet, M. and Sali, E., Visual features of intermediate complexity and their use in classification, *Nat Neurosci*, 5 (2002) 682-7.
[38] Valdes-Sosa, M., Bobes, M.A., Rodriguez, V. and Pinilla, T., Switching attention without shifting the spotlight object-based attentional modulation of brain potentials, *J Cogn Neurosci*, 10 (1998) 137-51.
[39] Van Petten, C. and Senkfor, A.J., Memory for words and novel visual patterns: repetition, recognition, and encoding effects in the event-related brain potential, *Psychophysiol*, 33 (1996) 491-506.
[40] VanRullen, R. and Thorpe, S.J., The time course of visual processing: from early perception to decision-making, *J Cogn Neurosci*, 13 (2001) 454-61.

Chapitre 1 : conclusion générale

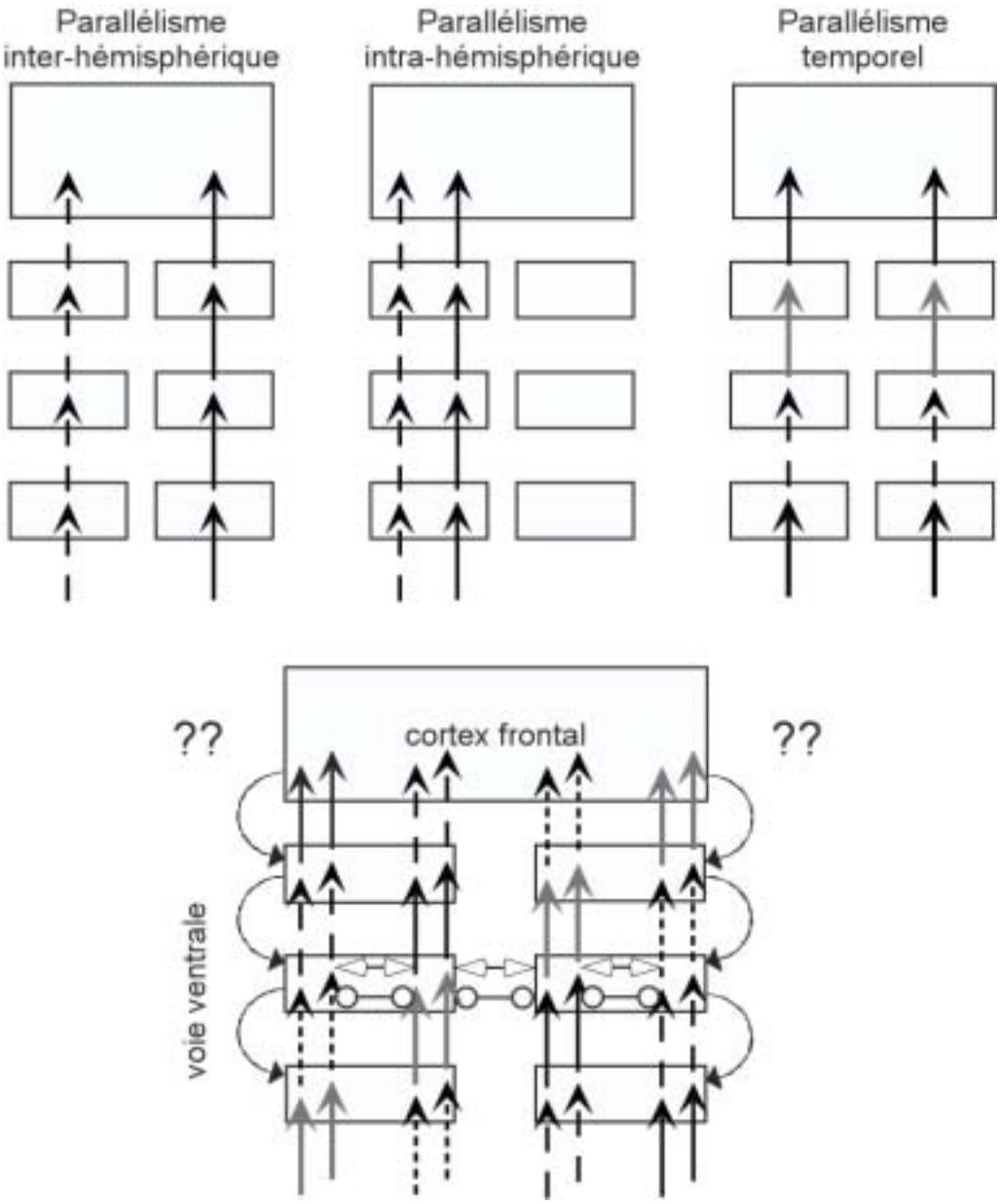
Le chapitre 1 constitue une synthèse du fonctionnement du système visuel mettant tout particulièrement l'accent sur le parallélisme des mécanismes de traitement des objets dans les scènes naturelles. De nombreux modèles proposent que dans les scènes naturelles, les objets soient traités de manière sérielle, l'un après l'autre. Cette conclusion s'appuie sur de nombreuses données en psychologie expérimentale et en neurosciences cognitives. Cependant, les recherches récentes dans ce domaine ont conduit au développement de modèles alternatifs dans lesquels les objets seraient traités en parallèle puis entreraient en compétition afin d'être sélectionnés au niveau comportemental. L'hypothèse défendue dans ce premier chapitre va plus loin : d'une part, les capacités de traitement en parallèle du système visuel seraient largement sous-estimées dans la plupart des modèles actuels ; d'autre part, la compétition entre les représentations d'objets apparaissant simultanément dans le champ visuel pourrait avoir lieu principalement dans des aires décisionnelles, par exemple de type frontal, et pas seulement dans le système visuel lui-même. Une telle hypothèse est soutenue par la première expérience de cette thèse, qui suggère un traitement en parallèle de deux scènes naturelles au niveau des aires visuelles, la compétition apparaissant plutôt au niveau frontal. Ce cas de parallélisme pourrait être particulier, chaque scène étant prise en charge par un hémisphère. Lorsque le traitement concerne des images présentées dans des quadrants, les performances diminuent et la mise en évidence d'un traitement en parallèle est moins claire, comme le montre l'article 2. Lorsque le système visuel doit faire face à 4 scènes naturelles différentes présentées de manière simultanée, la tâche devient particulièrement difficile. Mais dans cette situation, le système est vraiment poussé à l'extrême car les scènes sont le plus souvent non congruentes les unes par rapport aux autres, avec des descriptions sémantiques, des contextes et des prises de vue pouvant varier considérablement d'une image à l'autre. L'activation des populations neuronales dans la condition expérimentale présentant 4 scènes différentes n'a qu'une infime probabilité d'occurrence dans la vie courante. En revanche, le parallélisme pourrait être plus important avec des objets tous compatibles au sein d'un même contexte global. Une dissociation entre activités neuronales occipitales et frontales suggère à nouveau que la plus grande partie de la compétition pourrait avoir lieu dans ces dernières aires plutôt que dans les premières.

Le traitement visuel parallèle n'est pas limité aux objets, il comprend aussi le traitement du contexte, celui-ci nécessitant l'intégration en parallèle de nombreux éléments de la scène. Enfin, les représentations des objets et celles du contexte pourraient interagir en parallèle, l'identification du contexte contraignant l'interprétation des représentations d'objets. L'article 3 montre en effet que la catégorisation du contexte chez des sujets humains est à la fois précise et rapide, et ne met pas en jeu les indices de couleur de manière cruciale. Par contre il suggère aussi que les objets dans les scènes naturelles pourraient être traités plus rapidement que le contexte. Cependant, l'article 4 montre que le temps de traitement respectif des objets et du contexte serait fortement contraint par la diagnosticité des cibles, de telle sorte qu'en fonction de la tâche et de la nature des stimuli il pourrait y avoir une forte influence du contexte sur la catégorisation des objets. De plus, des données récentes suggèrent que si un premier traitement catégoriel d'un objet pourrait être fait rapidement, peut-être sans influence du contexte, le traitement plus fin de l'objet semble nécessiter plus de temps, il pourrait ainsi largement bénéficier d'un traitement en parallèle du contexte de présentation (Macé et al., 2002 ; Macé et al., en préparation).

Pour conclure, la revue de la littérature suggère fortement un traitement complexe et simultané de plusieurs objets dans les scènes naturelles. Il reste néanmoins à déterminer combien d'objets peuvent être traités simultanément et jusqu'à quel niveau d'intégration. Pour l'instant, différents types de parallélismes ont été proposés. Il pourrait ainsi y avoir un parallélisme inter-hémisphérique (Luck et al., 1994 ; Rousselet et al., 2002), un parallélisme intra-hémisphérique (Rousselet, Thorpe & Fabre-Thorpe, en préparation ; Chelazzi et al., 1998) et un parallélisme de type traitement en 'pipeline', des objets différents étant pris en charge par des niveaux hiérarchiques successifs de la voie ventrale (Keysers et al., 2001). Ces trois cas de parallélisme sont illustrés dans une figure de synthèse ci-dessous. A partir de ces différentes formes de parallélisme, on peut imaginer un mode de fonctionnement du système visuel dans des conditions plus réalistes, mais toujours en parallèle. Néanmoins, d'autres contraintes sont à prendre en compte, comme la compétition intra-hémisphérique, qui se met en place rapidement entre les représentations d'objets apparaissant dans le même quadrant visuel. Une compétition inter-hémisphérique pourrait aussi exister, même si cela ne semble pas être le cas dans la voie ventrale (Chelazzi et al., 1998 ; mais voir Murray et al., 2001). Dans un modèle parallèle réaliste, il faut aussi prendre en compte l'existence d'interactions feedback très rapides à tous les niveaux de traitement. Finalement, toutes les propriétés d'un objet ne sont pas analysées à la même vitesse.

Nous avons d'abord accès à une représentation grossière avant d'accéder à une représentation plus fine (Sugase et al., 1999). Ainsi, en fonction du degré de finesse requis par la tâche, le traitement en parallèle au niveau temporel pourrait être limité par la nécessité d'interactions prolongées entre plusieurs niveaux de traitement.

Compte tenu de ces contraintes supplémentaires, les 3 formes principales de parallélisme mises en avant ici pourraient interagir avec d'autres facteurs discutés précédemment : 1) l'influence du contexte ; 2) la nature de la tâche ; 3) les propriétés physiques des objets, notamment leur forme ; 4) des facteurs spatiaux, tout particulièrement le biais fovéal. Tous ces facteurs pourraient également interagir entre eux de manière plus ou moins linéaire. L'étude du parallélisme dans le système visuel est donc une aventure qui ne fait que commencer.



Parallélisme au sein du système visuel. Différentes études suggèrent l'existence d'un parallélisme inter-hémisphérique, intra-hémisphérique et temporel. Les flèches de différents niveaux de gris, types de pointillés et épaisseurs de traits symbolisent le traitement progressif de différents objets. L'empilement vertical de rectangles symbolise la progression hiérarchique au sein de la voie ventrale dans chaque hémisphère. Les voies ventrales de chacun des deux hémisphères convergent vers différentes zones du cortex frontal. Le modèle du bas représente une situation plus réaliste mais beaucoup plus difficile à tester dans laquelle chaque voie ventrale traite en parallèle différents objets à la fois dans l'espace et dans le temps. Les flèches descendantes représentent les influences rapides en retour d'une aire de plus haut niveau sur une aire de plus bas niveau. Les flèches horizontales représentent des interactions inhibitrices et excitatrices intra- et inter-hémisphériques. Enfin, le traitement d'un objet est symbolisé non par une seule flèche mais par 2 flèches décalées indiquant que certaines caractéristiques sont analysées plus rapidement que d'autres. Ce modèle ne capture bien sûr pas tous les aspects de la perception des scènes naturelles dans des conditions réalistes, mais donne un aperçu de la complexité du sujet.

Chapitre 2

Les visages ont-ils un statut particulier au sein du système visuel ?

La synthèse proposée au chapitre 1 fournit un certain nombre d'éléments clés permettant de mieux comprendre les fondements neuronaux de la perception rapide des scènes naturelles. Cependant, parmi les très nombreux objets présents dans les scènes naturelles, les visages d'êtres humains pourraient avoir un statut particulier. Cette hypothèse découle de deux ensembles de résultats semblant montrer (1) que les visages seraient pris en charge par une substrat neuronal anatomiquement distinct de celui prenant en charge les autres objets et (2) que le traitement des visages présenterait des caractéristiques fonctionnelles différentes de celui des autres objets. Ce statut particulier au sein du système visuel pourrait avoir pour conséquence un traitement des visages plus rapide que celui des autres objets. Etant donné l'importance des caractéristiques temporelles du signal dans le traitement des scènes naturelles, un traitement plus rapide des visages aurait des conséquences importantes pour les modèles de la perception visuelle.

Le présent chapitre vise à explorer les bases expérimentales sur lesquelles reposent la supposée spécificité des visages. Il ne s'agit donc pas ici de traiter de la perception et de la reconnaissance des visages en général ; ce sujet à été largement couvert dans Itier (2002), auquel le lecteur intéressé par plus détails peut se référer. Comme le précédent, ce chapitre n'a aucunement la prétention d'être exhaustif. Je n'entrerai donc pas dans les détails du débat actuel spécificité/expertise (e.g. Carmel & Bentin, 2002 ; Kanwisher, 2000 ; Tarr & Gauthier, 2000 ; Rossion & Gauthier, 2002), mais à partir des idées exposées au chapitre 1, je tenterai de proposer des explications alternatives à certains arguments de la littérature en faveur d'une spécificité des visages.

1. Y a t'il une ségrégation anatomique des mécanismes de traitement des visages ?

La spécificité des visages repose sur un certain nombre de résultats expérimentaux qui tendent à montrer qu'il existerait une zone corticale exclusivement dédiée au traitement des visages. Cette idée repose sur des arguments issus de la neuropsychologie, de l'imagerie fonctionnelle et de l'électro- et magnéto-encéphalographie.

1.1 Données neuropsychologiques

L'étude de patients présentant des lésions de la voie ventrale a révélé des cas exceptionnels. Certains patients sont atteints de prosopagnosie, c'est-à-dire qu'ils sont incapables de reconnaître des visages alors que leurs performances pour les autres objets seraient intactes (Farah et al., 1995). D'autres patients peuvent présenter le patron inverse, étant incapable de reconnaître les objets mais ayant des capacités de reconnaissance des visages inaltérés (Moscovitch et al., 1997). Cette double dissociation est le critère fondamental utilisé en neuropsychologie pour mettre en évidence des 'modules' de traitement, par exemple visuels. Bien que ce critère soit séduisant, il n'y a en réalité aucune raison de postuler l'existence d'un module ou d'une zone spécifique de traitement pour expliquer une double dissociation. Différents arguments expérimentaux et théoriques sont abordés point par point ci-dessous (voir aussi Dunn, 2003, Dunn & Kirsner, 2003 pour des arguments cliniques détaillés).

Les cas « purs » n'existent pas

Un argument majeur à l'encontre de l'idée selon laquelle il existerait un module de traitement des visages est l'absence de cas purs de prosopagnosie. La prosopagnosie est souvent associée à d'autres agnosies, même si elles peuvent être plus légères que le trouble de la perception des visages (voir références dans Moscovitch et al., 1997). Dans les cas prétendus purs, le problème qui se pose est le même que lorsqu'on essaye de déterminer la sélectivité d'un neurone : le résultat dépend toujours de la palette, forcément restreinte, des stimuli et des tâches utilisés. Il est très probable qu'aucun cas de prosopagnosie pure ne serait rapporté dans la littérature si les expérimentateurs avaient varier systématiquement les stimuli et les tâches qu'ils employaient pour tester des patients. Il a par exemple été montré que certains prosopagnosiques présentent un déficit plus large de discrimination fine entre stimuli autres que des visages. Certains patients sont aussi incapables d'apprendre à distinguer les individus de nouvelles catégories d'objets (Gauthier, Behrmann & Tarr, 1999 ; Laeng & Caviness, 2001).

Un autre problème souvent associé aux études de cas neuropsychologiques est le manque de puissance statistique. Lors de leur évaluation neuropsychologique, les patients réalisent généralement peu d'essais par test, ce qui augmente la probabilité de ne pas

trouver d'effet significatif lorsque la performance est proche du seuil de la chance. Or, beaucoup d'études rapportent des cas de prosopagnosie associés à des performances non significatives mais parfois bien supérieures à 50% dans des tests de reconnaissance de visages (Stone & Valentine, 2003). Ces cas supposés de prosopagnosie pourraient donc s'avérer être beaucoup plus partiels que ne le laissent croire les analyses statistiques.

Des modules plastiques

La neuropsychologie ne rapporte pas seulement l'existence d'un trouble de la perception des visages, mais l'existence de dizaines de troubles différents pour toutes sortes de catégories visuelles plus ou moins abstraites (Farah, 1990 ; Humphreys & Forde, 2003). Cette multitude de troubles pourrait être compatible avec l'existence d'une multitude de modules pour chaque catégorie d'objets, mais une explication alternative bien plus simple peut être proposée.

La voie ventrale est organisée comme une succession de réseaux compétitifs et auto associatifs. Ces réseaux, par leurs propriétés intrinsèques d'apprentissage, tendent à former naturellement des amas de neurones ayant des propriétés de réponse très semblables (Polk & Farah, 1995 ; Rolls & Deco, 2002 ; Tanaka, 2003). Des règles d'apprentissage synaptique relativement simples conduisent ainsi à la parcellisation catégorielle de la voie ventrale.

Cette parcellisation de la voie ventrale en amas de neurones codant pour des catégories différentes serait très proche d'une organisation modulaire si elle était stricte. Or ce n'est absolument pas le cas. Les neurones codant pour des propriétés semblables ont seulement tendance à être regroupés, comme nous l'avons vu au chapitre 1. Au sein d'une colonne corticale, les réponses neuronales restent assez hétérogènes, variant progressivement selon un continuum (Tanaka, 2003).

En opposition avec cet argument, certaines données suggèrent que la reconnaissance des visages pourrait s'effectuer par le biais de neurones « grand-mères », ces neurones dont l'activation permet de reconnaître une personne, comme par exemple le neurone Bill Clinton décrit par Kreiman et al. (2000, 2002). Par contre, d'autres ont rapporté des données qui montrent que les représentations des visages seraient relativement distribuées, constituant un avantage indéniable sur un système local puisque

les capacités mnésiques d'un système distribué varient de manière exponentielle avec le nombre de neurones disponibles (Rolls & Deco, 2002). Une autre conséquence d'un modèle distribué, associée au fait que la sélectivité des neurones n'est jamais stricte, est qu'un groupe de « neurones visages » peut potentiellement encoder un objet non visage dans la mesure où il partage suffisamment de propriétés communes de base avec les visages. De manière similaire, la réponse d'une petite population de neurones visages pourrait très bien permettre de discriminer entre plusieurs objets qui ne sont pas des visages (Riesenhuber & Poggio, 2002). Il a également été suggéré que la catégorisation visuelle pourrait dépendre de manière cruciale de neurones qui sont seulement grossièrement spécialisés (Thomas et al., 2001).

De manière générale, il est toujours très difficile d'établir avec certitude la spécificité des réponses fournies par une cellule. Par exemple, on ne sait pas qu'elle aurait été sa réponse pour l'ensemble, toujours très important, des stimuli non testés. De plus, une fois la sélectivité d'une cellule établie, il faudrait pouvoir définir dans quelle mesure la réponse de cette cellule participe aux décisions perceptives de l'animal lorsqu'il se trouve dans son environnement naturel. Sans ce lien, il reste difficile d'appréhender clairement comment l'espace des catégories visuelles est projeté sur la surface du cortex.

Des patients virtuels aux patients fantômes

Si l'organisation de la voie ventrale n'est pas modulaire au sens strict, il reste à expliquer pourquoi certaines lésions sont capables d'entraîner une prosopagnosie. La catégorie des visages d'êtres humains est très homogène et pourtant nous sommes capables d'effectuer des discriminations très fines entre différents visages et d'en reconnaître de très nombreux. Cette impressionnante capacité implique, d'un point de vue computationnel, que de nombreux neurones soient sensibles (de différentes manières) aux visages. Il n'est donc pas étonnant que de nombreux neurones répondent aux visages dans la voie ventrale. Nous avons également vu que les neurones répondant à des propriétés semblables tendent à être regroupés dans les mêmes colonnes corticales (Tanaka, 2003) non pas sous la pression de facteurs génétiques codant pour des modules, mais par le biais de contraintes dues aux mécanismes de plasticité synaptique (Rolls & Deco, 2002).

Une telle organisation rend donc très probable le fait qu'une lésion survenant au hasard dans la voie ventrale affecte de façon marquée les représentations liées aux visages, sans qu'il y ait pour autant besoin d'invoquer l'existence d'un module. Ceci est sans doute vrai pour n'importe quelle catégorie d'objets après un long apprentissage. Il semble en effet que dès qu'un singe apprend à faire des discriminations fines au sein d'une catégorie d'objets, on trouve ensuite de nombreux neurones codant pour ces objets et ayant des propriétés de réponses très voisines de celles observées dans le cas des « neurones visages » (Logothetis & Sheinberg, 1996).

Suivant cette logique, Cheng & Tarr (2003) ont proposé récemment un modèle simple de la voie ventrale simulant des mécanismes de catégorisation. Leur réseau de neurones est capable d'apprendre à discriminer toutes sortes de catégories à différents niveaux de classification et fournit des réponses semblables à celles obtenues chez les sujets humains. Après apprentissage de différentes catégories telles que des visages et autres objets, le réseau présente au niveau de sa couche de sortie des neurones sélectifs aux différentes catégories apprises. Des lésions effectuées au hasard dans ce modèle non modulaire de reconnaissance d'objets conduit à la formation de 'patients' simulés présentant différents troubles. Certains de ces 'patients virtuels' présentent une prosopagnosie alors que d'autres se comportent comme des agnosiques pour les objets (Cheng & Tarr, 2003 ; voir aussi Plaut, 1995 ; Polk & Farah, 1995). La majorité de ces 'patients virtuels' présentent cependant des déficits intermédiaires. Ceci suggère fortement que les dissociations rapportées entre la reconnaissance des visages et celle des objets dans la littérature neuropsychologique pourraient être attribuées à un biais d'échantillonnage dans lequel les cas intermédiaires sont très peu décrits (Schiller, 1997) ou bien encore parce que les patients compensent leurs déficits par des stratégies alternatives. Les patients 'fantômes' atteints de troubles intermédiaires pourraient bien représenter la grande majorité des cas.

1.2 Données IRMf et TEP

Dans une méta analyse de 17 études en TEP et en IRMf, Farah et Aguirre (1999) ont comparé la localisation de 84 foyers d'activation associés à des catégories comme des visages, des mots et autres objets. Le résultat est saisissant : contrairement aux

prédictions modulaires, ces foyers d'activation étaient distribués sur l'ensemble de la partie postérieure du cerveau correspondant au système visuel, sans qu'aucune organisation par catégorie n'apparaisse. Notamment, les visages ne présentaient pas plus de tendance à être regroupés dans une zone particulière que les autres catégories et aucun signe de latéralisation hémisphérique n'était apparent. Farah et Aguirre (1999) soulignent pourtant que la plupart des études en imagerie fonctionnelle ne permettant pas d'établir un lien causal entre l'activité d'une région et la reconnaissance visuelle, l'absence d'organisation claire pour les visages pourrait être due à l'activation d'aires qui ne participent pas de manière essentielle à la reconnaissance. Ceci pourrait expliquer le contraste entre l'existence de lésions focales chez des patients perturbant spécifiquement la reconnaissance des visages et une activité largement distribuée chez des sujets contrôles en imagerie fonctionnelle. Cette vaste distribution pourrait donc être un artefact. Cependant, on peut aussi appliquer le raisonnement inverse. La majorité des humains pourraient présenter un système unique de reconnaissance de toutes les catégories d'objets et les cas rapportés de prosopagnosie associées à des lésions focales correspondraient à des exceptions, des cas extrêmes sur un continuum dans une population présentant des degrés divers de ségrégation.

La grande diversité des foyers d'activation pourrait aussi être une conséquence du moyennage des données individuelles. Si chaque sujet présente une forte ségrégation mais que l'analyse porte seulement sur l'activité moyenne de la population de sujets, alors les différences inter individuelles, souvent très importantes, ont toutes les chances de masquer une possible organisation discrète de la voie ventrale. Ceci semble effectivement être le cas dans la mesure où des études ayant directement comparé la reconnaissance de visages par rapport à la reconnaissance d'autres objets séparément pour chaque sujet ont rapporté des foyers d'activation bien délimités pour les visages et d'autres catégories d'objets.

Dans ces études, la présentation de visages d'êtres humains se traduit par une activité plus importante au niveau du gyrus fusiforme (de manière prépondérante dans l'hémisphère droit) par rapport à de nombreux objets contrôles (Kanwisher et al., 1997 ; McCarthy et al., 1997 ; Puce et al., 1995). La logique appliquée dans de telles études consiste à soustraire le signal enregistré pour toute une gamme d'objets contrôles au

signal enregistré pendant la présentation de visages. La zone présentant une activité plus importante serait ainsi spécifique des visages. Certains parlent même de 'module' des visages, i.e. cette zone corticale serait de plus isolée fonctionnellement des neurones codant pour des objets autres que des visages. Cette logique est acceptée par une grande partie de la communauté scientifique et l'existence de la 'Fusiform Face Area' (FFA), la zone corticale fonctionnelle définie par la méthode soustractive (correspondant à une partie postérieure latérale du gyrus fusiforme), est souvent citée comme une preuve de l'existence d'un module des visages (Kanwisher, 2000).

Cependant, le raisonnement qui conduit à cette hypothèse modulaire a été remis en cause par une autre école de pensée. En effet, cette hypothèse postule qu'une réponse plus importante dans une aire corticale pour une catégorie que pour une autre constitue la preuve de la sélectivité de cette aire à cette catégorie. Pourtant, il est très difficile voire impossible de déterminer les opérations précises qui sont réalisées par les réseaux de neurones activés sur la base de données IRMf (Op de Beeck et al., 2001). Ainsi, une interprétation alternative des résultats d'IRMf montrant des foyers d'activations 'spécifiques' des visages serait que toutes les aires corticales visuelles de haut niveau contiennent des neurones avec les mêmes sélectivités mais avec une répartition différente de ces sélectivités d'un groupe de neurones à un autre (Op de Beeck et al., 2001). Ce débat est cependant loin d'être tranché.

Notons que le raisonnement modulaire a été étendu à d'autres catégories, comme par exemple les corps humains (Downing et al., 2001), la forme des objets (dans le 'Lateral Occipital Complex', ou LOC, voir revue dans Grill-Spector et al., 2001), les bâtiments (Aguirre et al., 1998), ou encore la structure spatiale des scènes visuelles qui activerait spécifiquement une région baptisée la PPA (pour 'Parahippocampal Place Area' ; Epstein & Kanwisher, 1998). On pourrait multiplier les exemples et obtenir, à partir des résultats de dizaines d'expériences ayant contrasté les activations pour différentes catégories d'objets, un vaste réseau de 'modules' se chevauchant plus ou moins.

Cependant, l'existence de la FFA a reçu des interprétations différentes. Il a notamment été montré qu'une activation de celle-ci est possible, bien qu'inférieure à celle obtenue pour des visages, lors de la présentation d'objets autres que des visages. En

revanche, la force de cette activité dépendrait de l'expertise du sujet avec la catégorie à discriminer (Gauthier, Tarr et al., 1999, Gauthier, Skudlarski et al., 2000). Elle pourrait refléter l'accès à des représentations fines, au niveau individuel (Gauthier, Tarr et al., 2000). La voie ventrale, plutôt que d'être parcellisée en termes de catégories, pourrait l'être en termes de types de mécanismes mis en jeu. Ainsi, si des mécanismes particuliers sont nécessaires pour la reconnaissance d'objets au niveau individuel (e.g. traitement configural) et que ceux-ci sont implémentés dans une zone particulière de la voie ventrale, il s'ensuit logiquement que les visages, plus que d'autres catégories, recruteraient cette zone, donnant ainsi l'impression d'un module spécifique de la catégorie visage (Gauthier, 2000).

Une telle approche peut cependant paraître paradoxale dans la mesure où elle entérine l'existence de la FFA, mais utilise les outils qui ont permis de la mettre en évidence pour démontrer que sa caractérisation fonctionnelle est inexacte. Il reste cependant à déterminer si cet effet de l'expertise est un effet général non spécifique, indépendant des mécanismes visuels mis en jeu, ou s'il correspond à la modification par l'expertise de mécanismes spécifiques, dont certains seraient spécifiques du traitement des visages. Les expériences de Gauthier et de ses collaborateurs sont très stimulantes, mais n'apportent en définitive aucun élément décisif permettant de démontrer que les différences entre les visages et les objets sont dues à une expertise poussée pour les visages ou à d'autres propriétés particulières aux visages. D'autres arguments plus convaincants en imagerie fonctionnelle permettent de remettre en cause l'existence d'un module des visages.

Notamment, les objets autres que des visages activent des zones corticales supposées être spécifiques des visages et ces derniers semblent activer diverses régions supposées être spécifiques d'autres catégories d'objets. Ces zones ne remplissent donc pas le critère d'encapsulation informationnelle qui définit en partie un module (Fodor, 1983). Ce point est particulièrement clair dans un article récent de Haxby et al. (2001) qui propose une stratégie différente pour aborder la question de la spécialisation fonctionnelle au sein de la voie ventrale. Ces chercheurs ont enregistré l'activité en IRMf associée à la présentation de visages et de plusieurs catégories d'objets (chats, maisons, bouteilles, ciseaux, chaussures et chaises). Pour chaque sujet, la soustraction de l'activité

moyenne associée à toutes les catégories de l'activité pour chacune de ces catégories révèle à chaque fois un réseau d'activation qui semble spécifique de chaque catégorie. Néanmoins, à la différence des travaux sur la FFA et la PPA par exemple, chaque catégorie n'est pas seulement caractérisée par un foyer de forte activation mais également par un réseau d'activations moins fortes. La représentation corticale de chaque catégorie semble ainsi largement distribuée, chevauchant en grande partie les représentations corticales des autres catégories testées. De plus, le foyer principal d'activation ne semble pas crucial pour la catégorisation d'un objet. En effet, Haxby et al. (2001) ont montré que même en ne prenant pas en compte les voxels les plus actifs pour une catégorie donnée, l'activité dans le réseau distribué des voxels moins actifs permet de prédire avec un taux supérieur à 90% quelle catégorie d'objets était soumise au sujet. Une analyse supplémentaire a également permis de montrer que l'activité des voxels qui sont maximalement activés pour une catégorie donnée permet d'effectuer une discrimination entre les autres catégories d'objets. Ces résultats s'appliquaient sans exception aux visages. L'existence de zones corticales spécifiques des visages pourraient donc bien être un artefact de la méthode soustractive qui ne prend en compte que les voxels les plus activés par une catégorie par rapport aux autres. Cette conclusion est en accord avec de nombreux enregistrements intracrâniens réalisés chez des patients épileptiques rapportant la présence de groupes de neurones répondant à des visages et à d'autres catégories d'objets en de très nombreux points du système visuel (Allison et al., 1994, 1999 ; McCarthy et al., 1999 ; Puce et al., 1999).

L'expérience d'Haxby et al. (2001) pose de manière plus fondamentale la question de la correspondance entre l'activité d'une population de neurones enregistrée en imagerie fonctionnelle et l'activité unitaire des neurones qui forment cette population. Si tous les neurones d'une population font exactement la même chose, alors l'activité de cette population nous renseigne sur les traitements réalisés dans cette zone du cerveau ; dans le cas contraire, l'activité de la population de neurones ne nous renseigne pas sur les traitements réalisés. Comme nous l'avons vu au Chapitre 1 et ci-dessus, même si les neurones ayant des propriétés semblables tendent à être regroupés, cette ségrégation est loin d'être absolue, des neurones voisins présentent des différences de sélectivité pouvant être substantielles et ceci à une échelle spatiale bien inférieure à la résolution atteinte

actuellement en IRMf. Il n'y a donc pas de relation directe entre l'activité d'une population de neurones et les propriétés individuelles des neurones qui composent cette population (Op de Beeck et al., 2001).

1.3 Données issues de l'électro- et de la magnéto-encéphalographie

Contrairement à l'imagerie cérébrale, l'électro- et la magnéto-encéphalographie fournissent une mesure plus directe de l'activité cérébrale. Elles pourraient ainsi être plus à même de révéler l'existence d'une ségrégation fonctionnelle du traitement des visages. Cependant, leur faible résolution spatiale empêche de tirer des conclusions définitives à ce sujet. Certains résultats ont pourtant été interprétés comme des preuves en faveur de la spécificité des visages.

Plusieurs composantes ont été décrites en potentiels évoqués qui semblent présenter une grande sélectivité pour les visages. La VPP, enregistrée en regard des électrodes fronto-centrales (Jeffreys, 1996) et la N170, enregistrée en regard des électrodes occipito-temporales (Bentin et al., 1996 ; George et al., 1996) sont les plus étudiées. Tout comme dans le cas des études en IRMf, une amplitude du signal plus importante en réponse aux visages par rapport à d'autres catégories d'objets sert de critère de spécificité (Bentin et al., 1996). Ce même critère a été appliqué aux enregistrements magnéto-encéphalographiques, révélant une M170, par analogie à la N170 (Linkenkaer-Hansen et al., 1998 ; Liu et al., 2000, 2002).

La signification d'un plus grand signal en réponse à des visages, qui a été particulièrement étudié au niveau de la N170, n'est pas plus évidente qu'elle ne l'est pour l'imagerie fonctionnelle. Selon l'hypothèse modulaire, la N170 est un marqueur neuronal d'un système de détection spécifique aux visages, fonctionnellement différent de la composante observée en réponse aux autres types d'objets (Carmel & Bentin, 2002). De manière alternative, l'hypothèse de l'expertise propose que plusieurs facteurs interviennent dans la différence d'amplitude entre les visages et les autres objets, notamment le niveau de catégorisation et le degré d'expertise visuel (Rossion, Curran & Gauthier, 2002 ; mais voir Bentin & Carmel, 2002). Des données présentées dans l'article 7 de cette thèse suggèrent aussi que la N170 pour des stimuli à l'endroit devrait recevoir une interprétation plus large (Rousselet, Macé & Fabre-Thorpe, sous presse). En soit, une

amplitude plus importante de la N170 pour les visages par rapport aux objets ne peut pas servir d'argument pour conclure à l'existence d'un module. Le même type de réflexion proposée plus haut en ce qui concerne l'imagerie fonctionnelle peut s'appliquer ici. Le point principal est qu'il est difficile d'établir l'origine de ces signaux et donc d'affirmer si oui ou non les mêmes populations de neurones participent à l'activité enregistrée au niveau du scalp. La même population de neurones répondant à différents objets pourrait fournir un signal plus important pour des visages parce que plus de neurones sont recrutés ou qu'ils déchargent plus fortement. De manière alternative, il est possible que la N170 pour les visages n'ait pas la même origine corticale que celle enregistrée pour les objets, certains ayant ainsi proposé de parler d'une N1 pour les objets, la N170 étant spécifique des visages (Bentin et al., 1996 ; Carmel & Bentin, 2002). Cette dernière idée est soutenue par la différence de topographie entre la N1 et la N170, la première étant beaucoup plus occipitale que la seconde, plus temporale et latérale. L'utilisation de cartes fonctionnelles d'activité dérivées de la théorie des champs de Lehman suggère également une différence qualitative entre la N1 des objets et la N170 des visages (Itier & Taylor, sous presse). Dans la mesure où ces questions ne sont pas résolues, il n'est pas possible de conclure quant à la spécificité des visages d'après le seul enregistrement de la N170. Pour la même raison, les effets de l'expertise sur la N170 ne peuvent pas servir de contre argument car il faudrait pour cela démontrer que les mêmes populations de neurones sont mises en jeu par les visages et les objets et spécifier plus précisément à quel niveau de traitement interviennent les effets d'expertise.

Un modèle intéressant suggère qu'une partie du signal de la N170 pour les visages pourrait provenir de régions beaucoup plus latérales du cortex temporal par rapport à la N1/N170 pour les objets (Bentin et al., 1996 ; McCarthy et al., 1999). La plus grande amplitude pour les visages par rapport aux objets pourrait provenir de ce générateur supplémentaire situé, semble t'il, dans la région du sulcus temporal supérieur (Itier & Taylor, sous presse). Cette région semble d'ailleurs particulièrement impliquée dans l'analyse des mouvements biologiques, de la direction du regard et de la discrimination des intentions d'autrui, basée sur des indices faciaux (voir revue dans Allison et al., 2000). De récentes analyses de sources convergent vers cette hypothèse (Shibata et al., 2002 ; Watanabe et al., 2003). Si cela s'avère être le cas, il pourrait être

vain de chercher à réconcilier les signaux enregistrés pour les visages et les objets. Cela expliquerait aussi pourquoi l'amplitude de la N170 est fortement corrélée avec la présence des yeux (Schyns et al., sous presse). Cependant, la véritable nature des traitements réalisés par les neurones activés dans cette zone et qui donnerait lieu à une partie du signal à l'origine de la N170 reste à déterminer. Il sera aussi particulièrement intéressant d'évaluer la spécificité des traitements sous-jacents. Ainsi, même si l'origine des signaux est en partie différente, toute interprétation sur les opérations neuronales effectuées dans ces divers réseaux restera spéculative sans l'apport de données complémentaires issues d'autres techniques d'investigations.

Malgré la difficulté méthodologique à déterminer la spécificité des mécanismes neuronaux sous-jacents à la N170, des études récentes suggèrent que cette onde correspond à un stade d'encodage structural pré-catégoriel, correspondant à la catégorisation d'un visage en tant que visage avant les processus de reconnaissance au niveau individuel (Bentin et al., 1996 ; Carmel & Bentin, 2002 ; Eimer, 2000a ; Sagiv & Bentin, 2001). Ce type d'interprétation est à resituer dans le cadre des modèles par description structurale (Biederman, 1987 ; Marr, 1982). Ceux-ci consistent typiquement en une phase d'extraction des différentes parties d'un objet puis à la formation d'une description 3D complète (encodage structural) avant l'appariement de cette description au catalogue des représentations d'objets stockées en mémoire, correspondant au stade de reconnaissance. Un modèle identique a été proposé pour la reconnaissance des visages (Bruce & Young, 1986).

Cette hypothèse structurale possède elle aussi ses faiblesses. Tout d'abord, attribuer à une onde une quelconque spécificité, que ce soit en termes de catégorie ou de traitement, n'est pas véritablement raisonnable. Les potentiels évoqués résultent d'une sommation grossière d'activités électriques cérébrales qui rend impossible une mise en correspondance aussi précise. Certains vont pourtant jusqu'à décrire l'existence de capteurs en MEG qui seraient spécifiques des visages (Liu et al., 2002). De plus, l'association entre encodage structural et N170 est rendu encore plus difficile par le manque de plausibilité biologique des modèles par description structurale (Rolls & Deco, 2002). Il n'y a donc aucun argument biologique robuste permettant de décrire la N170 comme le reflet d'un hypothétique stade d'encodage structural. Au contraire, les

représentations complexes pourraient se construire selon un processus interactif de structuration des formes visuelles mettant en jeu des représentations fines et grossières de manière simultanée (Bullier, 2001 ; Ullman, 1995).

Cependant, une piste récente provient de deux études de Tanaka et ses collaborateurs (Tanaka & Curran, 2001 ; Tanaka et al., 1999) qui montrent que l'amplitude de la N1/N170 enregistrée pour les objets augmente avec l'expertise visuelle des sujets et leur niveau de catégorisation. Ce résultat suggère qu'à la latence de la N170, certains mécanismes neuronaux permettant la discrimination des objets au niveau individuel pourraient déjà être à l'œuvre (Rossion & Gauthier, 2002), comme le suggèrent également les effets de familiarité et de répétition (Guillaume & Tiberghien, 2001 ; Itier & Taylor, 2002 ; Jemel et al., 2003). De manière alternative, ces effets pourraient aussi s'expliquer en partie par des modulations attentionnelles non spécifiques. Il n'est cependant pas possible à l'heure actuelle d'aller beaucoup plus loin dans l'interprétation de ces résultats.

2 Le traitement des visages est-il différent de celui des objets ?

La première partie de ce chapitre était consacrée à l'évaluation de certains arguments en faveur d'une ségrégation anatomique et fonctionnelle du traitement des visages. Malgré des résultats expérimentaux qui semblent parfois très convaincants, une analyse détaillée de la littérature, appuyée par des considérations théoriques, montre qu'il n'y a pas d'argument réellement convaincant en faveur de cette hypothèse.

Pourtant, à un autre niveau d'analyse, il apparaît que la reconnaissance des visages requiert des jugements intra catégoriels très fins. Ceci suggère qu'un système de traitement aux propriétés fonctionnelles uniques pourrait être spécifiquement nécessaire pour cette catégorie. Cette hypothèse est étayée par de nombreux résultats dans la littérature, le plus frappant étant l'effet d'inversion, mis en évidence tout d'abord au niveau comportemental et plus récemment au niveau de la N170.

2.1 L'effet d'inversion au niveau comportemental

L'exemple le plus frappant en faveur de l'hypothèse d'un système fonctionnel spécifique des visages est l'effet d'inversion. En effet, la présentation de visages à l'envers entraîne, par rapport à des visages à l'endroit, une diminution de la vitesse et de la précision des réponses dans une tâche de reconnaissance (Yin, 1969 ; Diamond & Carey, 1986), un effet dont l'ampleur est plus importante que pour les autres catégories d'objets. Il a été suggéré que l'inversion perturberait le traitement configural des visages, rendant difficile l'extraction des relations spatiales entre les traits tels que le nez, la bouche, les yeux, les contours du visages. Ces deux types d'information – traits et relations spatiales – seraient mis en jeu par les mécanismes d'identification des visages, la reconnaissance des objets dépendant quant à elle de l'analyse des traits (pour plus de détails lire Itier 2002).

Cette dernière conclusion peut paraître très surprenante quand on sait, comme nous l'avons vu au Chapitre 1, que les mécanismes de reconnaissance des objets implémentés dans la voie ventrale dépendent de manière cruciale du codage des relations spatiales entre les éléments qui composent un objet. La distinction entre les visages et les autres objets en ce qui concerne le traitement spatial n'est donc pas absolue, il est plus raisonnable d'envisager que le poids des facteurs spatiaux est plus important pour les visages par rapport aux objets. La nécessité d'effectuer des distinctions fines entre stimuli dont l'organisation spatiale est très proche, comme c'est le cas pour les visages, expliquerait cette différence. Il y aurait ainsi un traitement configural beaucoup plus marqué pour les visages que pour les objets discriminés de manière plus grossière. Une forte expertise pour les visages, biaisant la sélectivité spatiale de certaines populations neuronales, pourrait donc être à l'origine de l'effet d'inversion. Cette idée est soutenue par le fait qu'un effet d'inversion pour des objets autres que des visages est présent chez des sujets après un entraînement intensif à la réalisation de discriminations fines au niveau individuel entre les membres d'une même catégorie (Diamond & Carey, 1986 ; Gauthier & Tarr, 1997). Le seul effet d'inversion n'est donc pas suffisant pour permettre de conclure à la spécificité des visages. De manière plus générale, la littérature sur les effets d'inversion ignore la nature flexible et adaptative du système visuel. Les effets d'inversion, et plus généralement les effets de rotation dans l'espace, ne sont pas absolus,

leur amplitude dépend non seulement de l'expertise visuelle pour une certaine catégorie d'objets (Perrett et al., 1998 ; Rossion & Gauthier, 2002), mais aussi de la tâche demandée aux sujets et de la présence dans les stimuli des informations pertinentes permettant de réaliser cette tâche (e.g. Schyns, 1998).

L'existence d'un traitement des visages fondamentalement différent de celui des autres objets ne dépend pas seulement de l'existence de données en faveur de cette hypothèse, mais aussi du modèle du système visuel utilisé pour interpréter ces données. Dans le cadre d'un modèle simple et non-modulaire du traitement visuel, il est tout à fait possible d'expliquer un plus grand effet d'inversion pour les visages sans avoir recours à des mécanismes spécifiques. Pour cela, il est nécessaire de préciser un peu plus les mécanismes à l'œuvre dans la voie ventrale.

La perception visuelle est par essence catégorielle, elle correspond à la classification non optimale d'un vecteur d'activité rétinienne en entrée, permettant à l'animal de répondre de façon appropriée et donc viable à toute une gamme de stimuli présentant des caractéristiques communes. Plus précisément, la discrimination d'un objet semble dépendre d'une accumulation d'informations devant atteindre un certain seuil pour qu'une réponse soit déclenchée (Ditterich et al., 2003 ; Hanes & Schall, 1996 ; Perrett et al., 1998). Le taux d'accumulation d'indices en faveur de la présence de tel ou tel stimulus dans l'environnement dépendrait directement de la sélectivité des neurones codant pour le stimulus, et serait donc liée à l'expertise du sujet pour une catégorie d'objets. Des enregistrements unitaires chez le singe suggèrent que quelle que soit l'orientation d'un stimulus tel qu'un visage, des décharges ont lieu à la même latence à l'échelle d'une population de neurones de IT (Ashbridge et al., 2000 ; Perrett et al., 1998). Néanmoins, les réponses des neurones de IT sont en moyenne plus fortes et plus nombreuses pour les stimuli ayant une orientation à l'endroit. Cette sélectivité est acquise par apprentissage, reflétant les propriétés physiques des stimuli rencontrés dans la vie de tous les jours (Ashbridge et al., 2000 ; Karnath et al., 2000). Cependant, elle n'est pas non plus absolue, des neurones pouvant répondre à toutes sortes d'orientations autres que strictement à l'endroit. Ceci est dû à la variabilité des orientations de la plupart des objets de l'environnement se traduisant par une relative tolérance des neurones de IT qui présentent rarement une sélectivité très stricte. Ainsi, à l'échelle d'une population de

neurones codant par exemple des visages, le seuil de discrimination est atteint plus vite pour des visages à l'endroit qu'à l'envers tout simplement parce que cette population répond de manière beaucoup plus vigoureuse à des visages à l'endroit (Perrett et al., 1998). Un effet plus important de l'inversion pour les visages par rapport aux autres objets peut se comprendre si on considère les besoins du système. Effectuer des discriminations fines entre des visages implique le recrutement de très nombreux neurones rendus sensibles par apprentissage à des configurations très précises d'éléments. Parce ce que ces discriminations fines ont lieu principalement pendant que nous discutons avec quelqu'un, elles sont contraintes par le facteur gravitationnel (sauf pour le cas exceptionnel des astronautes !), de telle sorte que la population de neurones sensibles aux visages présente une sélectivité exacerbée à l'orientation. Un tel biais n'est pas présent pour les objets pour lesquels nous n'avons pas développé d'expertise pour deux raisons : (1) la non nécessité d'effectuer des discriminations fines n'a pas biaisé la sélectivité à l'orientation des neurones sensibles à ces catégories ; (2) la catégorisation plus grossière de ces objets n'implique pas nécessairement des descriptions complètes des stimuli, elle peut s'effectuer sur la base d'indices moins complexes présentant une plus faible sensibilité à l'orientation. Ce cadre de réflexion permet d'expliquer les effets d'inversion sans jamais avoir recours à des mécanismes spécifiques pour les visages. Il devrait logiquement s'appliquer à d'autres phénomènes comportementaux, comme les effets configuraux mis en évidence par Farah et al. (1995, 1998). Il sera intéressant dans le futur de voir comment des simulations neuronales du fonctionnement de la voie ventrale intégrant ces principes simples parviendront à expliquer des différences qui semblent parfois si importantes au niveau comportemental.

2.2 L'effet d'inversion au niveau électrophysiologique

L'effet d'inversion au niveau comportemental ne constitue pas un élément suffisant pour permettre de conclure à l'existence de mécanismes spécifiques des visages. L'inversion affecte également certains marqueurs neuronaux de la perception visuelle, l'effet le plus frappant se situant probablement au niveau de la N170.

En effet, la N170 est d'amplitude plus grande et/ou de latence plus tardive pour des visages à l'envers par rapport à des visages à l'endroit (Bentin et al., 1996 ; Eimer,

2000b ; Itier & Taylor, 2002 ; Rossion et al., 1999 ; 2000). Cet effet d'inversion n'est pas présent pour les autres objets dans certaines études (Bentin et al., 1996 ; Rebai et al., 2001 ; Rossion et al., 2000), mais a été mis en évidence dans d'autres expériences où il est toujours d'amplitude plus réduite que celui obtenu pour les visages (Eimer, 2000b ; Itier et al., 2003 ; Rossion et al., soumis), ce que confirment des données présentées dans l'article 7 de cette thèse (Rousselet, Macé & Fabre-Thorpe, soumis). Une telle dissociation pourrait indiquer que certains des mécanismes sous-jacents à la N170 seraient particulièrement sensibles aux visages. L'augmentation d'amplitude avec l'inversion pourrait être due à un recrutement additionnel de neurones dans la voie ventrale (Rossion et al., 1999) en accord avec une étude en IRMf montrant que les visages à l'envers recrutent une zone corticale plus importante que les visages à l'endroit (Haxby et al., 1999). En revanche, des études récentes suggèrent que l'effet d'inversion serait dû à l'activité des mêmes générateurs avec un décours temporel perturbé, sans faire intervenir de zones corticales supplémentaires (e.g. Watanabe et al., 2003). Personne n'a encore fourni d'explication convaincante de l'effet d'inversion au niveau de la N170. Il n'y a pas pour l'instant de raison de penser que cela implique nécessairement des mécanismes « spécifiques » aux visages. Si l'amplitude de la N170 peut être mise en relation avec un nombre de décharges neuronales (ce qui reste à prouver), on peut aussi imaginer que l'effet d'inversion serait dû à un retard de stabilisation d'un groupe de neurones particulièrement sensible aux visages. Comme les visages semblent être traités par défaut au niveau individuel, contrairement aux objets (Tanaka, 2001), on peut très bien imaginer qu'une portion du cortex de la voie ventrale soit plus sensible à ces codages fins au niveau individuel. Par défaut, ce réseau aurait la propriété de trouver un attracteur correspondant à un individu, et pas seulement à la catégorie visage. Un tel mécanisme d'identification étant sensible à l'orientation, un stimulus inversé activerait la même machinerie, mais le réseau ne parviendrait pas à se stabiliser aussi rapidement sur une solution (voir Rolls & Deco, 2002, pour comprendre le fonctionnement de tels réseaux). La mise en jeu de nombreuses connexions auto-associatives, dues à la moindre spécificité d'un visage à l'envers par rapport à un visage à l'endroit, serait à l'origine d'une bouffée supplémentaire d'activité de cette population de neurones à la recherche d'un attracteur, expliquant l'augmentation d'amplitude de la N170. Cette hypothèse est

pour l'instant étayée par un seul résultat expérimental montrant un petit effet d'inversion en réponse à des non visages pour lesquels les sujets ont développé une expertise (Rossion et al., 2002). De manière générale, cette hypothèse prédit que l'effet d'inversion devrait être proportionnel à la capacité du sujet à effectuer des discriminations fines entre les membres d'une catégorie. Un tel système devrait aussi être sensible à la nature de la tâche et à diverses manipulations attentionnelles, autant de facteurs dont l'influence sur l'amplitude de l'effet d'inversion n'a pas encore été systématiquement quantifiée.

3. Le traitement des visages est-il spécifiquement plus rapide que celui des objets ?

L'existence supposée d'un module des visages a plusieurs conséquences possibles pour les modèles du traitement visuel, l'une des plus importantes étant la forte possibilité que les visages soient traités plus rapidement que d'autres catégories d'objets. Nous allons maintenant évaluer les données en faveur d'une plus grande vitesse de traitement pour les visages. Celles-ci proviennent principalement des études de potentiels évoqués. Nous verrons aussi que la neuropsychologie de l'attention fournit des indices en faveur de cette idée.

3.1 Données électro- et magnéto-encéphalographiques

De nombreuses études ont rapporté une large gamme de temps de traitement dont certains seraient compatibles avec l'idée d'un avantage temporel pour les visages par rapport aux autres catégories d'objets.

Toute une série d'articles qui ont fait grand bruit ont rapporté des différences significatives à des latences inférieures à 100 ms entre des objets et des visages, entre différents types de visages ou encore entre des visages vus dans différentes conditions expérimentales (Braeutigam et al., 2001 ; Debruille et al., 1998 ; George et al., 1997 ; Mouchetant-Rostaing et al., 2000a, 2000b ; Pizzagalli et al., 1999 ; Seeck et al., 1997). Une telle rapidité remettrait profondément en cause les modèles actuels de la perception visuelle et constituerait une preuve quasi irréfutable de l'avantage computationnel des visages sur les autres objets. Cependant, tous les effets rapportés dans ces expériences peuvent s'expliquer par des différences physiques non contrôlées entre conditions

expérimentales ou par des mécanismes non spécifiques d'amorçage, ce que ne nient pas certaines de ces études. Ce sujet étant largement discuté par VanRullen (2000 ; VanRullen & Thorpe, 2001b), il ne sera pas détaillé ici. De manière générale, et ceci reste valable pour l'ensemble des travaux discutés ci-après, y compris les miens, ces études font l'amalgame entre significativité et discrimination. En effet, s'il y a par exemple une différence significative entre les signaux associés à deux catégories d'objets à une latence donnée, cela n'a pas du tout comme contrepartie directe que la discrimination entre ces deux catégories d'objets soit effectuée à cette latence. Cette considération méthodologique implique que les résultats montrant des traitements « hyper rapides » des visages soient considérés avec beaucoup de précaution. C'est une chose de montrer une différence significative entre deux courbes, c'en est une autre de découvrir la latence à laquelle le traitement correspondant à la tâche du sujet est effectué.

D'autres études, dont certaines en MEG, ont rapporté des latences qui sont plus compatibles avec les enregistrements unitaires chez le singe, vers 110-130 ms, c'est-à-dire à la latence de l'onde P1, M1 en MEG (Halgren et al., 2000 ; Itier & Taylor, 2002, 2003 ; Linkenkaer-Hansen et al., 1998 ; Schendan et al., 1998 ; Taylor et al., 2001). Quatre de ces études rapportent notamment des différences entre des visages présentés à l'endroit et des visages présentés à l'envers, un effet d'inversion apparaissant donc avant la N170 (Itier & Taylor, 2002, 2003 ; Linkenkaer-Hansen et al., 1998 ; Taylor et al., 2001). Le point intéressant ici est que les différences observées ne semblent pas dues à des différences physiques entre les stimuli étant donné que la présentation des visages dans les deux orientations est contrebalancée entre les sujets. Ceci n'élimine cependant pas la possibilité que ces effets soient dus à des différences physiques non contrôlées entre le groupe de visages à l'endroit et le groupe de visage à l'envers vus par chacun des sujets. Par contre, les études de Taylor et al. (2001) et de Linkenkaer-Hansen et al. (1998) ne présentaient pas ce biais. Néanmoins, la présentation des yeux (zone de fort contraste) dans la partie haute ou basse du champ visuel pourrait expliquer ces variations au niveau du scalp par une différence d'activité de certains des générateurs de la P1 présentant une organisation rétinotopique. On pourrait par contre argumenter que des différences physiques ne sont pas nécessairement de bas niveau, elles pourraient par exemple représenter l'encodage différentiel de propriétés de haut niveau par des

populations différentes de neurones de la voie ventrale. Ceci est compatible avec l'idée selon laquelle la discrimination de stimuli isolés commencerait avant 150 ms (Vogel & Luck, 2000) et avec des résultats d'analyses de sources montrant que la voie ventrale pourrait être entièrement parcourue en 100-120 ms (Di Russo et al., 2002 ; Foxe & Simpson, 2002 ; Martinez et al., 2001). L'hypothèse d'un accès à l'identité des visages dans la fenêtre temporelle de la N170 (Rossion & Gauthier, 2002) rend également plausible l'accès à un traitement grossier des visages (comme leur détection) en moins de 150 ms. Des enregistrements chez le singe ont également suggéré un traitement en deux étapes des visages dans le cortex inféro-temporal, présentant une première bouffée d'activité vers 90-120 ms correspondant à une catégorisation grossière (visage d'humain / visage de singe) suivie d'une activité plus spécifique vers 150-170 ms qui pourrait par exemple discriminer entre plusieurs identités ou émotions (Sugase et al., 1999). Il faut cependant rester prudent car aucune des études en PEV citées plus haut n'a apporté de preuve directe d'un traitement explicite des visages en moins de 150 ms. De plus, il n'y a aucun argument pour dire que ce type de traitement rapide serait spécifique des visages. La littérature chez le singe rapporte en effet des latences de décharge neuronale très précoces pour une vaste gamme d'objets autres que des visages. Le critère pour l'obtention systématique de telles décharges rapides pourrait être l'exposition prolongée de l'animal à la vue de ces objets. Donc, même si des expériences futures démontraient un avantage temporel pour les visages, il n'y aurait aucune raison de penser que cet avantage soit spécifique des visages *per se*.

D'autres études suggèrent un début de traitement plus lent des visages, dans une fenêtre temporelle de traitement de 150-180 ms qui correspond aux latences auxquelles la N170 est généralement enregistrée (Bentin et al., 1996 ; Carmel & Bentin, 2002 ; Eimer, 2000a ; Sagiv & Bentin, 2001). Ce résultat est soutenu par certaines études qui ont rapporté que la N170 n'est pas modulée par la familiarité des visages, remettant en cause l'idée d'un traitement de l'identité dans la fenêtre temporelle de la N170 (Bentin & Deouell, 2000 ; Eimer, 2000b, 2000c). Cette évaluation est également en accord avec les études montrant que des objets dans des scènes naturelles commencent à être catégorisés vers 150 ms (Thorpe et al., 1996 ; VanRullen & Thorpe, 2001b ; Fabre-Thorpe et al., 2001) voire plus de 170 ms (Johnson & Olshausen, 2003). Le temps nécessaire pour

catégoriser des visages dans les scènes naturelles n'a cependant pas encore été évalué. J'ai réalisé deux études pendant cette thèse qui portent directement sur ce sujet, elles sont décrites dans les articles 5, 6 à la suite de ce chapitre (Rousselet, Macé & Fabre-Thorpe, 2003; Rousselet, Macé, Thorpe & Fabre-Thorpe, en préparation). En l'état actuel des connaissances, les études réalisées précédemment sur la catégorisation des objets dans les scènes naturelles suggèrent donc indirectement qu'il n'y aurait pas d'avantage temporel dans le traitement des visages par rapport à d'autres catégories d'objets bien apprises comme des animaux et des moyens de transport.

Cependant, des résultats issus de la littérature sur la N170 sont utilisés de manière indirecte par certains chercheurs pour affirmer que les visages sont traités très rapidement. Notamment, le fait que la N170 ne soit pas modulée par la familiarité du visage (Bentin & Deouell, 2000 ; Eimer, 2000b, 2000c), ni par la tâche donnée aux sujets (Carmel & Bentin, 2002 ; Séverac-Cauquil et al., 2001) et soit engendrée par des dessins schématiques comme par des photographies de visage (Sagiv & Bentin, 2001) pourrait suggérer un traitement plus automatique des visages. Ces résultats sont controversés. L'absence de modulation de la N170 par la familiarité des visages est remise en cause (Guillaume & Tiberghien, 2001 ; Itier & Taylor, 2002 ; Jemel et al., 2003 ; Rossion & Gauthier, 2002). De plus, l'amplitude de la N170 peut être dans certains cas modulée par le degré d'attention allouée aux visages (Eimer, 2000b, 2000d), ainsi que par des influences descendantes/cognitives (Bentin & Golland, 2002 ; Bentin et al. 2002).

Mais la notion d'automaticité est employée de manière abusive dans la littérature sur les visages. Le traitement des visages est qualifié d'automatique par exemple lorsque l'attention ou la tâche ne modifie pas l'amplitude ou la latence de la N170. Cependant, il reste à définir ce que l'on entend par automaticité. En outre, des travaux supplémentaires devront préciser la nature des mécanismes de traitement des visages dont la N170 pourrait être la signature. La notion d'automaticité et son association avec l'idée d'un traitement rapide sont ancrés dans les esprits en grande partie suite aux travaux sur la recherche visuelle (Chapitre 1). Selon la nomenclature établie originellement par Treisman, certains stimuli 'pop-out' sont traités très rapidement par des mécanismes préattentifs et automatiques alors que d'autres stimuli sont traités plus lentement par des mécanismes attentionnels. Par un raisonnement circulaire,

pratiquement toutes les études qui rapportent des effets très précoces pour les visages (avant 100 ms), déclarent que leur traitement est automatique (donc rapide). La notion d'automaticité pose aussi problème d'un point de vue neurophysiologique. En effet, quand un patron de photons frappe la rétine, la cascade de décharges neuronales qui lui succède est irrépressible et tout aussi « automatique » quelle que soit la nature de cette stimulation.

Le fait que l'amplitude de la N170 soit modulée par l'emploi de tâches relativement difficiles (Eimer, 2000b, 2000d) suggère une autre explication. Il est en effet possible que par défaut les visages soient traités de manière plus approfondie que les autres objets. La N1/N170 associés à ces derniers est en effet facilement modulable par les demandes de la tâche et par l'expertise (Tanaka et al., 1999 ; Tanaka & Curran, 2001). Cette différence entre objets et visages pourrait trouver son origine dans la manière dont sont traités par défaut les visages, par un accès direct à l'identité, alors qu'une catégorisation relativement grossière des objets est suffisante dans notre vie quotidienne (Tanaka, 2001). Ce point de vue renforce l'idée selon laquelle certains mécanismes neuronaux - dont la N170 constitue en partie la signature - pourraient être intéressés par des aspects des visages comme leur identité.

Pour conclure, les données disponibles actuellement ne montrent pas d'avantage temporel pour le traitement des visages par rapport aux autres objets. Cependant cette question reste à approfondir. De manière plus générale, la diversité des latences de discrimination rapportées dans cette section pose la question de la fiabilité des données acquises en électro- et en magnéto-encéphalographie. S'il ne fait aucun doute que la résolution temporelle de ces techniques est très bonne, c'est peut-être l'utilisation de ces outils dans l'évaluation des contraintes temporelles qui pèsent sur le traitement visuel des objets qui doit être revue.

3.2 Données neuropsychologiques

Pour finir cette brève revue sur les arguments en faveur d'un traitement plus rapide des visages, je voudrais évoquer brièvement quelques résultats issus de travaux de neuropsychologie de l'attention. Dans le trouble d'extinction visuelle (voir Chapitre 1,

page 45), les visages semblent avoir un poids compétitif plus important que des mots ou des formes abstraites (Driver & Vuilleumier, 2001 ; Vuilleumier, 2000). Ceci se traduit par le fait que le taux d'extinction diminue considérablement lorsqu'un visage est présenté dans l'hémichamp controlatéral à la lésion par rapport à des objets contrôles. Le fait que les visages semblent attirer plus l'attention, ou avoir un poids compétitif plus important que d'autres catégories d'objets est souvent associé, en relation avec la littérature attentionnelle sur la tâche de recherche visuelle, à un traitement automatique, donc à une capacité de traitement plus rapide.

On peut cependant fournir une explication plus simple à ce phénomène dans le cadre du modèle parallèle de la compétition biaisée (Desimone & Duncan, 1995 ; Rolls & Deco, 2002). Dans la mesure où beaucoup plus de neurones sont certainement dédiés au traitement des visages dans la voie ventrale pour permettre leur discrimination fine, la présentation d'un visage dans l'hémichamp controlatéral à la lésion produit un signal dont l'importance est plus à même de contrebalancer le désavantage compétitif introduit par la lésion. D'autre part, les sets de stimuli contrôles utilisés dans les expériences de Vuilleumier sont relativement pauvres et mériteraient d'être enrichis d'objets de la vie de tous les jours pour que les comparaisons soient adéquates. Certains résultats sont néanmoins particulièrement intrigants. Par exemple, la valence émotionnelle des stimuli semble importante puisqu'un visage expressif est plus compétitif (plus souvent perçu) qu'un visage neutre, mais il en est de même d'une araignée par rapport à une forme abstraite. De plus, dans une des expériences, le taux d'extinction d'un visage normal était plus faible que celui d'un visage dont les traits internes étaient réarrangés, suggérant que l'effet est sensible à la configuration normale du visage. Cependant, le taux d'extinction pour le visage normal était supérieur à celui obtenu pour des mots ou des formes simples dans une autre expérience. Il est donc difficile ici de tirer des conclusions claires. Les expériences de Vuilleumier suggèrent bien l'existence d'un traitement de haut niveau dans la voie ventrale en l'absence de perception consciente (Chapitre 1), mais échouent à démontrer l'existence d'une spécificité dans le traitement des visages.

4. Conclusion

La revue de la littérature sur le traitement visuel des visages présentée dans ce chapitre était bien évidemment partielle. Son objectif était d'évaluer la pertinence des arguments expérimentaux en faveur d'un traitement qualitativement différent des visages par rapport aux autres objets. Une telle différence serait très importante au niveau théorique puisqu'elle obligerait à aménager une place à part pour les visages dans les modèles de la perception visuelle. Il apparaît que la supposée spécificité des visages ne repose pas sur des bases irréfutables. Si de réelles différences existent entre visages et objets elles semblent être de nature quantitative plutôt que qualitative. Une nouvelle perspective s'ouvre alors, celle qui consiste à comprendre la nature des modifications imposées aux populations de neurones de la voie ventrale par les contraintes perceptives de nos interactions sociales.

Pour finir, il est étonnant de constater qu'un ensemble de résultats ayant conduit certains chercheurs à supposer l'existence d'un module pour les visages soit également présent pour une autre catégorie comme les mots. En effet, la perception visuelle des mots est associée à (1) des troubles spécifiques en neuropsychologie (Farah, 1990) ; (2) des activités focales en imagerie fonctionnelle (Farah & Aguirre, 1999) ; (3) des potentiels corticaux importants (Allison et al., 1994) ; (4) une N170 de grande amplitude (Bentin et al., 1999) ; (5) un effet d'inversion sur la N170 (Rossion et al., sous presse) ; (6) la capacité à effectuer des discriminations très fines pour distinguer entre plusieurs exemplaires de cette catégorie... Il n'est pas pour autant nécessaire de déclarer l'existence d'un module des mots, ni de déclarer l'existence d'un module chaque fois qu'une catégorie d'objet satisfait à l'un de ces critères. En définitive, cette course aux modules et à la spécificité ne résout rien, elle ne fait que reporter d'un niveau d'explication à un autre des différences que nous n'arrivons pas à comprendre. L'alternative à cette démarche consiste, à partir de la réalité neuronale telle que nous la connaissons, à proposer des modèles simples de fonctionnement du système visuel qui pourront toujours être complexifiés au besoin.

Article 5

Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes.

Rousselet, G.A., Macé, M.J.-M. & Fabre-Thorpe M.

Journal of Vision 3, 440-455, 2003.

Résultats comportementaux de 48 sujets adultes dans deux expériences visant à tester et à comparer la vitesse de catégorisation de visages d'êtres humains, d'animaux et de visages d'animaux dans des scènes naturelles présentées à l'endroit et à l'envers.

Seules les 2 premières pages de l'article ont été insérées dans cette thèse. La version complète est disponible gratuitement en format pdf à l'adresse http://journalofvision.org/3/6/5/.

L'article est suivi de la reproduction d'un poster illustrant une partie de ce travail présenté à la European Conference on Visual Perception (ECVP) en 2002 à Glasgow, U.K.

Introduction

Le chapitre 2 a présenté deux points de vue contradictoires à propos du traitement des visages. Le premier considère que les visages d'êtres humains sont une catégorie à part qui serait prise en charge par des mécanismes spécifiques, voire un 'module'. Par contre, le second propose l'existence de mécanismes s'appliquant à toutes les catégories d'objets. La réponse des populations neuronales sous-jacentes présenterait cependant des biais différents en fonction des contraintes perceptives induites par chaque catégorie. Dans les deux cas, les visages d'humains constituent une catégorie spéciale, la différence étant que la première interprétation impose d'aménager aux visages une place à part dans les modèles du système visuel.

Certaines études électrophysiologiques décrites dans le chapitre 2 suggèrent notamment que les visages pourraient être analysés très rapidement, parfois en moins de 100 ms. Pourtant, des expériences réalisées précédemment dans l'équipe ont montré que la catégorisation d'objets autres que des visages (animaux, aliments, moyens de transport) dans les scènes naturelles est très rapide et précise (Delorme et al., 2000 ; Thorpe et al., 1996 ; VanRullen & Thorpe, 2001a, 2001b)

et semble dépendre de mécanismes particulièrement optimisés (Fabre-Thorpe et al., 2001). La comparaison avec les travaux réalisés sur les visages humains est cependant difficile dans la mesure où la grande majorité des études dans le domaine a mis en œuvre des visages apparaissant presque toujours en position centrale, à la même taille et sur un fond uniforme.

Le but des deux expériences dont les résultats sont exposés dans les articles 5, 6 et 7 était d'évaluer la vitesse de traitement des visages d'humains dans des scènes naturelles très variées. La tâche des sujets était dans tous les cas d'effectuer une catégorisation go/no-go d'objets cibles dans des scènes naturelles apparaissant brièvement à l'écran. Ce type de tâche n'avait jamais été mis en œuvre auparavant pour tester le traitement des visages. La première expérience mettait en jeu des visages à des échelles de taille très différentes, les photographies allant du gros plan jusqu'à des plans plus éloignés montrant une ou plusieurs personnes en entier. Les visages n'étaient pas systématiquement centrés et pouvaient apparaître en différents endroits dans les photographies. Cette catégorie était comparée à celle des animaux, eux aussi apparaissant avec des tailles et des positions différentes dans les images. La catégorie des animaux a déjà fait l'objet de nombreuses études au préalable, notamment décrites au chapitre 1. La seconde expérience mettait en jeu uniquement des gros plans de visages humains et des gros plans de visages d'animaux, afin de mieux contrôler l'homogénéité des stimuli.

Finalement, la littérature rapporte que les visages d'humains sont plus difficiles à traiter à l'envers qu'à l'endroit, alors que les autres objets ne souffriraient que peu de ce biais. Ceci pourrait refléter la mise en jeu de mécanismes différents pour les deux catégories. Ce phénomène est encore en grande partie incompris et n'a jamais été testé dans le cadre des scènes naturelles. Afin de mieux caractériser cet effet d'inversion dans chacune des expériences, la moitié des images étaient présentées à l'endroit et l'autre moitié à l'envers. L'article 5 décrit les données comportementales associées à ces deux expériences (Rousselet, Macé & Fabre-Thorpe, 2003). L'analyse des temps de réaction permet en effet une première évaluation de l'efficacité et du temps de traitement des différentes catégories testées. L'analyse des données électrophysiologiques permet une évaluation plus directe de ces facteurs. L'article 6 est consacré exclusivement à l'étude des activités différentielles enregistrées à la surface du scalp pendant ces expériences (Rousselet, Macé, Thorpe & Fabre-Thorpe, en préparation) ; finalement l'article 7 porte sur l'étude de la N170 enregistrée dans la deuxième expérience (Rousselet, Macé & Fabre-

Thorpe, sous presse) ; les résultats sur la N170 dans la première expérience font l'objet d'analyses complémentaires et seront donc décrites dans un article ultérieur.

Résultats

Les résultats s'articulent en trois points majeurs.

1) Les sujets humains sont capables de catégoriser très rapidement et très efficacement toutes les catégories de stimuli testés, que les images apparaissent à l'endroit ou à l'envers.
2) Les visages d'êtres humains à différentes échelles ainsi que les gros plans ne sont pas catégorisés plus vite ou plus précisément que des animaux ou des visages d'animaux.
3) Un effet d'inversion était présent pour les 4 catégories de stimuli testés. Il était cependant plus fort pour les stimuli humains.

Discussion

La très bonne performance des sujets confirment les résultats obtenus antérieurement dans l'équipe, à savoir que la catégorisation des objets dans les scènes naturelles est d'une redoutable efficacité et repose sans doute sur des mécanismes essentiellement parallèles et vers l'avant de traitement de l'information.

Les résultats des deux expériences montrent aussi que les visages d'humains ne semblent pas bénéficier d'un avantage computationnel qui leur permettrait d'être catégorisés plus rapidement ou plus efficacement que d'autres catégories d'objets telles que des animaux et des visages d'animaux. Par extension, la catégorisation rapide des visages d'êtres humains est également comparable à celle de la catégorie non biologique des moyens de transport, VanRullen & Thorpe (2001a) ayant montré un traitement similaire entre ceux-ci et les animaux.

L'effet d'inversion pour les animaux montre que celui-ci n'est pas spécifique des visages humains, en accord avec la littérature comportementale. Ce qui semble spécifique de ces derniers est l'amplitude de l'effet d'inversion. Pourtant, comme cela a été suggéré dans le chapitre 2 et dans l'article 5, une telle différence n'implique pas nécessairement la mise en jeu de mécanismes particuliers. Une autre explication a été proposée dans le cadre d'un modèle simple de fonctionnement de la voie ventrale dans lequel les réponses de populations de neurones sont biaisées par les contraintes imposées par nos interactions avec les visages d'humains par rapport à d'autres catégories d'objets. Les représentations dans la voie ventrale sont très flexibles. Les

neurones développent des capacités de codage à long terme qui prennent en compte les contraintes de l'environnement dans lequel l'individu évolue, lui permettant d'effectuer les tâches visuelles dont il a besoin. Ceci implique que les résultats obtenus ici dépendent sans aucun doute de la tâche demandée aux sujets. L'emploi d'une tâche plus spécifique pourrait avoir révélé des différences plus importantes entre stimuli humains et animaux. Pour autant, le modèle simple proposé pourrait aussi s'appliquer à ces situations. Cela reste à tester.

Finalement, le faible effet d'inversion observé pour tous les stimuli testés dans ces deux expériences suggère que les représentations mises en œuvres ne seraient pas d'une complexité très élevée, comme cela a déjà été discuté lors de la présentation des articles 3 et 4. Je voudrais ajouter ici quelques nouveaux arguments en faveur de cette idée. Tout d'abord, la catégorisation rapide des animaux dans les scènes naturelles peut être réalisée à des contrastes tellement bas que seul le système magnocellulaire peut encore être activé. Ce système est impliqué notamment dans le traitement visuel des basses fréquences spatiales (Macé, Fabre-Thorpe & Thorpe, 2002). D'autre part, des expériences réalisées par Bacon-Macé montrent que la catégorisation des animaux est toujours possible dans des conditions de fort masquage rétrograde où la perception consciente des stimuli est à la fois limitée et fugace (Bacon-Macé et al., soumis ; Thorpe et al., 2002). Qu'il s'agisse des expériences de masquage ou de contraste, il est bien souvent possible de répondre sur un « blob » ayant la forme d'un animal sans avoir besoin d'une analyse détaillée des parties ou de la texture par exemple. Cet argument est pour l'instant essentiellement subjectif. Une expérience prochaine devrait tester l'hypothèse du « blob » en utilisant des scènes naturelles filtrées à différentes fréquences spatiales (projet de Macé & Fabre-Thorpe). Cette expérience pourrait révéler que la catégorisation des animaux peut s'effectuer sur la base d'éléments relativement grossiers. Il pourrait en être de même de la catégorisation des stimuli humains. Dans ce cas, il sera intéressant de déterminer quelles sont les caractéristiques les plus importantes qui permettent de faire la distinction entre ces catégories. L'idée selon laquelle des éléments relativement simples pourraient permettre de réaliser des tâches en apparence complexes peut paraître surprenante, elle nous montre en tout cas que les mécanismes de la catégorisation visuelle sont loin d'être totalement élucidés.

Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes

Guillaume A. Rousselet Centre de Recherche Cerveau et Cognition, CNRS-UPS UMR 5549, Toulouse, France

Marc J.-M. Macé Centre de Recherche Cerveau et Cognition, CNRS-UPS UMR 5549, Toulouse, France

Michèle Fabre-Thorpe Centre de Recherche Cerveau et Cognition, CNRS-UPS UMR 5549, Toulouse, France

Object categorization can be extremely fast. But among all objects, human faces might hold a special status that could depend on a specialized module. Visual processing could thus be faster for faces than for any other kind of object. Moreover, because face processing might rely on facial configuration, it could be more disrupted by stimulus inversion. Here we report two experiments that compared the rapid categorization of human faces and animals or animal faces in the context of upright and inverted natural scenes. In Experiment 1, the natural scenes contained human faces and animals in a full range of scales from close-up to far views. In Experiment 2, targets were restricted to close-ups of human faces and animal faces. Both experiments revealed the remarkable object processing efficiency of our visual system and further showed (1) virtually no advantage for faces over animals; (2) very little performance impairment with inversion; and (3) greater sensitivity of faces to inversion. These results are interpreted within the framework of a unique system for object processing in the ventral pathway. In this system, evidence would accumulate very quickly and efficiently to categorize visual objects, without involving a face module or a mental rotation mechanism. It is further suggested that rapid object categorization in natural scenes might not rely on high-level features but rather on features of intermediate complexity.

Keywords: rapid visual categorization, human performance, natural scenes, human faces, animals and animal faces, inversion effect, mental rotation, configural processing

Introduction

Recent biologically plausible models of object visual processing have emphasized that much of the computation underlying scene categorization might rely on essentially parallel feed-forward mechanisms (Riesenhuber & Poggio, 2000; Thorpe & Imbert, 1989; VanRullen, Gautrais, Delorme, & Thorpe, 1998; Wallis & Rolls, 1997). These suggestions are supported by the finding that in humans, a differential brain activity develops between target and distractor trials from 150 ms in various categorization tasks using natural images (Thorpe, Fize, & Marlot, 1996; Rousselet, Fabre-Thorpe, & Thorpe, 2002). This processing time seems to correspond to an optimum, because it cannot be speeded up even with highly familiar natural images (Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001). Moreover, when considering the number of processing steps between the retina and the high-level visual cortical areas of the ventral pathway, this 150-ms delay challenges most models of visual processing because it appears compatible only with a first feed-forward wave of information processing (Thorpe & Fabre-Thorpe, 2001). Thus, this delay appears as the minimal processing time from which discriminability between two categories of stimuli can develop. However, even if the human visual system is able to extract a great deal of information in under 150 ms, visual perception does not end up after a first pass through the visual system that might not even allow access to a conscious representation (Dehaene & Naccache, 2001; Thorpe, Gegenfurtner, Fabre-Thorpe, & Bulthoff, 2001); in many cases, reaching a decision will require more time consuming detailed analysis.

In parallel, growing evidence suggests that faces may have a special computational status (Farah, Wilson, Drain, & Tanaka, 1998; Kanwisher, 2000; but see Tarr & Gauthier, 2000) that would allow them to be processed more efficiently and even faster than any other class of objects. However, the precise speed of face processing remains a controversial question. Indeed, very rapid categorization of isolated and relatively homogenous face stimuli has been reported in the literature, with brain activity onsets appearing as early as 50-80 ms poststimulus (George, Jemel, Fiori, & Renault, 1997; Mouchetant-Rostaing, Giard, Bentin, Aguera, & Pernier, 2000a, 2000b; Seeck et al., 1997). These findings have been disputed as other groups have reported early face processing in the 100-130-ms latency range (Debruille, Guillem, & Renault, 1998; Halgren, Raij, Marinkovic, Jousmaki, & Hari, 2000; Halit, de Haan, & Johnson,

DOI 10:1167/3.6.5 Received March 11, 2003; published July 31, 2003 ISSN 1534-7362 © 2003 ARVO

2000; Itier & Taylor, 2002; Linkenkaer-Hansen et al., 1998; Pizzagalli, Regard, & Lehmann, 1999; Schendan, Ganis, & Kutas, 1998; Yamamoto & Kashikura, 1999; Liu, Harris, & Kanwisher, 2002) or even later in the 150-200-ms latency range (Bentin, Allison, Puce, Perez, & McCarthy, 1996; Carmel & Bentin, 2002; Eimer, 2000; Jeffreys, 1996; Rossion et al., 2000; Taylor, Edmonds, McCarthy, & Allison, 2001).

However, the vast majority of experiments with faces used isolated, homogeneous, and well-centered stimuli. Such a bias in stimulus sets could explain early face selective brain activity that could be due either to a higher predictability of the expected stimuli that would speed up processing (Delorme, Rousselet, Macé, & Fabre-Thorpe, 2003) or to the bottom-up extraction of low-level physical properties from a set of homogenous stimuli (VanRullen & Thorpe, 2001b). Thus, the data obtained with isolated face stimuli may not necessarily apply to real-world situations. For instance, it is known from single-unit recordings in monkeys that the responses of neurons tuned to faces and other object categories are affected by the presence of other competing objects, and by the presence of a background (Chelazzi, Duncan, Miller, & Desimone 1998; Trappenberg, Rolls, & Stringer, 2002). Thus, it is interesting to investigate the functioning of the biological visual system in more realistic situations when faces are presented in the context of natural scenes. In order to obtained such a "realistic" estimate of face processing speed, we used a rapid go/no-go categorization task with briefly presented (20 ms) photographs of real-world scenes in which subjects had to react when the photograph contained a human face. Such a go/no-go design involves the simplest motor output possible, allowing subjects to respond as fast as they could with the minimal motor constraints. For comparison with another class of targets, subjects alternated between this face categorization task and an animal categorization task used in a series of earlier studies from our group.

The second issue we wanted to address concerned the characteristics of the object representations activated during rapid categorization tasks. These early representations could be specific to canonical presentations of the stimuli used in the tasks. Alternatively, they might rely on relatively view invariant representations. One way to address this issue is to analyze how processing is affected with inverted pictures. Indeed, face processing has been shown to be more sensitive to inversion than other object categories (Bentin et al., 1996; Rossion et al., 2000; Yin, 1969). This pattern of results has been taken as evidence that face perception relies on specific mechanisms dedicated to the processing of the configural information present in upright faces (Maurer, Le Grand, & Mondloch, 2002). To explain the additional time necessary to process inverted pictures, some models of object recognition postulate the existence of a normalization stage at which an object orientation must be aligned with a memory template before matching can take place (see review in Tarr & Bülthoff, 1998; Ullman, 1996). Such normalization stage might be associated with a time consuming mental rotation of misaligned objects (Jolicoeur, 1988; Tarr & Pinker, 1989; Vannucci & Viggiano, 2000). Here we wanted to assess whether this inversion effect would affect the rapid categorization of human faces or animals presented in the context of natural scenes. To address this last issue, half of the pictures (faces, animals, and other natural scenes), whether targets or distractors, were presented upside-down.

Behavioral performance was analyzed in subjects alternating between rapid categorization of human faces and of animals presented randomly, upright or inverted, in the context of natural scenes. The processing speed and the magnitude of the inversion effect were compared for human faces and animals in two experiments, in which the main difference was in the presentation scale of the targets.

Experiment 1

The first experiment was designed to compare directly the animal task used by our group in several previous experiments to a homologue human face task. In both tasks, target images were photographs of real-world scenes in which human faces or animals were shown at different scales, orientations, and positions (Figure 1). Because "face" stimuli did not contain isolated items, but faces in the context of human bodies embedded in natural scenes, we will refer in the remaining of the text to "human" pictures and "contextual face task."

Methods

Participants

The 24 adult volunteers in this study (12 women and 12 men; mean age 31 years, ranging from 19 to 53 years; 5 left-handed) gave their informed written consent. All participants had normal or corrected-to-normal vision.

Experimental procedure

Subjects were seated in a dimly lit room at 100 cm from a computer screen (resolution, 800 x 600; vertical refresh rate, 75 Hz) piloted from a PC computer. To start a block of trials, they had to place their finger on a response pad for 1 s. A trial was organized as follows: a fixation cross (0.1° of visual angle) appeared for 300-900 ms and was immediately followed by the stimulus presented during two frames (i.e., about 23 ms in the center of the screen). Participants had to lift their finger as quickly and as accurately as possible (go response) each time a target was presented and to withhold their response (no-go response) when the photographs did not contain a target. Responses were detected using infrared

Rapid categorization of human faces and animals in upright and inverted natural scenes

Guillaume A. Rousselet, Marc J-M. Macé, Caitlin R. Sternberg, Michèle Fabre-Thorpe, Simon J. Thorpe
CERCO, UMR 5549 CNRS-UPS, Toulouse, France emails: guillau/marc/mft/thorpe@cerco.ups-tlse.fr

Face studies almost always make use of isolated, centered and homogenous stimuli. **How do we process faces in natural scenes?**

➡ **2 experiments:**

- 48 subjects (24 women, 24 men), 24 subjects per experiment
- go/nogo categorization tasks:
  - target (50%): release button (<1s) = go response
  - non-target (50%): keep pressing button (1 s) = nogo response
- unique and brief presentation = 20 ms, ISI = 1600-2200 ms
- no a priori knowledge of number, position and types of targets
- ERP recordings with 32 channels

Experiment 1

Target pictures = human beings & animals at various scales in natural images (see pictures on the left)
Highly varied distractor pictures

2 **alternating** categorization tasks:

Task 1:	Targets = human faces
Task 2:	Targets = animals

2 series of each task alternating (counterbalanced)
96 trials per series
16 series

IN HUMAN FACE SERIES, 50% OF THE DISTRACTORS = ANIMAL PICTURES (AND VICE VERSA) ⟶ HIGH DISTRACTOR/TARGET SIMILARITY (PARTICULARLY IN EXP2) ⟵ IN ANIMAL FACE SERIES, 50% OF THE DISTRACTORS = HUMAN FACE PICTURES (AND VICE VERSA)

Experiment 2

Target pictures = close-up views of human faces & animal faces in natural images.

Distractor pictures = various close-up views of objects in natural scenes, many tricks. (see pictures on the right)

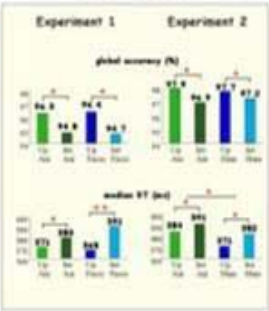
2 **block** categorization tasks:

Task 1:	Targets = human faces
Task 2:	Targets = animal faces

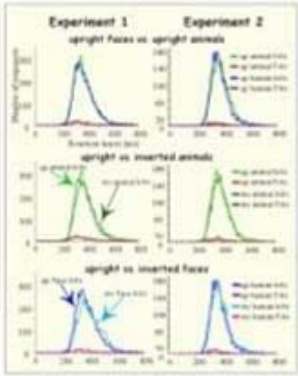
4 series with one task then 4 with the other (counterbalanced)
96 trials per series
8 series

1. Behavior

Global effects



RT distributions



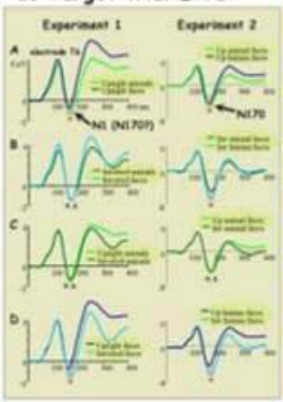
1. Behavior

Despite the high target unpredictability and the strong overlap between target and distractor features (see particularly pictures of exp2!) subjects were **highly accurate and fast**.

Exp1: Upright human faces vs. animals: same speed, same accuracy.
Exp2: Close-up views of human faces vs. animal faces: very weak temporal advantage for human faces.

Inversion effects were remarkably weak in all conditions.

2. Target trial ERPs

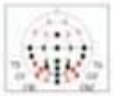


2) Target trial ERPs

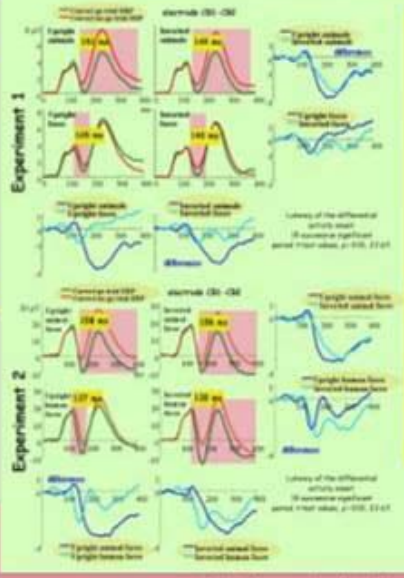
A) Faces did not elicit a large N1 in exp1, thus N170 is not the hallmark of face detection.
There was a clear N170 not specific to human faces in exp2, thus N170 might index mechanisms very sensitive to the *general face configuration* (at least with close-up views) independently of the nature of the component features.

C) No effect of inversion on the N1/N170 amplitude associated with animals & animal faces.

D) However, inversion clearly affected the N1/N170 amplitude for human faces.



3. Discrimination effects => latency at which target ERPs diverged from distractor ERPs (e.g. target human faces vs. distractor animal faces + other distractors)



3) Discrimination effects

Experiment 1
Contrary to behavioral results, **upright and inverted human faces were clearly discriminated earlier than animals at the cerebral level**.

Virtually no effect of inversion on animal categorization onset.
Clear impact of inversion on human face categorization onset.

Experiment 2
Again and even more dramatically, upright and inverted human faces were clearly discriminated earlier than animal faces.

Virtually no effect of inversion on the differential activity for either category.

Conclusions

1) Striking behavioral results despite:
brief presentation times,
relatively small face sizes (exp1),
high target/non-target similarity (exp2).
It seems that subjects did not rely only on the presence of features such as the eyes, mouth, nose, head, or a particular skin color.
=> might involve categorization mechanisms very sensitive to a large number of diagnostic features processed in parallel.

2) At the cerebral level, human faces were clearly categorized faster. This is surprising because animals have been shown to be processed very fast and in parallel in natural scenes (Thorpe et al. Nature 1996; Fabre-Thorpe J.Cog.Neurosci 2003; Rousselet et al. Nature Neurosci 2002).
=> might be explained by higher within category homogeneity and stronger expertise for human faces.

3) The pattern of behavioral & electrophysiological results is not consistent with the involvement of mental rotation mechanisms but might rather be explained by a lower rate of accumulation of evidence at the neuronal population level for inverted pictures.

Article 6

ERP studies of object categorization in natural scenes: in search for category specific differential activities.

Rousselet, G.A., Macé, M.J.-M., Thorpe, S.J. & Fabre-Thorpe, M.

(article en préparation)

Résultats électrophysiologiques de 48 sujets adultes dans deux expériences portant sur la vitesse de catégorisation de visages d'être humains, d'animaux et de visages d'animaux dans des scènes naturelles présentées à l'endroit et à l'envers. Le but des analyses présentées dans cet article consiste à identifier des marqueurs électrophysiologiques de l'activité neuronale susceptibles d'indexer la vitesse de traitement des différentes catégories d'objets testées.

Introduction

L'article 6 a pour objectif de fournir une évaluation électrophysiologique du temps de traitement des stimuli humains et animaux mis en jeu dans l'article 5. Un point important consistait à confirmer, à l'aide des potentiels évoqués, ce qui avait été mis en évidence au niveau comportemental : une vitesse de traitement similaire entre stimuli humains et animaux. Un second point consistait à évaluer l'impact de l'effet d'inversion sur le décours temporel du traitement, tel qu'il est indexé par les activités différentielles. D'autres questions découlaient directement de travaux antérieurs. Notamment, plusieurs études ont mis en évidence des activités différentielles très précoces pour des visages d'êtres humains (voir chapitre 2). Si ces différentielles ont parfois été interprétées comme le signe d'un traitement ultra rapide des visages d'humains, en moins de 100 ms, VanRullen & Thorpe (2001b) ont fourni des arguments suggérant que ces différences n'indexeraient pas l'activation de représentations de haut niveau, mais plutôt des différences physiques corrélées aux catégories visuelles mises en jeu. Cependant, leur interprétation reposait sur l'emploi de stimuli non humains, à savoir des animaux et des moyens de transport. Un des objectifs de la présente étude était donc d'évaluer plus directement le statut de ces différentielles précoces en réponse à des visages humains. Cela était possible par

l'utilisation d'un protocole identique à celui mis en place par VanRullen & Thorpe. En effet, au cours d'une série expérimentale d'une tâche donnée, par exemple la catégorisation de visages d'humains, la moitié des distracteurs étaient des cibles de l'autre tâche, par exemple des visages d'animaux dans l'expérience. Ce protocole de tâches alternées permet ainsi de calculer les potentiels évoqués par une catégorie de scènes naturelles vues soit comme cible, soit comme distracteur. En soustrayant ces deux types de potentiels, on obtient des courbes différentielles censées refléter uniquement le traitement de la catégorie cible, indépendamment des différences physiques, puisque les images cibles et distracteurs appartiennent à la même catégorie et sont même identiques sur l'ensemble des sujets testés. Par contre, les activités différentielles qui résultent de la soustraction des potentiels évoqués par les cibles et par les distracteurs au cours de la même série expérimentale reflètent à la fois le traitement des cibles et les différences physiques entre catégories cibles et distracteurs (par exemple les cibles sont des visages d'humains et les distracteurs sont des visages d'animaux ainsi que diverses scènes naturelles contrôles). On nommera les activités différentielles qui incluent différences physiques et différences liées à la tâche des AD1, et celles isolant les effets de la tâche des AD2. En utilisant cette stratégie, VanRullen & Thorpe (2001b) ont montré que les effets de type AD1 enregistrés à 150 ms pour la catégorisation animal/non-animal (Thorpe et al., 1996 ; Fabre-Thorpe et al., 2001) sont également présents dans les AD2. Par contre, les différences survenant avant 100 ms au niveau des AD1 disparaissaient totalement au niveau des AD2. Cette dichotomie suggère donc que des différences physiques sont à l'origine des effets très précoces alors que les effets à 150 ms sont bien liés à la tâche. La présente étude avait pour objectif de répliquer ces résultats et d'évaluer s'ils s'appliquaient aux visages d'humains.

Finalement, dans un article récent, Johnson & Olshausen (2003) ont fourni des interprétations différentes des AD1. Ils ont montré que leur latence est indépendante des temps de réaction des sujets, comme cela était le cas dans l'étude de Thorpe et al. (1996). Par contre les AD2 ont une latence plus tardive pour des temps de réaction plus lents. De plus, ils ont montré que les AD2 avaient une latence plus tardive que les AD1, classiquement rapportées à 150 ms dans les expériences de notre équipe. Ils ont donc conclu que les différences à 150 ms ne sont pas liées à la tâche. Pour tester cette hypothèse, l'article 6 comporte une analyse des deux types d'activités différentielles en fonction du temps de réaction des sujets.

Résultats

Les résultats sont très nombreux. En voici une présentation synthétique.

1) Au niveau des AD1, il y avait de nombreuses différences significatives avant 100 ms dans la première expérience (animaux et visages humains présentés dans une large gamme d'échelles de taille). Bien que toujours visibles sur les tracés électrophysiologiques, ces différences n'étaient presque plus significatives au niveau des AD1 de l'expérience 2, dans laquelle les stimuli étaient beaucoup plus homogènes. Dans tous les cas, il n'y avait aucune différence précoce au niveau des AD2.

2) Dans les deux expériences, les stimuli humains étaient caractérisés par une large AD1 à 120 ms. Une AD1 de faible amplitude était également présente à 120 ms pour les animaux et les visages d'animaux. Ces effets à 120 ms n'étaient pas présents au niveau des AD2.

3) Dans les deux expériences, les animaux étaient surtout caractérisés par de larges AD1 avec une latence de 150 ms. Dans la première expérience, les AD2 les plus précoces présentaient aussi une latence de 150 ms, mais avec une amplitude moins importante que les AD1.

4) Quels que soient les stimuli, les AD2 étaient toujours d'amplitudes plus faibles que les AD1.

5) De manière totalement inattendue, les AD2 calculées pour les stimuli humains dans la première expérience et pour les visages d'humains comme pour les visages d'animaux dans la seconde expérience se développaient avec des latences beaucoup plus tardives que les AD1.

6) Selon les conditions, l'effet d'inversion était très faible ou quasiment inexistant au niveau de la latence des activités différentielles. L'effet le plus important lié à l'inversion concernait la pente de l'activité différentielle, plus faible dans le cas des stimuli inversés.

7) L'analyse des AD1 et des AD2 montre que ces deux types d'activités différentielles varient de manière importante en fonction du temps de réaction des sujets, contrairement au résultat nul rapporté par Thorpe et al. (1996) et Johnson & Olshausen (2003).

Discussion

Par rapport aux résultats décrits par VanRullen & Thorpe, les résultats concernant la catégorisation des animaux dans l'expérience 1 confirme l'existence de différences précoces, avant 100 ms, qui ne semblent pas dépendre de la tâche puisqu'elles disparaissent quand les propriétés physiques des images sont égalisées (AD2). En ce qui concerne la catégorisation des

visages d'humains dans les expériences 1 et 2 et la catégorisation des visages d'animaux, les résultats sont plus compliqués. Un premier point important est la disparition presque complète des différences très précoces quand les propriétés physiques des cibles et des distracteurs sont égalisées. Il semble donc bien que les effets avant 100 ms pour les visages d'humains soient dus à des différences physiques de bas niveau (mais ce niveau reste à définir). Il faut cependant rester prudent dans l'interprétation de ce résultat, dans la mesure où certaines propriétés dites bas niveau pourraient se révéler utiles à la réalisation de la tâche. Deuxième point, dans les deux expériences, les stimuli humains sont caractérisés par de larges AD1 apparaissant vers 120 ms, alors que pour les animaux elles apparaissent vers 150 ms. Il est difficile de considérer que ce délai de 30 ms reflète une vraie différence de temps de traitement étant donné que les stimuli humains ne présentaient pas d'avantage sur les animaux au niveau des temps de réaction (Rousselet, Macé & Fabre-Thorpe, 2003). Les différences à 120 ms pourraient donc être due elles aussi à des différences physiques non contrôlées. Ce point de vue est renforcé par la disparition de ces effets au niveau des AD2, lorsque les signaux pour les mêmes catégories d'images mais avec des statuts différents sont comparés. Cependant, l'interprétation de ces signaux à 120 est compliquée par d'autres données. D'une part, des différences à 120 ms sont également présentes pour les animaux, mais avec une amplitude moins importante. Ceci pourrait être dû à la plus grande homogénéité des stimuli humains entre eux par rapport aux stimuli animaux, entraînant une meilleure sommation du signal au niveau des potentiels évoqués dans le premier cas par rapport au second. D'autre part, quand les AD1 sont recalculées en fonction des temps de réaction, les AD1 apparaissant vers 120-130 ms pour les stimuli humains correspondent aux temps de réaction les plus précoces, les réponses comportementales plus tardives étant associées à des latences plus longues d'AD1. On pourrait objecter que comme les différences avant 150 ms pour les stimuli humains et animaux disparaissent dans les AD2, celles-ci sont inintéressantes. Le problème est que les AD2 dans ces deux expériences, à part pour le cas des animaux dans la première expérience, ont des latences difficiles à interpréter et semblent extrêmement peu fiables pour prédire les durées de traitement dans les différentes tâches. Malgré des performances comportementales équivalentes dans les deux expériences entre humains et animaux, les AD2 dans la première expérience ont une latence plus tardive pour les humains par rapport aux animaux, alors que c'est l'inverse dans la seconde expérience. De plus, les AD2 pour les stimuli humains et pour les visages d'animaux présentent des latences dont les valeurs absolues sont très

importantes. Si la latence des AD2 indexait véritablement la vitesse de traitement des stimuli, alors on devrait s'attendre à ce que ces stimuli soient catégorisés avec des temps de réaction particulièrement longs, ce qui n'était pas du tout le cas.

Si tous ces résultats sont difficiles à interpréter, ils permettent cependant de conclure que la dichotomie entre AD1 et AD2 proposée par VanRullen & Thorpe (2001b) doit être considérée avec précaution. L'idée selon laquelle les AD1 ne présentent aucun intérêt (Johnson & Olshausen, 2003) est également à remettre en cause. Les AD2 très tardives pour les stimuli humains et pour les visages d'animaux pourraient signifier que ces stimuli sont traités par défaut plus en détails (c'est à dire jusqu'à une représentation plus fine) que d'autres stimuli, (comme des animaux), et ceci quel que soit leur statut (cible ou distracteur) dans la tâche exécutée. Ceci pourrait constituer une spécificité des visages au sens large (visages d'animaux inclus). Il est également possible d'interpréter ces données dans le cadre du modèle proposé dans l'article 2. Selon ce modèle, les AD1 à 150 ms seraient le reflet d'un mécanisme de sélection spatiale par lequel des aires corticales telles que V4 reçoivent des réactivations descendantes. Dans ce cas, certains stimuli pourraient être automatiquement sélectionnés qu'ils soient ou non la cible de la tâche, alors que d'autres stimuli seraient sélectionnés seulement s'ils présentent un intérêt pour la tâche en cours. Ainsi, les visages pourraient toujours être sélectionnés quand ils apparaissent, ils seraient donc systématiquement associés à une forte AD1, mais l'AD2 serait absente.

L'effet tardif de la tâche sur les AD2 nous rappelle aussi qu'une absence de différence entre deux conditions ne signifie pas une absence d'effet. Il faudrait réaliser d'autres expériences pour en être sûr. De manière générale, il est surprenant de trouver si peu de points communs entre les latences d'activités différentielles pour des catégories qui, au niveau comportemental, semblent être traitées à la même vitesse. La seule fenêtre de recouvrement concerne les AD1 apparaissant vers 120-140 ms pour les stimuli humains et animaux. En accord avec une partie de la littérature revue au chapitre 2, ces différences pourraient refléter un premier traitement grossier de certains objets tels que des visages dans une scène visuelle (Itier & Taylor, 2002, sous presse). L'amplitude plus importante de ces signaux pour les stimuli humains par rapport aux animaux pourrait refléter la mise en jeu additionnelle de populations de neurones dans des aires visuelles plus latérales dans le premier cas et pas dans le second (Itier & Taylor, sous presse). Cette piste prometteuse mériterait d'être explorée par de nouvelles expériences. Il faut souligner que cette première activité précoce pourrait ne pas être suffisante pour prendre une décision et déclencher

un acte moteur. Peut-être qu'un corrélat plus direct de la prise de décision n'est pas le début de l'activité différentielle mais se situe dans une période plus tardive, où l'activité est plus ample (Bacon-Macé et al., soumis). Cette hypothèse s'inscrit dans le cadre d'un modèle où les réponses comportementales sont déclenchées après l'accumulation d'un certain nombre de réponses au niveau d'une population neuronale (Ditterich et al., 2003 ; Hanes & Schall, 1996 ; Perret et al., 1998).

Finalement, l'inversion des scènes naturelles avait un effet relativement faible sur la latence des activités différentielles, affectant surtout leur pente, particulièrement dans le cas des AD1. Cet effet d'inversion sur la pente de l'activité différentielle renforce l'idée d'un modèle par accumulation de réponses neuronales, le nombre de neurones sélectifs à des objets à l'endroit étant plus important que ceux sélectifs à des objets inversés. Il reste à expliquer pourquoi l'amplitude du pic d'AD1 était plus importante pour les visages d'humains à l'envers.

ERP studies of object categorization in natural scenes: in search for category specific differential activities

Guillaume A. Rousselet, Marc J.-M. Macé, Simon J. Thorpe & Michèle Fabre-Thorpe

Introduction

Both behavioral and electrophysiological evidence can be used to provide information about the speed of visual processing. Behavioral data has the distinct advantage that its functional relevance is obvious: if an animal or human subject can make a behaviorally useful response to a particular type of visual stimulus in a certain time, it is clear that this information can be of survival value. Thus, the fact that humans can initiate go/no-go responses to the presence of an animal in a briefly flashed natural scene in as little as 230-250 ms puts a clear upper limit on the time required for visual processing (VanRullen & Thorpe, 2001a). And the fact that monkeys can perform the same sort of task with behavioral reactions that are even shorter (starting from 160-180 ms), imposes even more severe temporal constraints (Fabre-Thorpe et al., 1998). However, any behavioral reaction time measurement will include not just the time required for sensory processing, but also the time needed to initiate and execute the motor response. In such cases, electrophysiological measurements can be used to help determine the time course of the intervening processes. In animals, single unit recording can be used to determine precisely when individual neurons respond during a particular task and much can be learned from the time course of responses of neurons in regions such as inferotemporal cortex (e.g. Sheinberg & Logothetis 2001; Tanaka, 1996). There is also a limited amount of evidence from single unit recordings made in human patients during surgical procedures for the treatment of epilepsy, but the fact that such subjects are often heavily sedated means that the latencies obtained may well be abnormally long (e.g. Allison et al., 1999; Kreiman et al., 2000). One approach that has been used successfully in normal human subjects involves Event Related Potential (ERP) recording. By analyzing the averaged waveforms produced in response to images containing targets, and subtracting the average waveform produced in response to distractor images, one can obtain a difference waveform that can, in appropriate conditions, be used to determine the moment when responses to targets and distractors start to differ. The time at which the two waveforms start significantly to diverge provides an upper limit on the time necessary for the processing of targets to start.

In an early such study, Thorpe et al. (1996) found that the difference between the ERP to targets and distractors at frontal sites starts to show clear statistically significant effects from 150 ms following the onset of each trial.

However, interpreting these differential response functions is not without difficulties. In some conditions one can obtain statistical significant differences in the ERPs to two classes of image that could simply be due to low level differences in the physical properties of the images, and not to recognition per se. For example, suppose that one class of image was physically darker than the other one. This could easily produce differences in the neural responses in areas such as V1 that would be visible as significant effects occurring as remarkably short latencies. One way to avoid this potential confound is to change the target status of the images so that one can compare the ERP responses to the same images treated either as a target, or as a distractor. In such a case, the same physical images are compared and so any differences that are apparent cannot be due to

low level differences. This approach was first developed to study the effects of attention in the auditory system (e.g. Hillyard et al., 1973) and then in the visual system using relatively simple stimuli (e.g., Hillyard & Münte, 1984). VanRullen & Thorpe (2001b) extended this approach to natural scenes and showed that while early differential effects were abolished by such a manipulation, differential effects that started from 150 ms were still present after this procedure had been applied.

In VanRullen and Thorpe's experiment, there were two basic target categories – animals and means of transport. Subjects alternated between blocks in which animals were targets, and blocks in which the target category was "means of transport". In each condition, half the distractor images were targets from the other blocks and by carefully counterbalancing the experimental design, each individual image was treated as either a target or a distractor by different subjects.

In the present paper, we apply the same sort of analysis to another set of data, for which the behavioral results have been published previously (Rousselet et al., 2003). Two separate sets of data were used. In the first set of experiments, subjects had to decide either whether the image contained an animal, or whether the image contained a human face. The animals and faces could be at almost any size and position within a natural scene. In the second set, the subjects had to either respond to animal faces or human faces, but in this case the images were all relatively close up views of just the head region. As reported elsewhere, performance was exceptionally good, despite the wide range of stimuli used (Rousselet et al., 2003). Furthermore, it was found that inverting the images had remarkably little effect on performance, a point that is of major importance for understanding the nature of the underlying processing.

However, regarding the issue of ERP differential effects, the main conclusion from this study is that, particularly in the case of the face stimuli, the task dependent differences in ERP were surprisingly weak and of relatively long latency. This result, which contrasts strongly with the remarkably accurate behavioral responses of the subjects and their very short behavioral reaction times implies that strong task-dependent ERP differences are not required for performing such high level visual tasks. Instead, we argue that some of the very strong differential effects occurring from 135 ms onwards almost certainly reflect processing that is intimately related to the identification and recognition processes.

Methods

Forty-eight subjects volunteered in these two studies and gave their informed consent. All had normal or corrected to normal vision. Nine subjects participated in both experiments.

Task setup

Subjects were sat in a dimly lit room at 100 cm from a computer screen (resolution: 800 x 600, vertical refresh rate: 75 Hz) piloted from a PC computer. To start a block of trials, they had to place a finger on a response pad for one second, then a fixation cross (0.1° of visual angle) appeared for 300-900 ms and was followed by the stimulus presented for two frames, i.e. about 26 ms in the middle of the screen. Participants had to lift their finger as quickly and as accurately as possible (go response) each time a target was presented. Responses were detected using infrared diodes. Subjects had 1000 ms to respond after which their response was considered as a no-go response. This maximum response time delay was followed by a 300 ms black screen, before the fixation point was presented again for a variable duration, resulting in a random 1600-2200 ms inter-

trial interval. When the photographs contained no target, subjects had to keep their finger on the pad for at least 1000 ms (no-go response).

In experiment 1, an experimental session included 16 blocks of 96 trials and subjects alternated between two categorization tasks. In 8 of the blocks, the target was an animal and in the 8 other blocks, the target was a human face. Half of the subjects started with the animal categorization, the other half with the human face categorization and conditions alternated by blocks of two. In experiment 2, there were 8 blocks. In the first 4 blocks, the target was an animal face and in the other 4 blocks the target was a human face (counterbalanced). For both experiment, in each block, target and non-target trials were equally likely. Among the 48 non-targets, 24 were targets of the other categorization task. Thus, when performing a human face categorization task, on a 96 trial block, 48 pictures contained at least one human face, 24 non-target scenes contained animals, the last 24 non-targets being other types of natural scenes. Moreover, half of the targets and half of each of the non-target subsets were presented upright while the other half was presented inverted (rotation 180°). Each image was only seen once by a given subject, with one orientation (upright or inverted) and one status (target or non-target). Subjects had two training blocks of 48 images before starting the test session. Training pictures were not used during the test period.

Stimuli

We used photographs of natural scenes taken from a large commercial CD-ROM library (Corel Stock Photo Library). They were all horizontal photographs (768 by 512 pixels, sustaining about 19.9° by 13.5° of visual angle) and chosen to be as varied as possible. Animals included essentially mammals, but also birds, fish, and reptiles. Human faces were presented in real-world situations with views ranging from whole bodies at different scales to face close-ups and including Caucasian and non-Caucasian people. There was also a very wide range of non-target images that included outdoor and indoor scenes, natural landscapes (mountains, fields, forests, beaches...), street scenes, pictures of food, fruits, vegetables, plants, buildings, tools and other man-made objects, as well as some more tricky distracters (e.g. dolls, sculptures, statues… and non-target images containing humans for which the faces were not visible). In experiment 2, only close-up views of target objects were used and a special attempt was made to use many tricky distractors and "blob" objects appearing in positions similar to human and animal faces. Subjects had no *a priori* information about the presence, the size, the position or the number of targets in an image and trial unique presentation prevented learning.

EEG recording and analysis

A SynAmps amplifier system (Neuroscan Inc.) was used to record brain electrical activity with 32 electrodes mounted in an elastic cap (Oxford Instruments) in accordance with the 10-20 system with the addition of extra occipital electrodes (like CB1-CB2, which are referred as PO9-PO10 in the 10-10 system). The ground electrode was placed along the midline, ahead of Fz and impedance was systematically kept below 5 kΩ. Signals were digitized at a sampling rate of 1000 Hz (corresponding to a sample bin of 1 ms) and low-pass filtered at 40 Hz before analysis. Potentials were on-line referenced on electrode Cz and re-referenced off-line by subtracting the average of all signals from each individual signal. Baseline correction was performed using the 100 ms of pre-stimulus activity. Two artifact rejections were applied over the [-100 ms; +400 ms] time period, first on frontal electrodes with a criterion of [-80; +80 μV] to reject trials with eye movements, second on parietal

electrodes with a criterion of [-40; +40 μV] to remove trials with excessive activity in the alpha range. Only correct trials were averaged.

Significant differences between two conditions were assessed by performing paired *t*-tests at the $p<0.01$ level every ms at each scalp location. The time bin at which a significant value of *t*-test was reached and followed by at least 15 consecutive significant bins was taken as the onset latency of a differential activity between the two conditions. All values reported in the text met this criteria. For simplicity and because a special emphasis is placed on speed of processing in this paper, only the shortest differential activity onset latencies are reported.

Results

Subjects (n = 24, 12 women, 12 men, mean age 31) performed remarkably well in these tasks. A detailed analysis of the behavioral results has been published separately (Rousselet et al., 2003). In the first experiment, upright faces and animals were processed on average as efficiently (96.4% and 96.3%, respectively) and at the same speed (median reaction time: 368 ms and 371 ms, respectively).

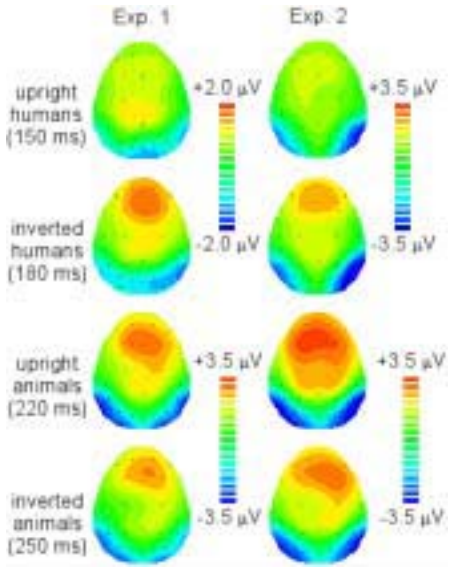


Figure 1. Two dimensional linear interpolation maps of the differential activities in each experiment and for each condition. The maps represent the signal recorded at the latency of the peak of the differential activities reported in figures 2 and 4, i.e. when the differential effect at CB1 and CB2 was maximal.

The time at which enough information was available about a given category was assessed from event-related potentials (ERP) on correct trials. Target ERPs were compared to distractor ERPs using a "running" *t*-test strategy in which differences were tested every millisecond on the whole set of scalp electrodes. Early and large differences were found over the entire set of 12 posterior electrodes over both hemispheres for the two conditions, with differential effects that were strongest at lateral occipito-temporal sites (Figure 1).

Regarding the differential activity signal, responses to upright target animals differed significantly from distractors as early as 148 ms (shortest differential activity onset, at least 15 consecutive paired *t*-test, 23 df, $p<0.01$), a result that constitutes a direct replication of previous studies performed in our laboratory (Fabre-Thorpe et al., 2001; Thorpe et al., 1996; VanRullen & Thorpe, 2001b) (Figure 2). However, differential effects when faces were targets started even earlier, with significant effects starting as early as 125 ms (Figure 2).

Furthermore, we investigated the effects of inversion on processing speed in such a task, a manipulation which is known to slow down particularly the identification of faces (Rossion & Gauthier, 2002). It appeared that both behavior and the onset of ERP differences were only very weakly affected by inversion. Inversion produced a global decrease of accuracy that was very similar for both faces and animals (<2%). Inverted pictures led to significantly longer RT than upright pictures but the inversion effect was reliably more pronounced for faces (+23 ms on median RT) than for animals (+9 ms on median RT). These weak effects at the behavioral level were confirmed by ERP results. The differential activity for inverted animals started virtually at the same latency (149 ms) but developed with a shallower slope and reached a lower amplitude than for upright animals (Figure 2). The differential activity onset for faces was delayed by inversion, being significant at 140 ms (+15 ms) (Figure 2).

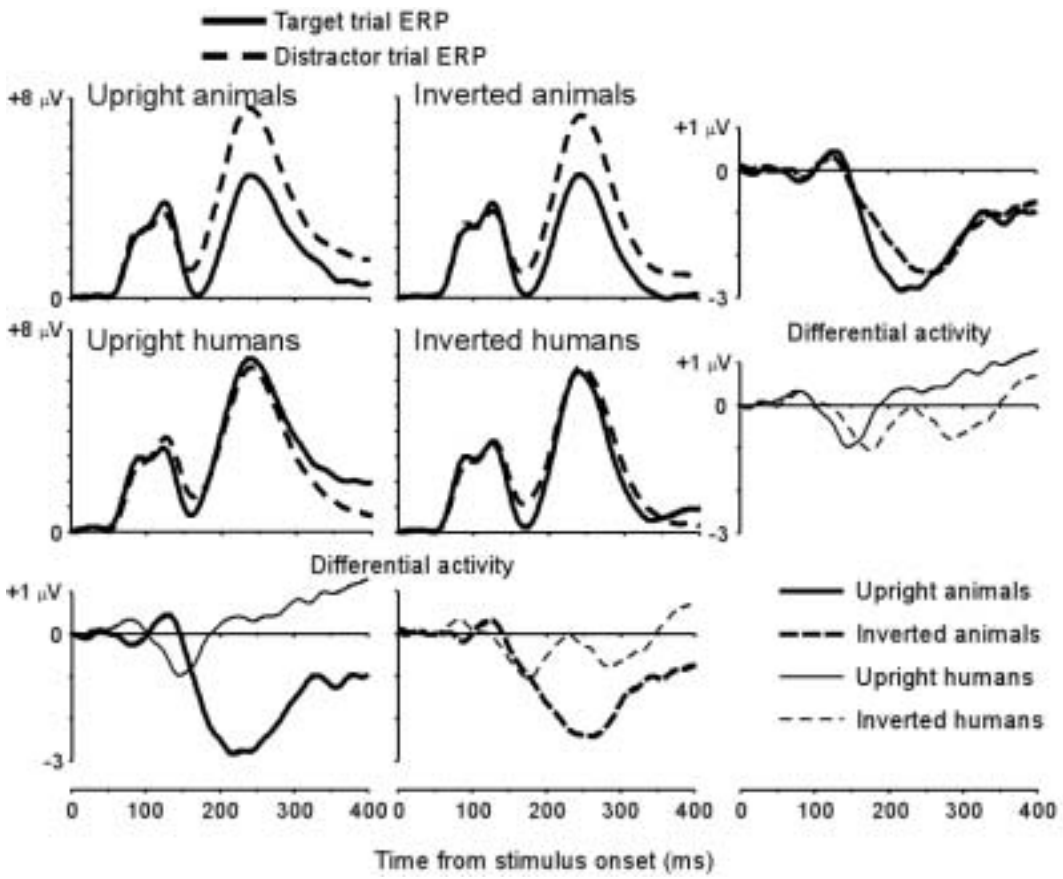


Figure 2. Comparison of the ERP associated with the processing of targets and distractors in experiment 1. Each graph represents the average signal recorded from occipital electrodes CB1 and CB2. These electrodes were chosen because it was there that the differential effects had the largest amplitude. For each target category, the ERPs are presented for upright and inverted stimuli. Target ERPs were computed from trials in which the indicated category was seen as target. Distractor ERP were computed from trials in which pictures with the same orientation as the target were seen as distractors. They include neutral distractors and pictures from the target category of the other task. The differential activities were computed by subtracting distractor trial ERPs from target trial ERPs separately for each category and each orientation. The two graphs on the right show the effect of inversion on the differential activities separately for both categories. The two graphs at the bottom allow the comparison of the differential activities associated with humans and animals separately for both orientation.

So far, these results seem to imply that face specific processing can start very shortly after stimulus presentation, as early as 120-130 ms, hence faster than the categorization of animals which seems to require an

additional 20-30 ms. Furthermore, this capacity relies on relatively view invariant representations as shown by the very weak inversion effects on processing efficiency. However, we did find small but reliable ERP differences before 100 ms (Figure 2). They appeared as early as 50-60 ms for upright and inverted faces (respectively 54 ms and 56 ms) and 70-90 ms for upright and inverted animals (respectively 83 ms and 73 ms). The onset latencies of all the significant differences for all the conditions in the two experiments are reported in Figure 3. We suspected these very early differences might be due to uncontrolled low-level differences between the sets of target and distractors images, as previously demonstrated for the categorization of natural images (Johnson & Olhausen, 2003; VanRullen & Thorpe, 2001b).

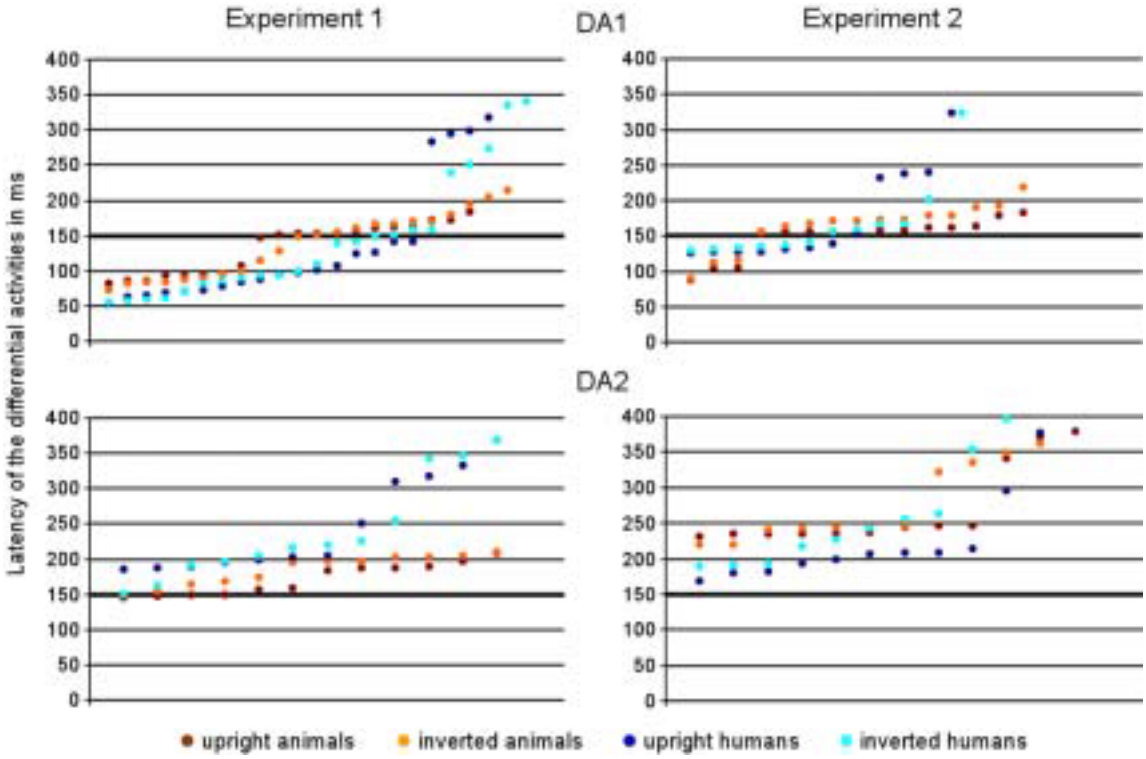


Figure 3. Latencies of the differential activities recorded in the two experiments from all 12 posterior electrodes. Each color disc represents one electrode. For each condition, the latencies were ordered from the shortest (left) to the latest (right) latencies. If an electrode presented a very early significant differential activity (<100ms), this value was taken into account and the t-test was applied to the subsequent bins to assess whether a second period of significant activity was present. The number of electrodes for each condition varies for this reason and also because some electrodes never reach the signification level of 15 consecutive steps with $p<0.01$ in some conditions. The two top panels report the latencies of the differential activities computed by subtracting the ERP associated with all distractors from the target ERP (named "type 1 differential activities"). The two bottom panels report the latencies of the task status differential activities, when physical differences were removed (named "type 2 differential activities").

Thus, a new experiment was designed in which subjects (n = 24, 12 women, 12 men, mean age 30, 9 of which participated in the first study) were required to categorize human faces and animal faces in pictures depicting close-up views of these targets. This manipulation was designed to decrease the physical differences between the two sets of target images. In order to further increase the similarity between targets and distractors and hence diminish the low-level differences, human faces were chosen to be as varied as possible and pictures that did not contain faces were chosen to contain many tricks like dolls, statues, flowers… At the behavioral

level, despite the greater target/distractor similarity, the use of close-up views led to excellent performance levels, with slightly higher accuracy and slightly longer reaction times for both categories compared to the first experiment (see Rousselet et al., 2003).

The stimulus manipulations in experiment 2 had several consequences at the ERP level. The key finding was that the very early differences recorded in experiment 1 for faces here disappeared completely (Figure 3). Upright animals were still associated with very early differential activities but the effects were restricted to occipital midline electrodes (shortest latency: 88 ms) (Figure 3). However, the large lateral occipito-temporal differential activities reported in experiment 1 were still present and even reached a higher amplitude in this second experiment (Figure 4). These differences appeared at about the same time as in the previous experiment reaching statistical significance respectively at 155 ms and 126 ms for upright animal and human faces (paired *t*-test, 23 df, $p<0.01$). Inversion had virtually no effect on these onset latencies, inverted animal faces being discriminated from distractors in 156 ms and inverted human faces in 130 ms. In addition, as reported in experiment 1, the slope of the activity was steeper for upright stimuli compared to inverted ones (Figure 4).

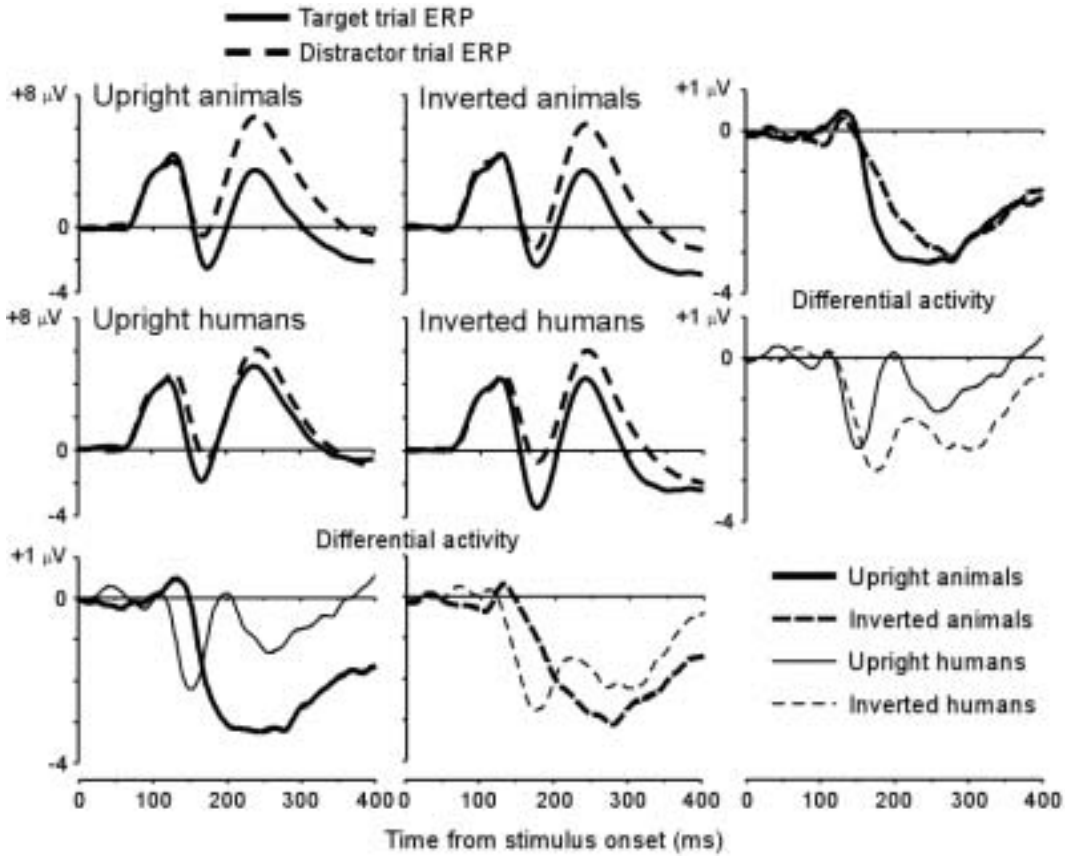


Figure 4. Comparison of the ERP associated with the processing of targets and distractors in experiment 2. Nomenclature as in Figure 1.

The second experiment directly demonstrated that very early differences recorded with human faces as targets were due to uncontrolled physical differences. However, some of these very early differences remained when animal faces were targets. For both experiments, we thus performed a subsequent analysis to assess more directly the sensitivity of the latency of the differential activity to the similarity between target and distractor

images. In addition to the differential activity reported above that took into account the ERP associated with all the distractors, two other kinds of differential activities were computed. In the first one, only neutral distractor ERPs were subtracted from target trial ERPs, whereas in the second one only ERP associated with distractors that formed the target category of the other task were used. Because natural scenes containing animals were probably more physically similar to those containing humans than neutral distractors, we reasoned that if the latency of the differential activity was affected by physical characteristics, then it should have an earlier onset in the first than in the second type of differential activity. The results confirmed this prediction.

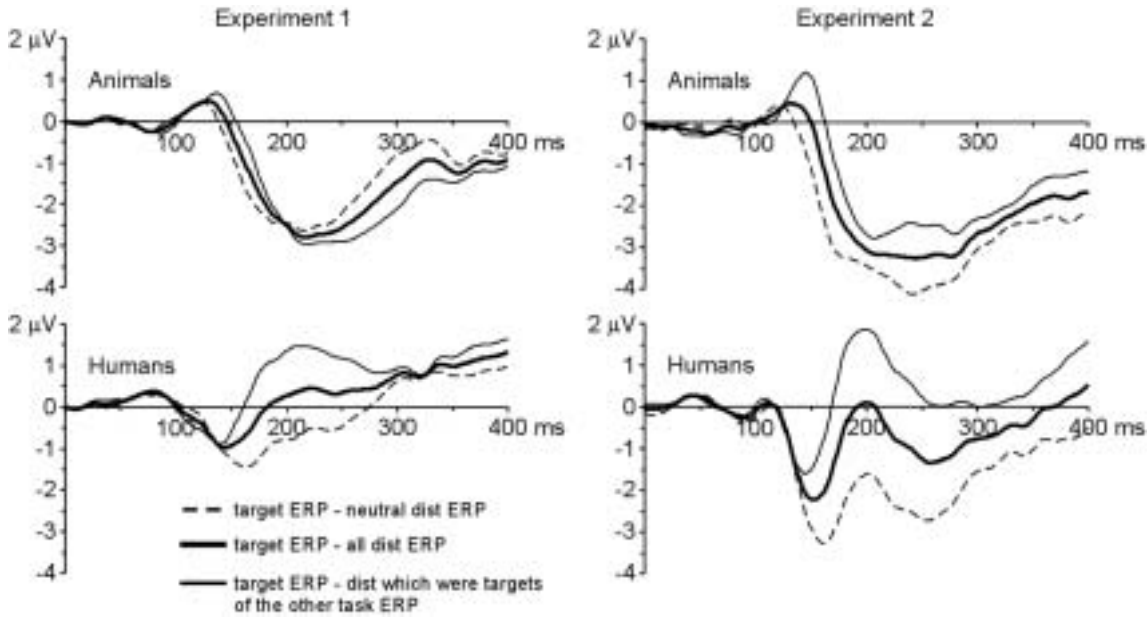


Figure 5. Effect of the visual similarity between targets and distractors on differential activities in the two experiments. Results are presented for upright trials only, inverted trials presented the same effect. Signals from electrodes CB1 CB2 were averaged. The thick curve represents the same differential activities reported in figure 1 and 2. It was computed by subtracting the ERP associated with all the distractors seen during the categorization of a target category from the ERP associated with the categorization of that target category. The thin dashed line represents the differential activity computed when only the ERP associated with the processing of the neutral distractors were subtracted from the target ERP. The thin continuous line represents the differential activity computed when only the ERP associated with the processing of the distractors that were targets of the other task were used (humans when animals were targets and vice versa).

As shown in Figure 5, the latency of the differential activity was directly influenced by the similarity between targets and distractors in both experiments. This was clear for the animal categorization task for upright stimuli (shortest differential activities in experiment 1: [target ERP – neutral distractor ERP] = 144 ms vs. [target ERP – all distractor ERP] = 148 ms vs. [target ERP – distractor that were target of the other task ERP] = 164 ms; experiment 2: 140 ms vs. 155 ms vs. 156 ms) as well as for inverted stimuli (experiment 1: 145 ms vs. 149 ms vs. 176 ms; experiment 2: 140 ms vs. 156 ms vs. 191 ms). However, the results for the human face task did not follow entirely this rule. In experiment 1, inverted faces led to increasingly delayed differential activity onsets with increasing physical similarity (123 ms vs. 140 ms vs. 146 ms) but this was not the case for upright faces which were associated with a paradoxical decrease of differential activity onset (131 ms vs. 125 ms vs. 108 ms). Even more striking were the results from experiment 2 which showed that both upright and inverted human faces were associated with differential activity onsets virtually insensitive to the physical similarities between targets

and distractors (upright human faces: 127 ms vs. 126 ms vs. 126 ms; inverted human faces: 130 ms vs. 130 ms vs. 126 ms). Note that this effect could also be seen in the animal task: when animal target ERPs were compared to human distractor ERPs, a significant 'bump' of differential activity appeared at about 120-130 ms post stimulus, just before the main differential activity onset at 150 ms (figure 5, top). A possible interpretation is that this early activity at 120-130 ms is related to the categorization of faces, independently of uncontrolled physical differences.

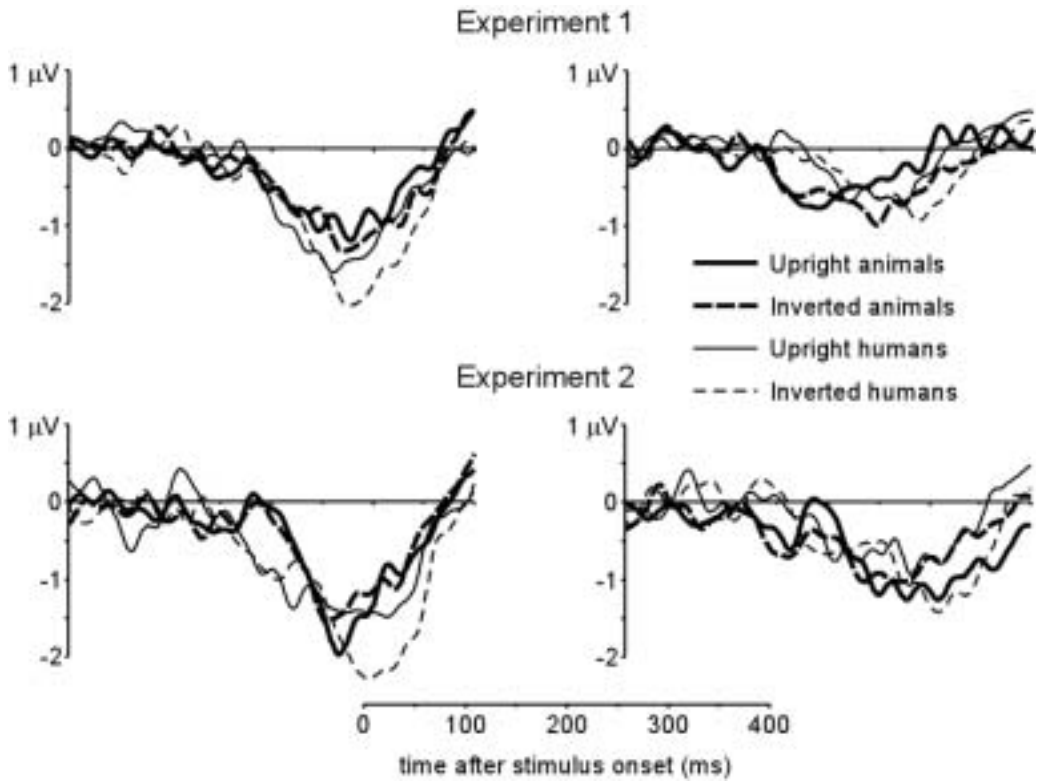


Figure 6. Differential activities showing the effects of task status independently of physical differences in the two experiments. These differential activities were computed by subtracting the ERP associated with a given category when seen as a distractor from the ERP associated with the same category when it was a target. Results are presented separately for the left hemisphere electrode CB1 (left) and for the right hemisphere electrode CB2 (right). Note that the shortest latencies reported in the text for the different categories were clearly lateralized in experiment 1. It appeared first at the electrodes situated over the right hemisphere in the animal task, appearing later over the left hemisphere (upright: 187 ms, inverted: 195 ms). The reverse pattern was observed in the human face task in which the earliest effects were lateralized to the left hemisphere, the first significant value in the right hemisphere being reached at 203 ms (upright) and 216 ms (inverted). Even if this pattern of lateralization is very interesting, it was not the aim of this experiment to tackle this sort of issue and we let it for future direct investigations. We thus concentrate on the shortest differential activity onsets for the different conditions independently of hemisphere effects. In experiment 2, no such pattern of lateralization was observed.

This hypothesis was tested by evaluating the processing speed of the different categories independently of their visual attributes. A new set of differential activities was computed in which target ERP for a given category and a given orientation was compared to ERP associated with the same category and the same orientation when it was seen as a distractor. This manipulation controlled for physical differences since across subjects the same pictures were seen as targets and as distractors. The only differences that remained were due to task status and should thus give us an estimate of the time required to access task related categorical information.

The results are depicted in Figures 3 and 6. Task related differential activities had a very small amplitude compared to those that included both category and task differences (Figure 6). In experiment 1, the animal task was found again to affect ERP at around 150 ms confirming a previous report that used this technique (Van Rullen & Thorpe, 2001b). This latency was almost unaffected by inversion (shortest upright latency: 145 ms, inverted: 149 ms; note that these two earliest effects were seen on the right hemisphere electrodes). Surprisingly, the human face task did not affect ERP before 185 ms (left hemisphere) for upright pictures. Task status had an earlier effect on inverted human ERP with a first significant activity at 151 ms (left hemisphere) post stimulus (this result contrasts with the absence of task status effect reported previously at the level of the N170 – the two signals do not necessarily have the same origin, Rousselet et al., in revision). The results from experiment 1 suggest that the early differential activities recorded for faces were unrelated to subject performance since they disappeared when physical properties between target and distractor ERP were equated. Only the large differential activities at 150 ms in the animal task seems to be related to the extraction of task related categorical information. However, results from experiment 2 cast doubt on this interpretation. Indeed, in experiment 2, the effects of task status on ERPs to both human faces and animal faces were all surprisingly late (Figures 3 and 6). In the first task the earliest differential activity was found at 168 ms and in the second task at 231 ms. This effect did not appear to suffer from inversion in the animal task, appearing even earlier for inverted pictures (219 ms), but inversion delayed the onset of the differential activity in the human face task (inverted pictures: 189 ms).

Discussion

By examining the averaged ERP responses in the various task conditions we were able to find clear and statistically significant differences between the responses to different stimulus classes at numerous electrode sites. Some of these differences had a distribution and a time course that was very similar to those seen in previous studies on rapid scene processing (Fabre-Thorpe et al., 2001; Johnson & Olshausen, 2003; Thorpe et al., 1996; VanRullen & Thorpe, 2001b). In this section we will discuss the various hypotheses that can be evoked to account for these differences.

A first point concerns the anatomical distribution of the differential responses. The original 1996 paper by Thorpe et al. concentrated on the differential signals observed at frontal recording sites which showed a clear enhanced negativity on no-go trials. This finding fitted with a number of other studies that had shown cortical negativity associated with no-go trials. Furthermore, in that original study, the fact that the temporal profile of the differential activity was essentially identical when calculated for trials with short reaction times and trials with long reaction times led the authors to speculate that the activity might be specifically related to response inhibition on no-go trials. However, more recent studies that have examined differential activity in forced choice tasks in which the subject has to make a response on every trial suggest that this explanation may be inadequate. For example, Johnson and Olshausen (2003) recently found a very similar pattern of differential effects at frontal sites when they compared a go/no-go and a forced choice response paradigm. Similar differential effects at frontal and parietal sites were also reported in a force-choice task by Antal et al. (2000). Such results are clearly inconsistent with the simple notion that the differences are caused by response inhibition per se.

Other results also argue against a response related explanation of the effect. In the original 1996 study, the restricted number of electrodes meant that very little data was available for more posterior sites close to the occipital cortex. Another explanation comes from the use of a linked ears reference in Thorpe et al. (1996, as

well as in Fabre-Thorpe et al., 2001), a method that tends to mask occipital activities in favor of frontal activities. In fact, later studies using an averaged reference showed that in parallel with the frontal differential activity, there was a clear differential activity with the opposite polarity at lateral occipito-temporal sites (Rousselet et al., 2002; VanRullen & Thorpe, 2001b). This bipolar arrangement can be seen clearly in Figure 1 of the present study. The close similarity between the onset times of these two different differential responses as well as source analysis using BESA is consistent with the idea that a considerable proportion of the differential responses at both frontal and occipito-temporal sites is produced by the same set of sources in occipitotemporal cortex (Delorme et al., accepted for publication; Fize et al., in revision). However, at least some of the later differential effects could depend specifically on activity in prefrontal areas.

What underlying processes could give rise to this differential activation in occipitotemporal areas? It is useful to distinguish at least three different potential causes, each characterizing activity at a particular level in the visual system. First, consider neurons at the earliest levels of the visual processing hierarchy, selective for relatively low level stimulus features such as contour orientation and the presence or absence of terminations. Suppose that we take a set of images from a given class (for example, photographs of human faces) and determine the average response of neurons in V1, and then do the same for another set of images from another class (for example, photographs of landscapes). If the images of landscapes contained a higher proportion of horizontal edges (for instance, because of the presence of a horizon), then a statistically significant difference between the average response to the two image classes might be present even though none of the neurons involved coded anything specific about either faces or landscapes. Attributing differential activity to a process related to categorization would in this case be an error.

Consider now what might happen if we were considering neurons at a later stage of visual processing that were selective to facial features. There is abundant evidence for such neurons from single unit recording studies in awake behaving monkeys where it is known that at least some neurons can respond selectively as a function of gaze direction (Perrett et al., 1992). Indeed, some reports suggest that the proportion of neurons selective for faces and facial features can reach as high as 20% in certain parts of the temporal lobe (Baylis et al., 1987). Clearly, if one was to measure the average response of this sort of population of neurons in response to the two different image categories (faces vs. landscapes), there could also be a strong difference in response. But in this case, the difference would have considerable significance for the task, because it would reflect the activity of populations of face selective cells that could clearly be involved in recognition and categorization.

Is there a way to distinguish between "interesting" and "uninteresting" differential activity? The methodology used by VanRullen and Thorpe and used again here provides one way of attacking the issue. By switching between two different target categories, the same images can be presented either as targets or distractors. If a difference still exists under these conditions, it is clear that no simple low level difference between the images could explain the effects because the two images sets are physically identical. The differential response curves plotted in Figure 6 show that all the experimental conditions produced effects with roughly the same form, but the point at which the effects became significant differed markedly. In the standard "animal/non-animal" task (experiment 1), clear differential effects emerged in the right hemisphere from 145 ms in the case of upright animals, and just slightly longer (149 ms) with inverted photographs (Figure 6, top right). This result thus reinforces the study by VanRullen and Thorpe (2001b), who also found significant effects with this type of analysis but at slightly longer latencies (156 ms). Together, such findings demonstrate clearly that

information related to the category must have started to be encoded by around 150 ms, as proposed by Thorpe et al. (1996).

However, the results for the other conditions were less clear. Thus, the comparison of responses to humans when targets with humans as non-targets in experiment 1 did not start to become significant until 185 ms. And in experiment 2, all the differential responses started later, with the earliest significant effects for animals not appearing until over 230 ms. This result is surprising because there is no obvious relationship between onset latency for this differential effect and the ability of the subjects to perform the task which was very similar in each case. The conclusion would seem to be that while this form of differential activity can (if successful) put a upper limit on the time required to extract a certain type of visual information, it does not necessarily provide a good predictor of when the subject will respond (it is an upper limit because there is always the possibility that earlier effects might not be captured by the ERP waveforms). If the differential activity was directly related to the decision process, one would expect that subjects would be as much as 80 ms slower at performing the task in experiment 2 than they were at detecting the presence of an animal in experiment 1, and yet this was very clearly not the case. Accuracy, mean reaction times and minimal reaction times were very similar for both tasks (Rousselet et al, 2003).

How could it be that subjects can perform the challenging visual task in experiment 2 without there being clear signs of task related activity in the ERP records? To understand this, consider again a population of "face-selective" cells in inferotemporal cortex. Let us suppose that these neurons have responses that are relatively "hard-wired" in that they will respond to the presence of a face essentially irrespective of the task being performed by the subject. In such a case, one can imagine that changing the target category for the subject might have little or no effect on the magnitude of the cumulative response of the neurons (no “type 2” differential activity would be observed). And yet, despite this, the neurons could still be perfectly well able to signal whether or not the scene contains a face. If the output of the neurons was being used to drive a decision mechanism (located perhaps in a brain area outside the visual processing pathways per se, such as prefrontal cortex), one could imagine that the subject could perform the task well without there being any clear sign of task-related differential effects in the visual system itself. On the other hand, a comparison of the responses to a wide set of distractor images with no faces present with the responses to images with a target present could well reveal clear differences because of the large number of face-selective cells that are activated.

Our suggestion is that with target categories such as faces that are processed very efficiently, there is no need for modulation of responsiveness within the visual system itself, with the result that no "type 2" differential effect would be visible.

Contrast this situation with an alternative processing model. Suppose that in order to reliably detect any one of a large number of animal forms, some form of top-down "priming" of neurons selective for particular animal features was required. The top-down priming would have the effect that the neurons would respond more strongly when the corresponding features were present, and this enhanced activation could be detected by a later decision stage. The increased response when a target was present in the scene might be visible at the level of the global ERP response because the amount of neural activity would be increased. However, in this case, changing the target category from "animal" to something else ("means of transport" as in the study by VanRullen & Thorpe (2001b), or "human" as in experiment 1 of the present study) would have the effect of removing the priming effect and revealing a “type 2” differential effect.

Note that both processing strategies would allow the subjects to perform the task reliably, but only when a top-down priming strategy is used would one expect to see changes in the responsiveness of neurons within the visual system as a function of the target class. Of course, in the absence of a task-dependent modulation, it is less easy to conclude that the differential activation seen between targets and distractors is necessarily related to higher level mechanisms related to categorization and recognition. However, it is interesting to note that with faces, the early type 1 differential effects tend to be considerably less long lasting than for animals, a result that might fit with the idea of a more hard-wired "automatic" processing in this case.

One criticism that has been raised concerning the relevance of the 150 ms type 1 differential effects is that there is no relation between the onset latency of the effect and behavioral reaction times. In the original Thorpe et al. (1996) study, it was shown that the differential activity at frontal sites has the same time course when the curves are plotted using average waveforms calculated for trials with fast reaction times as for slow trials. At the time, it seemed highly unlikely that the processing time required to analyze an image did not depend strongly on the nature of the image. But more recently, evidence has accumulated in favor of the view that processing time might in fact be relatively constant for many natural images. One argument comes from the study by Fabre-Thorpe et al. (2001) who found that the distribution of images with short reaction times was essentially random, as if the underlying neuronal mechanisms processed a relatively important part of the natural scenes with a fixed processing speed. Therefore, it is not impossible that high-level, categorization related, neuronal activity is actually reflected in ERP differential activity whose onset does not vary with RT.

In this context, we would like to argue that the differential activities recorded for humans and animals as early as 120-130 ms do not necessarily reflect low level physical differences, but might in fact be the signature of the early activation of high-level units coding for diagnostic properties in the image.

This is in contrast with the conclusion reached in the study by VanRullen & Thorpe (2001b). In their study, using the same animal categorization task as the one we used here, high-level representations were thought to be access not before 150 ms, as indexed by the latency of the first task status effects. However, it remains the possibility that what was attributed to low-level physical differences might in fact be high-level physical differences. Indeed, when the visual system is processing animals or humans, not exactly the same “high-level” neuronal populations are activated, which might be reflected early in ERP.

We must also leave open the possibility that task related top-down modulations, acting on high-level representations, cannot be captured by ERP recordings at the time they occur. Task effects at 150 ms might actually reflect later stages of visual processing.

Another point that must be considered in the present discussion is that “high-level” categorization does not necessarily imply that high-level representations are used to perform the task. It has been shown that “mid-level” representations can perfectly be used to perform this kind of classification, like the detection of faces in natural scenes (Ullman et al., 2002). Such “mid-level” representations might be used as diagnostic features in our task, allowing subjects to respond for the presence of high-level objects (Schyns, 1998). As this kind of features might well be processed in areas V4-TEO of the ventral pathway and activated by a feedforward wave of activation, this strengthen the hypothesis of a early “high-level” process of objects in natural scenes.

Furthermore, it has been suggested that the earliest evidence for coarse face processing might be found at around 120 ms (Itier & Taylor, 2002; Linkenkaer-Hansen et al., 1998). In keeping with this hypothesis, recent source analysis on ERP data have revealed that the fusiform gyrus, an area of the ventral pathway involved in

high-level object recognition, can be activated under 110 ms after stimulus onset (Di Russo et al., 2001; Martinez et al., 2001). It has also been suggested that such early activities might not be as "early" as generally thought because visual mechanisms in this time window might well be influenced by feedback from prefrontal cortex (Foxe & Simpson, 2002).

However, following this line of thinking, we do not mean that object categorization in natural scenes is achieved in 120-130 ms. Indeed, a significant difference between two ERP waveforms is not synonymous with the completion of the task by the visual system. What we mean is that by 120-130 ms after stimulus onset, it might well be that some objects are least coarsely categorized, or more generally speaking, that at the neuronal population level the categorization process has started.

In addition, this piece of data has also revealed that the fast categorization of objects in natural scenes is relatively unaffected by inversion. The shallower slope of differential activity recorded for inverted stimuli compared to upright ones reinforce the model of accumulation of evidence (Perrett et al., 1998) used previously in Rousselet et al. (2003) to explain how performance was affected by inversion in these tasks. This small effect of inversion suggests that the neuronal representations used to perform the task are relatively coarse, but this issue remains to be investigated more deeply. The data also suggest that stimuli like faces and humans form a very specific class of objects which are by default processed to a larger extent than other objects, for example animals in the present study (see discussion in Rousselet, Macé & Fabre-Thorpe, sous presse).

Relationship between RT and differential activity (draft section)

Differential activities were analyzed as a function of subjects' RT. For each subject, the RT histogram was divided into 3 equal parts. For each part, the corresponding target ERP were averaged separately. Then, distractor ERP were subtracted from target ERP to generate 3 types of differential activities corresponding to fast, medium and slow RT according to the RT distributions. In figures 7, 8, 9 and 10, these differential activities are reported as 1/3, 2/3, 3/3. Some of the best electrodes have been used to draw these figures. In each figure, the name of the electrode is indicated along with the onset of the differential activity in the 1/3, 2/3 and 3/3 conditions. The label "N.S." stands for non significant, which means that the t-test never exceeded criterion.

These preliminary analyses confirm that there is a clear relationship between behavioral RT and type 2 differential activity onset; it also demonstrates that such relationship exists in the case of type 1 differential activity, contrary to what was found by Thorpe et al.(1996) and Johnson & Olshausen (2003). Although these relationships exist, it must be noted that there is no direct mapping between RT latencies and differential activity onsets.

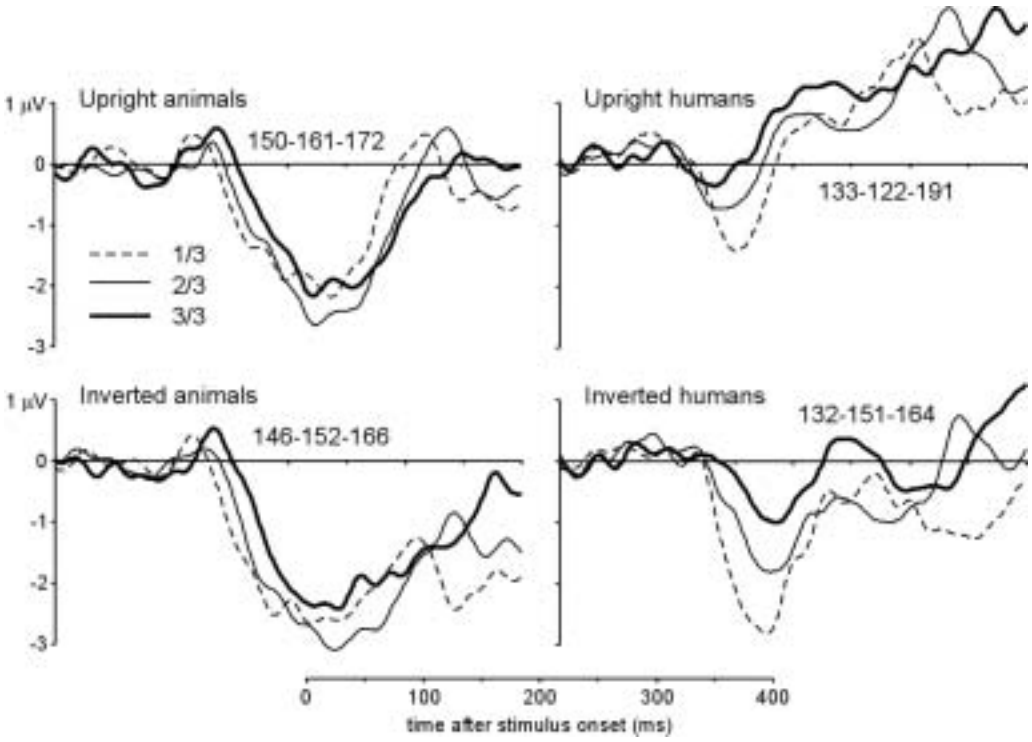


Figure 7. Type 1 differential activities as a function of RT in experiment 1 at electrode CB2.

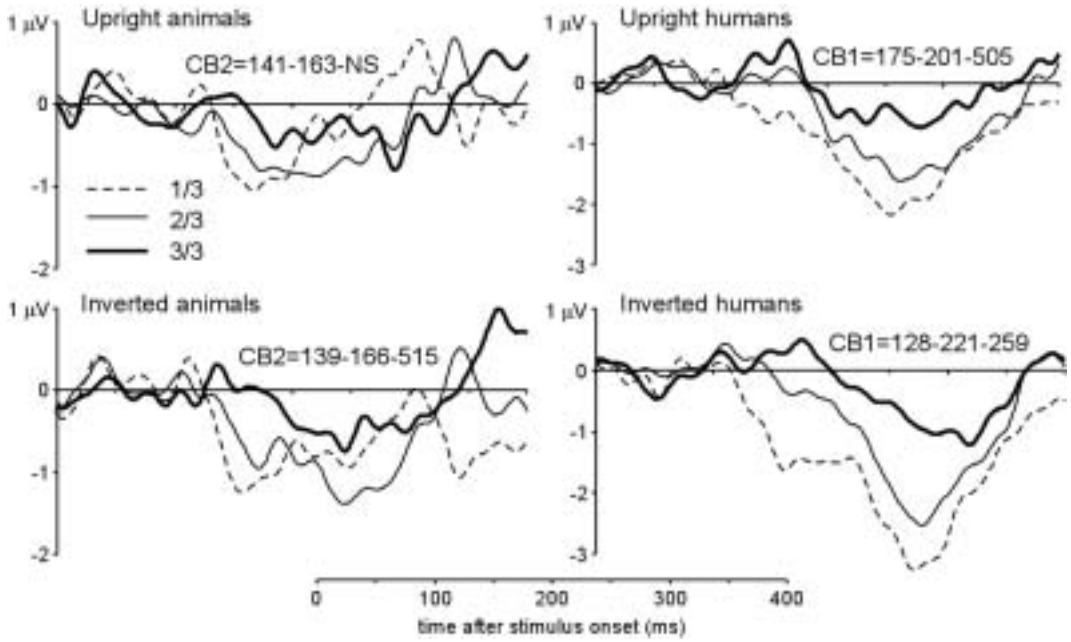


Figure 8. Type 2 differential activities as a function of RT in experiment 1.

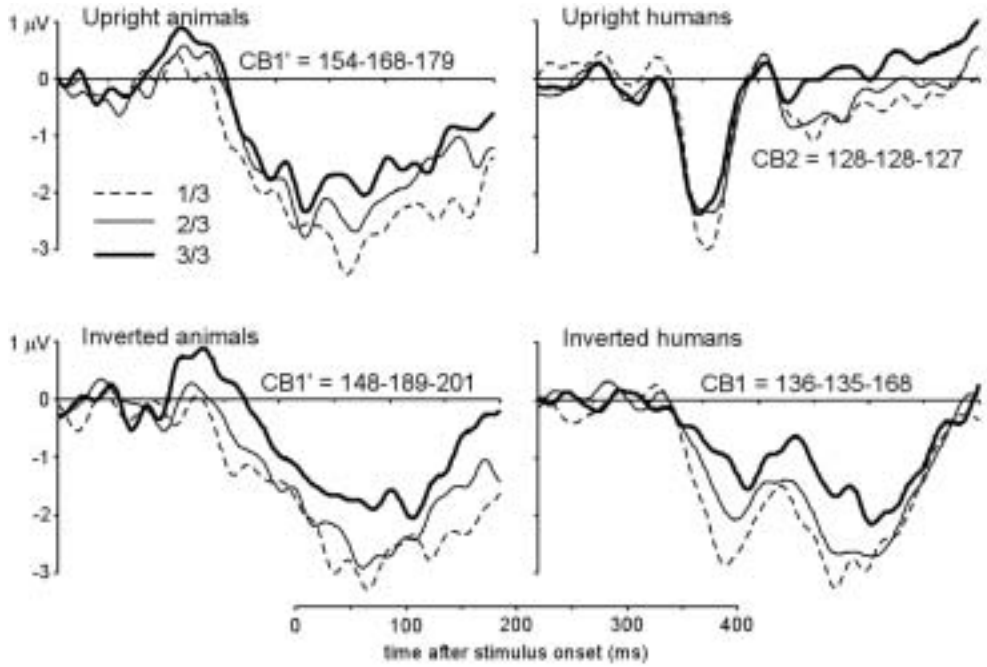


Figure 9. Type 1 differential activities as a function of RT in experiment 2.

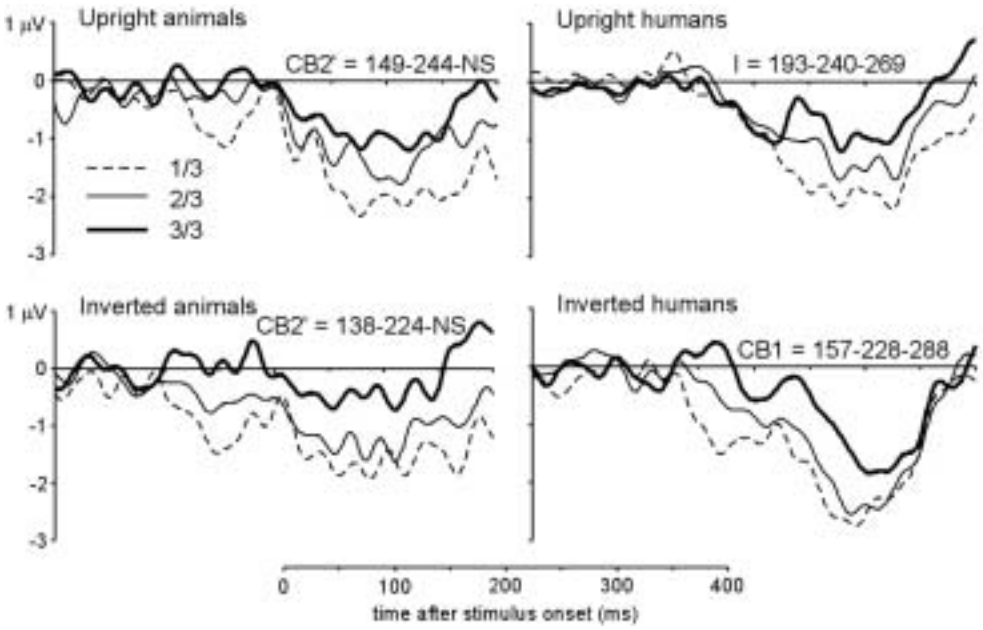


Figure 10. Type 2 differential activities as a function of RT in experiment 2.

Acknowledgements

Caitlin R. Sternberg and Anne-Sophie Paroissien are acknowledged for their help in running subjects in experiment 1 and 2 respectively. Thanks to Nadège M. Bacon for programming stimulus presentation in experiment 2. We also thank Rufin VanRullen for several brainstorming sessions about these data.

References

Allison, T., Puce, A., Spencer, D. D., & McCarthy, G. (1999). Electrophysiological studies of human face perception. I: Potentials generated in occipitotemporal cortex by face and non-face stimuli. *Cereb Cortex, 9*(5), 415-430.

Antal, A., Keri, S., Kovacs, G., Janka, Z., & Benedek, G. (2000). Early and late components of visual categorization: an event-related potential study. *Brain Res Cogn Brain Res, 9*(1), 117-119.

Baylis, G. C., Rolls, E. T., & Leonard, C. M. (1987). Functional subdivisions of the temporal lobe neocortex. *J Neurosci, 7*(2), 330-342.

Delorme, A., Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (accepted for publication). Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Cog Brain Res*.

Di Russo, F., Martinez, A., Sereno, M. I., Pitzalis, S., & Hillyard, S. A. (2002). Cortical sources of the early components of the visual evoked potential. *Hum Brain Mapp, 15*(2), 95-111.

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience, 13*(2), 171-180.

Fabre-Thorpe, M., Richard, G., & Thorpe, S. J. (1998). Rapid categorization of natural images by rhesus monkeys. *Neuroreport, 9*(2), 303-308.

Foxe, J. J., & Simpson, G. V. (2002). Flow of activation from V1 to frontal cortex in humans. A framework for defining "early" visual processing. *Exp Brain Res, 142*(1), 139-150.

Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science, 182*(108), 177-180.

Hillyard, S. A., & Münte, T. F. (1984). Selective attention to color and location: an analysis with event- related brain potentials. *Percept Psychophys, 36*(2), 185-198.

Itier, R. J., & Taylor, M. J. (2002). Inversion and Contrast Polarity Reversal Affect both Encoding and Recognition Processes of Unfamiliar Faces: A Repetition Study Using ERPs. *Neuroimage, 15*(2), 353-372.

Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. *Journal of Vision, 3*, 499-512.

Kreiman, G., Koch, C., & Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat Neurosci, 3*(9), 946-953.

Linkenkaer-Hansen, K., Palva, J. M., Sams, M., Hietanen, J. K., Aronen, H. J., & Ilmoniemi, R. J. (1998). Face-selective processing in human extrastriate cortex around 120 ms after stimulus onset revealed by magneto- and electroencephalography. *Neurosci Lett, 253*(3), 147-150.

Martinez, A., DiRusso, F., Anllo-Vento, L., Sereno, M. I., Buxton, R. B., & Hillyard, S. A. (2001). Putting spatial attention on the map: timing and localization of stimulus selection processes in striate and extrastriate visual areas. *Vision Res, 41*(10-11), 1437-1457.

Perrett, D. I., Hietanen, J. K., Oram, M. W., & Benson, P. J. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philos Trans R Soc Lond B Biol Sci, 335*(1273), 23-30.

Perrett, D. I., Oram, M. W., & Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. *Cognition, 67*(1-2), 111-145.

Rossion, B., & Gauthier, I. (2002). How does the brain process upright and inverted faces? *Behavioral and Cognitive Neuroscience Reviews, 1*(1), 62-74.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nat Neurosci, 5*(7), 629-630.

Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *Journal of Vision, 3*(6), 440-455.

Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (in revision). The N170 ERP component: specificity, effects of task status and inversion for animal and human faces in natural scenes. *Journal of Vision*.

Schyns, P. G. (1998). Diagnostic recognition: task constraints, object information, and their interactions. *Cognition, 67*(1-2), 147-179.

Sheinberg, D. L., & Logothetis, N. K. (2001). Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. *J Neurosci, 21*(4), 1340-1350.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu Rev Neurosci, 19*, 109-139.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520-522.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Comput. Neural Syst., 14*, 391-412.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nat Neurosci, 5*(7), 682-687.

VanRullen, R., & Thorpe, S. J. (2001a). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception, 30*(6), 655-668.

VanRullen, R., & Thorpe, S. J. (2001b). The time course of visual processing: from early perception to decision-making. *J Cogn Neurosci, 13*(4), 454-461.

Article 7

Animal and human faces in natural scenes: how specific to human faces is the N170 ERP component?

Rousselet, G.A., Macé, M.J.-M. & Fabre-Thorpe M.

(sous presse, *Journal of Vision*)

Résultats électrophysiologiques de 24 sujets adultes dans une expérience portant sur la catégorisation de visages d'être humains et d'animaux présentés en gros plan dans des scènes naturelles à l'endroit et à l'envers. Les résultats présentés dans cet article concernent la composante N170 des potentiels évoqués précoces. Cette composante qui semble tout particulièrement sensible aux visages est ici décrite pour la première fois dans le cadre des scènes naturelles, sans faire appel à des objets isolés. Elle est aussi décrite pour la première fois dans le cadre d'une tâche go/no-go.

Seules les 2 premières pages de l'article ont été insérées dans cette thèse. La version complète est disponible gratuitement en format pdf à l'adresse http://journalofvision.org/.

Introduction

La N170 est une composante des potentiels évoqués souvent considérée comme spécifique, ou très sensible aux visages (voir chapitre 2). Certains ont par exemple fait l'hypothèse quelle pourrait refléter un mécanisme d'encodage structural permettant la détection des visages (e.g. Carmel & Bentin, 2002 ; Eimer, 2000a). Cependant, la N170 est toujours décrite pour des visages isolés et aux propriétés physiques relativement homogènes. Le but de l'article 7 était de fournir une description de la N170 dans un contexte plus écologique, celui des scènes naturelles. Les données analysées correspondent à celles recueillies au cours de l'expérience 2 décrite dans les articles 5 et 6, dans laquelle des gros plans de visages d'humains étaient comparés à des gros plans de visages d'animaux. L'alternance entre deux tâches de catégorisation

permettait d'évaluer l'impact de la tâche sur la N170. Enfin, la présentation d'images à l'endroit et à l'envers permettait d'évaluer l'effet d'inversion dans les scènes naturelles.

Résultats

1) Il y avait une N170 très nette pour des visages présentés dans le contexte de scènes naturelles.

2) L'amplitude de la N170 n'était pas différente entre les visages d'humains et d'animaux vus à l'endroit. Par contre sa latence de pic était légèrement plus tardive pour les visages d'animaux.

3) L'inversion affectait de manière importante l'amplitude de la N170 dans le cas des visages d'humains, mais pas dans le cas des visages d'animaux. La latence de pic était plus grande pour les deux catégories de stimuli vus à l'envers. Dans le cas de scènes naturelles contrôles, qui servaient de distracteurs dans les deux tâches de catégorisation, l'inversion entraînait une augmentation de l'amplitude de la N170, mais pas de sa latence.

4) Le statut cible ou distracteur des stimuli n'avait aucun effet sur la N170.

Discussion

La N170 et les effets de l'inversion sur cette onde ne semblent pas spécifiques des visages d'êtres humains. Ce qui semble spécifique des visages humains est la force de l'effet d'inversion. La N170 pourrait donc recevoir une interprétation plus large, étant sensible à une large gamme de stimuli présentant un arrangement de type visage.

La grande similarité entre la N170 pour les visages d'humains et d'animaux pourrait recevoir une interprétation alternative. En effet, Schyns et al. (2003) ont récemment proposé que la N170 pourrait constituer une réponse automatique aux yeux, indépendamment de la tâche à effectuer. Cette explication pourrait s'appliquer aux présents résultats dans la mesure où les animaux utilisés avaient tous des yeux visibles. Cette hypothèse est particulièrement intéressante car des données récentes suggèrent que la N170 pour des visages aurait une origine corticale différente de celle de la N1 des objets (Itier & Taylor, sous presse). Elle pourrait ainsi refléter en grande partie l'activité de zones corticales intéressées par des éléments faciaux importants pour la communication sociale. Un autre élément en faveur de l'hypothèse de Schyns et al. (2003) est

fourni par une expérience de Bentin et al. (2002) montrant que la présentation de 2 points isolés engendre une N170 lorsqu'ils ont au préalable été présentés dans le contexte d'un visage, mais pas après la présentation d'objets contrôles. Il semble donc que quand deux points sont perçus comme des yeux, ils sont associés à une N170, alors que ce n'est pas le cas pour les mêmes stimuli perçus comme de simples points. Cependant il reste à expliquer pourquoi une N170 est toujours évoquée en réponse à des visages d'humains pour lesquels les yeux ont été enlevés (Eimer, 1998). Ceci s'accorde très bien avec l'hypothèse selon laquelle la N170 aurait deux générateurs principaux, l'un impliqué dans le traitement des visages, l'autre dans le traitement des yeux (Itier & Taylor, sous presse ; Taylor et al., 2001). Une autre interprétation du résultat de Eimer (1998) pourrait être que la région du visage dont les yeux sont absents constitue un stimulus suffisant pour déclencher des mécanismes mis habituellement en jeu par les yeux. De plus, une autre expérience d'Eimer (2000a) montre que l'amplitude de la N170 est considérablement réduite pour des vues de têtes humaines dont la région des yeux n'est pas visible. L'hypothèse d'une réponse automatique aux yeux, peut-être médiatisée par une région corticale latérale particulièrement intéressée par les stimuli faciaux, constitue donc une interprétation alternative valide des résultats de l'article 7. Cette hypothèse n'a pas été discutée dans l'article 7 qui se voulait très court. Elle sera par contre présentée dans un prochain article portant sur la N170 enregistrée au cours de la première expérience dans laquelle des visages d'humains et d'animaux à différentes échelles servaient de stimuli. Des analyses préliminaires de ces données suggèrent que l'amplitude de la N170 serait fortement modulée par la taille des visages. Etant donné que dans cette expérience les visages plus petits avaient aussi tendance à être plus excentrés, cet effet d'amplitude peut aussi être un effet d'excentricité. En effet, Eimer (2000d) a rapporté que des visages excentrés sont associés à une N170 beaucoup moins ample que des visages centrés. Un tel effet d'excentricité pourrait être mis en relation avec l'existence du biais fovéal mis en évidence par Levy et al. (2001). Il semble en effet que des objets que nous avons l'habitude d'analyser en détails, tels que des visages, ont une représentation corticale beaucoup plus sensible à des stimulations centrales que périphériques. La N170 pourrait ainsi être en grande partie le reflet de l'activité de populations de neurones particulièrement sensibles aux yeux dans la partie centrale du champ visuel. Ceci n'est bien sûr qu'une hypothèse. Si elle s'avère être correcte, il sera particulièrement intéressant de déterminer la spécificité des mécanismes sous jacents, étant donné que les résultats présentés dans l'article 7 indiquent qu'ils pourraient être

recrutés par des visages d'animaux très variés. Même si de nouvelles pistes apparaissent pour interpréter l'origine de la N170, le plus difficile à comprendre reste tout de même l'effet d'inversion, qui, par son importance, semble bel et bien spécifique des visages humains. Personne n'a encore fourni d'explication satisfaisante de ce phénomène. Il pourrait par exemple être très intéressant d'évaluer dans quelle mesure l'effet d'inversion est contraint par des facteurs spatiaux tels que le biais fovéal pour les visages. On pourrait imaginer qu'il disparaîtrait avec l'augmentation de l'excentricité. Cette hypothèse fera prochainement l'objet d'une expérience. Si un tel biais existe, cela pourrait fortement contraindre les hypothèses par rapport aux mécanismes sous-jacents.

Un autre point important mis en évidence dans la présente expérience est l'absence d'effet de la tâche sur la N170. Les tracés électrophysiologiques indiquent que cet effet est visible à des latences ultérieures, ce que confirment l'article 6 sur les activités différentielles. Comme cela a été fait dans l'article 6, on pourrait interpréter ce résultat dans le cadre d'un modèle où les stimuli faciaux, au sens large, sont par défaut analysés plus en détails que d'autres objets (par exemple des animaux dans l'expérience 1, associés à des activités différentielles liées à la tâche dès 150 ms).

Finalement, l'hypothèse présentée dans l'article 6 selon laquelle une catégorisation grossière des stimuli humains et animaux pourrait commencer vers 120 ms s'oppose au modèle qui associe N170 et détection des visages (e.g. Carmel & Bentin, 2002 ; Eimer, 2000a).

Animal and human faces in natural scenes: how specific to human faces is the N170 ERP component?

Guillaume A. Rousselet *Centre de Recherche Cerveau & Cognition, CNRS-UPS UMR 5549, Toulouse, France

Marc J.-M. Macé Centre de Recherche Cerveau & Cognition, CNRS-UPS UMR 5549, Toulouse, France

Michèle Fabre-Thorpe Centre de Recherche Cerveau & Cognition, CNRS-UPS UMR 5549, Toulouse, France

*Current address: Department of Psychology, McMaster University, Hamilton, ON, Canada

The N170 is an event-related potential component reported to be very sensitive to human face stimuli. This study investigated the specificity of the N170, as well as its sensitivity to inversion and task status when subjects had to categorize either human or animal faces in the context of upright and inverted natural scenes. A conspicuous N170 was recorded for both face categories. Pictures of animal faces were associated with a N170 of similar amplitude compared to pictures of human faces, but with delayed peak latency. Picture inversion enhanced N170 amplitude for human faces and delayed its peak for both human and animal faces. Finally, whether processed as targets or non-targets, depending on the task, both human and animal face N170 were identical. Thus, human faces in natural scenes elicit a clear but non-specific N170 that is not modulated by task status. What appears to be specific to human faces is the strength of the inversion effect.

Keywords: N170, event-related potentials, rapid visual categorization, natural scenes, human faces, animal faces

Introduction

Several studies using event-related potentials (ERPs) have isolated a component, the N170, which appears to reflect a stage of visual processing at which objects are categorized. This component is a negative potential peaking at around 150-170 ms over lateral occipito-temporal electrodes. It is generally larger and peaks earlier in response to human faces compared to many other object categories (Bentin, Allison, Puce, Perez, & McCarthy, 1996; Carmel & Bentin, 2002; George, Evans, Fiori, Davidoff, & Renault, 1996; Rossion et al., 2000; Sagiv & Bentin, 2001; Taylor, Edmonds, McCarthy, & Allison, 2001). The N170 is very sensitive to human faces and some authors have suggested that it reflects their early structural encoding before face recognition processes take place (e.g., Eimer, 1998, 2000a; Sagiv & Bentin, 2001). However, these conclusions are drawn from experiments that have mainly used central presentations of isolated and homogeneous stimuli (with the exception of Eimer, [2000b], for example, who used peripheral presentations). Here we report the results from an experiment in which we investigated whether a N170 can be found for faces in the more realistic context of natural scenes. To this end, subjects were requested to categorize as fast and as accurately as possible human faces in briefly flashed photographs of natural scenes. For comparison, they performed a control task in which they had to categorize animal faces under the same conditions. According to previous reports, a N170 of larger amplitude was expected in response to human faces compared to animal faces.

The N170 has also been found to be particularly affected by face inversion, contrary to other object categories. It is delayed for inverted faces compared to upright faces (Bentin et al., 1996; Eimer, 2000c; Itier & Taylor, 2002; Rebai, Poiroux, Bernard, & Lalonde, 2001; Rossion et al., 1999; Rossion et al., 2000). It is also delayed for faces with eyes removed (Eimer, 1998), during the analysis of single face components (Bentin et al., 1996; Jemel, George, Chaby, Fiori, & Renault, 1999), or when attention is directed to alphanumeric strings superimposed on the center of the face (Eimer, 2000c). N170 amplitude has been found to be larger in response to inverted than upright faces (Itier & Taylor, 2002; Rossion et al., 1999, 2000; Sagiv & Bentin, 2001). In relation with the behavioral literature, the effects of inversion on the N170 have been interpreted as reflecting the disruption of processing of the spatial relationships between face components (configural information; see

more details in Itier & Taylor, 2002; Maurer, Le Grand, & Mondloch, 2002; Rossion & Gauthier, 2002). Hence, normal face perception would rely on mechanisms dedicated to the processing of upright face configural information. However, an enhancement of N170 amplitude has also been found for inverted houses (Eimer, 2000c), and various categories of real world objects (Itier, Latinus, & Taylor, 2003); and an increase in latency has been reported for cars and words (Rossion, Joyce, Cottrell, & Tarr, in press), suggesting that the inversion effect might not be face specific (unlike results found by Rossion et al., 2000). In this study, we wanted to determine whether an inversion effect would occur with human and animal faces in natural scenes. To address this issue, half of the pictures were presented in an upright position, the other half were presented upside-down. According to some previous reports (Bentin et al., 1996; de Haan, Pascalis, & Johnson, 2002; Rebai et al., 2001; Rossion et al., 2000), an inversion effect was expected on the N170 for pictures containing a human face but not for those containing an animal face. However, a small inversion effect in response to animal faces was also possible given those found for various object categories (Eimer, 2000c; Itier et al., 2003; Rossion et al., in press).

Finally, there is a controversy in the literature about whether the N170 can be modulated by task requirements, for example, when faces are given a target task status versus a non-target task status. Among the few studies that investigated this aspect, some have reported that the N170 does not seem to be modulated by task requirements (Carmel & Bentin, 2002; Séverac-Cauquil, Edmonds, & Taylor, 2000). However, top-down effects have also been reported on the N170, indicating that the neural mechanisms indexed by the N170 are not totally immune from high-level control (Bentin & Golland, 2002; Bentin, Sagiv, Mecklinger, & von Cramon, 2002; Eimer, 2000b, 2000c). To investigate this issue in the present experiment, targets of a given task were used as non-targets in the other task. For example, when subjects performed the human face categorization task, half of the non-targets were pictures of animal faces (and vice versa). We were thus able to compare the N170 elicited by a given category of faces when processed either as target or as non-target.

To summarize, the present study was designed to assess the specificity of the N170 for human faces in natural scenes as well as its sensitivity to inversion and to task status in such a context.

Methods

Participants

Twenty-four participants were tested (12 women and 12 men, mean age 30 years, ranging from 19 to 51 years; 3 of them were left handed). They volunteered in this study and gave their written informed consent. All participants had normal or corrected-to-normal vision.

Experimental Procedure

Subjects sat in a dimly lit room at 100 cm from a computer screen (resolution, 800 x 600 pixels, vertical refresh rate, 75 Hz) controlled by a PC computer. To start a block of trials, they had to place their finger on a response pad for one second. A trial was organized as follows: a fixation cross (0.1° of visual angle) appeared for a 300-900 ms random duration and was immediately followed by the stimulus presented for two frames (i.e., about 23 ms in the middle of the screen). Participants had to lift their finger as quickly and as accurately as possible (go response) each time a target was presented. Responses were detected using infrared diodes. Subjects had 1000 ms to lift their finger, after which their response was considered a no-go response. A black screen remained for 300 ms following this maximum response time delay, before the fixation point was presented again for a variable duration, resulting in a random 1600-2200 ms inter-trial interval. When the photographs contained no target, subjects had to keep their finger on the pad for at least 1000 ms (no-go response).

Subjects alternated between two categorization tasks, processing either human faces or animal faces as targets. They were asked to respond as fast as possible while minimizing errors. Each task consisted of one block of four consecutive series of 96 trials each. Half of the subjects performed the human face task first, while the other half started with the animal face task. Before each task, subjects were given a 48-trial training session.

All series of pictures (Figure 1) contained 50% targets and 50% non-targets. Among non-targets, half were neutral non-targets that had to be processed as such in both tasks and half were targets of the other task (i.e., human faces when subjects performed the animal face task and animal faces when they performed the human face task. Moreover, half of the images for each condition were presented upright while the other half were presented upside-down (rotation 180°).

A given subject saw each image only once, with one orientation (upright or inverted) and one status (target or non-target), but the design was counterbalanced for all conditions across the set of subjects to allow all data comparisons without any bias over the group of subjects or the sets of images.

Stimuli

We used photographs of natural scenes taken from a large commercial CD-ROM library (Corel Stock Photo Libraries 1 and 2; e.g., see Figure 1). All photographs were horizontal (768 x 512 pixels, sustaining about 19.8°

Chapitre 2 : conclusion générale

Le chapitre 2 était consacré à l'évaluation de l'hypothèse selon laquelle des mécanismes spécifiques seraient mis en jeu pour le traitement des visages humains. Des arguments convaincants existent en effet montrant qu'il pourrait y avoir un 'module' de traitement des visages dans la voie ventrale. Ce module pourrait analyser les visages sur la base de mécanismes spécifiques, distincts de ceux mis en œuvre pour analyser d'autres catégories d'objets. Ceci pourrait avoir comme conséquence une analyse beaucoup plus rapide des visages. Cependant, des hypothèses alternatives permettent d'expliquer certains résultats en faveur d'un module sans avoir recours à des mécanismes spécifiques. Il sera passionnant dans un avenir proche d'évaluer dans quelle mesure ces hypothèses alternatives sont valides.

Le travail expérimental présenté dans les articles 5, 6 et 7 de cette thèse a permis de préciser dans quelle mesure les visages sont des objets spécifiques lorsqu'ils sont présentés dans le contexte de photographies de scènes naturelles. Tout d'abord, comme le montre l'article 5, cela ne semble pas être en termes de vitesse de traitement. Il semble en effet que toutes les catégories d'objets avec lesquelles nous sommes familiers puissent être analysées particulièrement rapidement, en mettant en jeu des mécanismes parallèles et essentiellement vers l'avant de traitement de l'information. Ce résultat est confirmé par l'analyse de données électrophysiologiques (article 6), suggérant que les effets très précoces, inférieurs à 100 ms, rapportés dans la littérature sur les visages, sont le reflet de différences physiques bas niveau. Cependant, il est suggéré que certains effets vers 120 ms pourraient être le reflet d'un traitement rapide et peu sophistiqué des visages d'humains et d'autres catégories d'objets bien apprises, tels que des animaux. Par contre, les visages au sens large, incluant des visages d'animaux, semblent être traités par défaut de manière plus détaillée que d'autres objets, comme le montrent à la fois l'analyse des activités différentielles et de la N170. De plus, contrairement à certaines hypothèses, la N170 ne semble pas spécifique des visages d'êtres humains mais semble être sensible à toutes sortes de stimuli présentant une configuration faciale. Cette sensibilité pourrait s'expliquer par la mise en œuvre d'aires corticales intéressées par des attributs visuels permettant la communication sociale, notamment la région des yeux. Finalement, seule l'ampleur de l'effet d'inversion au niveau de la N170 semble réellement spécifique des visages d'êtres humains. Il devient donc important d'élucider la nature exacte des mécanismes neuronaux sous-tendant cet effet.

Perspectives

Pour conclure, je voudrais aborder brièvement quelques points importants concernant le prolongement des travaux présentés dans cette thèse.

L'analyse de la littérature présentée au chapitre 1 m'a conduit à formuler l'hypothèse d'un traitement en parallèle des informations visuelles dans la voie ventrale. C'est également ce qu'indiquent les résultats de l'article 1 et dans une moindre mesure ceux de l'article 2. Mais une expérience cruciale reste à effectuer pour montrer l'existence d'un traitement réellement parallèle. L'expérience rapportée dans l'article 1 devrait être répliquée en ajoutant une condition dans laquelle 2 cibles apparaissent simultanément. En présence de deux images, l'une cible l'autre distracteur, une compétition pour gagner le contrôle de la réponse motrice semblait affecter l'amplitude de l'activité différentielle enregistrée en frontal. Avec deux cibles, la réponse motrice demandée étant la même, l'activité différentielle devrait apparaître à la même latence et avec la même amplitude que dans la condition "1 cible" au niveau des électrodes placées en regard de chacun des hémisphères. Cela constituerait un argument fort en faveur de l'hypothèse d'un parallélisme inter-hémisphérique. Ces expériences sur le parallélisme doivent aussi être poursuivies en intégrant des analyses de sources rigoureuses afin de tester la validité du modèle de sélection tardive. Dans le cadre de la seconde expérience sur le parallélisme comportant jusqu'à 4 images, il serait intéressant de tester quelques sujets sur un nombre très important d'essais afin d'augmenter le rapport signal sur bruit et de mettre en évidence d'éventuels effets d'apprentissage.

En ce qui concerne la vitesse de traitement de différentes catégories d'objets, il sera important dans de futures expériences de varier la diagnosticité des cibles de manière plus systématique afin d'évaluer son effet sur l'activité différentielle. Ceci pourrait être fait en variant les tâches demandées aux sujets et la nature relative des stimuli cibles et distracteurs. Une telle approche devrait à terme permettre d'élucider les incongruités soulevées dans l'article 6 et de mieux cerner les contraintes temporelles qui pèsent sur l'analyse des objets et du contexte, comme cela a été discuté dans les articles 3 et 4. Des expériences couplant cette approche à l'étude des facteurs spatiaux contraignant l'analyse des objets dans le champ visuel permettraient

enfin d'unifier les champs d'investigations abordés dans les chapitres 1 et 2, et pourraient nous renseigner sur la nature des mécanismes mis en jeu dans le traitement des visages.

Finalement, une des préoccupations principale de mon travail de thèse était la vitesse de traitement dans le système visuel. Il faut souligner que la forte énergie contenue dans les images flashées pourrait être à l'origine d'un raccourcissement de la vitesse de traitement dans les protocoles utilisés. Il serait très important de connaître précisément la différence pour le système visuel entre une image flashée et une image sur laquelle les yeux se « posent » après une saccade. L'utilisation d'images flashées trouve cependant sa justification dans sa similitude avec l'acte d'ouvrir les yeux sur une scène inconnue (ou d'allumer subitement la lumière dans une pièce sombre, ou de zapper devant sa télévision ou encore de feuilleter rapidement un magasine). D'autre part, un protocole dans lequel les sujets ont les yeux fermés et les ouvrent brusquement sur une scène déjà à l'écran est beaucoup plus difficile à mettre en œuvre, surtout pour établir une référence fiable du début de la stimulation, ce que fournit par contre l'usage d'images flashées. Il est déjà possible de tester des sujets dans des conditions plus naturelles en ayant recours à des appareils de suivis des mouvements oculaires. Il sera bientôt possible d'aller plus loin, lorsque des techniques telles que l'immersion dans des environnements 3D virtuels pourront être couplées avec l'enregistrement de l'EEG. Les expériences consisteront par exemple à se promener 'librement' dans une forêt et à détecter le plus rapidement possible la présence d'un animal.

Pour conclure, je voudrais souligner que les expériences réalisées au cours de cette thèse, même si elles ont vocation à nous renseigner sur le traitement des scènes naturelles, sont limitées par l'usage de photographies de scènes naturelles. En effet, même s'il existe des indices tridimensionnels dans les photographies, l'idéal serait de pouvoir tester des sujets dans des environnements 3D virtuels. Etant donné la vitesse de propagation de l'influx nerveux dans la voie magnocellulaire, particulièrement sensible aux indices 3D, les caractéristiques spatiales des objets telles que la profondeur et la disparité rétinienne pourraient fortement contraindre l'analyse des scènes visuelles. Plonger des sujets dans des environnements virtuels réalistes devraient permettre de poser de la manière la plus écologique possible des questions aujourd'hui très débattues telles que l'influence du contexte sur la perception des objets.

Bibliographie

Aguirre, G. K., & D'Esposito, M. (1999). Topographical disorientation: a synthesis and taxonomy. *Brain, 122*(9), 1613-1628.

Aguirre, G. K., Zarahn, E., & D'Esposito, M. (1998). An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. *Neuron, 21*(2), 373-383.

Allison, T., McCarthy, G., Nobre, A., Puce, A., & Belger, A. (1994). Human extrastriate visual cortex and the perception of faces, words, numbers, and colors. *Cerebral Cortex, 4*(5), 544-554.

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences, 4*(7), 267-278.

Allison, T., Puce, A., Spencer, D. D., & McCarthy, G. (1999). Electrophysiological studies of human face perception. I: Potentials generated in occipitotemporal cortex by face and non-face stimuli. *Cerebral Cortex, 9*(5), 415-430.

Anllo-Vento, L., Luck, S. J., & Hillyard, S. A. (1998). Spatio-temporal dynamics of attention to color: evidence from human electrophysiology. *Human Brain Mapping, 6*, 216-238.

Arguin, M., Cavanagh, P., & Joanette, Y. (1994). Visual feature integration with an attention deficit. *Brain & Cognition, 24*(1), 44-56.

Arguin, M., Joanette, Y., & Cavanagh, P. (1993). Visual search for feature and conjunction targets with an attention deficit. *Journal of Cognitive Neuroscience, 5*, 436-452.

Ashbridge, E., Cowey, A., & Wade, D. (1999). Does parietal cortex contribute to feature binding? *Neuropsychologia, 37*(9), 999-1004.

Ashbridge, E., Perrett, D. I., Oram, M. W., & Jellema, T. (2000). Effect of image orientation and size on objects recognition: responses of single units in the macaque monkey temporal cortex. *Cognitive Neuropsychology, 17*(1/2/3), 13-34.

Ashbridge, E., Walsh, V., & Cowey, A. (1997). Temporal aspects of visual search studied by transcranial magnetic stimulation. *Neuropsychologia, 35*, 1121-1131.

Baas, J. M. P., Kenemans, J. L., Böcker, K. B. E., & Verbaten, M. N. (2002). Threat-induced cortical processing and startle potentiation. *NeuroReport, 13*(1), 133-137.

Bacon-Macé, N., Macé, M. J.-M., Fabre-Thorpe, M., & Thorpe, S. J. (soumis). The time course of visual processing: backward masking and natural scene categorization.

Bar, M., & Aminoff, E. (2003). Cortical analysis of visual context. *Neuron, 38*(2), 347-358.

Barcelo, F., Suwazono, S., & Knight, R. T. (2000). Prefrontal modulation of visual processing in humans. *Nature Neuroscience, 3*(4), 399-403.

Barlow, H. B. (1972). Single units and sensation: a neuron doctrine for perceptual psychology? *Perception, 1*(4), 371-394.

Baylis, G. C., Rolls, E. T., & Leonard, C. M. (1987). Functional subdivisions of the temporal lobe neocortex. *Journal of Neuroscience, 7*(2), 330-342.

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience, 8*, 551-565.

Bentin, S., & Carmel, D. (2002). Accounts for the N170 face-effect: a reply to Rossion, Curran, & Gauthier. *Cognition, 85*(2), 197-202.

Bentin, S., & Deouell, Y. (2000). Structural encoding and identification in face processing: ERP evidence for separate mechanisms. *Cognitive Neuropsychology, 17*, 35-54.

Bentin, S., & Golland, Y. (2002). Meaningful processing of meaningless stimuli: the influence of perceptual experience on early visual processing of faces. *Cognition, 86*(1), B1-14.

Bentin, S., Mouchetant-Rostaing, Y., Giard, M. H., Echallier, J. F., & Pernier, J. (1999). ERP manifestations of processing printed words at different psycholinguistic levels: time course and scalp distribution. *Journal of Cognitive Neuroscience, 11*(3), 235-260.

Bentin, S., Sagiv, N., Mecklinger, A., Friederici, A., & von, C. Y. (2002). Priming visual face-processing mechanisms: electrophysiological evidence. *Psychological Science, 13*(2), 190-193.

Biederman, I. (1972). Perceiving real-world scenes. *Science, 177*, 77-80.

Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization*. Hillsdale, NJ: Erlbaum.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review, 94*(2), 115-147.

Biederman, I. (1988). Aspects and extensions of a theory of human image understanding. In Z. W. Pylyshyn (Ed.), *Computational processes in human vision: an interdisciplinary perspective* (pp. 370-428). Norwood (N.J.): Ablex.

Biederman, I., Glass, A. L., & Stacy, E. W., Jr. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology, 97*(1), 22-27.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*(2), 143-177.

Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W., Jr. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology, 103*(3), 597-600.

Blackmore, G. B., Nelson, K., & Trosciansko, T. (1995). Is the richness of our visual world an illusion? Transsaccadic memory for complex scenes. *Perception, 24*, 1075-1081.

Blake, R., & Logothetis, N. K. (2002). Visual competition. *Nature Reviews Neuroscience, 3*(1), 13-21.

Booth, M. C., & Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex, 8*(6), 510-523.

Braeutigam, S., Bailey, A. J., & Swithenby, S. J. (2001). Task-dependent early latency (30-60 ms) visual processing of human faces and other objects. *Neuroreport, 12*(7), 1531-1536.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3), 305-327.

Bullier, J. (2001). Integrated model of visual processing. *Brain Research Reviews, 36*(2-3), 96-107.

Bullier, J. (2003). Communications between cortical areas of the visual system. In L. M. Chalupa & J. S. Werner (Eds.), *The Visual neurosciences* (Vol. 1). Cambridge, MA: MIT Press.

Bullier, J., & Nowak, L. G. (1995). Parallel versus serial processing: new vistas on the distributed organization of the visual system. *Current Opinion in Neurobiology, 5*(4), 497-503.

Bundesen, C. (1998). A computational theory of visual attention. *Philosophical Transactions of the Royal Society of London series B: Biological Sciences, 353*(1373), 1271-1281.

Bunge, S. A., Hazeltine, E., Scanlon, M. D., Rosen, A. C., & Gabrieli, J. D. E. (2002). Dissociable Contributions of Prefrontal and Parietal Cortices to Response Selection. *NeuroImage, 17*, 1562–1571.

Carmel, D., & Bentin, S. (2002). Domain specificity versus expertise: factors influencing distinct processing of faces. *Cognition, 83*(1), 1-29.

Carrasco, M., Evert, D. L., Chang, I., & Katz, S. M. (1995). The eccentricity effect: target eccentricity affects performance on conjunction searches. *Perception & Psychophysics, 57*(8), 1241-1261.

Cave, K. R. (1999). The FeatureGate model of visual selection. *Psychological Research, 62*(2-3), 182-194.

Chelazzi, L. (1999). Serial attention mechanisms in visual search: a critical look at the evidence. *Psychological Research, 62*(2-3), 195-219.

Chelazzi, L., Duncan, J., Miller, E. K., & Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *Journal of Neurophysiology, 80*(6), 2918-2940.

Chelazzi, L., Miller, E. K., Duncan, J., & Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature, 363*(6427), 345-347.

Chelazzi, L., Miller, E. K., Duncan, J., & Desimone, R. (2001). Responses of neurons in macaque area V4 during memory-guided visual search. *Cerebral Cortex, 11*(8), 761-772.

Cheng, L., & Tarr, M. (2003). What can computational simulations tell us about the double dissociation between face and object recognition? *Journal of Cognitive Neuroscience supplement, C289*, p.118.

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences, 4*(5), 170-178.

Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance, 21*(1), 109-127.

Corbetta, M., Miezin, F. M., Shulman, G. L., & Petersen, S. E. (1993). A PET study of visuospatial attention. *Journal of Neuroscience, 13*, 1202-1226.

Corbetta, M., Shulman, G. L., Miezin, F. M., & Petersen, S. E. (1995). Superior parietal cortex activation during spatial attention shifts and visual feature conjunction. *Science, 270*(5237), 802-805.

Culham, J. C., & Kanwisher, N. G. (2001). Neuroimaging of cognitive functions in human parietal cortex. *Current Opinion in Neurobiology, 11*(2), 157-163.

Damasio, A. R. (1999). *Le Sentiment même de soi*. Paris: Odile Jacob.

Davis, G., Driver, J., Pavani, F., & Shepherd, A. (2000). Reappraising the apparent costs of attending to two separate visual objects. *Vision Research, 40*(10-12), 1323-1332.

Debruille, J. B., Guillem, F., & Renault, B. (1998). ERPs and chronometry of face recognition: following-up Seeck et al. and George et al. *Neuroreport, 9*(15), 3349-3353.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition, 79*(1-2), 1-37.

Dehaene, S., Naccache, L., Cohen, L., Le Bihan, D., Mangin, J.-F., Poline, J.-B., & Rivière, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience, 4*(7), 752-758.

Dehaene, S., Naccache, L., Le Clec, H. G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P. F., & Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature, 395*(6702), 597-600.

Delorme, A., Richard, G., & Fabre-Thorpe, M. (2000). Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans. *Vision Research, 40*(16), 2187-2200.

Delorme, A., Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (sous presse). Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Cognitive Brain Research*.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proceedings of the National Academy of Sciences of the United States of America, 93*(24), 13494-13499.

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of London series B: Biological Sciences, 353*(1373), 1245-1255.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review Neuroscience, 18*, 193-222.

Di Russo, F., Martinez, A., Sereno, M. I., Pitzalis, S., & Hillyard, S. A. (2002). Cortical sources of the early components of the visual evoked potential. *Human Brain Mapping, 15*(2), 95-111.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Eperimental Psychology: General, 115*, 107-117.

DiCarlo, J. J., & Maunsell, J. H. R. (2003). Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *Journal of Neurophysiology, 89*, 3264–3278.

Ditterich, J., Mazurek, M. E., & Shadlen, M. N. (2003). Microstimulation of visual cortex affects the speed of perceptual decisions. *Nature Neuroscience, 6*(8), 891-898.

Dobbins, A. C., Jeo, R. M., Fiser, J., & Allman, J. M. (1998). Distance modulation of neural activity in the visual cortex. *Science, 281*(5376), 552-555.

Donk, M. (1999). Illusory conjunctions are an illusion: the effects of target-nontarget similarity on conjunction and feature errors. *Journal of Experimental Psychology: Human Perception and Performance, 25*(5), 1207-1233.

Donk, M. (2001). Illusory conjunctions die hard: a reply to Prinzmetal, Diedrichsen, and Ivry (2001). *Journal of Experimental Psychology: Human Perception and Performance, 27*(3), 542-546.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science, 293*(5539), 2470-2473.

Driver, J., & Vuilleumier, P. (2001). Perceptual awareness and its loss in unilateral neglect and extinction. *Cognition, 79*(1-2), 39-88.

Duncan, J. (1998). Converging levels of analysis in the cognitive neuroscience of visual attention. *Philosophical Transactions of the Royal Society of London series B: Biological Sciences, 353*(1373), 1307-1317.

Duncan, J., Humphreys, G., & Ward, R. (1997). Competitive brain activity in visual attention. *Current Opinion in Neurobiology, 7*(2), 255-261.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96*(3), 433-458.

Duncan, J., Ward, R., & Shapiro, K. (1994). Direct measurement of attention dwell time in human vision. *Nature, 369*(6478), 313-314.

Dunn, J. C. (2003). The elusive dissociation. *Cortex, 39*, 177-179.

Dunn, J. C., & Kirsner, K. (2003). What can we infer from double dissociations? *Cortex, 39*, 1-7.

Eckstein, M. P. (1998). The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science, 9*(2), 111-118.

Edelman, G. M., & Tononi, G. (2000). *Comment la matière devient conscience*. Paris: Odile Jacob.

Edelman, S., & Intrator, N. (2000). (Coarse coding of shape fragments) + (retinotopy) approximately = representation of structure. *Spatial Vision, 13*(2-3), 255-264.

Edwards, R., Xiao, D., Keysers, C., Foldiak, P., & Perrett, D. (2003). Color sensitivity of cells responsive to complex stimuli in the temporal cortex. *J Neurophysiol, 90*(2), 1245-1256.

Egeth, H. E., Virzi, R. A., & Garbart, H. (1984). Searching for conjunctively defined targets. *Journal of Experimental Psychology: Human Perception and Performance, 10*(1), 32-39.

Eimer, M. (1998). Does the face-specific N170 component reflect the activity of a specialized eye processor? *Neuroreport, 9*(13), 2945-2948.

Eimer, M. (2000a). The face-specific N170 component reflects late stages in the structural encoding of faces. *Neuroreport, 11*(10), 2319-2324.
Eimer, M. (2000b). Effects of face inversion on the structural encoding and recognition of faces. Evidence from event-related brain potentials. *Cognitive Brain Research, 10*(1-2), 145-158.
Eimer, M. (2000c). Event-related brain potentials distinguish processing stages involved in face perception and recognition. *Clinical Neurophysiology, 111*, 694-705.
Eimer, M. (2000d). Attentional modulations of event-related brain potentials sensitive to faces. *Cognitive Neuropsychology, 17*(1/2/3), 103-116.
Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience, 17*(5), 1089-1097.
Elliffe, M. C., Rolls, E. T., & Stringer, S. M. (2002). Invariant recognition of feature combinations in the visual system. *Biological Cybernetics, 86*(1), 59-71.
Ellison, A., & Walsh, V. (1998). Perceptual learning in visual search: some evidence of specificities. *Vision Research, 38*(3), 333-345.
Enns, J. T., & Rensink, R. A. (1990). Influence of scene-based properties on visual search. *Science, 247*(4943), 721-723.
Enns, J. T., & Rensink, R. A. (1991). Preattentive recovery of three-dimensional orientation from line drawings. *Psychological Review, 98*(3), 335-351.
Epstein, R., DeYoe, E. A., Press, D. Z., Rosen, A. C., & Kanwisher, N. (2001). Neuropsychological evidence for a topographical learning mechanism in parahippocampal cortex. *Cognitive Neuropsychology, 18*, 481–508.
Epstein, R., Graham, K. S., & Downing, P. E. (2003). Viewpoint-specific scene representations in human parahippocampal cortex. *Neuron, 37*(5), 865-876.
Epstein, R., Harris, A., Stanley, D., & Kanwisher, N. (1999). The parahippocampal place area: recognition, navigation, or encoding? *Neuron, 23*(1), 115-125.
Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*(6676), 598-601.
Erickson, C. A., Jagadeesh, B., & Desimone, R. (2000). Clustering of perirhinal neurons with similar properties following visual experience in adult monkeys. *Nature Neuroscience, 3*(11), 1143-1148.
Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience, 13*(2), 171-180.
Fabre-Thorpe, M., Fize, D., Richard, G., & Thorpe, S. (1998). Rapid categorization of extrafoveal natural images: implications for biological models. In J. Bower (Ed.), *Computational Neuroscience: Trends in Research* (pp. 7-12). New-York: Plenum Press.
Fabre-Thorpe, M., Richard, G., & Thorpe, S. J. (1998). Rapid categorization of natural images by rhesus monkeys. *Neuroreport, 9*(2), 303-308.
Farah, J. M. (1990). *Visual agnosia: disorders of object recognition and what they tell us about normal vision*. Cambridge: MIT Press.
Farah, J. M., & Aguirre, G. K. (1999). Imaging visual recognition: PET and fMRI studies of the functional anatomy of human visual recognition. *Trends in Cognitive Sciences, 3*(5), 179-186.
Farah, M. J., Tanaka, J. W., & Drain, H. M. (1995). What causes the face inversion effect? *Journal of Experimental Psychology: Human Perception and Performance, 21*, 628-634.
Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review, 105*(3), 482-498.
Fernandez-Duque, D., & Thornton, I. M. (2000). Change detection without awareness: do explicit reports underestimate the representation of change in the visual system? *Visual Cognition, 7*(1/2/3), 323-344.
Fize, D., Boulanouar, K., Chatel, Y., Ranjeva, J. P., Fabre-Thorpe, M., & Thorpe, S. (2000). Brain areas involved in rapid categorization of natural images: An event-related fMRI study. *Neuroimage, 11*(6), 634-643.
Fize, D., Fabre-Thorpe, M., Richard, G., Doyon, B., & Thorpe, S. (en révision). Foveal vision is not necessary for rapid categorisation of natural images: a behavioural and ERP study.
Fodor, J. (1983). *The Modularity of mind: an essay on faculty psychology*. Cambridge, MA: MIT Press.
Földiák, P. (2002). Sparse coding in the primate cortex. In M. A. Arbib (Ed.), *Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press.
Foxe, J. J., & Simpson, G. V. (2002). Flow of activation from V1 to frontal cortex in humans. A framework for defining "early" visual processing. *Experimental Brain Research, 142*(1), 139-150.

Friedman, A., & Campell Polson, M. (1981). Hemispheres as independent resource systems: limited-capacity processing and cerebral specialization. *Journal of Experimental Psychology: Human Perception and Performance, 7*(5), 1031-1058.

Friedman-Hill, S. R., Robertson, L. C., & Treisman, A. (1995). Parietal contributions to visual feature binding: evidence from a patient with bilateral lesions. *Science, 269*(5225), 853-855.

Fries, P., Reynolds, J. H., Rorie, A. E., & Desimone, R. (2001). Modulation of oscillatory neuronal synchronization by selective visual attention. *Science, 291*(5508), 1560-1563.

Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Cognitive Brain Research, 16*(2), 123-144.

Gauthier, I. (2000). What constrains the organization of the ventral temporal cortex? *Trends in Cognitive Sciences, 4*(1), 1-2.

Gauthier, I., Behrmann, M., & Tarr, M. J. (1999). Can face recognition really be dissociated from object recognition? *Journal of Cognitive Neuroscience, 11*(4), 349-370.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience, 3*(2), 191-197.

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: exploring mechanisms for face recognition. *Vision Research, 37*(12), 1673-1682.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience, 2*(6), 568-573.

Gauthier, I., Tarr, M. J., Moylan, J., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). The fusiform "face area" is part of a network that processes faces at the individual level. *Journal of Cognitive Neuroscience, 12*(3), 495-504.

Gawne, T. J., Kjaer, T. W., & Richmond, B. J. (1996). Latency: another potential code for feature binding in striate cortex. *Journal of Neurophysiology, 76*(2), 1356-1360.

Gegenfurtner, K. R. (2003). Cortical mechanisms of colour vision. *Nature Reviews Neuroscience, 4*(7), 563 -572.

Gegenfurtner, K. R., & Kiper, D. C. (2003). Color Vision. *Annual Review of Neuroscience, 27*, 27.

George, N., Evans, J., Fiori, N., Davidoff, J., & Renault, B. (1996). Brain events related to normal and moderately scrambled faces. *Cognitive Brain Research, 4*(2), 65-76.

George, N., Jemel, B., Fiori, N., & Renault, B. (1997). Face and shape repetition effects in humans: a spatio-temporal ERP study. *Neuroreport, 8*(6), 1417-1423.

Gilbert, C., Ito, M., Kapadia, M., & Westheimer, G. (2000). Interactions between attention, context and learning in primary visual cortex. *Vision Research, 40*(10-12), 1217-1226.

Gottlieb, J. (2002). Parietal mechanisms of target representation. *Current Opinion in Neurobiology, 12*(2), 134-140.

Gray, C. M. (1999). The temporal correlation hypothesis of visual feature integration: still alive and well. *Neuron, 24*(1), 31-47, 111-125.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research, 41*(10-11), 1409-1422.

Grimes, J. (1996). On the failure to detect changes in scenes across saccades. In K. Akins (Ed.), *Perception* (pp. 89-110). New York: Oxford University Press.

Gross, C. G., Bender, D. B., & Rocha-Miranda, C. E. (1969). Visual receptive fields of neurons in inferotemporal cortex of the monkey. *Science, 166*(910), 1303-1306.

Guillaume, F., & Tiberghien, G. (2001). An event-related potential study of contextual modifications in a face recognition task. *Neuroreport, 12*(6), 1209-1216.

Halgren, E., Raij, T., Marinkovic, K., Jousmaki, V., & Hari, R. (2000). Cognitive response profile of the human fusiform face area as determined by MEG. *Cerebral Cortex, 10*(1), 69-81.

Hanes, D. P., & Schall, J. D. (1996). Neural control of voluntary movement initiation. *Science, 274*(5286), 427-430.

Hatfield, G. (1998). Attention in early scientific psychology. In R. D. Wright (Ed.), *Visual attention* (Vol. 8, pp. 3-25). Oxford (UK): Oxford University Press.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425-2430.

Haxby, J. V., Grady, C. L., Horwitz, B., Ungerleider, L. G., Mishkin, M., Carson, R. E., Herscovitch, P., Schapiro, M. B., & Rapoport, S. I. (1991). Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proceedings of the National Academy of Sciences of the United States of America, 88*(5), 1621-1625.

Haxby, J. V., Ungerleider, L. G., Clark, V. P., Schouten, J. L., Hoffman, E. A., & Martin, A. (1999). The effect of face inversion on activity in human neural systems for face and object perception. *Neuron, 22*(1), 189-199.

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision, 3*(1), 49-63.

He, Z. J., & Nakayama, K. (1992). Surfaces versus features in visual search. *Nature, 359*(6392), 231-233.

He, Z. J., & Nakayama, K. (1994). Perceiving textures: beyond filtering. *Vision Research, 34*(2), 151-162.

Heinze, H. J., Mangun, G. R., Burchert, W., Hinrichs, H., Scholz, M., Munte, T. F., Gos, A., Scherg, M., Johannes, S., Hundeshagen, H., & al., e. (1994). Combined spatial and temporal imaging of brain activity during visual selective attention in humans. *Nature, 372*(6506), 543-546.

Henderson, J., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50*, 243-271.

Henderson, J. M. (1992). Object identification in context: the visual processing of natural scenes. *Canadian Journal of Psychology, 46*(3), 319-341.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50*, 243-271.

Henderson, J. M., & Hollingworth, A. (2003). Eye movements and visual memory: detecting changes to saccade targets in scenes. *Perception & Psychophysics, 65*(1), 58-71.

Hilgetag, C. C., Théoret, H., & Pascual-Leone, A. (2001). Enhanced visual spatial attention ipsilateral to rTMS-induced 'virtual lesions' of human parietal cortex. *Nature Neuroscience, 4*(9), 953-957.

Hillyard, S. A., & Anllo-Vento, L. (1998). Event-related brain potentials in the study of visual selective attention. *Proceedings of the National Academy of Sciences of the United States of America, 95*(3), 781-787.

Hillyard, S. A., Vogel, E. K., & Luck, S. J. (1998). Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. *Philosophical Transactions of the Royal Society of London series B: Biological Sciences, 353*(1373), 1257-1270.

Hines, R. J., Paul, L. K., & Brown, W. S. (2002). Spatial attention in agenesis of the corpus callosum: shifting attention between visual fields. *Neuropsychologia, 40*(11), 1804-1814.

Hinkle, D. A., & Connor, C. E. (2002). Three-dimensional orientation tuning in macaque area V4. *Nature Neuroscience, 5*(7), 665-670.

Hochberg, J. (1986). Representation of motion and space in video and cinematic displays. In K. J. Boff & L. Kaufman & J. P. Thomas (Eds.), *Handbook of perception and human performance* (Vol. 1, pp. 22:21-22:64). New York: John Wiley & Sons.

Hollingworth, A. (2003). Failures of retrieval and comparison constrain change detection in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance, 29*(2), 388-403.

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General, 127*(4), 398-415.

Hollingworth, A., & Henderson, J. M. (1999). Object identification is isolated from scene semantic constraint: evidence from object type and token discrimination. *Acta Psychologica, 102*(2-3), 319-343.

Hollingworth, A., & Henderson, J. M. (2000). Semantic informativeness mediates the detection of changes in natural scenes. *Visual Cognition, 7*(1/2/3), 213-235.

Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance, 28*(1), 113-136.

Hollingworth, A., Schrock, G., & Henderson, J. M. (2001). Change detection in the flicker paradigm: the role of fixation position within the scene. *Memory & Cognition, 29*(2), 296-304.

Hollingworth, A., Williams, C. C., & Henderson, J. M. (2001). To see and remember: visually specific information is retained in memory from previously attended objects in natural scenes. *Psychonomic Bulletin & Review, 8*(4), 761-768.

Hopf, J. M., Luck, S. J., Girelli, M., Hagner, T., Mangun, G. R., Scheich, H., & Heinze, H. J. (2000). Neural sources of focused attention in visual search. *Cerebral Cortex, 10*(12), 1233-1241.

Hopf, J.-M., Boelmans, K., Schoenfeld, A. M., Heinze, H.-J., & Luck, S. J. (2002a). How does attention attenuate target–distractor interference in vision? Evidence from magnetoencephalographic recordings. *Cognitive Brain Research, 15*, 17-29.

Hopf, J.-M., & Mangun, G. R. (2000). Shifting visual attention in space: an electrophysiological analysis using high spatial resolution mapping. *Clinical Neurophysiology, 111*, 1241-1257.

Hopf, J.-M., Vogel, E., Woodman, G., Heinze, H.-J., & Luck, S. J. (2002b). Localizing Visual Discrimination Processes in Time and Space. *Journal of Neurophysiology, 88*, 2088–2095.

Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature, 394*(6693), 575-577.

Horowitz, T. S., & Wolfe, J. M. (2001). Search for multiple targets: remember the targets, forget the search. *Perception & Psychophysics, 63*(2), 272-285.

Humphreys, G., & Forde, E. (2003). *Hierarchies, similarity and interactivity in object recognition: On the multiplicity of 'category-specific' deficits in neuropsychological populations*. Available: http://psg275.bham.ac.uk/school_information/humphreysg/ghmanus2.htm [2003, 20 août].

Humphreys, G. W., & Riddoch, M. J. (2001). Detection by action: neuropsychological evidence for action-defined templates in search. *Nature Neuroscience, 4*(1), 84-88.

Husain, M., Shapiro, K., Martin, J., & Kennard, C. (1997). Abnormal temporal dynamics of visual attention in spatial neglect patients. *Nature, 385*(6612), 154-156.

Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 604-610.

Intraub, H. (1984). Conceptual masking: the effects of subsequent visual events on memory for pictures. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10*, 115-125.

Intraub, H. (1999). Understanding and remembering briefly glimpsed pictures: implications for visual scanning and memory. In V. Coltheart (Ed.), *Fleeting memories* (pp. 47-70). Cambridge, Massachusetts: MIT Press.

Intraub, H., & Richardson, M. (1989). Wide-angle memories of close-up scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(2), 179-187.

Itier, R. J. (2002). *Perception et reconnaissance des visages non familiers chez l'adulte et l'enfant : étude neurophysiologique du traitement de la configuration.* Manuscrit de thèse non publié, Paul Sabatier, Toulouse.

Itier, R. J., Latinus, M., & Taylor, M. J. (2003). Effects of inversion, contrast-reversal and their conjunction on face, eye and object processing: an ERP study. *Journal of Cognitive Neuroscience supplement, D292*, p154.

Itier, R. J., & Taylor, M. J. (2002). Inversion and Contrast Polarity Reversal Affect both Encoding and Recognition Processes of Unfamiliar Faces: A Repetition Study Using ERPs. *Neuroimage, 15*(2), 353-372.

Itier, R. J., & Taylor, M. J. (sous presse). N170 or N1? Spatiotemporal differences between object and face processing using ERPs. *Cerebral Cortex*.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*(3), 194-203.

Jeffreys, D. (1996). Evoked potential studies of face and object processing. *Visual Cognition, 3*, 1-38.

Jemel, B., Pisani, M., Calabria, M., Crommelinck, M., & Bruyer, R. (2003). Is the N170 for faces cognitively penetrable? Evidence from repetition priming of Mooney faces of familiar and unfamiliar persons. *Cognitive Brain Research, 17*(2), 431-446.

Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. *Journal of Vision, 3*, 499-512.

Joseph, J. S., Chun, M. M., & Nakayama, K. (1997). Attentional requirements in a 'preattentive' feature search task. *Nature, 387*(6635), 805-807.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience, 3*(8), 759-763.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*(11), 4302-4311.

Kanwisher, N., Yin, C., & Wojciulik, E. (1999). Repetition blindness for pictures: evidence for the rapid computation of abstract visual descriptions. In V. Coltheart (Ed.), *Fleeting memories* (pp. 119-150). Cambridge, Massachusetts: MIT Press.

Karayanidis, F., & Michie, P. T. (1996). Frontal processing negativity in a visual selective attention task. *Electroencephalogr Clin Neurophysiol, 99*(1), 38-56.

Karayanidis, F., & Michie, P. T. (1997). Evidence of visual processing negativity with attention to orientation and color in central space. *Electroencephalogr Clin Neurophysiol, 103*(2), 282-297.

Karnath, H.-O., Ferbera, S., & Bülthoff, H. H. (2000). Neuronal representation of object orientation. *Neuropsychologia, 38*, 1235-1241.

Kastner, S., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1998). Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science, 282*(5386), 108-111.

Kastner, S., & Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annual Review Neuroscience, 23*, 315-341.

Kelley, T. A., Chun, M. M., & Chua, K.-P. (2003). Effects of scene inversion on change detection of targets matched for visual salience. *Journal of Vision, 3*(1), 1-5.

Kenemans, J. L., Lijffijt, M., Camfferman, G., & Verbaten, M. N. (2002). Split-second sequential selective activation in human secondary visual cortex. *Journal of Cognitive Neuroscience, 14*(1), 48-61.

Keysers, C., & Perrett, D. I. (2002). Visual masking and RSVP reveal neural competition. *Trends in Cognitive Sciences, 6*(3), 120-125.

Keysers, C., Xiao, D. K., Foldiak, P., & Perrett, D. I. (2001). The speed of sight. *Journal of Cognitive Neuroscience, 13*(1), 90-101.

Kim, M.-S., & Cave, K. R. (1995). Spatial attention in visual search for features and feature conjunctions. *Psychological Science, 6*, 376-380.

Kimchi, R. (2000). The perceptual organization of visual objects: a microgenetic analysis. *Vision Research, 40*(10-12), 1333-1347.

Kinchla, R. A. (1992). Attention. *Annual Review of Psychology, 43*, 711-742.

Kleffner, D. A., & Ramachandran, V. S. (1992). On the perception of shape from shading. *Perception & Psychophysics, 52*(1), 18-36.

Kline, K., Amador-Garza, S., McAdams, C., Maunsell, J., & Sereno, A. (2003). Spatial and eye position modulation of neuronan activity in anterior inferior temporal and perirhinal cortices. *Journal of Cognitive Neuroscience supplement, E293*, p.188.

Kreiman, G., Fried, I., & Koch, C. (2002). Single-neuron correlates of subjective vision in the human medial temporal lobe. *Proceedings of the National Academy of Sciences of the United States of America, 99*(12), 8378–8383.

Kreiman, G., Koch, C., & Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience, 3*(9), 946-953.

Kubovy, M., Cohen, D. J., & Hollier, J. (1999). Feature integration that routinely occurs without focal attention. *Psychonomic Bulletin & Review, 6*(2), 183-203.

Laeng, B., & Caviness, V. S. (2001). Prosopagnosia as a deficit in encoding curved surface. *Journal of Cognitive Neuroscience, 13*(5), 556-576.

Lamme, V. A. (2003). Why visual attention and awareness are different. *Trends in Cognitive Sciences, 7*(1), 12-18.

Lange, J. J., Wijers, A. A., Mulder, L. J., & Mulder, G. (1998). Color selection and location selection in ERPs: differences, similarities and 'neural specificity'. *Biol Psychol, 48*(2), 153-182.

Lemaire, P. (1999). *Psychologie Cognitive*. Paris, Bruxelles: De Boeck Université.

Levy, I., Hasson, U., Avidan, G., Hendler, T., & Malach, R. (2001). Center-periphery organization of human object areas. *Nature Neuroscience, 4*(5), 533-539.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences of the United States of America, 99*(14), 9596-9601.

Linkenkaer-Hansen, K., Palva, J. M., Sams, M., Hietanen, J. K., Aronen, H. J., & Ilmoniemi, R. J. (1998). Face-selective processing in human extrastriate cortex around 120 ms after stimulus onset revealed by magneto- and electroencephalography. *Neuroscience Letters, 253*(3), 147-150.

Liu, J., Harris, A., & Kanwisher, N. (2002). Stages of processing in face perception: an MEG study. *Nature Neuroscience, 5*(9), 910-916.

Liu, J., Higuchi, M., Marantz, A., & Kanwisher, N. (2000). The selectivity of the occipitotemporal M170 for faces. *Neuroreport, 11*(2), 337-341.

Logothetis, N., & Sheinberg, D. (1996). Visual object recognition. *Annual Review Neuroscience, 19*, 577-621.

Logothetis, N. K. (1998). Single units and conscious vision. *Philosophical Transactions of the Royal Society of London series B: Biological Sciences, 353*, 1801-1818.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience, 19*, 577-621.

Luck, S. J., Chelazzi, L., Hillyard, S. A., & Desimone, R. (1997a). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of Neurophysiology, 77*(1), 24-42.

Luck, S. J., Fan, S., & Hillyard, S. A. (1993). Attention-related modulation of sensory-evoked brain activity in a visual search task. *Journal of Cognitive Neuroscience, 5*, 188-195.

Luck, S. J., Girelli, M., McDermott, M. T., & Ford, M. A. (1997b). Bridging the gap between monkey neurophysiology and human perception: an ambiguity resolution theory of visual selective attention. *Cognitive Psychology, 33*(1), 64-87.

Luck, S. J., & Hillyard, S. A. (1994a). Electrophysiological correlates of feature analysis during visual search. *Psychophysiology, 31*(3), 291-308.

Luck, S. J., & Hillyard, S. A. (1994b). Spatial filtering during visual search: evidence from human electrophysiology. *Journal of Experimental Psychology: Human Perception and Performance, 20*(5), 1000-1014.

Luck, S. J., & Hillyard, S. A. (1995). The role of attention in feature detection and conjunction discrimination: an electrophysiological analysis. *International Journal of Neuroscience, 80*(1-4), 281-297.

Luck, S. J., Hillyard, S. A., Mangun, G. R., & Gazzaniga, M. S. (1989). Independent hemispheric attentional systems mediate visual search in split-brain patients. *Nature, 342*(6249), 543-545.

Luck, S. J., Hillyard, S. A., Mangun, G. R., & Gazzaniga, M. S. (1994). Independent attentional scanning in the separated hemispheres of split-brain patients. *Journal of Cognitive Neuroscience, 6*(1), 84-91.

Luck, S. J., Vogel, E. K., & Shapiro, K. L. (1996). Word meanings can be accessed but not reported during the attentional blink. *Nature, 382*, 616-618.

Macé, M. J.-M., Fabre-Thorpe, M., & Thorpe, S. J. (2002). How robust is rapid visual categorization of natural images to large variations of contrast? *Journal of Cognitive Neuroscience supplement, A108*, 40.

Macé, M. J.-M., Thorpe, S. J., & Fabre-Thorpe, M. (en préparation). Category-level hierarchy: what comes first in vision.

Mack, A., & Rock, I. (1998). *Inattentional Blindness* (Vol. 1). Cambridge, Mass.: The MIT Press.

Maguire, E. A., Frith, C. D., & Cipolotti, L. (2001). Distinct neural systems for the encoding and recognition of topography and faces. *Neuroimage, 13*(4), 743-750.

Maki, W. S., Frigen, K., & Paulson, K. (1997). Associative priming by targets and distractors during rapid serial visual presentation: does word meaning survive the attentional blink? *Journal of Experimental Psychology: Human Perception and Performance, 23*, 1014-1034.

Mangun, G. R. (1995). Neural mechanisms of visual selective attention. *Psychophysiology, 32*(1), 4-18.

Marois, R., Chun, M. M., & Gore, J. C. (2000). Neural correlates of the attentional blink. *Neuron, 28*, 299–308.

Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.

Martinez, A., DiRusso, F., Anllo-Vento, L., Sereno, M. I., Buxton, R. B., & Hillyard, S. A. (2001). Putting spatial attention on the map: timing and localization of stimulus selection processes in striate and extrastriate visual areas. *Vision Research, 41*(10-11), 1437-1457.

Mattingley, J. B., Driver, J., Beschin, N., & Robertson, I. H. (1997). Attentional competition between modalities: Extinction between touch and vision after right hemisphere damage. *Neuropsychologia, 35*(6), 867-880.

McCarthy, G., Puce, A., Belger, A., & Allison, T. (1999). Electrophysiological studies of human face perception. II: Response properties of face-specific potentials generated in occipitotemporal cortex. *Cerebral Cortex, 9*(5), 431-444.

McCarthy, G., Puce, A., Gore, J. C., & Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience, 9*, 605–610.

McElree, B., & Carrasco, M. (1999). The temporal dynamics of visual search: evidence for parallel processing in feature and conjunction searches. *Journal of Experimental Psychology: Human Perception and Performance, 25*(6), 1517-1539.

McLeod, P., Driver, J., & Crisp, J. (1988). Visual search for a conjunction of movement and form is parallel. *Nature, 332*(6160), 154-155.

Mel, B., & Fiser, J. (2000). Minimizing binding errors using learned conjunctive features. *Neural Computation, 12*(2), 247-278.

Mendez, M. F., & Cherrier, M. M. (2003). Agnosia for scenes in topographagnosia. *Neuropsychologia, 41*, 1387–1395.

Missal, M., Vogels, R., Li, C. Y., & Orban, G. A. (1999). Shape interactions in macaque inferior temporal neurons. *Journal of Neurophysiology, 82*(1), 131-142.

Missal, M., Vogels, R., & Orban, G. A. (1997). Responses of macaque inferior temporal neurons to overlapping shapes. *Cerebral Cortex, 7*(8), 758-767.

Moore, T., & Armstrong, K. M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature, 421*(6921), 370-373.

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science, 229*(4715), 782-784.

Moscovitch, M., Winocur, G., & Behrmann, M. (1997). What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of Cognition Neuroscience, 9*(5), 555-604.

Mouchetant-Rostaing, Y., Giard, M. H., Bentin, S., Aguera, P. E., & Pernier, J. (2000a). Neurophysiological correlates of face gender processing in humans. *European Journal of Neuroscience, 12*(1), 303-310.

Mouchetant-Rostaing, Y., Giard, M. H., Delpuech, C., Echallier, J. F., & Pernier, J. (2000b). Early signs of visual categorization for biological and non-biological stimuli in humans. *Neuroreport, 11*(11), 2521-2525.

Mumford, D. (1991). On the computational architecture of the neocortex. I. The role of the thalamo-cortical loop. *Biological Cybernetics, 65*(2), 135-145.

Murphy, T. D., & Eriksen, C. W. (1987). Temporal changes in the distribution of attention in the visual field in response to precues. *Perception & Psychophysics, 42*(6), 576-586.

Murray, E. A., & Richmond, B. J. (2001). Role of perirhinal cortex in object perception, memory, and associations. *Current Opinion in Neurobiology, 11*(2), 188-193.

Murray, M. M., Foxe, J. J., Higgins, B. A., Javitt, D. C., & Schroeder, C. E. (2001). Visuo-spatial neural response interactions in early cortical processing during a simple reaction time task: a high-density electrical mapping study. *Neuropsychologia, 39*, 828-844.

Naccache, L., Blandin, E., & Dehaene, S. (2002). Unconscious masked priming depends on temporal attention. *Psychological Science, 13*(5), 416-424.

Nagy, A. L., & Sanchez, R. R. (1990). Critical color differences determined with a visual search task. *Journal of the Optical Society of America A, 7*(7), 1209-1217.

Nakamura, K., Kawashima, R., Sato, N., Nakamura, A., Sugiura, M., Kato, T., Hatano, K., Ito, K., Fukuda, H., Schormann, T., & Zilles, K. (2000). Functional delineation of the human occipito-temporal areas related to face and scene processing. A PET study. *Brain, 123*(9), 1903-1912.

Nakayama, K., & Silverman, G. H. (1986). Serial and parallel processing of visual feature conjunctions. *Nature, 320*(6059), 264-265.

Nakayama, K. I. (1990). The iconic bottleneck and the tenuous link between early visual processing and perception. In C. Blakemore (Ed.), *Vision: Coding and efficiency* (pp. 411-422). Cambridge, UK: Cambridge University Press.

Nobre, A. C., Coull, J. T., Walsh, V., & Frith, C. D. (2003). Brain Activations during Visual Search: Contributions of Search Efficiency versus Feature Binding. *NeuroImage, 18*(1), 91-103.

Nowak, L., & Bullier, J. (1997). The timing of information transfer in the visual system. In K. S. Rockland & J. H. Kaas & A. Peters (Eds.), *Extrastriate visual cortex in primates* (Vol. 12, pp. 205-241). New York: Plenum Press.

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology, 34*(1), 72-107.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology, 41*(2), 176-210.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*(3), 145–175.

Op De Beeck, H., & Vogels, R. (2000). Spatial sensitivity of macaque inferior temporal neurons. *The Journal of Comparative Neurology, 426*(4), 505-518.

Op de Beeck, H., Wagemans, J., & Vogels, R. (2001). Can neuroimaging really tell us what the human brain is doing? The relevance of indirect measures of population activity. *Acta Psychologica, 107*(1-3), 323-351.

Oram, M. W., & Foldiak, P. (1996). Learning generalisation and localisation: Competition for stimulus type and receptive field. *Neurocomputing, 11*(2), 297-321.

Oram, M. W., Xiao, D., Dritschel, B., & Payne, K. R. (2002). The temporal resolution of neural codes: does response latency have a unique role? *Philosophical Transactions of the Royal Society of London series B: Biological Sciences, 357*(1424), 987-1001.

O'Regan, J. K. (1992). Solving the "real" mysteries of visual perception: the world as an outside memory. *Canadian Journal of Psychology, 46*(3), 461-488.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*(5), 1011-1031.

O'Regan, J. K., Rensink, R. A., & Clark, J. J. (1999). Change blindness as a result of 'mudsplashes'. *Nature, 398*, 34.

O'Regan, K. (1992). Solving the "real" mysteries of visual perception: the world as an outside memory. *Canadian Journal of Psychology, 46*, 461-488.

Palmer, J. (1998). Attentional effects in visual search: relating search accuracy and search time. In R. D. Wright (Ed.), *Visual attention* (Vol. 8, pp. 348-388). Oxford (UK): Oxford University Press.

Panzeri, S., Schultz, S. R., Treves, A., & Rolls, E. T. (1999). Correlations and the encoding of information in the nervous system. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 266*(1423), 1001-1012.

Pashler, H. E. (1998). *The Psychology of attention*. Cambridge: MIT Press.

Perrett, D. I., Oram, M. W., & Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. *Cognition, 67*(1-2), 111-145.

Perrett, D. I., Rolls, E. T., & Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Experimental Brain Research, 47*(3), 329-342.

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., Jr., Miller, G. A., Ritter, W., Ruchkin, D. S., Rugg, M. D., & Taylor, M. J. (2000). Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria. *Psychophysiology, 37*(2), 127-152.

Pisella, L., Gréa, H., Tilikete, C., Vighetto, A., Desmurget, M., Rode, G., Boisson, D., & Rossetti, Y. (2000). An 'automatic pilot' for the hand in human posterior parietal cortex: toward reinterpreting optic ataxia. *Nature Neuroscience, 3*(7), 729-736.

Pizzagalli, D., Regard, M., & Lehmann, D. (1999). Rapid emotional face processing in the human right and left brain hemispheres: an ERP study. *Neuroreport, 10*(13), 2691-2698.

Plaut, D. C. (1995). Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology, 17*, 291-321.

Polk, T. A., & Farah, M. J. (1995). Brain localization for arbitrary stimulus categories: a simple account based on Hebbian learning. *Proceedings of the National Academy of Sciences of the United States of America, 92*(26), 12370-12373.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32*(1), 3-25.

Potter, M. C. (1975). Meaning in visual search. *Science, 187*(4180), 965-966.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning, 2*(5), 509-522.

Potter, M. C. (1999). Understanding sentences and scenes: The role of conceptual short-term memories. In V. Coltheart (Ed.), *Fleeting memories* (pp. 13-46). Cambridge, Massachusetts: MIT Press.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology, 81*, 10-15.

Potts, G. F., Liotti, M., Tucker, D. M., & Posner, M. I. (1996). Frontal and inferior temporal cortical activity in visual target detection: Evidence from high spatially sampled event-related potentials. *Brain Topography, 9*, 3-14.

Potts, G. F., O'Donnell, B. F., Hirayasu, Y., & McCarley, R. W. (2002). Disruption of neural systems of visual attention in schizophrenia. *Arch. Gen. Psychiatry, 59*, 418-424.

Potts, G. F., & Tucker, D. M. (2001). Frontal evaluation and posterior representation in target detection. *Cognitive Brain Research, 11*, 147-156.

Prinzmetal, W., Diedrichsen, J., & Ivry, R. B. (2001). Illusory conjunctions are alive and well: a reply to Donk (1999). *Journal of Experimental Psychology: Human Perception and Performance, 27*(3), 538-541.

Puce, A., Allison, T., Gore, J. C., & McCarthy, G. (1995). Face-sensitive regions in human extrastriate cortex studied by functional MRI. *Journal of Neurophysiology, 74*(3), 1192-1199.

Puce, A., Allison, T., & McCarthy, G. (1999). Electrophysiological studies of human face perception. III: Effects of top-down processing on face-specific potentials. *Cerebral Cortex, 9*(5), 445-458.

Rafal, R., Danziger, S., Grossi, G., Machado, L., & Ward, R. (2002). Visual detection is gated by attending for action: Evidence from hemispatial neglect. *Proceedings of the National Academy of Sciences of the United States of America, 99*(25), 16371-16375.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79-87.

Rebai, M., Poiroux, S., Bernard, C., & Lalonde, R. (2001). Event-related potentials for category-specific information during passive viewing of faces and objects. *International Journal of Neuroscience, 106*(3-4), 209-226.

Rees, G., Wojciulik, E., Clarke, K., Husain, M., Frith, C., & Driver, J. (2000). Unconscious activation of visual cortex in the damaged right hemisphere of a parietal patient with extinction. *Brain, 123*(8), 1624-1633.

Reinagel, P. (2001). How do visual neurons respond in the real world? *Current Opinion in Neurobiology, 11*(4), 437-442.

Rensink, R. A. (2000a). The dynamic representation of scenes. *Visual Cognition, 7*(1/2/3), 17-42.

Rensink, R. A. (2000b). Seeing, sensing, and scrutinizing. *Vision Research, 40*, 1469–1487.

Rensink, R. A. (2002). Change detection. *Annual Review of Psychology, 53*, 245–277.

Rensink, R. A., & Enns, J. T. (1995). Preemption effects in visual search: evidence for low-level grouping. *Psychological Review, 102*(1), 101-130.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8*, 368-373.

Reynolds, J. H., Chelazzi, L., & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience, 19*(5), 1736-1753.

Reynolds, J. H., & Desimone, R. (1999). The role of neural mechanisms of attention in solving the binding problem. *Neuron, 24*(1), 19-29, 111-125.

Reynolds, J. H., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron, 26*(3), 703-714.

Riesenhuber, M., & Poggio, T. (1999). Are cortical models really bound by the "binding problem"? *Neuron, 24*(1), 87-93, 111-125.

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology, 12*(2), 162-168.

Robertson, L. C. (2003). Binding, spatial attention and perceptual awareness. *Nature Reviews Neuroscience, 4*, 93-102.

Rolls, E. T., Aggelopoulos, N. C., & Zheng, F. (2003). The receptive fields of inferior temporal cortex neurons in natural scenes. *Journal of Neuroscience, 23*(1), 339-348.

Rolls, E. T., & Deco, G. (2002). *Computational neuroscience of vision*. New York: Oxford University Press.

Rolls, E. T., Tovee, M. J., & Panzeri, S. (1999). The neurophysiology of backward visual masking: information analysis. *Journal of Cognitive Neuroscience, 11*(3), 300-311.

Rosenbluth, D., & Allman, J. M. (2002). The effect of gaze angle and fixation distance on the responses of neurons in V1, V2, and V4. *Neuron, 33*(1), 143-149.

Rossion, B., Curran, T., & Gauthier, I. (2002). A defense of the subordinate-level expertise account for the N170 component. *Cognition, 85*(2), 189-196.

Rossion, B., Delvenne, J. F., Debatisse, D., Goffaux, V., Bruyer, R., Crommelinck, M., & Guerit, J. M. (1999). Spatio-temporal localization of the face inversion effect: an event- related potentials study. *Biological Psychology, 50*(3), 173-189.

Rossion, B., & Gauthier, I. (2002). How does the brain process upright and inverted faces? *Behavioral and Cognitive Neuroscience Reviews, 1*(1), 62-74.

Rossion, B., Gauthier, I., Goffaux, V., Tarr, M. J., & Crommelinck, M. (2002). Expertise training with novel objects leads to left-lateralized face like electrophysiological responses. *Psychological Science, 13*(3), 250-257.

Rossion, B., Gauthier, I., Tarr, M. J., Despland, P., Bruyer, R., Linotte, S., & Crommelinck, M. (2000). The N170 occipito-temporal component is delayed and enhanced to inverted faces but not to inverted objects: an electrophysiological account of face-specific processes in the human brain. *Neuroreport, 11*(1), 69-74.

Rossion, B., Joyce, C. A., Cottrell, G. W., & Tarr, M. J. (sous presse). Early lateralization and orientation tuning for face, word and object processing in the visual cortex. *Neuroimage*.

Rousselet, G. A., & Fabre-Thorpe, M. (soumis). How long to get to the "gist" of real-world natural scenes? *Visual Cognition*.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience, 5*(7), 629-630.

Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *Journal of Vision, 3*(6), 440-455.

Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (en préparation). N170 evoked by faces in natural scenes: specificity, effects of size, task status and inversion.

Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (sous presse). Animal and human faces in natural scenes: how specific to human faces is the N170 ERP component? *Journal of Vision*.

Rousselet, G. A., Macé, M. J.-M., Thorpe, S. J., & Fabre-Thorpe, M. (en préparation). ERP studies of object categorization in natural scenes: in search for category specific differential activities.

Rousselet, G. A., Thorpe, S. J., & Fabre-Thorpe, M. (2003). Taking the MAX from neuronal responses. *Trends in Cognitive Science, 7*(3), 99-102.

Rousselet, G. A., Thorpe, S. J., & Fabre-Thorpe, M. (sous presse). Processing of one, two or four natural scenes in humans: the limits of parallelism. *Vision Research*.

Sagiv, N., & Bentin, S. (2001). Structural encoding of human and schematic faces: holistic and part- based processes. *Journal of Cognitive Neuroscience, 13*(7), 937-951.

Sato, N., Nakamura, K., Nakamura, A., Sugiura, M., Ito, K., Fukuda, H., & Kawashima, R. (1999). Different time course between scene processing and face processing: a MEG study. *Neuroreport, 10*(17), 3633-3637.

Schall, J. D. (1997). Visuomotor areas of the frontal lobe. In K. S. Rockland & J. H. Kaas & A. Peters (Eds.), *Extrastriate visual cortex in primates* (Vol. 12, pp. 527-638). New York: Plenum Press.

Schall, J. D., & Thompson, K. G. (1999). Neural selection and control of visually guided eye movements. *Annual Review of Neuroscience, 22*, 241-259.

Schendan, H. E., Ganis, G., & Kutas, M. (1998). Neurophysiological evidence for visual perceptual categorization of words and faces within 150 ms. *Psychophysiology, 35*(3), 240-251.

Schiller, P. H. (1997). Past and present ideas about how the visual scene is analyzed by the brain. In K. S. Rockland & J. H. Kaas & A. Peters (Eds.), *Extrastriate visual cortex in primates* (Vol. 12, pp. 59-90). New York: Plenum Press.

Schyns, P. G. (1998). Diagnostic recognition: task constraints, object information, and their interactions. *Cognition, 67*(1-2), 147-179.

Schyns, P. G., Jentzsch, I., Johnson, M., Schweinberger, S. R., & Gosselin, F. (2003). A principled method for determining the functionality of ERP components. *NeuroReport, 14*(13), 1665-1669.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time and spatial scale dependant scene recognition. *Psychological Science, 5*, 195-200.

Schyns, P. G., & Oliva, A. (1997). Flexible, diagnosticity-driven, rather than fixed, perceptually determined scale selection in scene and face recognition. *Perception, 26*(8), 1027-1038.

Seeck, M., Michel, C. M., Mainwaring, N., Cosgrove, R., Blume, H., Ives, J., Landis, T., & Schomer, D. L. (1997). Evidence for rapid face recognition from human scalp and intracranial electrodes. *Neuroreport, 8*(12), 2749-2754.

Sereno, A. B., & Kosslyn, S. M. (1991). Discrimination within and between hemifields: a new constraint on theories of attention. *Neuropsychologia, 29*(7), 659-675.

Séverac-Cauquil, A., Edmonds, G. E., & Taylor, M. J. (2000). Is the face-sensitive N170 the only ERP not affected by selective attention? *Neuroreport, 11*(10), 2167-2171.

Shadlen, M. N., & Movshon, J. A. (1999). Synchrony unbound: a critical evaluation of the temporal binding hypothesis. *Neuron, 24*(1), 67-77, 111-125.

Shafritz, K. M., Gore, J. C., & Marois, R. (2002). The role of the parietal cortex in visual feature binding. *Proceedings of the National Academy of Sciences of the United States of America, 99*(16), 10917–10922.

Shapiro, K., Driver, J., Ward, R., & Sorenson, R. E. (1997a). Priming from the attentional blink: a failure to extract visual tokens but not visual types. *Psychological Science, 8*, 95-100.

Shapiro, K., & Terry, K. (1998). The attentional blink: the eyes have it (but so does the brain). In R. D. Wright (Ed.), *Visual attention* (Vol. 8, pp. 306-329). Oxford (UK): Oxford University Press.

Shapiro, K. L., Arnell, K. M., & Raymond, J. E. (1997b). The attentional blink. *Trends in Cognitive Sciences, 1*(8), 291-296.

Shapiro, K. L., & Luck, S. J. (1999). The attentional blink: a front-end mechanism for fleeting memories. In V. Coltheart (Ed.), *Fleeting memories* (pp. 95-118). Cambridge, Massachusetts: MIT Press.

Sheinberg, D. L., & Logothetis, N. K. (1997). The role of temporal cortical areas in perceptual organization. *Proceedings of the National Academy of Sciences of the United States of America, 94*(7), 3408-3413.

Sheinberg, D. L., & Logothetis, N. K. (2001). Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. *Journal of Neuroscience, 21*(4), 1340-1350.

Shevelev, I. A., Novikova, R. V., Lazareva, N. A., Tikhomirov, A. S., & Sharaev, G. A. (1995). Sensitivity to cross-like figures in the cat striate neurons. *Neuroscience, 69*(1), 51-57.

Shibata, T., Nishijo, H., Tamura, R., Miyamoto, K., Eifuku, S., Endo, S., & Ono, T. (2002). Generators of visual evoked potentials for faces and eyes in the human brain as determined by dipole localization. *Brain Topography, 15*(1), 51-63.

Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J., & Davis, J. (1995). Visual cortical mechanisms detecting focal orientation discontinuities. *Nature, 378*(6556), 492-496.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*, 1193-1216.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception, 28*, 1059-1074.

Simons, D. J., Chabris, C. F., Schnur, T., & Levin, D. T. (2002). Evidence for preserved representations in change blindness. *Consciousness and Cognition, 11*(1), 78-97.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences, 1*(7), 261-267.

Singer, W. (1999). Neuronal synchrony: a versatile code for the definition of relations? *Neuron, 24*(1), 49-65, 111-125.

Smid, H. G. O. M., Jakob, A., & Heinze, H.-J. (1999). An event-related brain potential study of visual selective attention to conjunctions of color and shape. *Psychophysiology, 36*, 264-279.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs, 74*, 1-29.

Steinman, S. B. (1987). Serial and parallel search in pattern vision? *Perception, 16*(3), 389-398.

Stone, A., & Valentine, T. (2003). Perspectives on prosopagnosia and models of face recognition. *Cortex, 39*, 31-40.

Sugase, Y., Yamane, S., Ueno, S., & Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature, 400*(6747), 869-873.

Suzuki, W. A., & Amaral, D. G. (1994). Topographic organization of the reciprocal connections between the monkey entorhinal cortex and the perirhinal and parahippocampal cortices. *Journal of Neuroscience, 14*(3), 1856-1877.

Suzuki, W. A., Miller, E. K., & Desimone, R. (1997). Object and place memory in the macaque entorhinal cortex. *Journal of Neurophysiology, 78*(2), 1062-1081.

Swadlow, H. A. (2003). Fast-spike interneurons and feedforward inhibition in awake sensory neocortex. *Cerebral Cortex, 13*(1), 25-32.

Tallon-Baudry, C., & Bertrand, O. (1999). Oscillatory gamma activity in humans and its role in object representation. *Trends in Cognitive Sciences, 3*(4), 151-162.

Tanaka, J., Luu, P., Weisbrod, M., & Kiefer, M. (1999). Tracking the time course of object categorization using event-related potentials. *Neuroreport, 10*(4), 829-835.

Tanaka, J. W. (2001). The entry point of face recognition: evidence for face expertise. *Journal of Experimental Psychology: General, 130*, 534-543.

Tanaka, J. W., & Curran, T. (2001). A neural basis for expert object recognition. *Psychological Science, 12*(1), 43-47.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience, 19*, 109-139.

Tanaka, K. (2003). Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities. *Cerebral Cortex, 13*(1), 90-99.

Tarr, M. J., & Gauthier, I. (2000). FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience, 3*(8), 764-769.

Taylor, M. J., Edmonds, G. E., McCarthy, G., & Allison, T. (2001). Eyes first! Eye processing develops before face processing in children. *Neuroreport, 12*(8), 1671-1676.

Thomas, E., Van Hulle, M. M., & Vogels, R. (2001). Encoding of categories by noncategory-specific neurons in the inferior temporal cortex. *Journal of Cognitive Neuroscience, 13*(2), 190-200.

Thorpe, S. J., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520-522.

Thorpe, S. J., & Imbert, M. (1989). Biological constraints on connectionist models. In R. Pfeifer & Z. Schreter & F. Fogelman-Soulié & L. Steels (Eds.), *Connectionism in perspective* (pp. 63-92). Amsterdam: Elsevier.

Thorpe, S. J. (1990). Spike arrival times: A highly efficient coding scheme for neural networks. In R. Eckmiller & G. Hartman & G. Hauske (Eds.), *Parallel processing in neural systems* (pp. 91-94). North-Holland: Elsevier.

Thorpe, S. J., Bacon, N., Rousselet, G. A., Macé, M. J.-M., & Fabre-Thorpe, M. (2002). Rapid categorisation of natural scenes: feedforward vs feedback contribution evaluated by backward masking. *Perception supplement, 31*, p150.

Thorpe, S. J., & Fabre-Thorpe, M. (2001). Seeking categories in the brain. *Science, 291*(5502), 260-263.

Thorpe, S. J., Gegenfurtner, K. R., Fabre-Thorpe, M., & Bulthoff, H. H. (2001a). Detection of animals in natural images using far peripheral vision. *European Journal of Neuroscience, 14*(5), 869-876.

Thorpe, S. J., Delorme, A., & Van Rullen, R. (2001b). Spike-based strategies for rapid processing. *Neural Networks, 14*(6-7), 715-725.

Torralba, A. (2003). Contextual priming for object detection. *International Journal of Computer Vision, 53*(2), 153-167.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems, 14*, 391-412.

Touryan, J., & Dan, Y. (2001). Analysis of sensory coding with complex stimuli. *Current Opinion in Neurobiology, 11*(4), 443-448.

Trappenberg, T. P., Rolls, E. T., & Stringer, S. M. (2002). Effective Size of Receptive Fields of Inferior Temporal Visual Cortex in Natural Scenes. In T. G. Dietterich & S. Becker & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press.

Treisman, A. (1998a). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London series B: Biological Sciences, 353*(1373), 1295-1306.

Treisman, A. (1998b). The Perception of features and objects. In R. D. Wright (Ed.), *Visual attention* (Vol. 8, pp. 26-54). Oxford (UK): Oxford University Press.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: evidence from search asymmetries. *Psychological Review, 95*(1), 15-48.

Treisman, A., & Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance, 16*(3), 459-478.

Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology, 14*(1), 107-141.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97-136.

Triesch, J., Ballard, D. H., Hayhoe, M. M., & Sullivan, B. T. (2003). What you see is what you need. *Journal of Vision, 3*(1), 86-94.

Ullman, S. (1995). Sequence seeking and counter streams: a computational model for bidirectional information flow in the visual cortex. *Cerebral Cortex, 5*(1), 1-11.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5*(7), 682-687.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle & M. A. Goodale & R. J. W. Mansfield (Eds.), *Analysis of visual behavior*. Cambridge: MIT Press.

Vandenberghe, R., Duncan, J., Dupont, P., Ward, R., Poline, J. B., Bormans, G., Michiels, J., Mortelmans, L., & Orban, G. A. (1997). Attention to one or two features in left or right visual field: A positron emission tomography study. *Journal of Neuroscience, 17*, 3739-3750.

VanRullen, R. (2000). *Une première vague de potentiels d'action, une première vague idée de la scène visuelle. Rôle de l'asynchronie dans le traitement rapide de l'information visuelle.* Manuscrit de thèse non publié, Université Paul Sabatier, Toulouse.

VanRullen, R. (sous presse). Visual Saliency and Spike Timing in the Ventral Visual Pathway. *Journal of Physiology*.

VanRullen, R., Gautrais, J., Delorme, A., & Thorpe, S. (1998). Face processing using one spike per neurone. *Biosystems, 48*(1-3), 229-239.

VanRullen, R., & Koch, C. (2003). Competition and selection during visual processing of natural scenes and objects. *Journal of Vision, 3*(1), 75-85.

VanRullen, R., Reddy, L., & Koch, C. (sous presse). Visual search and dual-tasks reveal two distinct attentional resources. *Journal of Cognitive Neuroscience*.

VanRullen, R., & Thorpe, S. (1999). Spatial attention in asynchronous neural networks. *Neurocomputing, 26-27*, 911-918.

VanRullen, R., & Thorpe, S. J. (2001a). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception, 30*(6), 655-668.

VanRullen, R., & Thorpe, S. J. (2001b). The time course of visual processing: from early perception to decision-making. *Journal of Cognitive Neuroscience, 13*(4), 454-461.

VanRullen, R., & Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vision Research, 42*(23), 2593-2615.

Varela, F., Lachaux, J.-P., Rodriguez, E., & Martinerie, J. (2001). The Brainweb: phase synchronization and large-scale integration. *Nature Reviews Neuroscience, 2*, 229-239.

Verghese, P., & Nakayama, K. (1994). Stimulus discriminability in visual search. *Vision Research, 34*(18), 2453-2467.

Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science, 287*(5456), 1273-1276.

Vinje, W. E., & Gallant, J. L. (2002). Natural stimulation of the nonclassical receptive field increases information transmission efficiency in V1. *Journal of Neuroscience, 22*(7), 2904-2915.

Vogel, E. K., & Luck, S. J. (2000). The visual N1 component as an index of a discrimination process. *Psychophysiology, 37*(2), 190-203.

von der Malsburg, C. (1999). The what and why of binding: the modeler's perspective. *Neuron, 24*(1), 95-104, 111-125.

Vuilleumier, P. (2000). Faces call for attention: evidence from patients with visual extinction. *Neuropsychologia, 38*(5), 693-700.

Vuilleumier, P., & Rafal, R. D. (2000). A systematic study of visual extinction. Between- and within-field deficits of attention in hemispatial neglect. *Brain, 123*, 1263-1279.

Vuilleumier, P., Sagiv, N., Hazeltine, E., Poldrack, R. A., Swick, D., Rafal, R. D., & Gabrieli, J. D. (2001). Neural fate of seen and unseen faces in visuospatial neglect: a combined event-related functional MRI and event-related potential study. *Proceedings of the National Academy of Sciences of the United States of America, 98*(6), 3495-3500.

Vuilleumier, P., Schwartz, S., Clarke, K., Husain, M., & Driver, J. (2002). Testing Memory for Unseen Visual Stimuli in Patients with Extinction and Spatial Neglect. *Journal of Cognitive Neuroscience, 14*(6), 875–886.

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology, 51*(2), 167-194.

Walsh, V., Ashbridge, E., & Cowey, A. (1998). Cortical plasticity in perceptual learning demonstrated by transcranial magnetic stimulation. *Neuropsychologia, 36*, 45-49.

Walsh, V., Ellison, A., Ashbridge, E., & Cowey, A. (1999). The role of the parietal cortex in visual attention—hemispheric asymmetries and the effects of learning: a magnetic stimulation study. *Neuropsychologia, 37*(2), 245-251.

Wang, Y., Fujita, I., & Murayama, Y. (2000). Neuronal mechanisms of selectivity for object features revealed by blocking inhibition in inferotemporal cortex. *Nature Neuroscience, 3*(8), 807-813.

Watanabe, M., Tanaka, H., Uka, T., & Fujita, I. (2002). Disparity-selective neurons in area V4 of macaque monkeys. *Journal of Neurophysiology, 87*(4), 1960-1973.

Watanabe, S., Kakigi, R., & Puce, A. (2003). The spatiotemporal dynamics of the face inversion effect: a magneto- and electro-encephalographic study. *Neuroscience, 116*, 879-895.

Wojciulik, E., & Kanwisher, N. (1999). The generality of parietal involvement in visual attention. *Neuron, 23*, 747-764.

Wolfe, J. M. (1994). Visual search in continuous, naturalistic stimuli. *Vision Research, 34*(9), 1187-1195.

Wolfe, J. M. (1998). Visual search. In H. Pashler (Ed.), *Attention* (pp. 13-73). Hove (UK): Psychology Press Ltd.

Wolfe, J. M. (1999). Inattentional amnesia. In V. Coltheart (Ed.), *Fleeting memories* (pp. 71-94). Cambridge, Massachusetts: MIT Press.

Wolfe, J. M. (2001). The level of attention: Mediating between the stimulus and perception. In L. Harris (Ed.), *Levels of Perception: a Festschrift for Ian Howard*.: Springer Verlag.

Wolfe, J. M. (2003). The level of attention: Mediating between the stimulus and perception. In L. Harris & M. Jenkin (Eds.), *Levels of Perception*. New York, NY: Springer-Verlag.

Wolfe, J. M., & Bennett, S. C. (1997). Preattentive object files: shapeless bundles of basic features. *Vision Research, 37*(1), 25-43.

Wolfe, J. M., & Cave, K. R. (1999). The psychophysical evidence for a binding problem in human vision. *Neuron, 24*(1), 11-17, 111-125.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 419-433.

Wolfe, J. M., Friedman-Hill, S. R., Stewart, M. I., & O'Connell, K. M. (1992). The role of categorization in visual search for orientation. *Journal of Experimental Psychology: Human Perception and Performance, 18*(1), 34-49.

Wolfe, J. M., & Gancarz, G. (1996). Guided Search 3.0: A model of visual search catches up with Jay Enoch 40 years later. In V. Lakshminarayanan (Ed.), *Basic and clinical applications of vision science* (pp. 189-192). Dordrecht, Netherlands: Kluwer Academic.

Wolfe, J. M., Klempen, N., & Dahlen, K. (2000). Post-attentive vision. *Journal of Experimental Psychology: Human Perception and Performance, 26*(2), 693-716.

Wolfe, J. M., Oliva, A., Butcher, S. J., & Arsenio, H. C. (2002). An Unbinding Problem? The disintegration of visible, previously attended objects does not attract attention. *Journal of Vision, 2*, 256-271.

Wolfe, J. M., O'Neill, P., & Bennett, S. C. (1998). Why are there eccentricity effects in visual search? Visual and attentional hypotheses. *Perception & Psychophysics, 60*(1), 140-156.

Woodman, G. F., & Luck, S. J. (1999). Electrophysiological measurement of rapid shifts of attention during visual search. *Nature, 400*(6747), 867-869.

Yantis, S. (1998). Objects, attention, and perceptual experience. In R. D. Wright (Ed.), *Visual Attention* (pp. 187-214). New York, NY: Oxford University Press.

Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum Press.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*, 141-145.

Résumé

Cette thèse porte sur le traitement rapide des informations visuelles contenues dans les scènes naturelles. Elle s'articule en deux chapitres constitués chacun d'une revue de la littérature et d'articles présentant des travaux expérimentaux réalisés au cours de celle-ci.

Le chapitre 1 s'intéresse tout d'abord au degré de parallélisme dans le traitement des scènes naturelles. Contrairement aux modèles sériels qui postulent que les objets sont analysés l'un après l'autre, une revue détaillée de la littérature suggère une grande part de parallélisme dans le traitement visuel. Les deux premiers articles de cette thèse portent sur la catégorisation d'objets dans les scènes naturelles et suggèrent que l'interférence entre représentations d'objets aurait lieu principalement au niveau décisionnel, probablement dans les aires frontales. La seconde partie du chapitre 1 s'intéresse au parallélisme de traitement qui permet d'extraire le sens du contexte général d'une scène. L'article 3 décrit l'efficacité du système visuel à extraire rapidement le sens global d'une scène et suggère que celui-ci pourrait interagir en parallèle avec la catégorisation des objets. L'article 4 tente de mieux cerner la participation des facteurs visuels ascendants et descendants dans l'analyse des scènes naturelles.

Parmi toutes les catégories, les visages humains pourraient être traités de façon très particulière. Le chapitre 2 discute certains arguments en faveur d'une spécificité des mécanismes impliqués. Des explications alternatives y sont proposées permettant d'envisager un modèle unique de traitement visuel pour toutes les catégories d'objets. L'article 5 montre qu'au niveau comportemental les visages d'êtres humains dans des scènes naturelles ne sont pas traités plus rapidement que d'autres catégories d'objets familiers. L'article 6 tente de déterminer le temps de traitement de ces stimuli au niveau électrophysiologique. Plusieurs hypothèses sont discutées. L'article 7 montre que la N170 n'est pas aussi spécifique des visages d'êtres humains que communément admis. Ce qui semble leur être spécifique est l'ampleur de l'effet d'inversion au niveau comportemental et électrophysiologique. Tous ces résultats sont discutés dans le cadre des modèles actuels du traitement visuel.

Title

Rapid visual categorization of natural scenes: limits of parallelism and specificity of faces. A behavioral and electrophysiological study in humans.

Summary

This thesis focuses on the fast processing of visual information in natural scenes. It hinges on 2 chapters both containing a review of the literature and research papers describing experimental work completed during the thesis.

Chapter 1 addresses first the degree of parallelism in the processing of natural scenes. In opposition with serial models postulating that objects are analyzed one after the other by the visual system, the detailed review of the literature suggests a large part of parallelism is present in visual processing. Interference between object representations would occur mainly at the decisional level, probably within frontal areas. The first two papers of this thesis address the question of object categorization in natural scenes and present data in favor of this hypothesis. The second part of chapter 1 focuses on parallel processing which allows us to extract the meaning of the general context of a scene (background). Paper 3 describes the efficiency of the visual system in extracting the global meaning of a scene in a rapid manner and suggests that it might interact in parallel with the categorization of objects. Paper 4 attempts to clarify the involvement of bottom-up and top-down visual factors in the analysis of natural scenes.

Among all categories, human faces could be processed in a very specific way. Chapter 2 discusses some arguments in favor of the specificity of underlying mechanisms. Alternative explanations are suggested, allowing us to consider a unique model of visual processing for all object categories. Paper 5 shows that at the behavioral level human faces in natural scenes are not processed faster than other categories of familiar objects. Paper 6 tries to determine the processing time of these stimuli at the electrophysiological level. Several hypotheses are discussed. Paper 7 shows that the N170 is not as specific to human faces as commonly thought. What seems to be specific to human stimuli is the magnitude of the inversion effect at the behavioral and electrophysiological levels. All these results are discussed in the context of current models of visual processing.

mots-clés : perception visuelle, catégorisation, paradigme go/no-go, scènes naturelles, parallélisme, visages, potentiels évoqués