Traitrements cognitifs mis en jeu dans la perception visuelle de scènes complexes et conséquences sur l'indexation automatique d'images

Jingqiang Li

Université de Haute Bretagne Rennes 2

U.F.R. de Sciences Humaines

N° attribué par la bibliothèque ∟∟∟∟∟∟∟∟∟∟∟

THESE

pour obtenir le grade de

Docteur en Psychologie de L'Université de Rennes 2

Nouveau régime

Spécialité Psychologie Expérimentale

présentée et soutenue publiquement
par

Jingqiang LI

Le 23 Mai 2008

Traitements cognitifs mis en jeu dans la perception visuelle de scènes complexes et conséquences sur l'indexation automatique d'images

---

Sous la co-direction de Jean-Pierre Gaillard et Alain LIEURY

JURY

Mr Jean-Pierre ROSSI, Professeur, Université de Paris-Sud (Paris XI), Président
Mr Simon THORPE, Directeur de recherche CNRS, Toulouse, Rapporteur
Mme Nathalie PORTOLAN, Ingénieur de recherche, France Télécom R&D
Mr Jean-Pierre GAILLARD, Maître de conférences, Université Rennes 2, Co-directeur
Mr Alain LIEURY, Professeur émérite, Université Rennes 2, Co-directeur

Li, Jingqiang. Traitements cognitifs mis en jeu dans la perception visuelle de scènes complexes et conséquences sur l'indexation automatique d'images, 2008

Remerciements

Je voudrais tout d'abord remercier Jean-Pierre Gaillard pour sa patience et sa disponibilité. Merci de m'avoir si bien guidé dans mon travail et de m'avoir fait confiance tout au long de ces années.

Je remercie Alain Lieury pour ses conseils et ses encouragements.

Je remercie le professeur Jean-Pierre Rossi et le professeur Simon Thorpe d'avoir accepté d'être membres du jury.

Je remercie Bernard Marquet et Nathalie Portolan pour m'avoir accueilli dans le laboratoire Design, Recherche et Relation Client (DRC), ce qui m'a permis de comprendre des méthodologies et de mettre en lien la recherche et son application. Je remercie également tous les membres du laboratoire DRC avec qui j'ai passé trois ans formidables.

Je remercie tous les membres du laboratoire (LPE) de Rennes pour leur bonne humeur.

Je remercie très sincèrement les relecteurs (-trices) : Hugo, Dorothée, Amaël, Dominique, Cécile, Laëtitia, Séverine, Florence, Gaël, Stéphane… pour leurs remarques pertinentes à la fois sur mon travail de recherche et mon niveau en français.

Enfin, je remercie infiniment Yi Fan pour sa patience, sa bonne humeur ainsi que pour son soutien inconditionnel tout au long de ces trois dernières années.

Sommaire

Introduction Générale

Quoi de plus simple, en apparence, que de percevoir un paysage de montagne, de forêt, un bord de mer, lors d'une promenade ! Chez l'être humain, la reconnaissance visuelle des scènes complexes est généralement rapide, automatique et fiable. Cette simplicité contraste avec la difficulté, d'une part, à modéliser en psychologie de la vision, les processus de reconnaissance visuelle, et, d'autre part, à produire en vision par ordinateur, des algorithmes de reconnaissance, simples, efficaces et robustes. Une des pistes actuellement à l'étude pour aborder ce problème fondamental est d'analyser les traitements cognitifs mis en jeu dans la perception visuelle de scènes complexes.

Ce travail de thèse s'intègre dans des recherches menées conjointement par trois groupes de travail du Laboratoire de Psychologie Expérimentale (EA N°1285, CRPCC) de L'Université de Rennes 2 et deux laboratoires de France Telecom Recherche & Développement (DRC et Pôle Image, site de Cesson Sévigné).

Le projet de recherche vise à étudier les processus de perception des scènes naturelles en s'appuyant sur un triptyque : psychologie cognitive, sciences de l'ingénieur et neurosciences. Mon travail de thèse a donc été consacré à étudier les facteurs principaux utilisés par les sujets humains lors de la perception de scènes visuelles complexes. Un algorithme de traitement d'images a été conçu par France Telecom R&D afin de modéliser le traitement des scènes complexes. Ce travail tente de reproduire ce fonctionnement, d'une part, et s'attache d'autre part à améliorer l'architecture en termes d'ergonomie de l'algorithme de traitement d'images développé par France Telecom R&D[1].

Ce travail de thèse s’est ensuite intéressé à la question de la perception des scènes visuelles, et s’est plus particulièrement axé sur le rôle joué par les contours et la structuration spatiale des différentes zones de luminance dans la catégorisation de scènes naturelles complexes. Dans un premier temps, les arguments théoriques qui témoignent de l’importance que revêtent les propriétés

[1] En ce qui concerne l'amélioration de l'aspect ergonomique de cet algorithme, les propositions d'amélioration ne sont pas discutées dans cette thèse, les droits étant réservés à France Telecom R&D.

perceptives dans le traitement des scènes visuelles seront présentés. Parmi ces propriétés perceptives, les contours et le contraste de luminance sont corrélés significativement avec les mouvements oculaires (Mannan, Ruddock & Wooding, 1996, 1997). Ces deux types d'informations jouent un rôle différent lors de la perception d'une scène naturelle. L'un permet au système visuel d'accéder à l'information portant sur les objets. L'identification est alors basée sur une approche locale. L'autre permet d'avoir une structure grossière de la scène. L'identification est alors basée sur une approche globale. La question que nous nous posons est de savoir comment ces deux types d'informations sont utilisés par les sujets humains dans une tâche d'identification de scènes visuelles.

Cinq expérimentations auront pour but d'étudier finement les processus utilisés par les sujets humains en fonction de ces deux facteurs. La présentation des données expérimentales et leur discussion seront présentées à la suite de chacune d'entre elles.

Chapitre 1 Le système visuel

Pour définir ce qui est important dans une scène, il est nécessaire de savoir comment notre système visuel capte, traite et analyse les images qu'il reçoit de notre environnement. Avant d'entreprendre une revue de la littérature scientifique dans ce domaine, ce chapitre présente de manière non exhaustive les fonctionnalités du système visuel, de la rétine jusqu'aux premières aires du cortex visuel primaire.

Cette présentation succincte nous permettra de mieux comprendre, par la suite, les différentes études expérimentales qui ont été effectuées.

La vision est un ensemble de mécanismes physiques, neurophysiologiques et psychologiques complexes par lesquels les stimuli lumineux sont transformés en sensations. Depuis longtemps, elle fait l'objet de nombreuses études et expériences qui couvrent domaines différents, par exemple en sciences de l'ingénierie (traitement d'images), neurobiologie, psychophysique et psychologie expérimentale.

Ainsi, les chercheurs essayent de répondre à la question : Quels sont les mécanismes de la perception visuelle du monde extérieur?

Nous présentons donc, par la suite, les fonctionnalités des cellules ou neurones selon l'ordre de réception des stimuli lumineux.

1. L'œil

Dans cette partie de présentation, nous ne rentrerons pas dans les détails de l'anatomie et de la physiologie de l'œil. Nous nous intéressons aux fonctionnalités des cellules rétiniennes en jeu dans la perception et pour lesquelles nous commencerons par les photorécepteurs de la rétine, qui se situent dans le fond de l'œil (Figure 1.1).

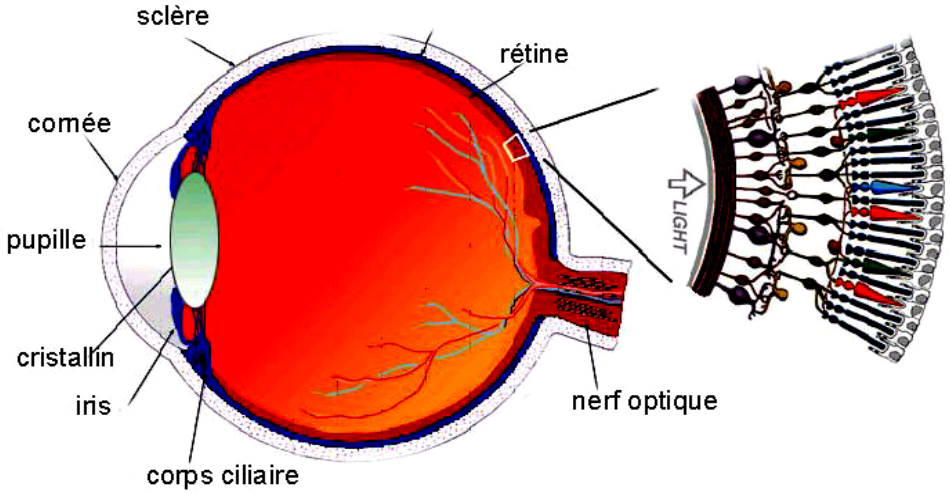


Figure 1.1. Œil et rétine humaine.
(http://coursenligne.u-strasbg.fr/pages.jsp?idRub=471&idsite=142).

1.1. La rétine

La rétine tapisse le fond de l'œil. C'est le lieu de transduction du message lumineux venant de l'extérieur en signaux nerveux envoyés au cerveau. Elle est formée d'un épithélium pigmentaire et d'une partie neurale. L'épithélium pigmentaire est constitué d'une couche de cellules qui contiennent de la mélanine ; ce sont elles qui absorbent la lumière. La partie neurale est composée de trois couches distinctes de neurones (Figure 1.2) :

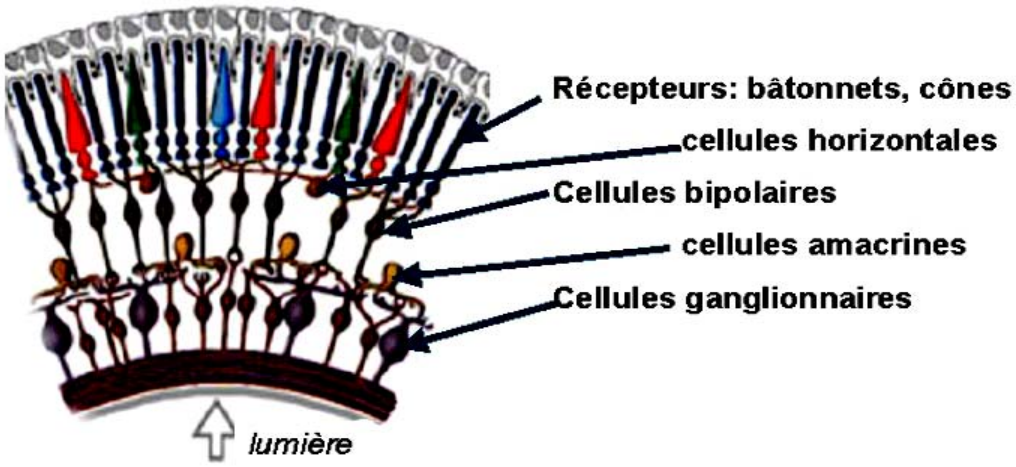


Figure 1.2. Structure de la rétine.
(http://coursenligne.u-strasbg.fr/pages.jsp?idRub=471&idsite=142).

- la couche des photorécepteurs ;
- la couche des cellules horizontales et bipolaires ;
- la couche des cellules ganglionnaires.

Ces trois couches de neurone sont séparées par deux zones de contacts synaptiques :

- la couche plexiforme externe ;
- la couche plexiforme interne.

Trois autres types de neurones sont contenus dans la rétine : les cellules horizontales et les cellules amacrines qui forment deux circuits horizontaux, et, les cellules interplexiformes (Buser & Imbert, 1987).

Une fois la structure générale de l'œil décrite, nous allons présenter plus précisément les trois grandes couches de cellules nerveuses composant la rétine.

1.1.1. Les photorécepteurs

La couche des photorécepteurs est la couche la plus profonde, par rapport à l'arrivée de la lumière. Deux types de photorécepteurs différents se distinguent : les cônes et les bâtonnets. Leur répartition sur la rétine est non uniforme et divise celle-ci en trois zones :

- **la fovéa** : partie centrale circulaire de la rétine qui fait environ 1.5 mm soit 5 degrés dans le champ visuel. Cette partie se caractérise par une grande densité de cônes, les bâtonnets étant complètement absents du centre de la fovéa (2 degrés) ;
- **la para fovéa** : également appelée « tache jaune ». Cette partie circulaire occupe un diamètre d'environ 6 mm, soit 15 à 20 degrés dans le champ visuel. Lorsque l'on quitte la fovéa les bâtonnets commencent à apparaître et leur rapport au nombre de cônes ne cesse d'augmenter ;
- **la périphérie** : le nombre de cônes devient presque nul.

A ces trois régions vient s'ajouter la papille optique. Elle correspond à l'endroit où se rejoignent tous les axones des cellules ganglionnaires qui quittent la rétine à destination des centres supérieurs (aires du cortex). Ils forment ainsi le nerf optique. Cette zone est dépourvue de cellules nerveuses à l'exception de ces axones. Elle constitue donc un point aveugle de la rétine, qui correspond à un trou dans le champ visuel. Elle est également appelée « tache aveugle »

La rétine comporte environ 130 millions de cellules photosensibles différentes, portant des noms reflétant leur forme : cônes et bâtonnets.

- **Les cônes**

Les cônes sont environ au nombre de 5 millions par rétine chez l'homme (Bullier, 1997). Ils sont dédiés à la vision diurne (photopique) et aux couleurs. Les cônes sont de trois types selon leur courbe de sensibilité spectrale (Schnapf, Kraft, Nunn, & Baylor, 1988). Les cônes "L" (Long, sensibles au rouge), sont accordés autour d'une longueur d'onde centrée sur 564 nm, les "M" (Medium, sensibles au vert) à 533 nm et les "S" (Short, sensibles au bleu-violet) à 437 nm (Alleyson, 1999) (Figure 1.3).

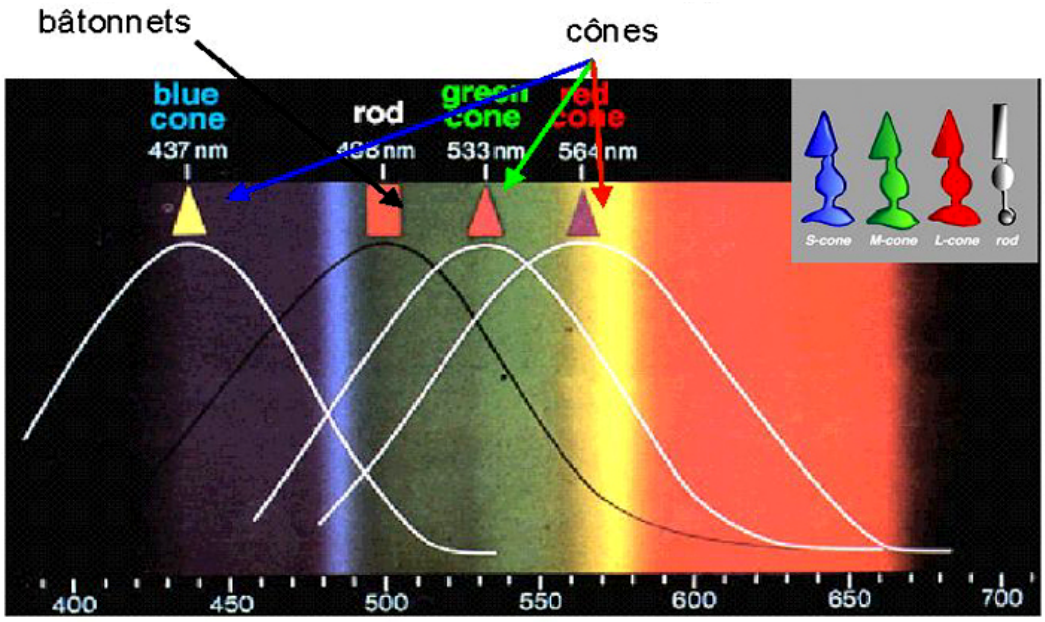


Figure 1.3. Sensibilité différentielle des différents types de récepteurs à la longueur d'onde (http://coursenligne.u-strasbg.fr/pages.jsp?idRub=471&idsite=142).

- **Les bâtonnets**

Il y a environ 120 millions de bâtonnets par rétine chez l'homme (Bullier, 1997). Ils sont actifs en vision nocturne (scotopique) et ils ne permettent pas la vision des couleurs. La courbe de sensibilité spectrale des bâtonnets se situe entre celle des cônes S et M (pic à 505 nm). Il n'existe qu'une sorte de bâtonnets.

Un des rôles principaux de la rétine est de permettre au système visuel de coder le signal lumineux. Ce codage de la lumière en potentiel membranaire est appelé fonction de transduction des photorécepteurs. Ils adaptent leur réponse à la dynamique de l'intensité lumineuse captée (Tableau 1.1).

Tableau 1.1

Propriétés des deux systèmes issus des cônes (vision photopique) et des bâtonnets (vision scotopique) (Extrait de la thèse de A. Chauvin, 2003)

Système photopique		**Système scotopique**
Récepteurs	Cônes	Bâtonnets
Nombre de photorécepteurs	5 millions	120 millions
Photo pigments	3 opsines différentes	Rhodopsine
Sensibilité	Faible, vision diurne	Élevée, vision nocturne
Localisation dans la rétine	Fovéa et Para-fovéa	Hors de la fovéa
Taille du champ récepteur	Petit en fovéa et s'agrandit en périphérie	Plus grand

1.1.2. La couche plexiforme externe

Cette couche est le siège d'interactions entre les photorécepteurs, les cellules horizontales et les cellules bipolaires. Les photorécepteurs délivrent leur signal à la fois aux cellules horizontales et aux cellules bipolaires.

- **Les cellules bipolaires**

Les cellules bipolaires, comme tous les neurones de la rétine sauf les cellules ganglionnaires, transmettent l'influx nerveux non pas avec des potentiels d'action mais sous la forme de simples potentiels gradués. On parle tout de même de réponse ON lorsqu'une dépolarisation amène une augmentation de la relâche de neurotransmetteurs et de réponse OFF quand une hyperpolarisation diminue la quantité de neurotransmetteurs relâchés.

- **Les cellules horizontales**

Ces cellules, sont en contact avec les synapses entre les récepteurs et les cellules bipolaires. Les cellules horizontales modulent la réponse des photorécepteurs. Elles connectent plusieurs photorécepteurs et sont connectées entre elles par l'intermédiaire de synapses électriques. De ce fait, leur champ récepteur est bien plus large que celui des photorécepteurs. Cette architecture permet un lissage de l'information transmise par les cônes. Ainsi les cellules horizontales porteraient une information de luminance locale moyenne qui rentrerait en jeu dans l'adaptation du photorécepteur à la luminance. Chez le primate, elles ont une action sur les bipolaires, antagoniste de celle des cônes.

- **Les cellules amacrines**

Ces cellules sont en contact avec les cellules bipolaires et ganglionnaires. Elles sont impliquées dans des mécanismes de modulation du gain de la réponse des bipolaires et des ganglionnaires. De plus, elles sont impliquées dans l'adaptation des champs récepteurs des cellules ganglionnaires en fonction de l'intensité moyenne et par conséquent du rapport signal sur bruit (Beaudot, 1994). Enfin, certaines jouent un rôle dans la détection du mouvement.

1.1.3. La couche des cellules ganglionnaires

Les cellules ganglionnaires ont le même type de champs récepteurs circulaires à opposition centre-périphérie que les cellules bipolaires. Les cellules ganglionnaires sont de trois types :

- les cellules **Y** (ou parasol, ou alpha), les cellules **X** (ou midget ou bêta) et enfin les cellules **W** (ou gamma). Elles se projettent respectivement sur les couches magnocellulaire, parvocellulaire et koniocellulaire du corps géniculé latéral (CGL). Les cellules alpha ont un corps cellulaire et un champ récepteur plus grand que les deux autres types de cellules, elles sont très sensibles aux changements temporels (réponse phasique). Ces neurones ne codent pas le contraste de couleur mais sont sensibles à des faibles niveaux de contraste de luminance (Figure 1.4).

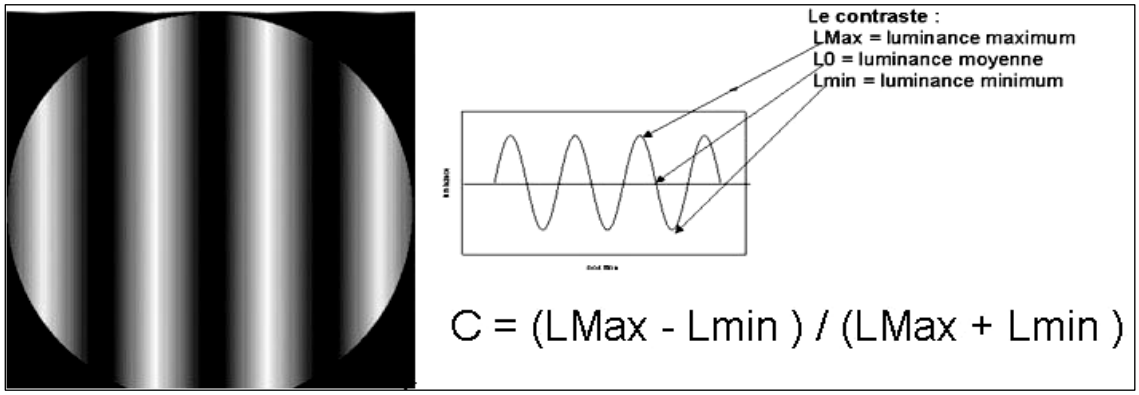


Figure 1.4. Illustration du contraste de luminance
(http://coursenligne.u-strasbg.fr/pages.jsp?idRub=471&idsite=142).

- Les cellules bêta ont un petit corps cellulaire et un petit champ récepteur. Elles sont activées par un ou plusieurs cônes, peu sensibles aux changements temporels (réponse tonique). Ces cellules sont responsables de la vision des couleurs. Parmi les cellules bêta et alpha, on distingue deux grands types fonctionnels de neurones : les neurones à centre ON et les neurones à centre OFF (Imbert, 1983). Les neurones à centre ON (respectivement OFF) augmentent leur fréquence de décharge lors de

l'augmentation (respectivement de la diminution) de contraste dans le centre du champ récepteur. Lorsque le changement de contraste fait intervenir non seulement le centre du champ récepteur mais également sa périphérie les neurones ON et OFF ne donnent pratiquement pas de réponse. Ils signalent donc le changement local de contraste.

- Enfin, les cellules gamma ont des propriétés et des morphologies mal connues. Elles interviendraient dans le codage des couleurs (opposition bleu/jaune) (voir la thèse de N. Guyader, 2004). De plus, le caractère ON ou OFF d'une cellule bipolaire est transmis à la cellule ganglionnaire qui lui est connectée. La plupart des cellules ganglionnaires ne sont pas très sensibles aux stimuli lumineux qui touchent à la fois le centre et la périphérie du champ récepteur. Ainsi, une obscurité totale ou un éclairage uniforme leur fait émettre peu de potentiels d'action. Ces cellules sont toutefois très sensibles aux différences d'éclairement survenant à l'intérieur du champ récepteur, comme lorsqu'une zone d'ombre ou de lumière balaie leur champ récepteur d'un côté à l'autre par exemple.

Les cellules ganglionnaires sont les seules à transmettre le signal nerveux sous forme de potentiels d'action. Considérant que ce sont leurs axones qui forment le nerf optique et transmettent donc l'information à de grandes distances de la rétine, la génération de potentiels d'action dans ces cellules prend alors tout son sens. Ces potentiels d'action sont d'ailleurs générés de façon spontanée et c'est donc leur fréquence de décharge qui est amplifiée ou diminuée par l'apparition de lumière dans leur champ récepteur.

Grâce à cette organisation en couches, l'information est véhiculée vers les centres de traitement corticaux en plusieurs étapes.

La rétine est la première structure impliquée dans le traitement de l'information visuelle. Comme nous l'avons vu, elle se compose de capteurs qui transforment l'intensité et la chrominance de la lumière incidente en un signal électrique. Elle effectue les premiers traitements de l'information visuelle. Ces

traitements sont spatio-temporels, ce qui se traduit notamment par une sensibilité aux variations rapides de luminosité.

1.2. Le champ récepteur (CR)

On considérera le champ récepteur d'une cellule visuelle comme étant la région de l'espace visuel dans laquelle un stimulus approprié engendre des potentiels d'action, et on inclura dans la caractérisation du champ récepteur les propriétés de ce stimulus. Lorsqu'un stimulus est présenté à l'extérieur du CR, la cellule enregistrée ne produit pas de réponse significative. Les travaux de Hartline (1938), ainsi que ceux de Kuffler (1953) chez le chat et de Barlow (1953) (cité par Guyader, 2004) chez la grenouille, ont permis la caractérisation précise des propriétés des CR des cellules ganglionnaires de la rétine. Kuffler montra que ceux-ci sont concentriques, avec une zone centrale (sélective à la lumière, ON, ou à l'obscurité, OFF) et une zone périphérique antagoniste (OFF ou ON), ce qui rend ces cellules particulièrement sensibles aux différences d'éclairement survenant à l'intérieur de leur champ récepteur. Les champs récepteurs ont une forme grossièrement circulaire pouvant être divisée en deux zones concentriques (Figure 1.5).

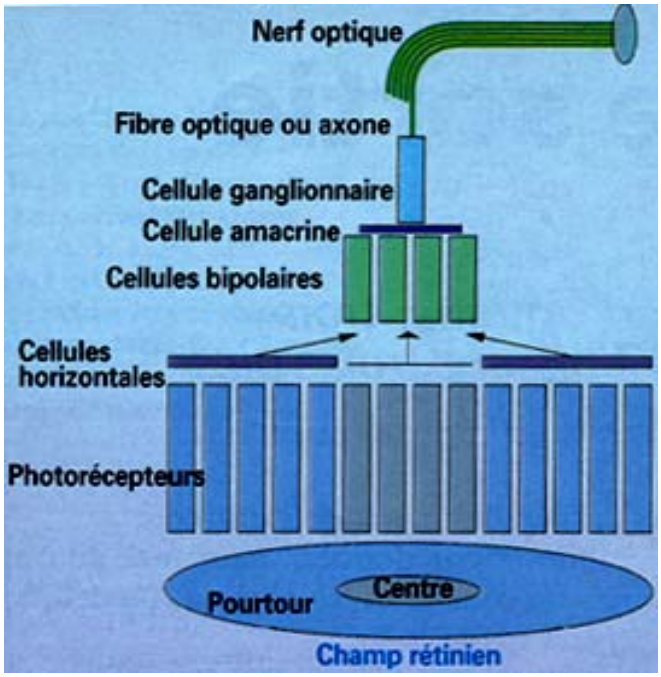


Figure 1.5. Schéma d'un champ récepteur.

(http://www.bioinformatics.org/oeil-couleur/dossier/images/champs-recepteurs.png).

Plusieurs travaux ont porté sur la taille des champs récepteurs des cellules de la rétine (Jung & Spillmann, 1970 ; Westheimer, 1967). Les recherches s'accordent sur la petite taille du centre des champs récepteurs, de l'ordre de 4-5 arcs min, dans la fovéa dans des conditions de vision photopique (diurne). La taille du champ récepteur augmente (jusqu'à 2 degrés) avec l'excentricité rétinienne et est plus grande dans des conditions de vision photopique que scotopique (nocturne).

Par la suite, et en particulier grâce aux travaux de Hubel et Wiesel (1981), l'étude des champs récepteurs des cellules visuelles a été étendue à l'exploration du cortex.

Nous allons maintenant revenir au traitement du signal lumineux. Après avoir été capté par les photorécepteurs de la rétine, le signal lumineux est transmis aux différentes cellules de la rétine décrites ci-dessous.

2. De la rétine au cortex

2.1. Le cheminement du signal visuel

Les axones des cellules ganglionnaires de la rétine convergent pour former le nerf optique. La grande majorité des fibres du nerf optique (90%) se projette sur le corps genouillé latéral (CGL). Le chiasma optique est une zone où s'effectue le croisement des voies visuelles (Figure 1.6). Les informations provenant d'une part de l'hémisphère temporal de l'œil droit et d'autre part de l'hémisphère nasal de l'œil gauche se rejoignent sans se mélanger. C'est à partir des deux corps genouillés latéraux (CGL) que l'information de chaque œil est projetée de manière alternée sur les aires visuelles corticales.

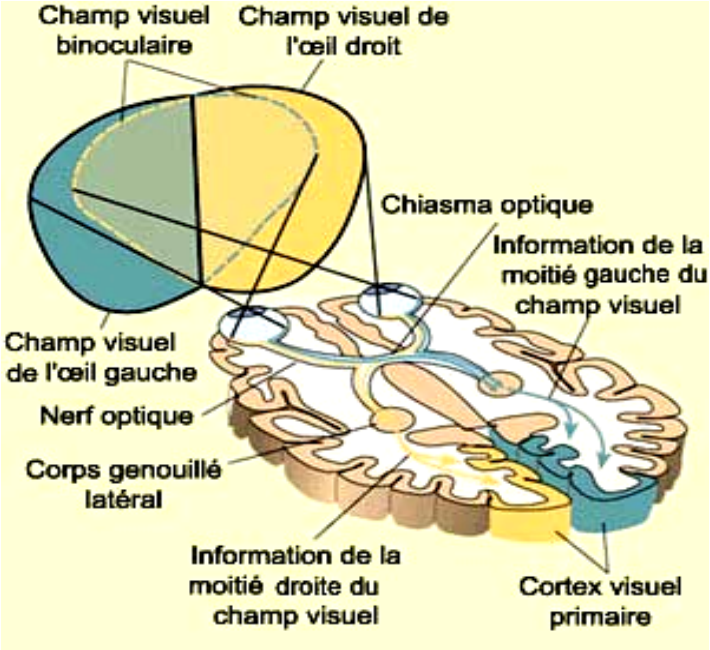


Figure 1.6. Projections des fibres rétiniennes sur le cortex visuel primaire. (http://lecerveau.mcgill.ca/flash/a/a_02/a_02_cr/a_02_cr_vis/a_02_cr_vis.html).

2.2. Les corps genouillés latéraux (CGL)

Le corps genouillés latéral (CGL) est une structure bilatérale du thalamus. Nous avons donc un CGL droit et un CGL gauche. Chaque CGL a 6 couches : la couche 1 étant la plus profonde et la 6, la plus superficielle (Figure 1.7). Chaque couche d'un CGL est rétinotopique, c'est-à-dire que deux cellules proches l'une de l'autre dans le CGL ont des champs récepteurs proches l'un de l'autre.

- Les couches 1 et 2 sont magnocellulaires, les autres sont parvocellulaires.

- Les couches 2, 3 et 5 sont ipsilatérales (c'est-à-dire qu'elles reçoivent des axones de cellules ganglionnaires dans l'œil de même latéralisation) ; les autres couches sont controlatérales.

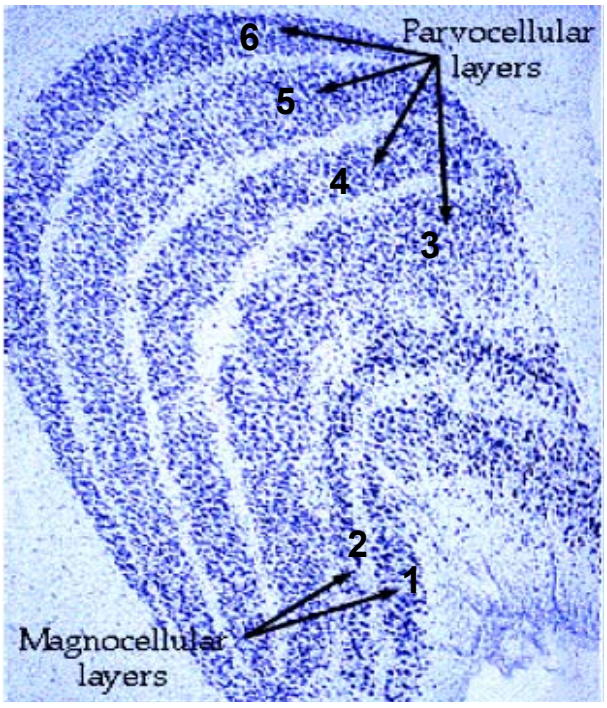


Figure 1.7. La structure du CGL (extrait de la thèse de A. Chauvin, 2003).

3. Le cortex visuel primaire

Les cellules rétiniennes se projettent, par l'intermédiaire des voies décrites ci-dessus, au niveau du cortex occipital, de chaque côté de la scissure calcarine, dans l'aire dite « V1 ». Elles s'y disposent selon une cartographie, dite rétinotopique (organisation similaire à celle de la rétine). Le phénomène essentiel lié à cette répartition est « l'amplification fovéale ». La surface fovéale sur la rétine, correspondant à un très faible pourcentage du champ visuel, est reliée à près de la moitié de la surface corticale ; le champ visuel périphérique, beaucoup plus grand, n'est relié qu'à l'autre moitié. En d'autres termes, on peut dire que chaque récepteur rétinien de la fovéa est en liaison directe avec certainement plusieurs centaines de neurones centraux, alors qu'un champ récepteur périphérique ne dispose que de quelques-uns d'entre eux. Ceci explique les performances relatives de notre fonction visuelle et notamment la rapide dégradation de l'acuité visuelle dès que l'on s'écarte, même assez peu, des conditions de vision centrale.

3.1. L'aire visuelle primaire V1

L'aire V1 du cortex visuel primaire est, actuellement, la mieux connue des aires corticales. L'organisation des neurones de l'aire V1 est distribuée dans six couches superposées, numérotées classiquement de 1 à 6 (Figure 1.8).

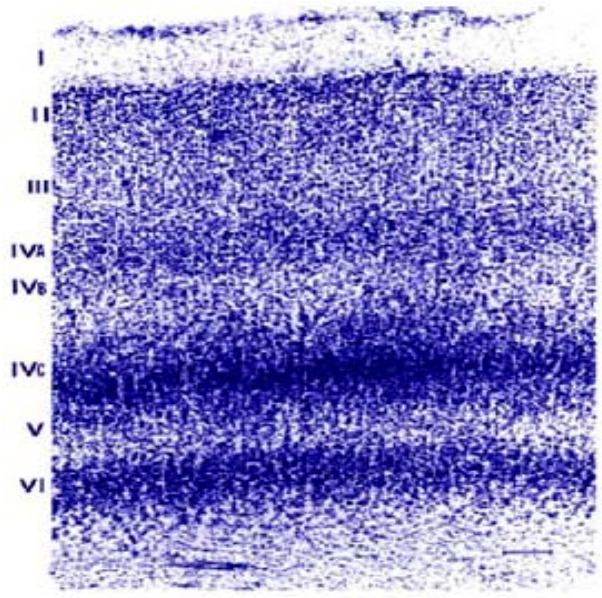


Figure 1.8. La structure laminaire du cortex V1 (extrait de la thèse de A. Chauvin, 2003).

- Couche 1 : la couche la plus externe, composée de dendrites, d'axones en majorité inhibiteurs et de cellules gliales.
- Couches 2 et 3 (supragranulaires) : composées de neurones excitateurs qui projettent leurs axones vers les aires extra striées V2, V3.
- Couche 4 (granulaire) : elle est divisée en 3 sous couches 4a, 4b et 4cα et 4cβ. Les cellules du CGL se projettent sur la couche 4cα pour les magnocellulaires et 4cβ pour les parvocellulaires, alors que les koniocellulaires se projettent sur les couches 2 et 3.
- Couche 5 et 6 (infragranulaires) : composées de neurones excitateurs se projetant sur les CGL et le colliculus supérieur (voir la thèse de Chauvin 2003, pour de plus amples détails).

Les fibres nerveuses issues des CGL se terminent principalement dans la couche 4. Cette couche présente une alternance de bandes de neurones successivement connectées à l'œil droit puis à l'œil gauche. Cette alternance de dominance oculaire a été mise en évidence par Hubel et Wiesel (1974). La classification en couches de V1 est directement liée avec les types d'afférents et les différentes aires de projection. Par exemple, les neurones issus des couches 2, 5, 6 se projettent respectivement sur les aires V2 et V3, et les CGL.

Les neurones de chacune des couches communiquent (émettent) surtout dans la direction verticale. Ainsi, les neurones de la couche 4 connectent préférentiellement les neurones des couches 3 et 5 (Figure 1.8). Les neurones inclus dans une même verticale ont donc des propriétés similaires, telle qu'une sélectivité à une même orientation du stimulus visuel ou encore à une même dominance oculaire (Hubel & Wiesel, 1974).

Les neurones sensibles à une même orientation forment une colonne corticale (Figure 1.9). Un ensemble de colonnes juxtaposées, couvrant toutes les orientations, définit une hypercolonne dont le centre se trouve au milieu d'une bande de dominance oculaire. Ce centre (blob de cytochrome oxydase) constitue un point d'ancrage de la distribution des orientations optimales des neurones. La plupart des neurones situés dans les blobs n'ont pas de sélectivité à l'orientation. Les pourtours de ces blobs sont communément appelés les inters blobs ; leurs neurones sont, eux, très sensibles à l'orientation.

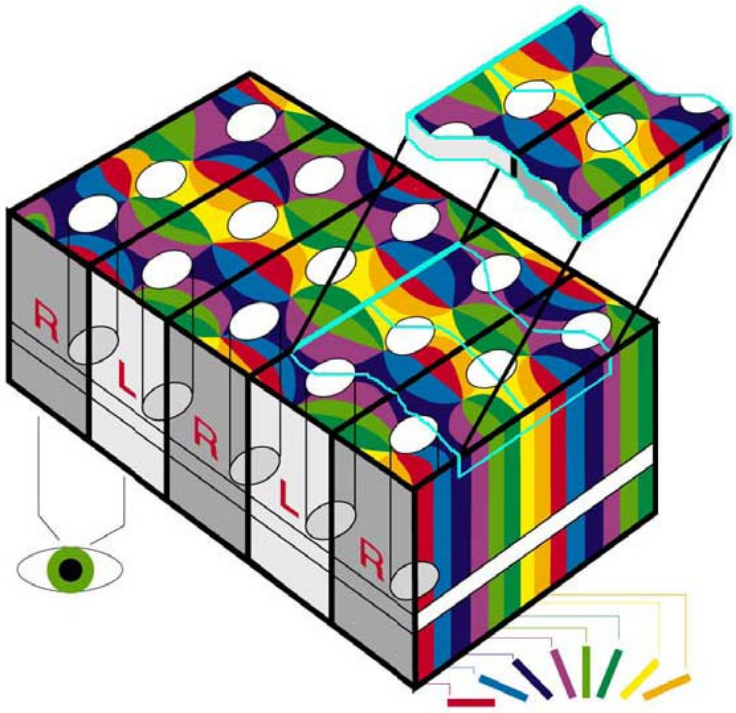


Figure 1.9. La structure d'une hypercolonne
(http://www.weizmann.ac.il/brain/images/ImageGallery.html).

- Une hypercolonne mesure environ 1 mm^3.
- Une hypercolonne comprend deux colonnes de dominance oculaire (œil droit et gauche). Les champs récepteurs des cellules d'une hypercolonne occupent des positions correspondantes sur les rétines.
- L'ensemble des hypercolonnes est rétinotopique (c'est-à-dire que 2 hypercolonnes adjacentes ont des champs récepteurs adjacents sur les rétines). Mais il y a grossissement à la fovéa.
- Chaque colonne de dominance oculaire possède 6 couches.
- Dans chaque colonne de dominance oculaire l'orientation préférée des cellules est encodée dans l'angle du rayon de l'hypercolonne.
- Dans chaque colonne de dominance oculaire la fréquence spatiale préférée des cellules est encodée dans la longueur du rayon de l'hypercolonne.

Les quelques 6000 hypercolonnes que compte V1, sont incroyablement peu nombreuses en comparaison des 130 millions de photorécepteurs. L'avantage de cette organisation en colonnes d'orientations consisterait à représenter sur une même surface en deux dimensions du cortex, un nombre élevé de variables : la représentation spatiale du champ visuel, la dominance oculaire, l'orientation, la fréquence spatiale, la couleur, le mouvement, et les cibles (ou les afférences) vers d'autres modules. Il semblerait par conséquent que la nature ait fait le choix d'une diminution de la résolution spatiale du champ visuel au profit de l'intégration d'un ensemble d'attributs élémentaires indispensables pour des traitements de hauts niveaux. Cette diminution de la résolution spatiale est néanmoins compensée par la sur-résolution fovéale qui permet de consacrer la moitié de la surface de V1 à seulement 10 degrés de notre champ visuel, celui-ci comptant un angle de vue 16 fois plus grand.

Les champs récepteurs des cellules du cortex visuel primaire ont des tailles comprises entre un et sept degrés d'angle visuel, avec une taille moyenne à 2,7 degrés (Felleman, 1981).

3.2. Les différentes cellules du cortex visuel

Hubel et Wiesel (1962, 1968) ont montré que V1 est caractérisée par deux grands types de cellules : les cellules simples et les cellules complexes. Un troisième type de cellule a également été mis en évidence : les cellules hypercomplexes. Ces trois types de cellules sont présents dans V1 mais également dans les autres aires visuelles (Tableau 1.2).

- Les cellules simples répondent à des signaux lumineux type barres de lumière.
- Les cellules complexes répondent à des barres lumineuses en mouvement. On suppose souvent que ces champs récepteurs sont formés par la combinaison non linéaire de sous unités assimilables aux

cellules simples (Hubel & Wiesel, 1962; Movshon, Thompson, & Tolhurst, 1978).

- Les cellules hypercomplexes, quant à elles, sont sensibles à la longueur des stimuli (Movshon, 1978).

Tableau 1.2

Sensibilités des deux grands types de cellules : les cellules simples et les cellules complexes en fonction des propriétés du stimulus visuel. (Chauvin, 2003)

Sensibilité	Cellule simple	Cellule complexe
Orientation	Oui	Oui
Taille	Oui	Oui
Position	Oui	Non
Couleur	Oui	Non
Direction du mouvement	Oui	Non

4. Les voies visuelles

Le traitement cortical ne se limite pas à l'aire V1. En effet, on décrit chez le primate une trentaine d'aires corticales qui diffèrent par leur architecture, leur connectivité, leur organisation topographique et/ou leurs propriétés fonctionnelles (Bullier, 1998). Celles-ci ont été moins étudiées que V1 et leurs fonctions restent encore mal connues. Cependant, on distingue grossièrement deux grands systèmes corticaux de traitement de l'information, ou encore deux voies principales : une voie dite dorsale (V1-V2-V4-(AIT-PIT-CIT)) et une voie ventrale (V1-V2-V3-MT) (Figure 1.10). A mesure que l'on progresse dans la "hiérarchie" des aires visuelles selon chaque voie, on observe une augmentation de la taille des champs récepteurs et les cellules deviennent sensibles à des stimuli de complexité croissante (Maunsell & Newsome, 1987).

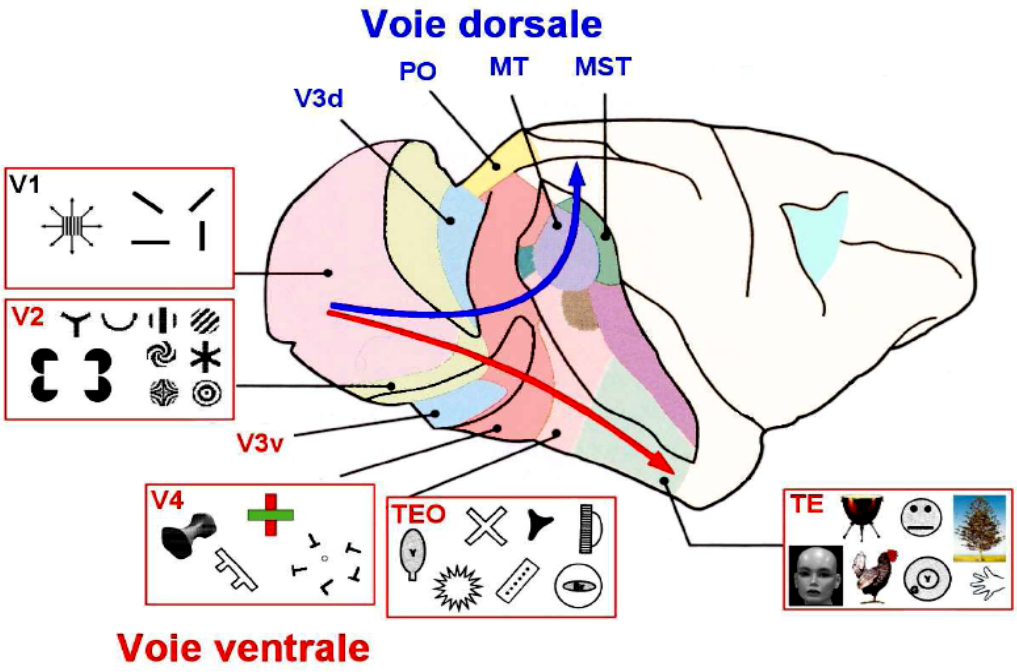


Figure 1.10. Représentation schématique de l'architecture du système visuel chez le macaque (Extrait de la thèse de M. Macé, 2006).

Abréviations :

V1-V4: Aires visuelles 1 à 4.
AIT : Inféro-temporal Antérieur.
PIT : Inféro-temporal Postérieur.
CIT : Inféro-temporal Central.
MT : Aire médiane temporale.
TEO : Aire temporo-occipitale.
PO : Pariéto-occipital (Occipito-pariétale).
TE : Aire temporale.
TO : Occipito-temporale.
MST : Aire médiane temporale supérieure.

Parmi les aires visuelles situées au-delà de VI, on distingue deux grands ensembles. Les unes comme les aires MT, MST et FST (Aire Temporale Supérieure) (Bullier, 1998) sont principalement reliées au cortex pariétal ; on dit alors qu'elles appartiennent à la voie dorsale ou occipito-pariétale.

D'autres aires, comme l'aire V4, sont principalement connectées au cortex inférotemporal et font parties de la voie ventrale ou occipito-temporale. La séparation entre ces deux voies se situe au niveau de V2.

4.1. La voie dorsale

La voie dorsale comprend les neurones M (magnocellulaires) reliés par les cellules ganglionnaires *parasols* qui collectent les signaux de plusieurs cellules bipolaires diffuses. Les neurones de cette voie sont principalement reliés aux aires du cortex pariétal. Ils sont impliqués dans la perception des objets en mouvement. Les neurones de cette voie répondent donc à des stimuli visuels se déplaçant à des vitesses rapides, même lorsqu'ils sont de faible contraste (pour une revue, voir la thèse de N. Guyader, 2004). Leur réponse concerne les basses fréquences spatiales de l'image rétinienne et les hautes fréquences temporelles[2]. La voie dorsale est appelée également la voie "où et comment" pour signifier qu'elle est spécialisée dans la perception du mouvement et la coordination visuo-motrice (Ungerleider & Haxby, 1994; Watanabee, 2003)

4.2. La voie ventrale

Une deuxième voie, la voie ventrale, comprend les neurones P (parvocellulaires) qui sont reliés aux cellules ganglionnaires *midget*, directement liées aux bipolaires de même nom et véhiculant le même signal (Hérault, 2001). Leurs signaux sont à large bande, du type passe-haut en fréquences spatiales (HF soit une image filtrée dont les termes de haute fréquence sont conservés), et du type passe-bas en fréquence temporelle (BF soit une image filtrée dont les termes de basse fréquence sont conservés). Leur réponse est à variable, spatiale et temporelle, non séparée. Cette voie transmet les signaux chromatiques avec

[2] Une fréquence spatiale faible (basse fréquence) résulte d'une image peu contrastée où les tons de gris changent lentement. À l'opposé, une haute fréquence spatiale résulte d'une image très contrastée, où les tons de gris varient très rapidement. La plupart des images sont constituées à la fois de plages de hautes et de basses fréquences comme par exemple, un lac (basse fréquence) et un réseau routier (haute fréquence).

une réponse en opposition de couleur. La voie ventrale est également appelée voie "qui ou quoi" car elle est spécialisée dans la perception des formes et des objets (Livingstone & Hubel, 1987 ; Livingstone & Hubel, 1988).

4.3. Les autres voies visuelles

Il existe une troisième voie avec les neurones K (koniocellulaires) qui se situent entre les couches précédemment décrites. Ces neurones relient l'information des cellules ganglionnaires *gamma*, ils se projettent dans les blobs de cytochrome oxydase. Jusqu'à présent, la fonctionnalité de cette voie est encore mal connue.

4.4. La vitesse de traitement

Dans cette présentation de l'anatomie et de la neurobiologie du système visuel, nous ne rentrons pas dans le détail des temps de traitement nécessaires pour véhiculer l'information visuelle des photorécepteurs de la rétine au cortex visuel. De nombreuses études ont porté sur la rapidité de fonctionnement du système visuel et ont obtenu des résultats relativement précis quand au décours temporel de la transmission du signal dans une tâche perceptive.

Chez le primate, les informations rétiniennes sont envoyées jusqu'au cortex inférotemporal (la région impliquée dans la reconnaissance d'objets) par l'intermédiaire du nerf optique. Du point de vue d'un traitement unidirectionnel de ces informations, elles passent par quelques étapes principales, par exemple, le CGL, le thalamus, V1 puis V2, V4 et enfin le cortex Inférotemporal. D'après les données électrophysiologiques chez le singe, un minimum de 10 ms serait nécessaire pour que l'information soit disponible d'une étape à l'autre (Nowak, Munk, Girard, & Bullier, 1995). Chez l'homme, ces latences seraient un peu allongées, entre 10 et 15 ms (Bacon, 2002).

La sensibilité de l'activation des neurones varie entre eux, certains neurones semblant répondre différemment pour certains types de stimuli. De plus, le système semble travailler plutôt avec une vitesse et un nombre d'étapes optimisées. Comme nous l'avons vu, les neurones M (magnocellulaire) véhiculent les basses fréquences spatiales de l'image rétinienne et les hautes fréquences temporelles sont activées 20 ms avant que les neurones P (Parvocellulaire), quant à eux, véhiculent les hautes fréquences spatiales et les basses fréquences temporelles. Tracées à partir de données neurophysiologiques, les courbes de la Figure 1.11 représentent les temps de réponse des différentes cellules du système visuel à partir de la présentation d'un stimulus visuel.

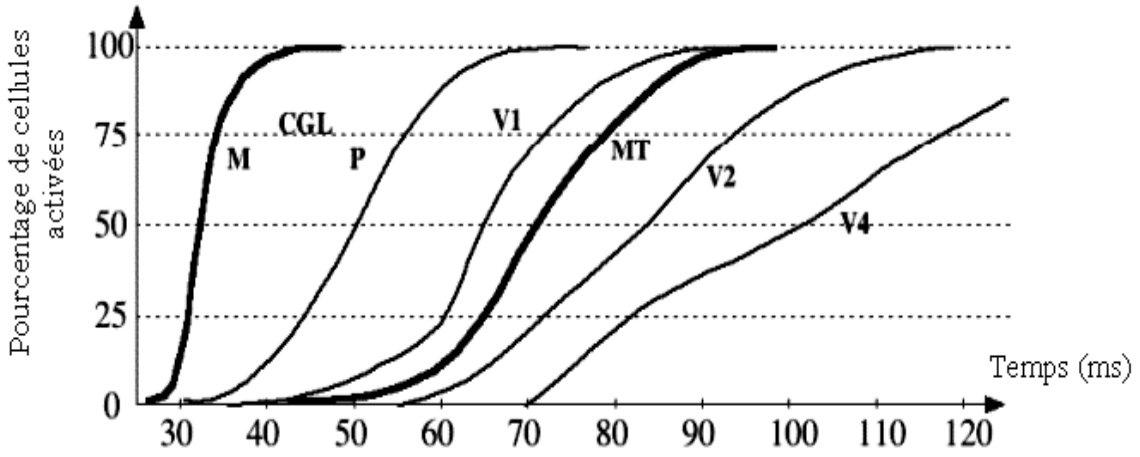


Figure 1.11. Décours temporel des réponses des cellules du cortex visuel. Histogramme cumulé des réponses des différentes cellules du système visuel en fonction du décours temporel du traitement de l'information. L'origine des temps correspond à la présentation du stimulus visuel (Schmolesky *et al.*, 1998, extrait de la thèse de N. Guyader, 2004).

La précédence de l'information magnocellulaire est à mettre en correspondance avec le principe de traitement "*de l'image grossière au trait détail*" (coarse-to-fine) de l'information visuelle. La voie magnocellulaire véhicule les basses fréquences ce qui permet d'effectuer une première analyse de l'image : la structuration spatiale, les grandes formes jusqu'aux traits détails plus fins qui seraient véhiculés par la voie parvocellulaire.

5. Conclusion

Comme nous venons de le voir, le signal lumineux est traité par plusieurs types de neurones et subit donc plusieurs modifications depuis les photorécepteurs de la rétine jusqu'aux aires du cortex visuel. La rétine est la première structure impliquée dans le traitement de l'information visuelle. Comme nous l'avons vu, elle se compose de capteurs qui transforment l'intensité et la chrominance de la lumière incidente en un signal électrique. Elle effectue les premiers traitements de l'information visuelle. La dernière couche de cellules de la rétine, les cellules ganglionnaires véhiculent par le nerf optique l'information au cortex visuel primaire, *via* les corps genouillés latéraux.

Arrivée dans le cortex visuel, l'information est véhiculée vers les centres de traitement corticaux en plusieurs étapes, en raison de l'organisation en couches du cortex visuel. De plus, à mesure que l'on progresse dans la « hiérarchie » des aires visuelles, on observe une augmentation de la taille des champs récepteurs et les cellules deviennent sensibles à des stimuli de complexité croissante. Cet acheminement de l'information visuelle s'effectue par deux voies distinctes : la voie dorsale et la voie ventrale. Bien que reliées entre elles et connectées à plusieurs endroits dans le cortex occipital temporal et frontal (Bullier, 1998), il apparaît que les aires de la voie ventrale et de la voie dorsale traitent des aspects différents du signal visuel. Le signal subit donc différents traitements et est décomposé en plusieurs attributs. Il est également, en fonction des attributs extraits, véhiculé à différentes vitesses par les deux voies de traitement de l'information visuelle.

Nos connaissances sur le système visuel sont donc considérables mais ne suffisent pas à expliquer complètement les mécanismes intimes de la perception visuelle. Les travaux sur la rétine et sur le cortex visuel se poursuivent et apportent chaque jour des nouveautés qu'il faut ensuite tenter d'intégrer dans des hypothèses de fonctionnement.

Chapitre 2 Mouvements oculaires et propriétés des informations fixées

Les mouvements oculaires sont coordonnés pour des finalités que l'on peut identifier : voir avec précision, avoir une vue d'ensemble, suivre une cible en mouvement, bouger la tête sans perdre son orientation dans l'espace. Ces mécanismes sont à la disposition de l'organisme pour prendre connaissance de son environnement, sélectionner des objets d'intérêt et continuer à les percevoir bien que l'environnement bouge ou que lui-même se déplace. Ils nécessitent des fonctions cognitives pour, par exemple, suivre les cibles en mouvement et programmer des gestes en rapport avec ces cibles. Des perceptions visuelles fiables permettent aussi de développer des représentations internes du monde extérieur (images visuelles) et des fonctions cognitives ("schémas de situation de base" ou "Situations Cognitives de Référence" selon Rossi, 2005), qui s'appliquent à ces perceptions en cours. A un stade ultérieur, l'organisme humain a développé des capacités à manipuler ces représentations internes même en l'absence de stimulations perceptives. Dans l'exploration d'une scène visuelle, certes les caractéristiques des stimuli lumineux attirent le regard, mais le traitement cognitif en cours ne produit pas seulement une perception de la scène, il prend le commandement de l'exploration elle-même.

Dans ce chapitre, les résultats apportés par des études concernant les mouvements oculaires seront présentés selon deux axes : l'un porte sur l'endroit où le sujet regarde, et l'autre porte sur le temps de fixation pour certains types de propriétés informationnelles.

1. Importance des points de fixations dans l'exploration des scènes

1.1. Définition des scènes naturelles

En faisant une synthèse de propositions de différents auteurs (Henderson & Hollingworth, 1999b), nous définirions la scène visuelle comme étant la somme, dans une vue sémantiquement cohérente du monde réel, des éléments du

contexte et des objets organisés régulièrement. C'est une composition spatio-temporelle d'éléments dans un contexte. Cette définition fournit plusieurs points importants : premièrement, une scène naturelle doit être un paysage réel, d'un point de vue "écologique", c'est ce que l'on perçoit du monde réel. Deuxièmement, la scène naturelle doit contenir des éléments naturels ou artéfactuels (objets fabriqués par l'homme). Enfin, elle a normalement un contexte qui met en relation les objets dans une scène.

Dans cette recherche, les scènes naturelles que nous utilisons sont des images à deux dimensions, codées sur 8 bits, constituant un échantillon représentatif de l'ensemble des vues rencontrées par un homme : ce sont des photographies de paysages venant de la base de données COREL[3].

Les scènes naturelles se placent au milieu du triptyque symbolique, géométrique et écologique. En premier lieu, elles sont des objets numériques, et donc assimilables à une distribution 2D de luminance[4], un agencement de régions à plusieurs échelles et réductibles à un point dans un espace multidimensionnel. Toutefois, une scène est comme un objet sémantique. Chaque image est globalement une représentation de plage, de montagne, de mer, de désert ou de ville. Ainsi, comme toute représentation, une scène possède un contenu propre. De plus, les scènes naturelles sont, par définition, des représentations réalistes de l'environnement dans lequel nous évoluons, et en ce sens, elles sont "écologiques" (Chauvin, 2003).

1.2. Importance des points de fixations

Notre champ de vision est très large (220 degrés, sauf pour les personnes atteintes de handicap visuel). Cependant, cette vision est imprécise, excepté aux

[3] Fondée en 1985, Corel Corporation (www.corel.com) est l'une des principales sociétés spécialisées dans les outils de création de contenu, la gestion des processus métier et les solutions XML pour les entreprises.

[4] La luminance (L) d'une source est le rapport entre l'intensité lumineuse (I) émise dans une direction donnée et la surface apparente (S) (projection de la surface dans la direction considérée) de la source lumineuse dans cette direction. La luminance s'exprime en cd/m².

alentours du point de fixation. La vision périphérique nous permet de reconnaître la situation générale sans les détails. Pour voir les détails, nos mouvements oculaires complètent notre vision du monde avec une vitesse de 3 fixations par seconde environ, ceci s'appelle la vision fovéale. Le champ fovéal est la seule partie de l'œil qui permet une acuité visuelle maximale. Il reçoit des informations provenant d'une zone qui correspond à un angle visuel de 2° environ. L'acuité est jugée médiocre en dehors de cette zone. L'observateur oriente le regard par un perpétuel mouvement de l'œil pour diriger l'axe fovéal vers la partie de l'image retenue afin d'avoir une analyse fine.

Un degré au centre du champ visuel représente 20 mm^2 de surface corticale (Schiller, 1997; Dougherty *et al*., 2003) alors qu'à dix degrés d'excentricité, il ne représente plus que 2 mm^2, soit dix fois moins. Comme chez l'humain, la moitié de la surface de V1 est dévolue aux dix degrés centraux du champ visuel. Un objet perçu de manière périphérique (> 10°) sera traité par un nombre beaucoup plus faible de neurones que le même objet perçu au centre. Ainsi la résolution, la forme et la quantité d'information propre à l'objet dépendent de sa position relative par rapport à la fovéa.

Lorsque le sujet regarde une scène, chaque fixation permet de positionner la fovéa sur une région particulière dans la scène, puis de générer un nouveau point de vue et une nouvelle perspective. Etant donné le nombre de positions et de régions possibles pour une scène, l'exploration constitue un échantillonnage spatio-temporel particulier de l'image. Si des points fixés sont particuliers, cela n'est pas dû au hasard, mais du au fait qu'ils sont choisis pour leur intérêt par l'observateur.

1.2.1. Région informative

La première étude portant sur la localisation des fixations sur une scène est apportée par Buswell (cité par Henderson & Ferreira, 2004, avec la technique de la caméra de cinéma). Celui-ci a demandé à 200 sujets d'examiner 55 ouvres d'art

(tableaux d'architecture, de sculpture, etc.). Buswell trouve une régularité des informations fixées, les regards se concentrant sur des zones porteuses d'informations. En effet, les sujets fixent de façon préférentielle les personnages par rapport aux informations contextuelles. De plus, ils ont tendance à grouper des fixations sur des "*régions informatives*", les fixations du regard ne se distribuent pas donc aléatoirement dans une scène.

Le terme "*région informative*" définit une zone de l'image dont certaines propriétés physiques ou sémantiques sont singulières ou particulières par rapport au reste de l'image (Chauvin, 2003).

Deux autres études (Antes, 1974; Mackworth & Morandi, 1967) mesurent *l'informativité* d'une région lors de l'exploration de scènes naturelles (tableaux et carte géographique). Selon Loftus & Mackworth (1978, cité par Chauvin, 2003), *l'informativité* d'un objet est défini par l'inverse de la probabilité d'occurrence d'un objet dans une scène étant donné le reste de la scène et l'histoire de l'observateur. Dans cette étude, deux photographies couleurs sont chacune divisées en 64 zones carrées. Un groupe de participants évalue *l'informativité* de chacune de ces zones en se basant sur le principe de la facilité à reconnaître chaque zone. Un autre groupe indépendant de participants examine les images complètes en donnant leur préférence des zones. Leurs mouvements oculaires sont enregistrés. Les auteurs calculent les densités de fixations par zone et la position des premières fixations. L'hypothèse sous jacente est que plus le nombre de fixations dans une zone donnée est grand, plus *l'informativité* de cette zone est importante. Les résultats montrent que les régions estimées les plus informatives sont les régions les plus fixées et inversement, les régions considérées relativement non informatives sont peu fixées.

Il est important de noter que le matériel expérimental utilisé par Mackworth et Morandi (1967) est constitué d'images visuellement simples et peu familières (tableaux et carte géographique). Les sujets peuvent facilement trouver des régions informatives dans la périphérie de ces images.

En utilisant des images de scènes plus complexes, Antes (1974) apporte une preuve supplémentaire, les régions estimées informatives attirent plus de fixations que les autres jugées comme étant non informatives.

Une étude plus récente (Henderson & Ferreira, 2004) montre que des régions uniformes et vides dans une scène sont peu fixées, les sujets concentrant leurs fixations sur des régions informatives de la scène (Figure 2.1).

Figure 2.1. Les fixations (carrés blancs) d'un sujet dans une tâche de mémorisation d'une scène : les fixations se localisent sur des objets ; les régions uniformes et vides sont peu fixées (Henderson & Ferreira, 2004).

1.2.2. Objectif et spécificité des tâches

Yarbus (1967) demande aux sujets d'examiner des dessins de scènes et des œuvres d'art selon des objectifs différents. Quand les sujets ont pour consigne d'estimer l'âge de chaque personne dans le tableau "le visiteur inattendu", Yarbus observe que les observateurs ont tendance à fixer leur regard sur les visages des personnages peints (Figure 2.2). D'après la recherche de Yarbus, nous pouvons introduire une autre définition concernant *l'informativité,* celle-ci peut être définie

comme l'intérêt, pour une région jugée porteuse d'informations utiles ou essentielles, selon le but du sujet. Yarbus en conclut que les mouvements oculaires sont déterminés par les caractéristiques de la tâche et par le but du sujet.

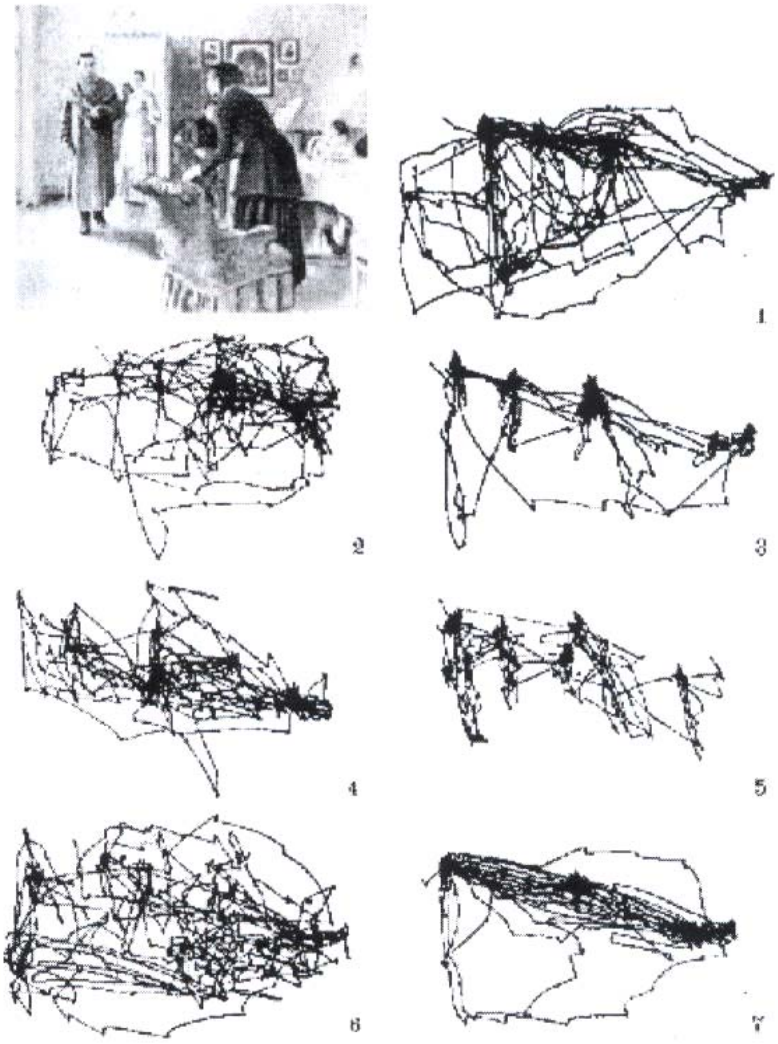


Figure 2.2. Sept enregistrements d'exploration d'un tableau "Le visiteur inattendu" par le même sujet. (1) exploration libre; pour les six autres enregistrements le sujet doit; (2) estimer les circonstances matérielles de la famille, (3) estimer l'âge de chaque personne, (4) deviner ce que faisait les personnages avant l'arrivée du "visiteur inattendu". (5) mémoriser les vêtements portés par les personnages, (6) mémoriser la position des personnes et des objets dans la pièce. (7) estimer le temps passé par le "visiteur inattendu" loin de la famille (les personnages) (Yarbus, 1967).

Henderson et Hollingworth (1999b) proposent plusieurs types de facteurs descendants (*top-down*) pouvant influencer les mouvements oculaires, comme, la mémoire épisodique. Ainsi, le fait de savoir que la clé de ma voiture vient d'être posée sur la droite de l'ordinateur me fait orienter mon regard dans cette direction pour prendre la clé, puisque je dois partir pour faire des courses.

Un autre exemple montre que les fixations oculaires ne dépendent pas de la saillance visuelle de la scène, mais du souvenir de cet objet. Elles sont dirigées à l'endroit où un objet est absent si les sujets se rappellent qu'un objet était précédemment à cet endroit (Altmann & Kamide, 2004; Bock, Irwin, & Davidson, 2004; Griffin, 2004; Ryan, Althoff, Whitlow, & Cohen, 2000; Spivey, Richardson, & Fitneva, 2004).

Une tâche spécifique détermine aussi des mouvements oculaires spécifiques. Land, Mennie et Rusted (1999) et Hayhoe, Shrivastava, Mruczek, et Pelz (2003) le montrent dans leurs études sur la fabrication du thé et de sandwichs : les sujets ont une forte tendance à fixer un objet associé à la tâche. Par exemple, ils fixent le couteau juste avant de le prendre.

Pelz et Canosa (2001) montrent que dans une tâche de lavage à la main, les sujets produisent parfois les fixations en amont, fixations liées aux futures actions. Ces mouvements oculaires anticipatifs sont liés aux buts plutôt qu'à la prépondérance des propriétés visuelles de l'environnement actuel. De plus, les mouvements oculaires possèdent d'autres modèles particuliers pendant des situations complexes et dynamiques telles que la conduite (Land & Lee, 1994; Liu, 1998).

En conclusion, les zones correspondantes aux points de fixations nous permettent de classer les propriétés de fixations en deux grandes catégories : les propriétés sémantiques (effet mémoire), d'une part, et les propriétés physiques (saillance), d'autre part.

2. Propriétés des informations fixées

2.1. Propriétés sémantiques

2.1.1. Fixations précoces et sémantique de la scène

Dans cette partie, nous allons nous intéresser à une question importante dans l'exploration d'une scène, à savoir, de quel type d'informations dépend le traitement précoce? Autrement dit, les premières fixations oculaires sont-elles influencées par des propriétés physiques d'une région locale ou par la sémantique de la scène (l'identité de la scène)? Cette question implique la relation interne entre l'objet et le reste de la scène. D'un point de vue écologique (c'est-à-dire dans le cas d'une scène visuelle du monde réel), la signification de l'objet est liée au contenu de la scène : la relation fonctionnelle entre objet et scène est donc déterminée naturellement (voir aussi chapitre 3, point 1.2, p 59). En se basant sur ce principe, différentes méthodologies portant sur les caractéristiques des fixations précoces se sont développées.

Le paradigme le plus courant consiste à manipuler la congruence entre un objet prédéfini et la scène dans laquelle se trouve cet objet.

Antes (1974) montre que les objets non congruents sont souvent fixés immédiatement après la première saccade dans la scène. Quant à l'amplitude moyenne de la saccade autour de l'objet cible, elle est comprise entre 6.5 degrés et 8 degrés d'angle visuel. Cette observation suggère que les sujets pourraient déterminer la cohérence sémantique entre l'objet et le reste de la scène en se basant sur l'information périfovéale obtenue avec une simple fixation.

Loftus et Mackworth (1978) manipulent des objets prédéfinis en utilisant des dessins, ces derniers possédant des caractéristiques semblables à de vraies scènes. Chaque objet cible est soit congruent ou non avec le reste de la scène. Par exemple, un tracteur apparaît dans une scène de ferme ou dans une scène sous-marine (Figure 2.3). Dans leur expérience, les participants doivent regarder

chaque dessin pendant quatre secondes dans une épreuve de reconnaissance. D'après eux, si la signification d'objet influence la commande de mouvements oculaires, alors l'objet sémantiquement non congruent devrait attirer plus de fixation que l'objet sémantiquement congruent. Les résultats concernant les zones de fixations sont intéressants. En premier lieu, les objets non congruents ont tendance à être fixés plus tôt que les objets congruents. De même, les régions informatives (les zones où se trouve l'objet non-congruent) attirent beaucoup de fixations par rapport à celles qui ne sont pas informatives.

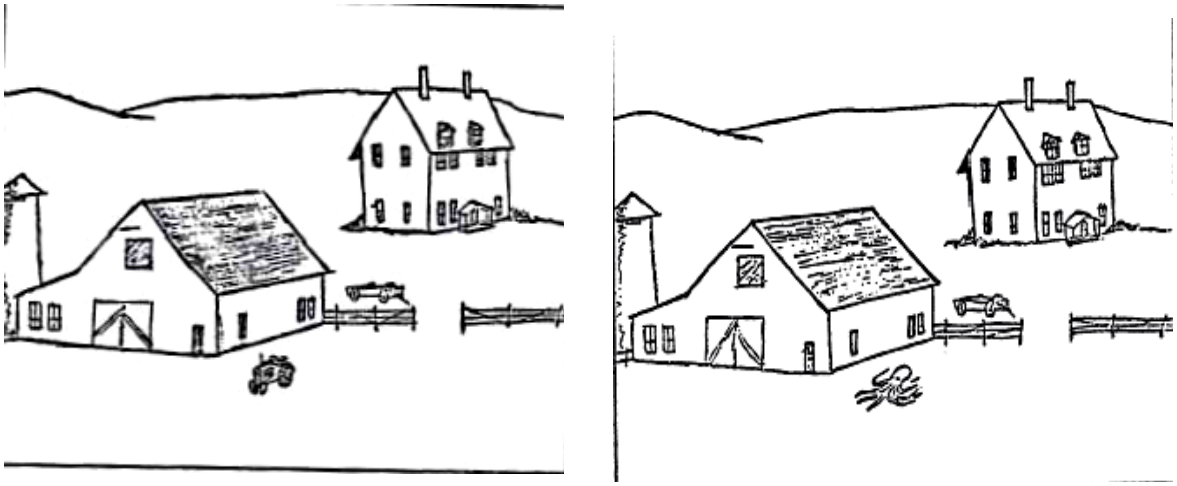


Figure 2.3. Scène naturelle avec un objet congruent (le tracteur) et un objet incongru (la pieuvre) (Loftus & Mackworth, 1978).

Des études plus récentes ne retrouvent pas cet effet d'attraction sémantique dès le début de l'exploration (De Graef, Christiaens, & d'Ydewalle, 1990; Henderson & Hollingworth, 1998; Henderson, Weeks, & Hollingworth, 1999). Par exemple, De Graef et ses collaborateurs demandent à des sujets de rechercher la présence d'un "non objet" (forme dépourvue de sens mais proche de celle d'un objet) ou d'un "objet ressemblant" non identifiable, dans un dessin aux traits de scènes naturelles. En reproduisant le paradigme de Loftus et Mackworth (1978), De Graef n'observe pas les mêmes résultats : les objets cibles ne sont pas fixés plus précocement lorsqu'ils sont non congruents avec la scène que lorsqu'ils sont congruents. En examinant les premières fixations, De Graef constate que les sujets n'ont pas tendance à fixer plus les objets non congruents que les objets congruents concernant les huit premières fixations dans la scène, et par la suite, ils fixent de manière préférentielle les objets non congruents.

Enfin, Henderson et ses collaborateurs (1999) conduisent deux expériences pour tenter de répliquer les résultats de Loftus et MacWorth (1978). Des objets cibles congruents ou non congruents sont créés indépendamment pour chacun des 24 dessins aux traits complexes et, qui font références à des photographies d'environnements normaux. Par exemple, une ferme et une scène de cuisine contiennent respectivement une poule et un mixeur comme objets-cibles (Figure 2.4).

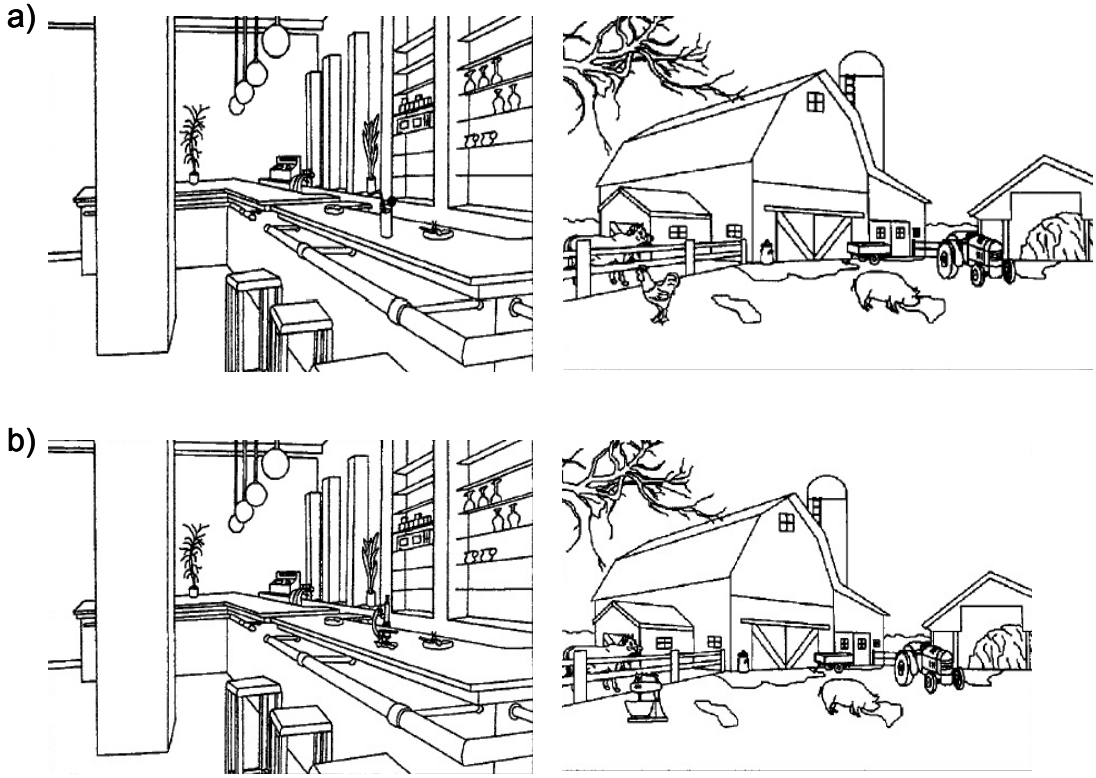


Figure 2.4. Exemple de scènes comportant un objet congruent (a) et un objet non congruent (b) (Hollingworth & Henderson, 1998; Hollingworth & Henderson, 1999; Henderson, Weeks, & Hollingworth, 1999).

Dans un premier temps, les auteurs se placent dans les conditions de l'expérience originale utilisées par Loftus et MacWorth (1978), en utilisant toutefois des stimuli plus complexes. En effet, les dessins de Loftus et MacWorth ne comportaient que peu d'objets qui étaient très espacés. Un premier groupe de sujets doit évaluer si tous les objets des vingt-quatre dessins aux traits de scènes naturelles sont congruents avec la scène. Les pourcentages de bonne réponse,

respectivement 89,5% et 89,3% pour les conditions congruentes et non congruentes montrent que les stimuli sont bien construits. Un second groupe, dont les mouvements oculaires sont enregistrés, observe les mêmes 24 images pendant quinze secondes. Contrairement à Loftus et MacWorth (1978), les auteurs se dotent d'un panel de mesures reflétant : l'analyse sémantique extra fovéale (nombre de fixations pour atteindre la cible) et l'influence de la congruence sémantique sur la densité de fixation fovéale. Concernant l'analyse extra fovéale, Henderson et ses collaborateurs n'observent pas que les objets non congruents sont fixés plus tôt, contrairement à ce qui est rapporté par Loftus et MacWorth (1978). Quant à la densité de fixation, les auteurs trouvent que le nombre de fixations est plus important pour les objets non congruents que pour les objets congruents (pour une revue, voir Chauvin, 2003).

Pourquoi les études récentes ne retrouvent-elles pas les mêmes types de résultats que ceux de Loftus et de Mackworth (1978) ?

D'abord, nous pouvons supposer que l'anomalie sémantique et les "distinctivités perceptives" ont pu être confondues (Antes & Penland, 1981; De Graef, Christiaens, & d'Ydewalle, 1990; De Graef, 1992; Rayner & Pollatsek, 1992). Toutes les scènes utilisées par Henderson, Weeks, & Hollingworth, (1999) ont été créées de la même manière et avec le même style que celles utilisées par Loftus et Mackworth (1978), et les objets cibles ont été dessinés indépendamment du contexte pour avoir une relation "non-naturelle" entre ces objets prédéfinis et la scène.

Ensuite, ces différences de résultats sont probablement dues à la caractéristique du matériel, les scènes utilisées par De Graef, Christiaens, & d'Ydewalle (1990) et par Henderson, Weeks, & Hollingworth (1999) étant issues de photographies de scènes normales, celles-ci sont visuellement plus complexes et incluent plus de contours que celles employées par Loftus et Mackworth (1978). L'identification périphérique des objets est sans doute plus difficile dans les expériences de De Graef, Christiaens, & d'Ydewalle (1990) et Henderson, Weeks, & Hollingworth (1999) qu'elle ne l'est dans les premières expériences de Loftus et Mackworth (1978). Une autre hypothèse pourrait être l'absence de mesure de

l'excentricité de l'objet non congruent par rapport à l'objet principal déterminant la sémantique de la scène.

Cet ensemble des données suggère que les fixations précoces ne se porteraient pas plus tôt sur les objets non congruents que sur les objets congruents. Autrement dit, le traitement précoce d'une scène ne semble pas être influencé par la sémantique de scène. Il existe d'autres types d'informations "intéressantes" qui "attirent" les mouvements oculaires.

2.1.2. Position du regard après les premières fixations

La sémantique de la scène ne semblerait pas influencer les fixations précoces. Cependant, elle devient de plus en plus prégnante au cours de l'exploration. Cette évolution s'observe par la probabilité de fixer à nouveau une région qui a été fixée auparavant. Des études portant sur la mesure des densités de fixations, définie par le nombre de fixations dans une région indiquée avec un temps donné, montrent que les sujets tendent à regrouper leurs fixations sur des régions informatives, mais en des moments différents (Antes, 1974; Buswell, 1935 ; Mackworth & Morandi, 1967; Yarbus, 1967).

Donc, un facteur important pour caractériser les mouvements oculaires consiste à étudier les saccades oculaires. Le temps de fixation dans une région prédéfinie est-il le même pour le début et la fin de l'exploration?

2.1.3. Temps de fixation et propriétés sémantiques

Dans les premiers travaux concernant les points de fixation, le temps de fixations n'est pas pris en compte (Antes, 1974; Buswell, 1935; Mackworth & Morandi, 1967; Yarbus, 1967). A l'heure actuelle, il paraît pertinent de prendre en compte le temps de fixation dans une région donnée car il donne des explications complémentaires sur le mécanisme de traitement des informations visuelles. De plus, le temps total de fixation dans une région est corrélé avec le nombre de

fixations dans une région. Ainsi, le temps total de fixation (la somme des durées de toutes les fixations) est plus long non seulement pour des régions physiquement saillantes mais aussi pour des régions sémantiquement informatives (De Graef, Christiaens, & d'Ydewalle, 1990; Friedman, 1979; Henderson, Weeks, & Hollingworth, 1999).

Dans l'objectif de comprendre plus finement le mécanisme de traitement de l'information, la manière de mesurer certaines variables doit être prise en compte. Par exemple, il semble pertinent de mesurer le temps de chaque fixation temporo-spatiale dans une région plutôt que la somme de toutes les fixations, ou de mesurer le temps de la première et la N^{ième} fixation ainsi que le temps total pour un même objet ou une même région, etc. (Henderson, Weeks, & Hollingworth, 1999). Cependant, aucune revue de la littérature spécialisée ne permet de fournir des résultats détaillés concernant ces mesures.

L'influence des propriétés sémantiques sur le temps de fixation pendant l'exploration de scènes a été étudiée plus largement. Loftus et Mackworth (1978) constatent que le temps de fixations est plus long pour un objet informatif que pour un objet non-informatif. Friedman (1979) prouve lui aussi que le temps de fixation est corrélé avec la probabilité d'apparition de cet objet dans la scène, les objets les moins probables sont fixés plus longtemps. Par la suite, De Graef, Christiaens, & d'Ydewalle (1990) divisent le temps total d'exploration de scène en deux parties : pendant la deuxième partie, le temps de la première fixation sur un objet congruent est plus court que sur un objet non congruent. Enfin, Henderson, Weeks, & Hollingworth (1999) constatent que la première et la deuxième fixation ainsi que le temps total de fixation sont plus longs pour les objets non congruents que pour les objets congruents. Hollingworth (2003) montrent la précédence (état de ce qui précède ou fait de précéder) du temps de fixations sur un objet non congruent quelque soit la phase d'exploration. D'après De Graef, Christiaens, & d'Ydewalle (1990), le temps total et le temps de la première fixation sur un objet sont deux facteurs différents pour mesurer le degré de l'influence des propriétés sémantiques. Mais, la différence entre ces deux mesures n'est pas très claire.

2.2. Propriétés physiques

Les propriétés visuelles des informations fixées sont étudiées selon différents types de méthodologie.

La première méthode consiste à mesurer l'existence de similarités entre les explorations de différents sujets (Mannan, Ruddock, & Wooding, 1995) et à chercher à relier ces similarités aux propriétés physiques des scènes (Mannan, Ruddock, & Wooding, 1996, 1997a, 1997b).

La seconde s'attache à mesurer la valeur des sorties d'un algorithme en chaque point de fixation, puis ces valeurs sont comparées à d'autres modes de sélection des points : aléatoire, mouvements oculaires enregistrés sur d'autres images, changement de repère Einhäuser & König, 2003; Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000; Reinagel & Zador, 1999).

Enfin, les études portant sur les modèles de cartes de saillances prédisent que certains types de traits élémentaires attirent la localisation des fixations (Itti, Gold, & Koch, 2001; Itti & Koch, 2000; Itti & Koch, 2001; Itti, Koch, & Niebur, 1998; Koch & Ullman, 1985; Niebur, Itti, & Koch, 2002;).

2.2.1. Contours, bords et contrastes

Mannan et ses collaborateurs (Mannan, Ruddock, & Wooding, 1995, 1996, 1997a, 1997b) manipulent la nature d'images de scènes, de sorte que seulement un certain type d'informations visuelles reste disponible, puis mesurent les changements comportementaux induits par cette manipulation des propriétés physiques. Ils utilisent trois versions de onze images différentes : normales (N), filtrée passe-bas (BF) ou filtrée passe-haut (HF). La tâche des sujets est d'observer chaque image pendant trois secondes en vue de répondre à des questions. L'expérience se déroule en quatre sessions espacées de trois jours chacune. L'ordre des sessions est toujours le même : BF – HF – N – N. En

répétant les mesures d'une même image pour chaque sujet, les auteurs s'intéressent à l'évaluation de la similarité entre deux explorations d'une même image réalisées par un même sujet et par deux sujets différents. Ils espèrent ainsi mettre en évidence la part de sélection des points de fixations spécifiques aux scènes. Les auteurs construisent un indice de similarité pour comparer deux séquences de fixations. L'indice renvoie seulement à la similarité spatiale entre les séquences sans tenir compte de l'ordre des fixations.

En ce qui concerne l'exploration d'une même image par plusieurs sujets, les données montrent un fort degré de similarité, toujours plus élevé pour la première moitié du temps d'observation par rapport à l'ensemble de l'exploration. Quand à l'effet du filtrage des scènes, il n'affecte pas significativement les fixations sur une scène pendant la première seconde et demie d'exploration. En revanche, sur les trois secondes d'exploration, les durées de fixation et les longueurs des saccades augmentent. Ces mêmes chercheurs utilisent ces données pour les comparer aux prédictions d'un ensemble d'algorithmes d'extraction de "traits visuels". Ces trait visuels sont le "maximum et minimum de luminance locale", "maximum et minimum du contraste physiologique" (convolution avec un filtre DOG), contraste local de Michelson, densité de bords (filtre de Sobel), et hautes fréquences de l'image (voir la thèse de A. Chauvin, 2003). Ils montrent une similarité entre les algorithmes et les mouvements oculaires, uniquement pour l'extraction des bords, du contraste locale et des hautes fréquences.

2.2.2. Contrastes, variances et corrélations

Reinagel et Zador (1999) étudient la relation entre les propriétés statistiques des scènes naturelles et la structure du système visuel. Dans leur étude, un ensemble de 77 images, composées d'images scènes naturelles, de scènes artificielles, d'hommes ou d'animaux, est présenté à cinq participants, ayant pour consigne de les observer attentivement. Une fois les mouvements oculaires enregistrés, les auteurs mesurent la variance de la luminosité moyenne de l'image et la corrélation entre le centre estimé du regard et son voisinage. Ils

montrent que les propriétés des scènes naturelles telles qu'elles sont perçues par l'œil sont contraintes par le contenu de l'image : les régions sélectionnées sont des régions particulières.

Quand à la fonction de corrélation entre les points à la vision fovéale et leur voisinage, elle décroît plus rapidement lorsque l'analyse porte sur les patchs sélectionnés par les sujets par rapport à un échantillonnage aléatoire. Les travaux de Reinagel et Zador sont confirmés en partie par Krieger, Rentschler, Hauske, Schill, & Zetzsche (2000). Ces derniers s'intéressent aux statistiques de second ordre (variance et fonction de corrélation), et mesurent les bispectres[5]des scènes naturelles (pour une revue, voir Chauvin, 2003).

2.2.3. Carte de saillance

La partie suivante s'attache à décrire des modèles permettant de prédire les propriétés des régions sélectionnées pendant l'exploration d'une scène. Ces modèles sont majoritairement classés dans les modèles de l'attention qui se caractérisent par des mécanismes fondamentaux tels que les filtres sélectifs, spot, glue puis le focus attentionnel, et qui font une distinction entre traitements parallèles et automatiques *versus* traitements sériels et attentionnels.

Ces dernières années, grâce aux nouvelles découvertes portant sur le fonctionnement neuronal en l'aire cérébrale V1, la modélisation du système visuel devient une méthode crédible permettant de rendre compte des propriétés des informations fixées. Un point commun à l'ensemble de ces modélisations est la définition de "régions intéressantes" dans les scènes, autrement dit, la construction d'une représentation de l'environnement pondérée par l'intérêt exercé par chacune des régions. Cette représentation se définit sur le terme d'une "carte de saillance" qui est une représentation simplifiée de l'environnement, qui accentue les régions dont les propriétés diffèrent de celles des régions voisines.

[5] Le bispectre d'une fonction *f* est la transformée de Fourier de la triple corrélation :

$$g(x,y) = \iint f(t)f(t+x)f(t+y)\,\mathrm{d}t.$$

Jusqu'à présent, différents modèles de carte de saillance sont présentés dans la littérature. Cependant, quel que soit le contenu du modèle, ils sont tous caractérisés par quatre phases de traitements importantes :

1) l'extraction des traits ;

2) l'intégration des cartes de traits dans la carte de saillance ;

3) la sélection de la zone la plus saillante ;

4) la modulation pour une représentation centrale.

Nous présentons par la suite un modèle historique de Koch et Ullman (1984; 1985), et le plus cité, un modèle d'Itti, Koch et Niebur (1998).

2.2.3.1. Le modèle de Koch et Ullman (1984; 1985)

Un des modèles dont de nombreux travaux sont issus est le modèle de carte de saillance de Koch et Ullman (1984, 1985) (Figure 2.5). Une carte de saillance est un modèle imaginé par Koch et Ullman pour rendre compte d'une variété de phénomènes pré-attentifs associés au déplacement du focus attentionnel et de la dichotomie attentif / préattentif. Cette dichotomie est décrite par les modèles de la "Théorie d'Intégration des Traits" (FIT pour "*Features Integration Theory")* de Treisman et Gelade (1980) et "le modèle de recherche guidée" (*Guided Search Model)* de Wolfe et Gancarz (1996), qui supposent un premier traitement automatique sur l'ensemble du champ visuel suivi d'un traitement localisé déployé par le sujet (le focus attentionnel). Ce modèle décrit la recherche visuelle de manière suivante :

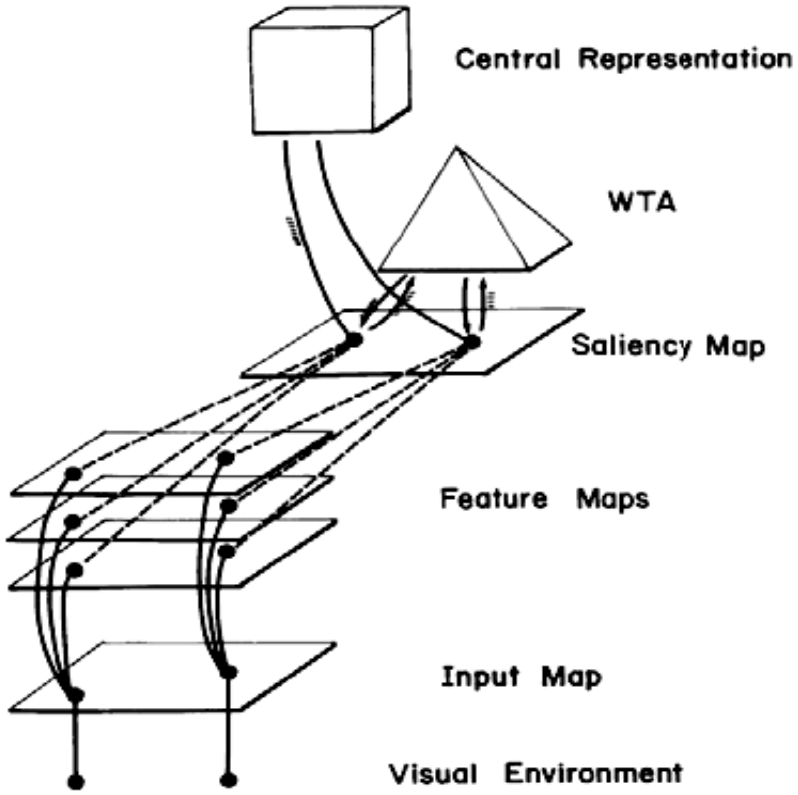


Figure 2.5. Modèle original de carte de saillance de Kock et Ullman (1984, 1985).

Un ensemble de traits sont extraits en parallèle pour générer des cartes de traits respectant la topologie de l'espace d'entrée. Des inhibitions latérales permettent d'isoler les formes différentes de leur voisinage. Ensuite, ces cartes de traits se combinent pour avoir une carte de saillance, qui est une représentation déformée de l'environnement accentuant les régions d'intérêt du champ visuel qui respecte également la topologie des entrées. Enfin, la dernière partie du modèle comporte une représentation centrale des objets de l'environnement contenant des détecteurs de traits indépendants de leurs localisations dans le champ visuel. Le lien entre la représentation centrale et la carte de saillance est l'objet d'un réseau "*winner take all*" (WTA), qui est une implantation neuronale de la fonction maximum. Ce réseau sélectionne les régions par ordre de saillance décroissante simulant ainsi le déplacement du "focus attentionnel". Si les caractéristiques d'un stimulus sont telles que sa projection sur la carte de saillance atteint un niveau de saillance maximale, alors il est sélectionné immédiatement et "saute aux yeux" (*pop-up*). En revanche, si d'autres stimuli sont plus saillants, alors le focus attentionnel s'y déplacera.

2.2.3.2. Le modèle de Itti, Koch et Niebur (1998)

Le modèle développé par Itti, Koch et Niebur (1998) est le plus largement cité dans la littérature et a donné lieu un grand nombre de publications (Itti, 2000; Itti & Koch, 2000; Itti & Koch, 2001; Niebur, Itti, & Koch, 2002;). Pourtant ce modèle n'est qu'une simple mise à jour des principes de Koch et Ullman (1984, 1985). Il se base sur l'idée d'une carte de saillance qui encode l'intensité du stimulus ou la saillance pour toute position de la scène visuelle. La carte de saillance reçoit des entrées du processus visuel primaire et permet une stratégie de contrôle efficace dans laquelle le focus attentionnel balaie simplement la carte de saillance dans un ordre décroissant de saillance. La Figure 2.6 montre comment l'image est encodée par les neurones, au travers de quelques mécanismes de détection des caractéristiques pré-attentionnelles, en cartes de contrastes pour chacune des caractéristiques. Au sein de chaque carte de caractéristiques, les neurones entrent en compétition spatiale de saillance. Les cartes de caractéristiques sont combinées ensuite pour obtenir la carte de saillance.

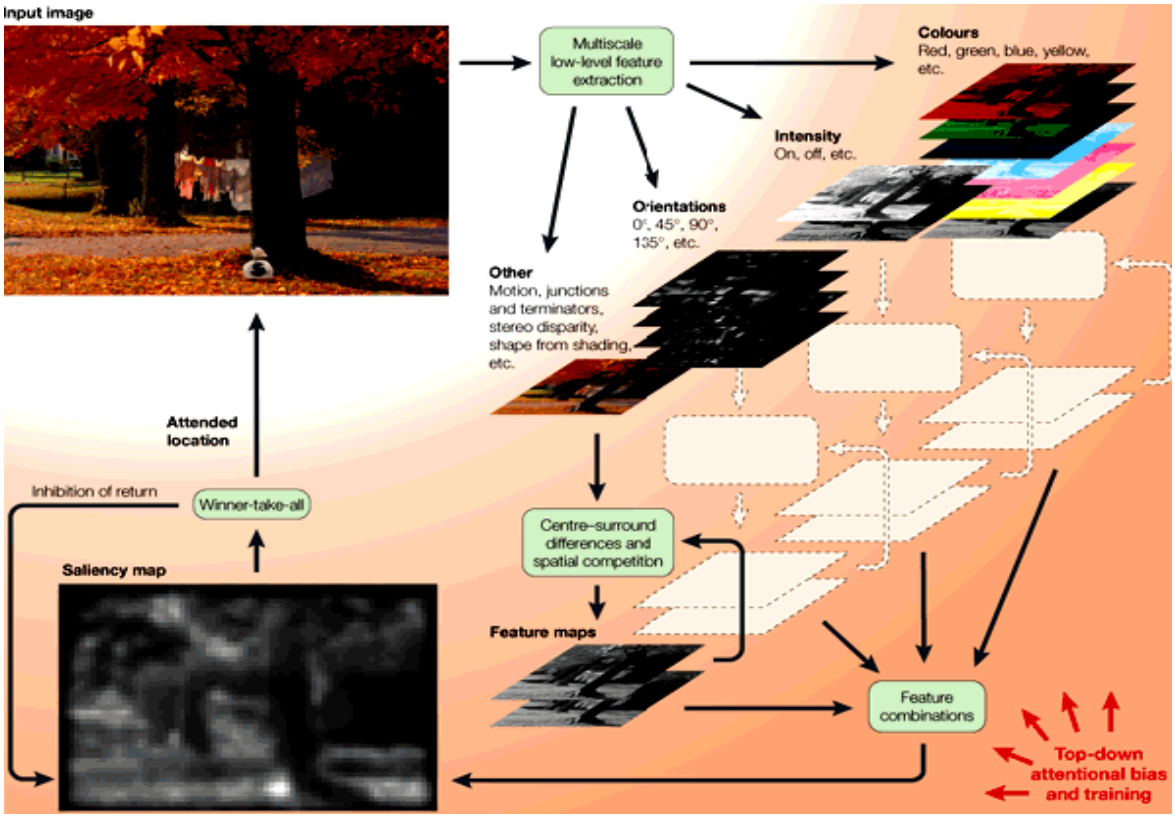


Figure 2.6. Système d'attention visuelle "*bottom-up*" (Koch & Ullman, 1985) modifié par Itti & Koch, 2001.

2.2.3.3. Problèmes et questions non résolus

1) Les propriétés des traits dans la carte de saillance ne sont pas déterminées explicitement

Le cadre de la carte de saillance est une approche très pertinente pour prédire ce qui commande les saccades oculaires, mais, il reste encore quelques points à améliorer.

Il est certain que des propriétés visuelles d'une scène influencent les points de fixations, pourtant il est difficile de déterminer exactement quelles propriétés du stimulus sont importantes, et comment ces propriétés se combinent pour influencer les lieux de fixation. Les propriétés des stimuli étudiés dans des expériences de mouvement oculaire incluent la couleur, le contraste, l'intensité, la symétrie, et la densité de bords. Un modèle de carte de saillance est souvent dérivé d'une combinaison linéaire sur l'ensemble de ces propriétés spécifiques. Cependant, jusqu'à présent, il n'y a pas encore d'évidence forte indiquant que ces propriétés spécifiques ont un rôle unique ou même central pour déterminer les saccades dans les scènes.

- Ces propriétés spécifiques sont faiblement corrélées avec les lieux réels de fixation dans les scènes (Parkhurst, Law, & Niebur, 2002).

- Seule la densité de contour et le contraste local sont prédicteurs des lieux de fixations (Mannan, Ruddock, & Wooding, 1997b; Reinagel & Zador, 1999).

2) Comment intervient le processus descendant?

Les modèles de carte de saillance indiquent les endroits saillants où se trouvent le plus grand nombre de fixations oculaires. Cependant, cette approche se concentre principalement sur les traitements ascendants pour les saccades sans préciser de manière explicite le processus descendant. La première étape de traitement dans tout modèle de type ascendant pour l'attention est le traitement

des premières caractéristiques visuelles. Elles sont calculées (traitées) de manière pré-attentive, de façon massivement parallèle à travers tout le champ de vision. D'un autre côté, l'attention peut clairement et vivement moduler, de façon descendante, les premiers traitements de la vision. En raison de la modulation par rétroaction qui influence le traitement des premières caractéristiques, les modèles doivent prendre en compte cet aspect non-négligeable. Toutefois, la saillance ne semble pas fortement corrélée au déplacement réel de la fixation, surtout lorsqu'il s'agit d'une tâche obligeant les sujets à devenir très actifs (Henderson, Weeks, & Hollingworth, 1999; Land & Hayhoe, 2001; Pelz & Canosa, 2001). Par conséquent, les facteurs sémantiques doivent être intégrés pour que le modèle soit complet.

3) La carte de saillance est elle unique ou multiple?

Pour résoudre le problème de la représentation multiple d'une vue au sein des nombreux réseaux neuronaux, beaucoup de modèles ascendants proposent que ces multiples représentations soient regroupées en une carte de saillance unique. Informatiquement, une représentation explicite de la saillance dans une carte dédiée renforce l'idée qu'une partie de la sélection spatiale puisse être réalisée au cours de la détection des caractéristiques pré-attentives. D'un point de vue "écologique", le mécanise de perception utilise-t-il une seule ou plusieurs cartes de saillance? En tout cas, les réponses pour ces questions ne sont toujours pas convaincantes.

2.2.4. Temps de fixation et propriétés physiques

L'étude du temps de fixation alloué aux propriétés physiques dans une image n'a pas été beaucoup développée jusqu'à présent. De plus, cet indicateur est souvent mesuré de manière qualitative. Cependant, certains faits sont déjà présentés dans la littérature : le temps de fixation est influencé par la luminance (Loftus, 1985) et le contraste d'une image de scène (Loftus, Kaufman, Nishimoto, & Ruthruff, 1992). Henderson et Hollingworth (1998) ont comparé les distributions du temps de fixation pour des photographies couleurs, des images couleurs

simplifiées de scènes et des dessins au trait noir et blanc. Ils n'ont pas trouvé de différence de temps de fixation pour ces trois types d'images.

Van Diepen et ses collègues ont utilisé un paradigme de "masque mobile" pour manipuler la qualité d'information visuelle disponible à la fixation indépendamment de l'information extra fovéale disponible (van Diepen, De Graef, & d'Ydewalle, 1995; van Diepen, Wampers, & d'Ydewalle, 1998). Les sujets ont pour objectif de chercher des non-objets dans des dessins de scènes de différents types, l'image à fixer étant soit normale, soit dégradée par réduction du contraste, ou par recouvrement d'un masque de bruit sur la région fixée. Le masque se déplace avec les yeux de sorte qu'il soit toujours centré sur le point de fixation, et la région à fixer se trouve dans le champ fovéal. Quand l'image est dégradée, la durée des premières fixations est plus longue qu'avec des images normales. Cette constatation suggère que le temps de fixation initiale est influencé par l'acquisition d'informations visuelles de la région fixée.

Des études en détection de changements dans les scènes montrent également l'évidence de l'influence des propriétés visuelles sur le temps de fixation. Dans le paradigme trans-saccadique, les changements apparaissant pendant l'exécution d'une saccade, les participants regardent des images de scènes complexes tout en essayant de détecter des changements dans ces scènes (Bridgeman, Hendry, & Stark, 1975; Bridgeman & Stark, 1979; Currie, McConkie, Carlson-Radvansky, & Irwin, 2000; Grimes, 1996; Henderson & Hollingworth, 1999a ; McConkie, 1990; McConkie & Currie, 1996). Ce paradigme permet d'étudier l'influence des changements de propriétés visuelles d'une région sur la durée de fixation. Les sujets ne détectent pas souvent le changement, mais le temps de fixation est plus long (Henderson & Hollingworth, 2003; Hollingworth & Henderson, 2002; Hollingworth, Williams, & Henderson, 2001; Hayhoe, Bensinger, & Ballard, 1998). Ces augmentations sont observées quand l'objet change de couleur et lors d'une rotation de l'objet de 90 degrés autour de son axe vertical. Les deux dernières manipulations modifient les détails visuels tout en préservant l'identité de l'objet. Ainsi, l'augmentation du temps de fixation suggère que les propriétés visuelles peuvent influencer le temps de fixation.

Pendant l'exploration de scènes, les sujets préfèrent fixer certains endroits "intéressants". Le choix de ces derniers dépendant non seulement des propriétés physiques et sémantiques de cette région, mais aussi de la spécificité de tâche ainsi que du but des observateurs. La première fixation ne semblerait pas être influencée par la sémantique de la scène. Cependant, au cours de l'exploration, les fixations se localisent préférentiellement sur les objets non congruents.

Le fait que la sémantique d'une scène influence postérieurement les mouvements oculaires, suggère que la thématique d'une scène (ferme, montage, plage etc.) peut-être appréhendée précocement.

Que représente l'identification d'une scène?

Comment une scène peut-elle être identifiée rapidement?

Le chapitre suivant à pour objectif d'apporter des réponses concernant ces deux questions.

Chapitre 3 Scène naturelle, structure et sens général

1. Scène naturelle, un objet d'étude différent

1.1. Complexité des scènes naturelles

"*Notre environnement visuel est d'une extraordinaire richesse. 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. Pourtant, de nombreuses études révèlent que les mécanismes visuels implémentés dans le cerveau sont par essence limités, incapables de saisir toute la richesse du monde extérieur à chaque instant*" (Rousselet, 2003).Par exemple, notre capacité à détecter un changement dans une image est remarquablement faible. Ce résultat provient d'un ensemble d'études concernant la "cécité au changement" (*change blindness*) : les sujets sont souvent incapables de détecter un changement parfois très important intervenant dans une image après une brève interruption de la présentation de la scène, ou seulement après une longue série d'alternances entre la scène d'origine et la scène modifiée (Rensink, 2002; Simons & Levin, 1997). Ces défaillances de détections supposent qu'une scène est représentée de manière implicite, soit peu détaillée et que seule son identité soit gardée. Sans attention focalisée portant sur l'objet cible avant modification, les changements ne sont pas détectés. Or, il est possible que notre système visuel soit en grande partie amnésique, comme le suggère Wolfe (1999), et que nous ayons recours à la richesse du monde qui nous entoure plutôt qu'à une hypothétique représentation interne de scène.

1.2. Invariance structurale

Une scène possède des régularités et des propriétés spécifiques concernant sa structure globale. Par exemple, dans une scène de cuisine (Figure 3.1), on trouve un plafond, puis des murs liés au plafond, etc. Ce sont des éléments nécessaires pour construire une scène de cuisine, la structure des cuisines ne variant pas beaucoup. Un autre exemple concerne une scène de

montagne, celle-ci est normalement caractérisée par une partie de ciel en haut, et une partie composée par des rochers ou des arbres, etc. Généralement, cette structure globale reste invariante dans la plupart du même type de scène.

Figure 3.1. Un exemple d'une scène de cuisine.

Une scène ne peut pas être considérée simplement comme la somme de tous les objets qu'elle contient. Elle a une structure qui est déterminée par la relation spatiale des objets qui s'y trouvent. En effet, cette relation est déterminée non seulement par la contrainte physique de l'univers (la gravité et l'espace), mais aussi par la contrainte sémantique imposée par l'identité et le fonctionnement des objets (Biederman, Mezzanotte, & Rabinowitz, 1982). Par exemple, deux objets ne peuvent pas occuper une même place en même temps, ou un extincteur ne se trouve pas au dessus d'une boite aux lettres (Biederman, Mezzanotte, & Rabinowitz, 1982).

1.3. Contexte

Le contexte de scène est un élément spécifique, qui caractérise les objets lui appartenant et qui ont souvent une cohérence sémantique. Les éléments du contexte varient en fonction du cadrage de la prise de vue. Par exemple, une scène prise à partir du point de vue de la porte d'un bureau, se composerait du sol,

du plafond, des murs comme des éléments contextuels, et d'une table, d'un bureau, d'une chaise, d'un téléphone, et d'autres objets bien positionnés. Cependant, si une scène est photographiée par quelqu'un assis sur une chaise devant la table d'un bureau, une partie de ce bureau pourrait être considérée comme contexte, l'ordinateur et le téléphone pourraient être considérés comme les objets composant la scène.

Selon la définition de Henderson et Hollingworth (1999), les objets sont associés avec le contexte d'une scène, ils ne sont généralement pas dissociés les uns des autres (Figure 3.2). Par exemple, qu'est-ce qui est caché dans les deux ellipses colorées? Il y aura une forte chance que l'on dise : des voitures et des personnes. Le contexte visuel d'une scène semble pouvoir activer les connaissances portant sur sa catégorie et sur la localisation des objets s'y trouvent. Il guide ainsi l'orientation de l'attention vers des cibles potentielles dans les tâches de recherche visuelle (Chun, 2000; Torralba, Oliva, Castelhano, & Henderson, 2006). Il influence alors 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 (Henderson & Hollingworth, 1999).

Figure 3.2. Illustration de la relation entre des objets et le contexte : image gauche est une image masquée par les deux ellipses colorées d'image originale (image droite).

1.4. Structure neuro-anatomique impliquée dans la perception de scènes

Tout au long de la voie du traitement de l'information, le champ visuel et les activités neuronales deviennent de plus en plus complexes, les attributs physiques correspondants à chaque degré du champ visuel déclencheraient des activations neuronales différentes (Figure 1.10 du chapitre 1). Selon cette hypothèse, les scènes pourraient être traitées différemment d'autres entités comme, par exemple, les lettres, les objets ou les visages. Cette dichotomie s'observe par les structures neuronales localisées dans les zones cérébrales différentes.

Lié spécifiquement aux scènes naturelles, le cortex parahippocampique (PPA, Parahippocampal Place Area) est activé par des stimuli tels que les objets, les visages et les maisons (Figure 3.3).

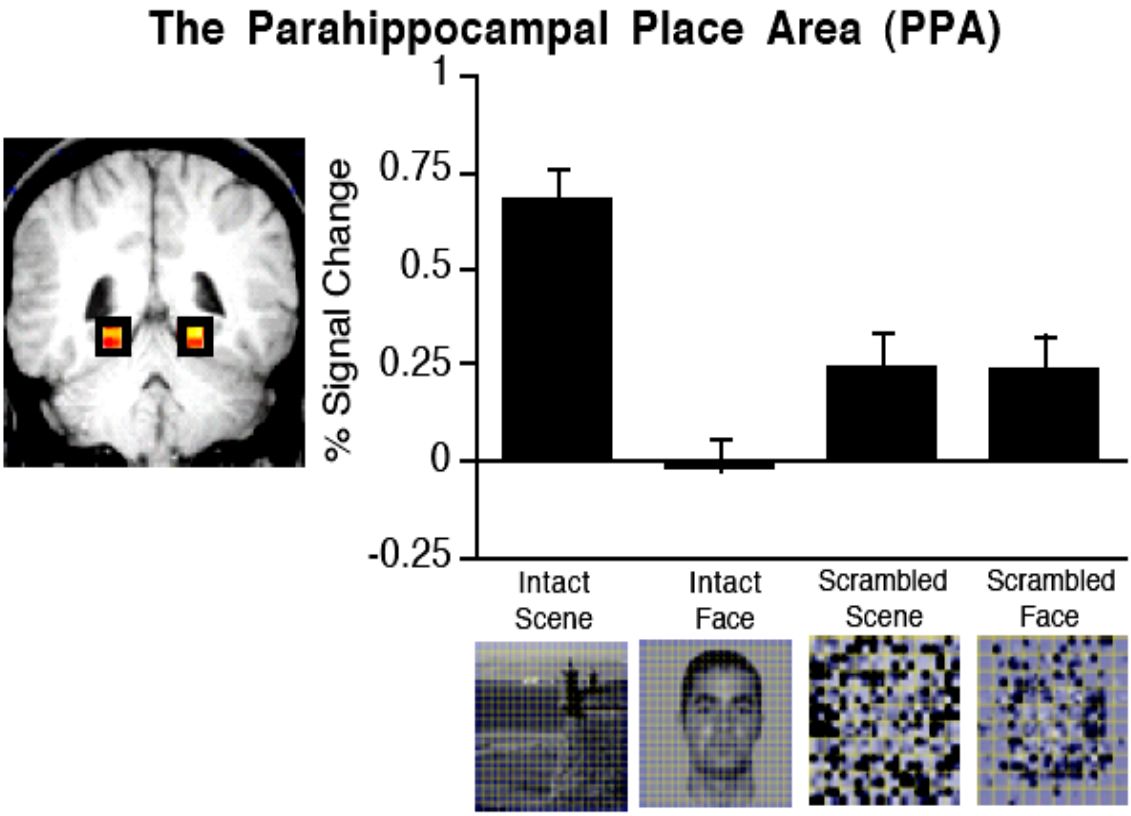


Figure 3.3. Caractéristique d'activation du cortex PPA (Chun, 2003).

De manière générale, le cortex PPA et le cortex pariétal semblent être préférentiellement activés pendant la perception d'une scène naturelle plutôt que pendant la perception d'un visage par exemple (Nakamura *et al*., 2000; Sato, 1999). Une activation différente du cortex PPA est constatée lors d'une étude concernant la perception d'une scène visuelle et d'un visage (O'Craven & Kanwisher, 2000).

De manière plus spécifique, le cortex PPA serait impliquée dans le traitement de la structure spatiale de la scène (Epstein, 2005; Epstein, Harris, Stanley, & Kanwisher, 1999; Epstein & Kanwisher, 1998; Epstein, DeYoe, Press, Rosen, & Kanwisher, 2001). Enfin, le cortex PPA semble sensible aux changements liés à la scène mais pas à ses éléments constitutifs (Epstein, Graham, & Downing, 2003).

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 PPA (Aguirre & D'Esposito, 1999; Epstein, DeYoe, Press, Rosen, & Kanwisher, 2001; Farah, 1990; Mendez & Cherrier, 2003).

Une autre étude complémentaire, montrant le rôle spécifique du cortex PPA dans la perception des scènes, a été présentée par Steeves, Humphrey, Culham, Menon, & Goodale (2002). Ces chercheurs ont examiné la performance d'un patient (D. F.) présentant une agnosie de forme visuelle (incapacité grave d'identification des objets) dans une étude en IRMf (imagerie par résonance magnétique fonctionnelle). Le patient D. F. devait réaliser une tâche d'identification de scènes, il pouvait classer celles-ci, en particulier lorsqu'elles étaient présentées en couleurs plutôt qu'en noir et blanc. Une activation de la région PPA a été alors observée par les chercheurs.

2. Sens général et identification de scènes

2.1. Sens général d'une scène

En quoi consiste l'identification d'une scène ? Comment une scène naturelle est-elle représentée ? Ces deux question sont, en effet, liées à l'identification du "sens général d'une scène" ("*gist of a scene*"). Cependant, le "sens général d'une scène" est une définition peu précise. D'une manière générale, le terme "sens général d'une scène" fait référence à sa signification et à l'appréhension de son niveau de catégorisation primaire (par exemple, s'il s'agit d'une ville, d'une chambre d'enfant). Il peut être présenté par un des aspects suivants (Figure 3.4) :

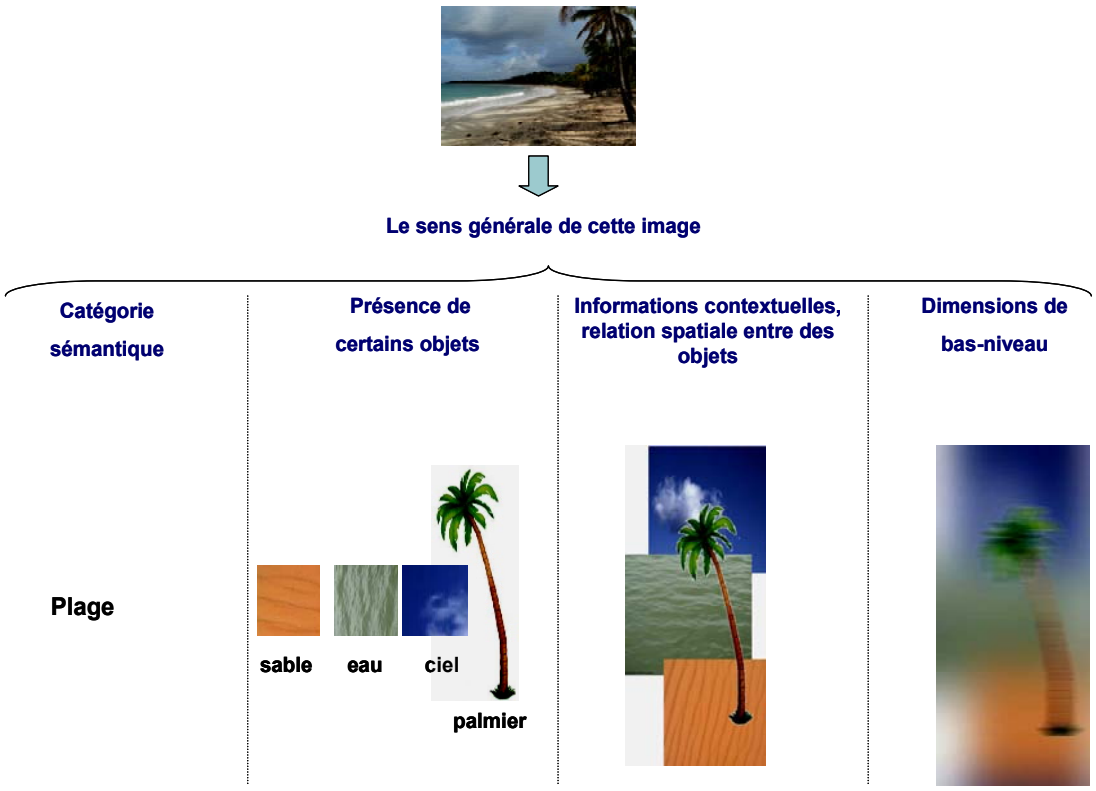


Figure 3.4. Illustration du sens général pour une scène de plage.

- catégorie sémantique (ville ou plage, par exemple), (Potter, 1975; Schyns & Oliva, 1994; Greene & Oliva, 2006; Fei Fei, Fergus, & Perona, 2004; Castelhano & Henderson, 2005) ;

- présence de certains objets, (Staub, Rado, & O'Connor, 2002; Potter, Staub, Rado, & O'Connor, 2002; Intraub, 1997; Fei Fei, Fergus, & Perona, 2004; Greene & Oliva, 2006; Wolfe, 1998) ;
- informations contextuelles et la relation spatiale entre des objets (Torralba & Oliva, 2003; Greene & Oliva, 2006) ;
- dimension de bas-niveau (Oliva & Torralba, 2001, 2003, 2006; Tversky & Hemenway, 1983).

Le sens général d'une scène peut être appréhendé extrêmement rapidement :

- en une seule fixation (Biederman, 1972; Intraub, 1981) ;
- en 20 ms (Bacon-Macé, Macé, Fabre-Thorpe, & Thorpe, 2005; Delorme, Richard, & Fabre-Thorpe, 2000; Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001; Rousselet, Joubert, & Fabre-Thorpe, 2005; Thorpe, Fize, & Marlot, 1996; VanRullen & Thorpe, 2001) ;
- entre 20 et 50 ms (Oliva, 2005 ; Potter, 1976, 1993, 1999; Potter, Staub, & O'Connor, 2004; Potter, Staub, Rado, & O'Connor, 2002 ; Wolfe, 1998).

Cependant, il est alors nécessaire que le traitement de l'image puisse perdurer au moins 150 à 200 ms après que l'image ait disparu du champ visuel pour que cette information puisse se consolider (Thorpe, Fize, & Marlot, 1996; VanRullen & Thorpe, 2001).

2.2. Méthodologies de l'étude du sens général d'une scène

Le sens général d'une scène s'inscrit dans la compréhension de celle-ci depuis de nombreuses années. Différentes méthodologies permettent de montrer qu'il peut être identifié rapidement.

Tout abord, le sens général d'une scène peut être déterminé en utilisant le paradigme de détection d'objet (Biederman, 1972; Biederman, Glass, & Stacey, 1973; Biederman, Rabinowitz, Glass, & Stacey, 1974; Biederman, Mezzanotte, & Rabinowitz, 1982; Boyce, Pollatsek, & Rayner, 1989; Hollingworth & Henderson, 1998; Murphy & Wisniewski, 1989). Dans ce paradigme, un mot désignant l'objet cible est montré, suivi d'une scène présentée brièvement puis d'un masque. Les participants doivent préciser si l'objet cible était présent ou non dans la scène. L'objet cible peut être cohérent ou non avec la scène (la cohérence peut être définie selon la sémantique de la scène ou de sa structure). Les résultats montrent que les participants sont influencés par la cohérence entre la cible et la scène. Par conséquent, le sens général de la scène a dû être acquis pour établir ce rapport.

L'étude des mouvements oculaires constitue la deuxième approche méthodologique. Elle montre que le sujet a tendance à chercher des informations intéressantes après une première fixation. (Mackworth & Morandi, 1967; Antes, 1974; Loftus & Mackworth, 1978). Le fait que les fixations dans une scène peuvent être immédiatement dirigées vers des régions périphériques informatives implique les hypothèses suivantes :

- Premièrement, les informations sur le sens général d'une scène seraient disponibles juste après la première fixation (Loftus & MacWorth, 1978).
- Deuxièmement, le rapport sémantique entre le sens général de la scène et des objets périphériques pourrait être compris rapidement. Ainsi, un grand nombre d'objets dans la périphérie semblent traités en parallèle ou avec un processus sériel extrêmement rapide.

Le troisième type de méthodologie et caractérisé par un ensemble d'expériences classiques rapportées par Potter (1999) en utilisant le paradigme "présentation visuelle sérielle rapide" (*Rapid Serial Visual Presentation*), et d'autres études du même type (Intraub, 1979, 1980, 1981). Les images sont présentées très rapidement les unes après les autres selon deux conditions différentes : une condition de détection et une condition de mémorisation. Les sujets doivent détecter si une scène cible prédéfinie est présentée ou non

(condition de détection), et répondre si une image vue auparavant a été présentée ou non (condition de mémorisation). Les résultats montrent que, dans les deux conditions, les sujets n'ont aucune difficulté à trouver une image-cible placée dans une séquence d'images avec des temps de présentation de 100 ms (Intraub, 1981; Potter, 1975, 1976). 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 ms sans masque (Biederman, 1972, 1981; Biederman, Glass, & Stace, 1973, Biederman, Rabinowitz, Glass, & Stacey, 1974).

Des résultats similaires sont observés dans une tâche de catégorisation de type go/no go (les participants doivent relâcher un bouton si la scène présentée contient la cible préalablement définie, par exemple un animal "go", ou, au contraire, maintenir le bouton pressé si la scène ne contient pas la cible "no go"). Dans ces études, les sujets devaient détecter la présence d'un animal dans des photographies de scènes naturelles présentées pendant 20 ms. Une telle tâche de catégorisation était réalisée à la fois très précisément (94% de bonne réponse) et rapidement (temps de réaction médian est de 445 ms).

Ces résultats ont été répliqués plusieurs fois selon différentes conditions expérimentales :

- scènes présentées en couleur (Thorpe, Fize, & Marlot, 1996) ;
- scènes présentées en noir et blanc (Delorme, Richard, & Fabre-Thorpe, 2000) ;
- la cible appartienne à une catégorie biologique ou non (VanRullen & Thorpe, 2001) ;
- sujets entraînés (Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001) ;
- tâche sollicitant des ressources attentionnelles (Fei Fei, VanRullen, Koch, & Perona, 2002, 2005).

Les conditions de la réussite de cette tâche indiquent que le sens général de la scène est appréhendé très précocement durant le traitement visuel de la scène (Bar, 2004; Biederman, 1972 ; Irwin & Yeomans, 1986; Oliva, 2005 ; Potter, 1976, 1993, 1999; Potter, Staub, & O'Connor, 2004; Potter, Staub, Rado, & O'Connor, 2002 ; Wolfe, 1998), probablement à partir du résultat des traitements préattentifs, qui calculent massivement, et en parallèle, chacune des dimensions de bas niveau de la scène (Treisman, 1998; Treisman & Gelade, 1980).

La partie suivante s'attache à présenter en détail comment le sens général est identifié rapidement.

2.3. Saisie d'informations et identification

Comme le montrent centaines données neuro-anatomiques, il existe un réseau largement distribué d'aires corticales, permettant divers aspects du traitement visuel des scènes naturelles. S'il est pour l'instant difficile d'envisager la façon selon laquelle chaque partie de ce réseau fonctionne, il existe cependant des données comportementales ainsi que des travaux de modélisation permettant de mieux comprendre comment le sens général d'une scène peut être déterminé très rapidement. Cette rapidité est due aux différents types d'informations utilisées et aux modes de traitement différents.

2.3.1. Objet typique

Le sens général d'une scène pourrait être déduit suite à un traitement très rapide d'un ou plusieurs objets typiques (Friedman, 1979) ainsi que leurs relations spatiales (De Graef, Christiaens, & d'Ydewalle, 1990). Par exemple, la présence d'un four seul, ou d'un fourneau et d'un réfrigérateur, suggère normalement une cuisine. Cette hypothèse suppose que l'identité d'un objet typique soit porteuse de l'identification de la scène.

Une telle proposition est basée sur un point de vue dominant de la perception de scènes : l'identification d'une scène serait l'aboutissement d'une série de traitements effectués dans la voie ventrale. Selon cette logique, les objets seraient systématiquement traités avant la scène.

Supposons que le traitement des informations relatives aux objets passe par les processus suivants. D'abord, l'image rétinienne est traduite en un ensemble d'informations visuelles primitives, par exemple les surfaces et les bords. Ensuite, ces informations primitives sont utilisées pour construire des descriptions structurales de la symbolique de l'objet dans la scène. Enfin, ces significations structurales construites sont appariées avec les représentations stockées en mémoire à long terme. Quand un appariement est trouvé, l'identification se produit, et l'information sémantique stockée en mémoire concernant ce type d'objet devient disponible (Figure 3.5). Un tel point de vue implique que les deux premières étapes doivent être considérées comme un traitement perceptif ayant pour objectif de traduire la stimulation rétinienne en description structurale, cette dernière étant compatible avec les représentations stockées en mémoire. L'étape d'appariement peut, cependant être considérée comme une interface entre perception et cognition, dans laquelle l'information perçue est associée avec les représentations en mémoire.

Nous pouvons distinguer les modèles de l'identification de l'objet dans une scène en trois catégories selon l'étape de son identification et le rôle joué par le contexte :

- la première catégorie suppose que les informations contextuelles influenceraient les deux premières étapes de l'identification d'objet ;
- la seconde les place au moment de l'interaction lors de l'étape d'appariement, quand des descriptions perçues sont appariées aux représentations en mémoire à long terme ;
- la troisième suggère que l'identification d'objet (étape d'appariement comprise) est isolée dans l'identification de scène.

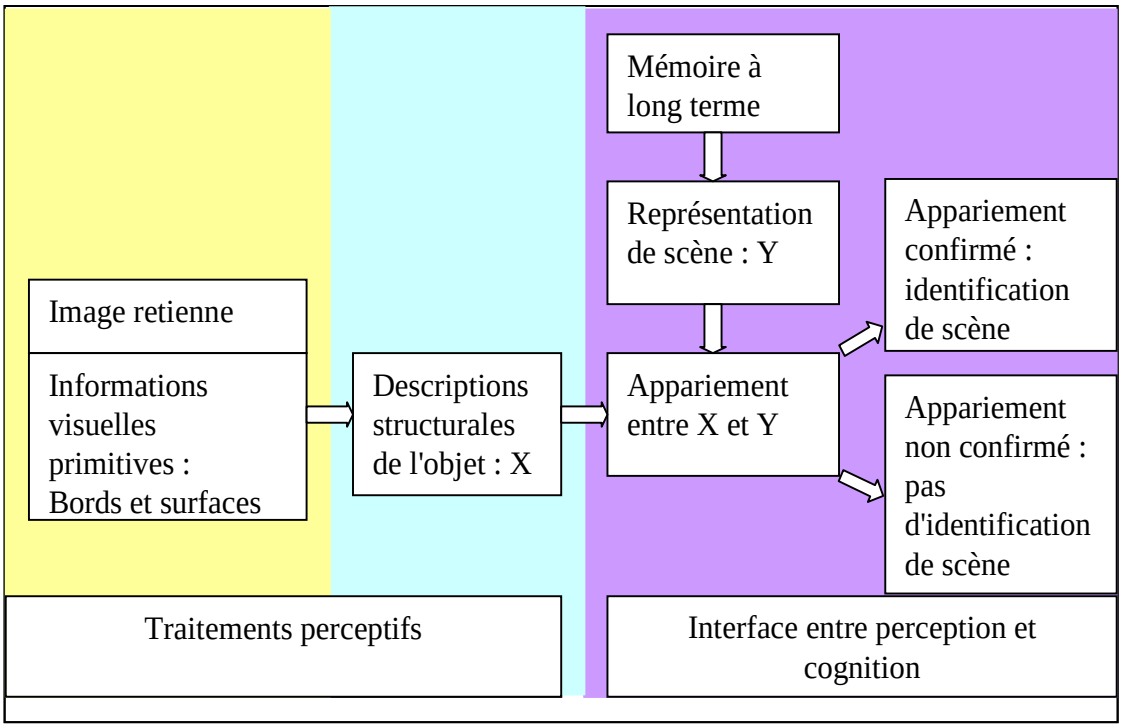


Figure 3.5. Illustration simplifie des trois étapes de l'identification de l'objet de l'image.

2.3.1.1. Le modèle de schéma perceptuel

Dans ce modèle, l'hypothèse sous-jacente serait qu'il existe en mémoire un schéma de scène (la représentation en mémoire d'une scène typique). Celui-ci est interactif avec l'analyse perceptuelle d'objet dans la scène (Biederman, 1981 ; Biederman, Mezzanotte, & Rabinowitz, 1982; Boyce, Pollatsek, & Rayner, 1989; Palmer, 1975). Le schéma de scène contient des informations sur les objets et leurs relations spatiales. L'activation précoce d'un schéma de scène faciliterait l'analyse d'objets du type "schéma-cohérent" et, peut-être, inhiberait l'analyse perceptuelle d'objets "schéma-incohérent" (Biederman, Mezzanotte, & Rabinowitz, 1982; Boyce, Pollatsek, & Rayner, 1989) Ce modèle prévoit que l'identification des objets cohérents à une scène soit facilitée par rapport aux objets incohérents.

2.3.1.2. Le modèle d'amorçage

Ce modèle suppose que l'influence contextuelle apparaisse quand la perception structurale de l'objet soit appariée avec une des représentations en mémoire à long terme (Bar & Ullman, 1996; Friedman, 1979; Friedman & Liebelt, 1981; Kosslyn, 1994; Palmer, 1975; Ullman 1996). L'activation d'un schéma de scène amorce les représentations stockées concernant les objets de type "schéma-cohérent". Cet amorçage peut être considéré comme une modulation du critère de quantité d'informations perceptuelles nécessaires pour choisir une représentation particulière d'objet afin de pouvoir procéder à un appariement. Friedman (1979) a prouvé que moins d'informations perceptuelles devront être codées, dans le choix d'une représentation d'objet amorcé, par comparaison à la représentation d'objet non amorcé. Cette conception ressemble au modèle précédent concernant la précédence de l'identification des objets cohérents par rapport à l'identification des objets incohérents. Cependant, elle diffère parce qu'elle propose que la connaissance de scène influence, seulement, le critère employé pour déterminer si un type particulier d'objet est présent, sans pour cela, influencer directement l'analyse perceptuelle de la symbolique de l'objet.

2.3.1.3. Le modèle fonctionnel d'isolement

Le modèle fonctionnel d'isolement propose que l'identification d'objets soit isolée de la connaissance de scènes (Hollingworth & Henderson, 1998). Ce modèle est conforme aux théories de l'identification d'objet (Biederman 1987; Bülthoff, Edelman, & Tarr, 1995; Marr & Nishihara 1978). Dans celles-ci, il est suffisant de distinguer des analyses visuelles ascendantes entre les catégories d'objets. Ces théories sont également conformes à celles proposant une division architecturale entre traitement perceptuel et traitement sémantique (Fodor, 1983; Pylyshyn, 1980, 1998). Le modèle fonctionnel d'isolement prévoit que les expériences concernant l'analyse perceptuelle des objets ne devraient trouver aucun effet de relation entre l'objet et la scène.

Cependant, la majeure partie des recherches sur les scènes naturelles est dédiée au traitement des objets dans les scènes, négligeant le traitement des scènes elles-mêmes dans leur ensemble. Cette question est importante dans la mesure où les informations grossières d'une scène, sa structure spatiale ou sa catégorie sémantique, pourraient servir à comprendre le sens général de la scène.

2.3.2. Propriétés de "scène-niveau"

Un point de vue alternatif émerge de plusieurs études ayant montré que la compréhension d'une scène peut s'effectuer très efficacement, même lorsque des dessins ou des photographies sont très brièvement présentées à des sujets humains (Biederman, 1981, 1988; Biederman, Mezzanotte, & Rabinowitz, 1982; Intraub, 1999; Oliva & Schyns, 1997, 2000; Potter, 1975, 1976; Potter & Levy, 1969; Schyns & Oliva, 1994). Le fait que l'identification du sens général d'une scène soit toujours possible, même dans des conditions de présentations très brèves, est souvent considéré comme une preuve que les mécanismes sous-jacents sont également rapides. Selon cette logique, certains suggèrent que l'identification des scènes peut être réalisée de manière concomitante, ou même avant l'identification des objets, et qu'elle soit basée sur des propriétés de "scène-niveau".

2.3.2.1. Fréquences spatiales

La première propriété de "scène-niveau" serait relative aux fréquences spatiales. Schyns et Oliva (1994) ont construit des scènes hybrides, constituées de la partie basse fréquence (BF) d'une scène et de la partie haute fréquence (HF) d'une autre (Figure 3.6). En s'appuyant sur la courbe de sensibilité au contraste, elles considèrent que les basses fréquences sont inférieures à deux cycles par degré d'angle visuel, et les hautes fréquences à six cycles par degré. Ces images contiennent des informations contradictoires en hautes et basses fréquences comme les stimuli hiérarchiques.

Figure 3.6. Images hybrides. L'image hybride gauche superpose une scène filtrée passe-haut (fréquence de coupure supérieure à six cycles par degré angulaire) et une scène autoroute filtrée passe-bas (fréquence de coupure inférieure à deux cycles par degré angulaire) ; l'image hybride droite superpose les mêmes scènes avec un pattern de filtrage inversé (Schyns et Oliva ,1994)

Dans l'expérience de Schyns et Oliva (1994), les sujets doivent dire si la scène "test" (ville, autoroute, vallée, salon), présentée seule et sans filtrage, se trouve ou non dans l'image hybride présentée antérieurement. Dans cette expérience, elles montrent que pour catégoriser une scène naturelle les basses fréquences suffisent. Ensuite, en alternant une expérience d'appariement et une procédure de sensibilisation (Oliva & Schyns, 1997), les auteurs montrent une dominance des basses fréquences sur les hautes fréquences pour une présentation rapide des images de 30 ms (63% [BF] vs 28% [HF]) et une dominance des hautes fréquences pour une présentation plus longue de 150 ms (18% [BF] vs 86% [HF]). De plus, ces auteurs montrent d'une part que ces résultats ne sont pas dus à un déficit de perception des hautes fréquences (Schyns & Oliva, 1994) et d'autre part, que la dominance des basses fréquences est modulable par les contraintes de la tâche (Schyns & Oliva, 1999).

Gaillard et Bourges (1999) ont réalisé une expérience comparant des scènes de type "campagne" à des scènes de type "ville" en manipulant le filtrage fréquentiel (BF et HF). Ils concluent que les images de type "campagne" contiennent plus de hautes fréquences spatiales que les images de type "ville". Celles-ci sont moins bien et plus lentement reconnues que celles de type "ville".

Cependant, en ce qui concerne les images de type "ville", comportant plus de basses fréquences que celles de type "campagne", celles-ci sont aussi bien et aussi rapidement reconnues quel que soit leur mode de filtrage.

La dominance globale n'est donc pas tout à fait confirmée dans cette expérience.

2.3.2.2. Couleur

Oliva et Schyns (2000) montrent que la couleur semble jouer un rôle dans la catégorisation rapide de scènes, du moins lorsqu'elle est considérée comme facteur caractéristique de sa catégorie.

Des sujets doivent dénommer ou vérifier la catégorie de chaque scène (plage ou montagne) présentée pendant 120 ms, soit avec leur couleur normale, une couleur anormale, ou en niveaux de gris. Pour les scènes dans lesquelles la couleur était caractéristique de sa catégorie, la présence de la couleur normale facilite la catégorisation. Par exemple, la plage avec des sables jaune et l'eau verte est mieux dénommée par rapport à la plage rouge et à l'eau bleu foncée. Pour les scènes dans lesquelles la couleur n'était pas caractéristique de sa catégorie, aucun effet de couleur n'a été trouvé. La couleur normale s'est également avérée être un facteur facilitant la catégorisation des scènes filtrées "passe-bas" (BF) quand la couleur était caractéristique de sa catégorie. Ce dernier résultat suggère qu'une organisation brute grossière caractéristique de la couleur peut être suffisante pour la catégorisation de scènes (Oliva & Schyns, 2000).

2.3.2.3. D'autres propriétés de "scène-niveau"

D'autres données expérimentales en psychophysique prouvent que la catégorisation de scènes est possible en se basant sur des informations perceptives. Les auteurs (Oliva, Torralba, Guérin-Dugué, & Hérault, 1999; Guérin-Dugué & Oliva, 2000; Oliva & Torralba 2001) montrent qu'il est possible d'associer à chaque catégorie de scènes un spectre prototypique ou une classe de spectres

représentatif d'une catégorie (Figure 3.7). Les filtres de Gabor échantillonnent le spectre d'énergie et permettent de projeter chaque image dans un espace multidimensionnel favorisant le regroupement par catégorie (voir la thèse de A. Chauvin, 2003).

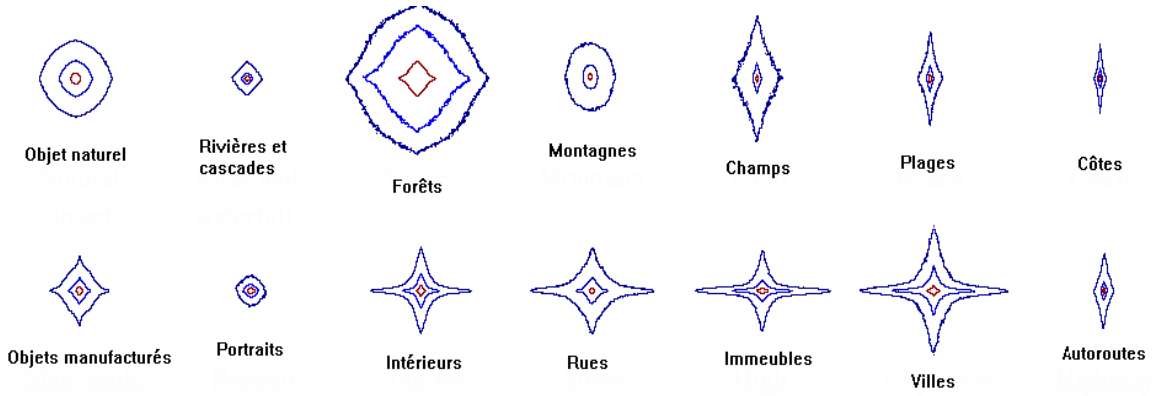


Figure 3.7. Spectre d'amplitude prototypique de plusieurs catégories de scènes naturelles (Chauvin, 2003).

Récemment, Oliva et Torralba (2001, 2003, 2006) ont émis l'hypothèse qu'une scène peut-être caractérisée par son niveau global. Une représentation holistique et "bas-dimensionnelle" est nommée "enveloppe spatiale". Celle-ci caractérise un ensemble de sept dimensions pouvant représenter une scène (Figure 3.8).

- **Etendue** : vue claire de la ligne d'horizon.
- **Profondeur** : degré de perception pour la profondeur de scène, par exemple, y a-t-il un plan rapproché au bout d'un mètre? Peut-on voir des kilomètres en arrière plan?
- **Camouflage** : peut-on cacher quelque chose?
- **Mouvement** : la scène se déplace-t-elle ou est-elle changeante? Certains types de scènes possèdent cette caractéristique, par exemple, une scène de coucher de soleil, ou une scène de paysage concernant le brouillard se levant, etc.
- **Navigation** : degré de difficulté à traverser l'environnement.
- **Température** : température physique de l'environnement extérieur, une scène de désert correspondrait à cette catégorie.

- **Expansion** : indices perceptifs dans l'image avec des lignes parallèles convergentes, ou point de vue plat sur une surface simple ?

Figure 3.8. Exemple de scènes caractérisant par certaines dimensions de "scène niveau" ainsi que le degré de chaque caractéristique du niveau le moins représentatif au plus représentatif (Greene & Oliva, 2006).

Cependant pour l'obtention de ces septe dimensions, il suffit de calculer directement l'information spectrale grossièrement codée dans une image de scène, sans tenir compte les objets ou d'autres primitives de formes (Oliva & Torralba, 2001, 2003, 2006). La modélisation de cette approche s'est avérée capable de classer un grand ensemble de scènes qui s'accordent bien avec la classification produite par les sujets humains.

2.3.2.4. Grandes formes volumétriques

D'autres théories postulent que les scènes pourraient être identifiées à partir d'indices visuels qui leur seraient propres. Par exemple, Biederman (1995) a proposé un modèle structural de la reconnaissance d'objets fondé sur l'usage de "géons" (formes primitives tridimensionnelles, Biederman, 1987) pouvant 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 les informations visuelles spécifiques des scènes indépendamment des informations sur les objets. Cette proposition n'a pas encore été testée de manière empirique.

Cependant, cette hypothèse semble aller l'encontre de la théorie courante (approche locale). D'une part, les modèles par description structurale ne sont pas plausibles biologiquement (Rolls & Leco, 2002). D'autre part, comme le font remarquer Henderson et 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 à un niveau grossier, puisse permettre son identification (Sanocki & Epstein, 1997).

Cette idée nous paraît intéressante, d'une part, si une scène pouvait être considérée comme un objet "particulier", alors la théorie de la reconnaissance de l'objet (Biederman, 1987, Gaillard, Boulliou & Gautier, 1996) pourrait s'appliquer à la reconnaissance de la scène. Inversement, si elle ne peut pas être considérée comme un objet, y-a-t-il des propriétés caractérisant les formes volumétriques qui puissent servir dans l'identification de la scène ?

2.3.3. Information contextuelle et connaissance de la tâche

L'identification du contexte pourrait 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 fourni 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 pourrait être considéré comme un autre facteur important influençant la performance, probablement en permettant aux experts d'accéder à des détails caractéristiques 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 capter le plus efficacement possible les propriétés diagnostiques des cibles de la tâche. Autrement dit, une catégorisation de haut niveau ne dépend pas forcément d'une représentation de haut niveau si des représentations de plus bas niveau permettent de catégoriser correctement un stimulus dans une tâche donnée (Ullman, Vidal-Naquet, & Sali, 2002).

2.4. Traitement d'information et identification

Nous venons de voir différents types d'informations saisies pouvant expliquer la compréhension rapide d'une scène.

Quels sont alors les mécanismes de traitements correspondants à ces informations saisies ? L'attention joue-t-elle un rôle important pour sélectionner ces informations ? Quelles stratégies sont alors utilisées par les sujets dans l'identification rapide de scènes ?

Par la suite, différents processus de traitement des informations au sein des scènes naturelles seront présentés.

2.4.1. Processus pré-attentif et attentif

La dichotomie entre traitements pré-attentifs et traitements nécessitant une attention focalisée est d'abord abordée dans la "Théorie d'Intégration des Traits " (FIT, pour *Feature Integration Theory*, Treisman & Gelade, 1980). Cette théorie stipule que les mécanismes visuels de la perception des scènes naturelles peuvent être 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 perceptifs tels que la couleur, la texture, les contours locaux, le mouvement, la taille, etc. Cette activité semblerait fonctionner de manière automatique. 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.

Dans une seconde étape, l'attention agit de manière sérielle afin d'assembler différents éléments de base constituant un objet complexe, puis d'en former une représentation de haut niveau. Les objets constitués par un ensemble d'éléments de base ne pourraient être représentés dans le système visuel sans faire appel à l'attention. Il existerait selon Treisman et 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 l'appariement (mise en place de liens dynamiques) entre la représentation de cette position dans la carte de contrôle et chacune des caractéristiques codées de cette position dans les différentes cartes d'éléments. Cet appariement permettrait d'assembler de manière explicite les éléments présents sur cette position en une représentation complexe tout en excluant les éléments se trouvant sur d'autres positions.

Avant la focalisation de l'attention, le monde visuel serait composé d'attributs physiques discrets, sans organisation spatiale. 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.

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.

2.4.2. Modèle sériel

Le modèle sériel a beaucoup apporté dans les études de recherche visuelle. Par exemple, dans une tâche de recherche visuelle, les sujets ont pour consigne de rechercher un objet cible prédéterminé parmi des distracteurs (un ensemble d'objets non cibles). La cible apparaît en général dans 50% des cas et les sujets répondent en pressant un bouton (ou une touche du clavier) pour indiquer qu'ils ont trouvé la cible et sur un autre bouton pour indiquer que la cible est absente. Dans ce type de tâche, le temps de réaction et parfois la réponse du sujet (bonne ou mauvaise réponse) sont mesurés. Ainsi, il a été montré que le temps nécessaire pour détecter une cible dépend du type de cible utilisé et du nombre de distracteurs. Quand les sujets doivent trouver une cible qui diffère des distracteurs selon une seule dimension, par exemple, chercher une barre horizontale parmi des barres verticales, leurs temps de réaction sont relativement courts et indépendants du nombre de distracteurs (Figure 3.9).

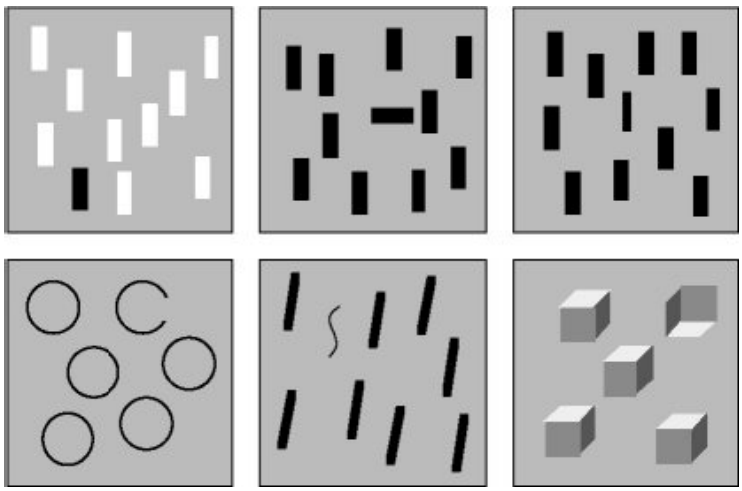


Figure 3.9. Eléments utilises comme distracteurs, tiré de Wolfe (2001).

En variant systématiquement le nombre de distracteurs présentés, il est possible de construire, une fonction de recherche décrivant l'évolution des temps de réaction en fonction du nombre de distracteurs dans une recherche de conjonction d'éléments (Figure 3.10). 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.

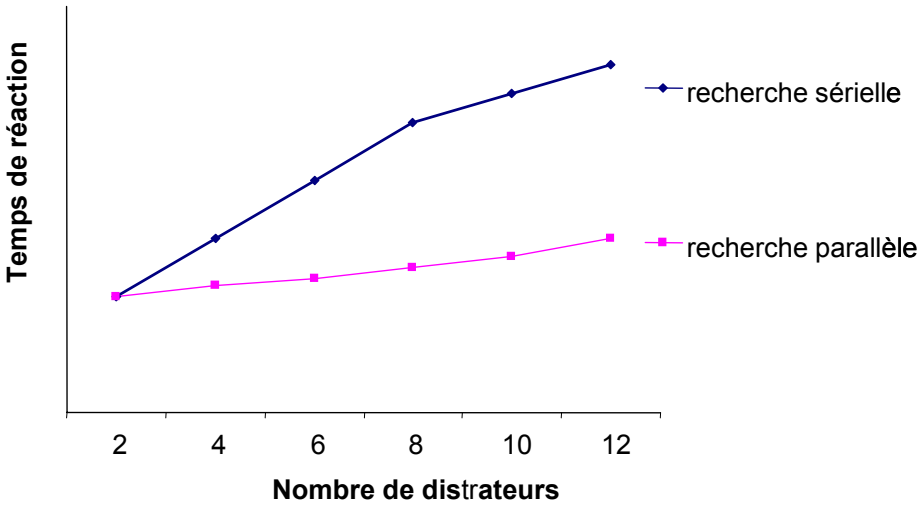


Figure 3.10. Illustration des fonctions de recherche avec deux courbes concernant la recherche sérielle et la recherche parallèle.

L'interprétation de ce type de résultats suppose que des mécanismes pré-attentifs interviennent dans l'objectif de traiter parallèlement l'ensemble des stimuli (Treisman & Gelade, 1980). En effet, l'attention spatiale serait nécessaire pour permettre aux sujets de se focaliser sur les éléments saillants. Lorsque la cible caractérise une seule dimension, elle semble "sauter aux yeux" (*pop out*) sans demander plus de temps de réaction au sujet. Par contre, quand la cible possède plusieurs dimensions : par exemple elle est définie par une conjonction d'élément eux-mêmes partagés par les distracteurs (chercher un **X** rouge parmi des **X** bleus et **A** rouges), les temps de réaction des sujets sont plus longs et augmentent avec le nombre de distracteurs (Figure 3.9).

2.4.3. Modèle hybride : guided search model (GSM)

Pour rendre compte de la simplicité de la théorie FIT, le modèle de recherche guidée (GSM, "*guided search model*", Wolfe, Cave, & Franze, 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. Elaboré dans le but de prédire et d'expliquer les performances humaines dans des tâches de recherche visuelle (dans lesquelles il s'agit de détecter la présence d'une cible définie par un ou plusieurs traits basiques), ce modèle distingue, tout comme le modèle de Treisman et Gelade (1980), deux niveaux dans le traitement d'un stimulus visuel : un niveau préattentif et un niveau attentif.

Le modèle GSM (Figure 3.11) possède la même structure que le FIT. Il a des cartes d'éléments simples (cartes de traits) et une carte d'activation similaire à la carte de contrôle de la FIT. Concrètement, ce modèle fonctionne de la manière suivante : Dans un premier temps, les différents traits et dimensions basiques du stimulus (couleur, orientation, contraste, etc.) sont traités simultanément et de manière préattentive à travers le champ visuel. Puis, ces traitements parallèles conduisent à la création de plusieurs cartes de traits, qui codent chacune pour les régions du dispositif visuel les plus actives (carte d'activation).

L'attention se dirigerait d'abord vers l'objet qui a produit l'activité la plus forte de la carte d'activation. Pour chaque position dans la carte d'activation (chaque objet du 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 les éléments encodés de cette carte et les éléments de la cible, spécifiés par un amorçage descendant. La carte d'activation classe tous les items du champ visuel par ordre, de la cible la plus probable à celle la moins probable. La recherche visuelle consisterait à parcourir cette liste, un item après l'autre jusqu'à ce que la cible soit trouvée. Dans ce modèle, les différents niveaux d'activation dépendent à la fois des processus de type bottom-up (la saillance perceptive des traits visuels), et des processus top-down (le but de la tâche).

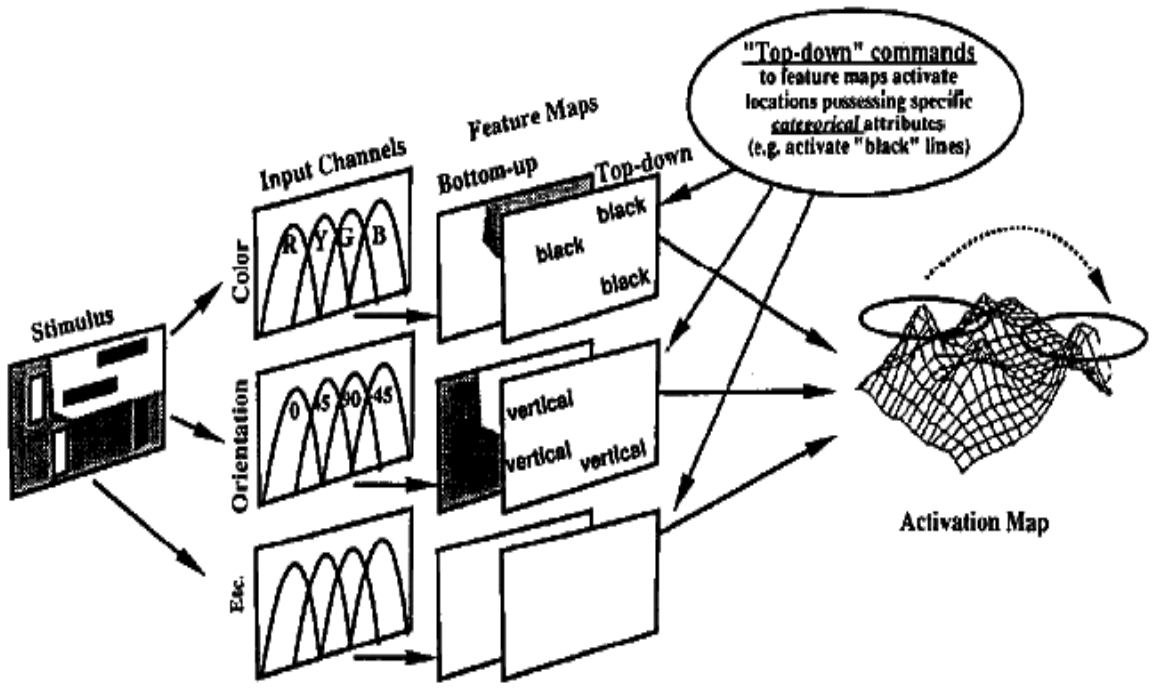


Figure 3.11. Le modèle GSM (guided search model), Wolfe, Cave, & Franze, 1989.

L'information visuelle pré-attentive bénéficie d'un statut particulier dans le modèle GSM. Les attributs élémentaires discrets envisagés par Treisman & Gelade (1980) sembleraient plutôt découpés en fichiers d'objets par le système visuel pré-attentif (Wolfe & Bennett, 1997). Selon cette hypothèse, avant la focalisation de l'attention, le système visuel pré-organiserait le flot d'informations visuelles du monde extérieur en objets potentiels, les éléments appartenant à chaque objet étant regroupés sous la forme d'un "fichier". Au sein de chaque fichier, les éléments qui composent un objet potentiel sont toujours discrets, ils ne sont pas reliés entre eux, c'est à dire que leur organisation spatiale ne serait pas prise en compte sans le déploiement de l'attention à cet endroit.

Cette absence totale de structuration spatiale pré-attentive a cependant été contestée récemment sur la base de nouvelles expériences de recherche visuelle. Si un tel processus de traitement représente la réalité du traitement de l'information chez l'homme, sans intervention de l'attention, les éléments ne permettront pas de former une représentation structurale de l'image.

2.4.4. Architecture triadique de la vision

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, 2000a, 2002).

Dans son architecture triadique ("*triadic architecture*"), Rensink inclut un système attentionnel de traitement des objets et un système non attentionnel à capacité limitée, dédié au traitement du sens général de la scène (sa catégorie) et de sa structure spatiale. Parce que ce système maintient en mémoire des informations à propos d'aspects stables de la scène, il pourrait servir à guider l'attention 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 (Figure 3.12).

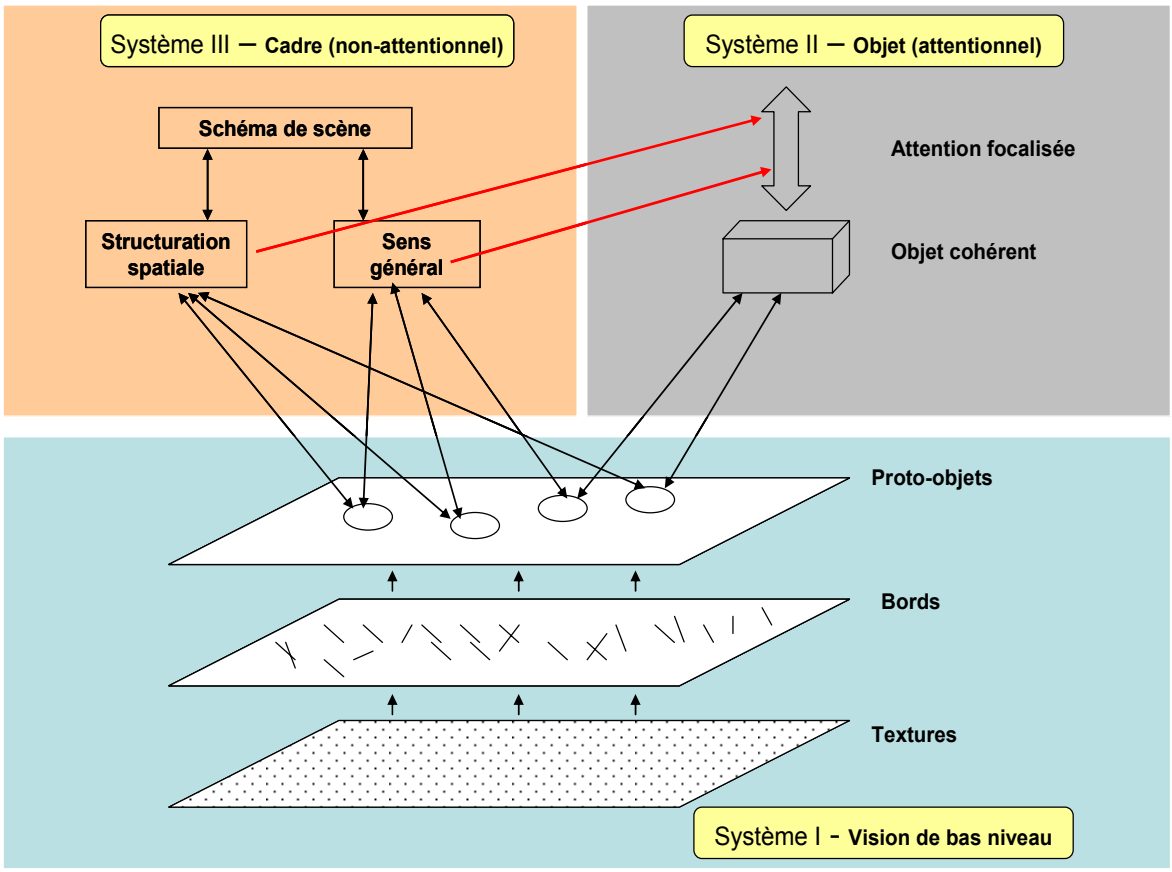


Figure 3.12. Architecture triadique de la vision selon Rensink (2000a, 2000b). (extrait et traduit de Rensink, 2000a).

Dans ce modèle, les premières étapes du traitement d'une scène visuelle sont similaires au système préattentif postulé par Treisman et Gelade (1980). Rensink propose en effet que, dès la présentation d'une scène visuelle, les éléments basiques de la scène sont massivement traités de manière automatique, parallèle et préattentive. Ces traitements permettent de fournir une représentation rétinotopique de la scène. Avant l'arrivée de l'attention, cette représentation serait composée de proto-objets (ou objets préattentifs) (Wolfe, 1999; Wolfe & Bennett, 1997), assemblages complexes de traits et de fragments correspondants à des structures localisées dans la scène.

Le deuxième système postulé par Rensink est un système attentionnel à capacité limitée. Dans ce modèle, les proto-objets sont organisés, traités et liés entre eux pour former des objets intégrés grâce à l'attention focalisée, ce qui leur permet de conserver leur identité et continuité dans le temps et dans l'espace (Wolfe, 1999, Rensink, 2000a). Ainsi, seule la focalisation de l'attention sur ces structures serait en mesure de leur conférer la cohérence spatio-temporelle indispensable à leur conversion en objet stable, et à leur survie identitaire. Cependant, une fois l'attention détournée de lui, l'objet intégré perdrait sa cohérence spatio-temporelle et retrouverait l'état labile des proto-objets, sans qu'il y ait de post-effet de l'attention (ce point est par ailleurs agréé par Wolfe, 1999), si ce n'est une trace mnésique plus conceptuelle que visuelle, stockée en mémoire à court terme.

Enfin, ces deux systèmes de représentation (système de bas niveau et système attentionnel à capacité limitée) s'accompagnent d'un troisième système non attentionnel. En effet, les informations visuelles véhiculées par les proto-objets fournissent le sens général et la structuration spatiale de la scène, qui, eux-mêmes activent le schéma de scène (des connaissances conceptuelles stockées en mémoire à long terme). L'ensemble de ces éléments permet de guider à nouveau l'attention pour se focaliser sur les objets d'intérêt.

2.4.5. Modèle parallèle

Une revue rapide de la littérature fournit des arguments plus directs en faveur des modèles parallèles de l'exploration de la scène.

Par exemple : l'attention se déploie d'objet en objet et non pas d'une position spatiale à une autre (Duncan, Humphreys, & Ward, 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, Driver, Pavani, & Shepherd, 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 ont des ressources limitées.

Les modèles parallèles à capacité limitée supposent que tous les items dans notre champ visuel sont traités en même temps par un mécanisme compétitif. Ce mécanisme est fondé sur des interactions inhibitrices entre les représentations activées par les différents éléments d'une scène (Duncan & Humphreys, 1989). Le mécanisme possède une quantité de "ressources attentionnelles" limitée, sa distribution dépend du nombre et des caractéristiques des items à traiter : plus il y a d'items, moins il y a de ressources sollicitées par item. Nous pouvons imaginer que dans un modèle parallèle strict, s'il y a de nombreux d'items à traiter, la ressource attentionnelle allouée à chaque item deviendrait alors très faible ou voire quasi nulle.

Pour rendre compte de cette hypothèse, des chercheurs utilisent la métaphore d'une "course" entre les représentations pour décrire le fonctionnement du système visuel. Certains items possèdent des caractéristiques particulières, qui peuvent être les plus rapidement catégorisés et ils gagnent donc cette course (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).

De plus, l'attention ne devrait pas ou très peu être sollicitée dans la compréhension rapide des scènes (Fei Fei, VanRullen, Koch, & Perona, 2002, 2005). Les traitements d'éléments perceptifs sembleraient agir automatiquement en parallèle avec le processus pré-attentif (Treisman & Gelade, 1980).

3. Structuration des zones de luminance et identification

3.1. Structuration des zones de luminance, une autre propriété de "scène-niveau"

Comme nous avons vu précédemment, une scène peut être déterminée par l'identification d'un ou plusieurs objets principaux (Friedman 1979) et, par leurs relations spatiales (De Graef, Christiaens, & d'Ydewalle, 1990). Par ailleurs, elle peut être également identifiée par l'information de scène-niveau sans vérification de l'identité des différents objets (Biederman, 1981, 1988; Schyns & Oliva ,1994). La plupart des recherches soutiennent la dernière hypothèse selon laquelle le traitement précoce de la scène est basé plutôt sur l'information globale de scène que sur l'information locale d'objets (Antes & Penland, 1981; Loftus, Nelson, & Kallman, 1983; Schyns & Oliva, 1994). Schyns et Oliva (1994) ont démontré que des scènes peuvent être identifiées par des basses fréquences parmi lesquelles les relations spatiales grossières de structures de la scène sont gardées. En outre, lorsque les sujets identifient une scène présentée très brièvement (50 ms), ils tendent à baser leur interprétation sur l'information de basses fréquences plutôt que sur l'information de hautes fréquences (Schyns & Oliva 1994).

Un point relatif concerne les représentations internes fonctionnelles dans l'identification de scènes. Biederman (1995) a proposé qu'un arrangement de primitives volumétriques, chacune représentant un objet dominant dans la scène, pouvant permettre l'identification rapide de scènes, indépendamment de l'identification locale d'objets. Selon ce point de vue, les scènes possèdent une représentation structurale d'un niveau beaucoup plus large que l'objet. Cette hypothèse nous paraît intéressante sur le fait qu'une scène ne peut pas être

forcement considérée par la somme de ces composants (les objets s'y trouvant), mais par plusieurs zones structurales ayant le même caractère physique. Par exemple, une scène de plage est composée par une zone d'eau, une zone de sable et une partie du ciel, etc. Notre hypothèse considèrerait que ces différentes zones peuvent-être regroupées par les propriétés physiques et sémantiques de celles-ci.

De plus, des études concernant les mouvements oculaires montrent que les fixations précoces sont influencées par les propriétés physiques de l'image, notamment par la densité des contours et le contraste local (Mannan, Ruddock, & Wooding, 1996, 1997; Reinagel & Zador, 1999) ainsi que la structure de la scène (Sanocki & Epstein, 1997; Castelhano & Henderson, 2003; Oliva & Torralba, 2003). Donc, ces deux types d'informations fusionneraient en une représentation de différentes zones de luminance lors de l'exploration d'une scène. En effet, la structure de la scène étant aussi identifiée très rapidement (De Graef, Christiaens, & d'Ydewalle, 1990). Nous pourrions imaginer qu'une image grossière ayant des différentes zones de luminance soit codée par le système visuel dans les traitements précoces de l'image. Cette image grossière ne devrait pas contenir des informations concrètes sur les objets mais des propriétés globales de la scène. La question est de savoir si ce type d'informations permet une identification de la scène, et s'il ne le permet pas, quel rôle joue t-il?

Des données neuropsychologiques nous montrent d'autres arguments en faveur de l'identification rapide de scènes, qui pourrait se baser sur les informations issues de la localisation d'objets. En effet, le système visuel se compose de deux voies ayant des rôles différents dans la perception de la scène, à savoir, la voie dorsale et la voie ventrale. La voie dorsale encode la localisation des objets de la scène, tandis que la voie ventrale est sollicitée dans l'identification des objets et de la scène. Comme le montre la littérature portant sur la perception des scènes, le processus d'identification de scène est *a priori* un processus parallèle de traitements des objets traités par la voie ventrale. Cependant, la rapidité de compréhension d'une scène implique que le traitement progressif tout au long de la voie ventrale soit mis en question. Les éléments perceptifs ne seraient donc pas traités de manière progressive et parallèle. La question est alors

de savoir si la voie dorsale permet d'identifier une image lorsque seule la structuration spatiale de l'image est disponible (ce type d'information ne contient pas d'informations sur l'identité d'objets mais sur leur localisation). Dans ce cas, la voie dorsale se trouverait prioritaire et considérée comme une voie rapide. La voie ventrale serait très peu (ou partiellement) sollicitée car ce type d'information ne montre aucun éléments pertinents sur l'identification d'objets. Le mécanisme portant sur les zones de luminances différentes n'ayant pas fait l'objet d'investigation expérimentale, cette étude nous a semblé pertinente.

3.2. Structuration des zones de luminance et mécanisme de pré-identification

Une revue de la littérature concernant le mécanisme de perception de scènes naturelles nous permet d'avoir un aperçu des différents points de vue. Ce pendant, bien que les découvertes récentes soient de plus en plus nombreuses et affinent certains points, elles ne sont pas toujours très claires. Le processus de traitement d'une scène est un traitement de haut niveau du fait de sa complexité par rapport aux objets. C'est un processus complexe, le système visuel organisant un lien dynamique entre les traitements ascendants (*bottom-up*) et les descendants (*top-down*).

Dans la compréhension rapide de la scène, les modèles de la carte de saillance ne semblent pas aptes à expliquer le mécanisme de traitement. Selon ces modèles, une carte de saillance est engendrée suite au calcul de l'importance des éléments perceptifs avec un temps nécessaire. Donc, pour avoir la carte de saillance, des traitements parallèles et sériels sont sollicités. Tandis que dans l'identification rapide de la scène, lorsque l'image est présentée très rapidement (20 ms), les traitements des éléments perceptifs pour construire la carte de saillance, sont contraints par le temps de présentation de l'image. Le temps de présentation est tellement court que le sujet n'a pas assez de temps pour percevoir les informations de l'image. Cela nous permet de penser que la carte de saillance de cette image caractériserait plutôt une image grossière, ou une image

ayant des propriétés globales de la scène. Dans le cas présent, cette carte ne caractériserait aucune information précise concernant les objets : l'identification de la scène dérive-t-elle donc d'une image grossière ? Si ce n'est pas le cas, cette carte de saillance servira-t-elle comme une étape de "pré-identification", ce qui permettrait d'accéder à l'étape ultérieure de l'activation sémantique de la scène ?

De plus, le processus de traitement impliqué dans l'identification rapide de scènes est principalement du type ascendant. Les éléments perceptifs étant traités plutôt en parallèle, certaines propriétés devraient être traitées plus rapidement en fonction de la sensibilité différente des activités neuronales. Par exemple, les éléments caractérisant la basse fréquence ou le fort contraste sont traités avec priorité (voie dorsale) par rapport à ceux caractérisant la haute fréquence ou le faible contraste (voie ventrale, les cellules parvocellulaire sont activées avec environ 20 ms de retard par rapport aux cellules magnocellulaire, Schmolesky *et al*., 1998). Donc, le traitement de ces éléments gagnerait la "course". Est-il vraisemblable que celui qui gagne la "course" est celui qui permet d'activer l'identification de la scène? Ou alors, le fait de gagner la "course" serait-il plutôt considéré comme une étape nécessaire permettant d'autres types de traitements complémentaires afin d'activer l'identification de la scène.

Dans les chapitres à venir, à travers diverses expériences, nous tenterons de répondre à plusieurs questions principales :

1) L'hypothèse de Biederman (1995) sur "la reconnaissance d'objets par ses composants" s'applique-t-elle à la reconnaissance d'une scène?

2) La structuration de zones de luminance permet elle de catégoriser une scène?

3) Quel est le traitement cognitif mis en jeu dans cette activité ?

Chapitre 4 Rôle des points d'intérêt--tâche de reconnaissance (expérience 1)

1. Introduction

Les expériences exposées dans les chapitres suivants (4, 5, 6, 7, 8) visent à étudier la représentation de scènes naturelles, et plus particulièrement les facteurs qui influencent la représentation en mémoire des informations portées par une scène visuelle complexe.

Les travaux sur le sens général d'une scène menés ces dernières années ont mis en évidence qu'une scène peut être identifiée rapidement en utilisant des informations pertinentes locales d'une part (Friedman, 1979; De Graef, Christiaens, & d'Ydewalle, 1990), et d'autres informations globales ayant des propriétés de "scène-niveau" d'autre part (Biederman, 1981, 1988; Biederman, Mezzanotte, & Rabinowitz, 1982; Intraub, 1999; Oliva & Schyns, 1997, 2000; Potter, 1975, 1976; Potter & Levy, 1969; Schyns & Oliva, 1994). Le sens général peut être représenté "qualitativement", par la sémantique (Fei Fei, Fergus, & Perona, 2004; Greene & Oliva, 2006; Potter, 1975; Renniger & Malik, 2002; Schyns & Oliva, 1994) ou "quantitativement", par des propriétés de bas-niveau (Greene & Oliva, 2006; Torralba & Oliva, 2002; Tversky & Hemenway, 1983).

Les mouvements oculaires dans la perception de scènes visuelles sont influencés, d'une part, par des éléments informatifs (Antes, 1974; Mackworth, 1978; Hollingworth, 2003), et d'autre part, par des éléments perceptifs (Koch & Ullman, 1985; Mannan, Ruddock, & Wooding, 1996, 1997a, 1997b; Itti, Koch, & Niebur, 1998; Itti & Koch, 2000; Itti, Gold, & Koch, 2001; Itti & Koch, 2001; Niebur, Itti, & Koch, 2002). Les éléments informatifs ont besoin d'une attention focalisée et ils sollicitent une intervention sémantique. Quant aux éléments perceptifs, ils sollicitent les traitements préattentifs sans avoir besoins d'une attention focalisée (Koch & Ullman, 1985; Itti, Koch, & Niebur, 1998; Itti & Koch, 2000; Itti, Gold, & Koch, 2001; Itti & Koch, 2001; Niebur, Itti, & Koch, 2002; Wolfe, Cave, & Franze, 1989; Wolfe & Gancarz, 1996).

Le mécanisme de traitement impliqué dans l'identification rapide du sens général d'une scène est considéré également comme un processus rapide. Ce

traitement ne semblerait pas être un processus strictement sériel réalisé par la voie ventrale. Il serait un traitement rapide portant sur des propriétés de "scène-niveau" permettant d'aboutir à l'identification de scène. Ces propriétés pourraient être caractérisées par une représentation grossière de la structuration spatiale des objets de scène, qui semblerait être traitée par la voie dorsale. Selon les modèles de carte de saillance (Itti, Koch, & Niebur, 1998; Itti & Koch, 2000; Itti, Gold, & Koch, 2001; Itti & Koch, 2001; Niebur, Itti, & Koch, 2002), une telle carte est fusionnée par des cartes de traits, elle est caractérisée par les éléments les plus saillants de scène.

La question que nous nous posons est donc de savoir comment cette carte de saillance est représentée en mémoire? Est- elle représentée par des propriétés de "scène-niveau"?

Avant de répondre à ces deux questions, nous présentons l'origine de la problématique de cette recherche par l'introduction d'un algorithme de traitement d'images conçu à France Telecom Recherche & Développement.

1.1. Algorithme de traitement des scènes complexes basé sur le traitement de points d'intérêt

France Telecom R&D a conçu un logiciel de traitement d'images en se basant sur l'approche locale. Ce logiciel cherche à déterminer les points d'intérêt d'images, c'est-à-dire les éléments les plus pertinents et représentatifs selon un traitement mathématique (l'analyse en ondelettes). Ainsi, une image peut être représentée par de nombreuses "signatures" caractérisant ces points d'intérêt. Les plus proches spatialement permettent de former une zone de saillance. Seules quelques zones de saillance sont révélatrices du contenu de l'image. Selon ce point de vue, une scène peut donc être représentée par ces zones d'intérêt.

La détection des points d'intérêt s'effectue par un détecteur d'ondelettes qui extrait les contours des régions de forte concavité et de forte convexité. Les points d'intérêt sont localisés sur les contours concaves (Figure 4.1).

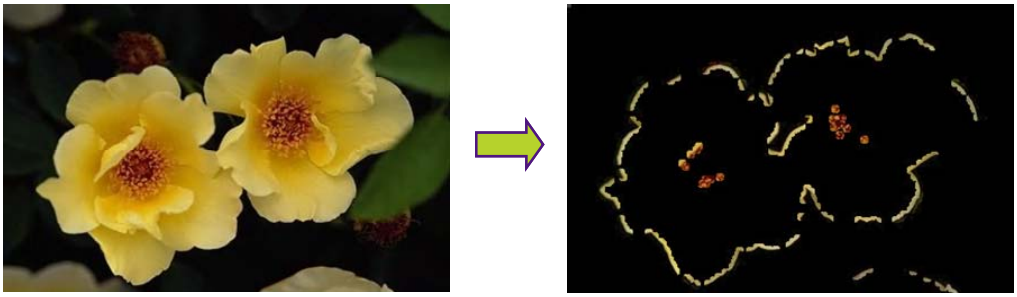


Image originale

Détecteur ondelettes :
les points d'intérêt peuvent être hiérarchisés selon leur importance
(par exemple de 1 à 500)

Figure 4.1. Illustration des points d'intérêt extraits par le détecteur ondelettes.

Le principe de cet algorithme s'appuie sur le fonctionnement du système perceptif visuel humain. Pour ce dernier, il existe un processus de codage et un processus de stockage. La détection des points d'intérêt constitue la première étape de traitement de l'algorithme, dite "extraction de points d'intérêt". Ensuite, ces points d'intérêt sont reliés entre eux en construisant des "chaînes fovéales" afin de hiérarchiser l'importance de chaque point d'intérêt, celles-ci représentant alors une image (codage). Enfin, les "chaînes fovéales" sont stockées dans la base de données (stockage) (Figure 4.2).

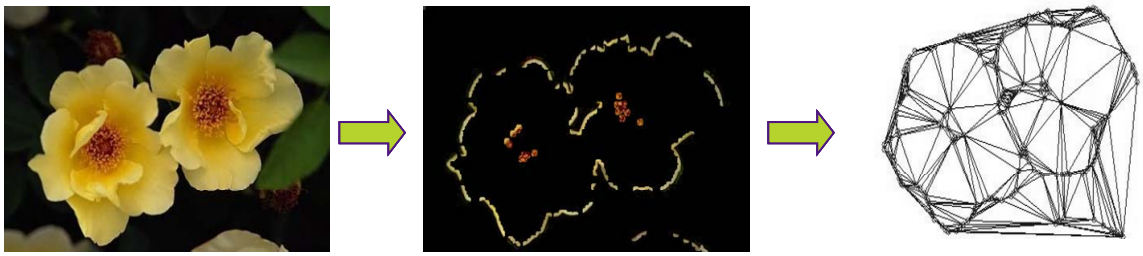


Image originale

Détecteur ondelettes :
les points d'intérêt peuvent être hiérarchisés selon leur importance
(par exemple de 1 à 500)

Construction des chaînes fovéales

Figure 4.2. Illustration de la construction de chaînes fovéales.

Lors de l'identification d'une nouvelle image (image requête), l'algorithme compare sa représentation (définie avec ses chaînes fovéales) avec les représentations appariées stockées dans la base de données, et ce, en proposant une liste d'images hiérarchisées selon un degré de similarité décroissante. L'image en tête de la liste est alors considérée comme l'image la plus proche de l'image requête.

1.2. Algorithme de traitement d'images et la théorie de "la reconnaissance d'un objet par ses composants"

Le principe de ce traitement est similaire à celui de la théorie de la reconnaissance d'objet par ses composants de Biederman (1987). Selon Biederman, la perception se focalise sur les propriétés non accidentelles, et la détection de ces propriétés non accidentelles se traduit par l'analyse des régions concaves. Toutes les régions concaves sont considérées comme des régions informatives (Figure 4.3).

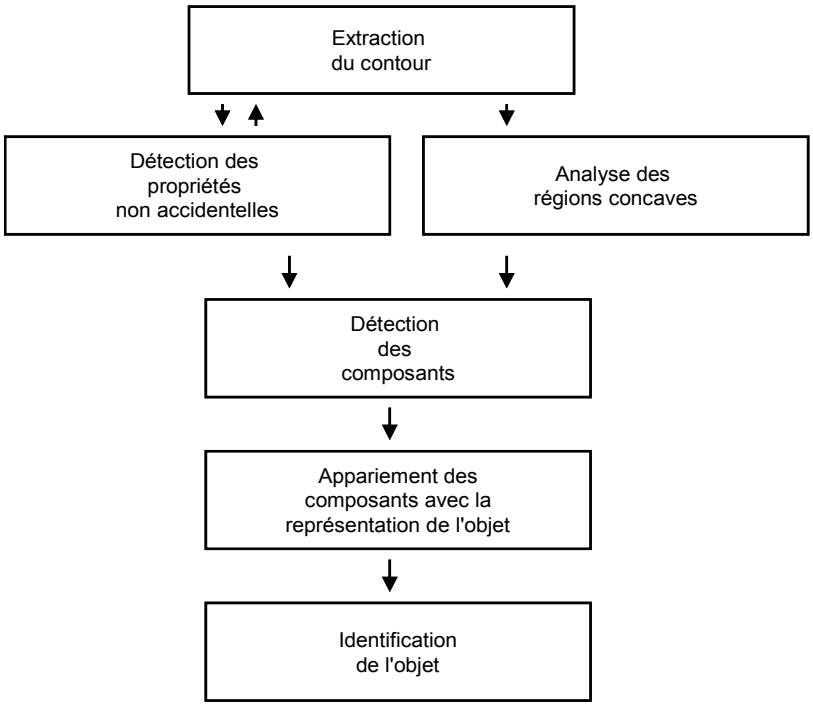


Figure 4.3. Processus de reconnaissance d'objets selon Biederman (1987).

La reconnaissance des objets repose sur la perception d'éléments géométriques de base à partir desquels l'objet peut être construit. Les primitives sont, selon Biederman (1987) (Figure 4.4), des volumes géométriques élémentaires, comme un parallélépipède, un cylindre, un tronc de cône, un élément torique, etc. (nommés géon par cet auteur). La perception des géons permettrait un accès à la représentation prototypique de l'objet. Par exemple, la reconnaissance du téléphone peut être activée par la reconnaissance de géon 1, 3 et 4 (Figure 4.4).

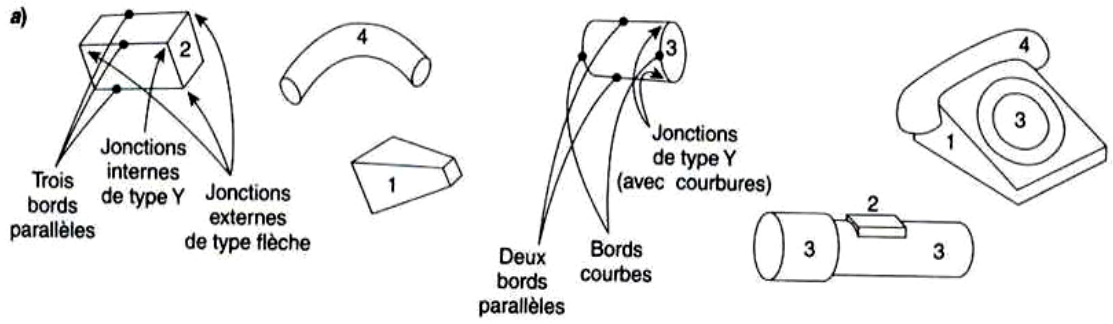


Figure 4.4. Illustration des géons selon Biederman (1995), extrait de Delorme et Flückiger (2003).

Le principe de conception de cet algorithme est assez proche la théorie de la reconnaissance des objets par ses composants (Biederman, 1987).

L'observation du comportement humain dans une tâche d'identification de scène utilisant les points d'intérêt prédéfinis par l'algorithme de traitement d'images de France Télécom R&D présente deux particularités.

- Cette observation nous permet de tester si la théorie de la reconnaissance de l'objet de Biederman (1987) peut s'appliquer à l'identification d'une scène.

- Elle permet d'étudier non seulement le processus humain d'identification de scènes visuelles complexes, mais aussi d'initier une étude ergonomique de cet algorithme.

1.3. Paradigme de l'expérience 1

L'expérience 1 vise à appréhender dans quelle mesure l'importance des points d'intérêt définis précédemment est visuellement pertinente lors du traitement d'une scène visuelle complexe.

Le paradigme utilisé est basé sur l'approche locale de la perception de scènes naturelles (Friedman, 1979; De Graef, Christiaens, & d'Ydewalle, 1990) supposant que des informations locales sont suffisantes pour reconnaître une scène.

1.3.1. Définition des "géons de scène"

Comme nous l'avons vu dans le chapitre précédent, la scène possède des caractéristiques intrinsèques. Par exemple, elle peut être caractérisée par des propriétés de "scène-niveau" ou des informations contextuelles. Toutes les particularités d'une scène impliquent probablement un processus d'identification différent par rapport à l'identification d'un objet. Ainsi, il est plus intéressant d'observer ces aspects particuliers chez les sujets humains.

C'est à partir de ce point de vue que nous avons divisés la présentation des points d'intérêt en trois catégories, sous forme de séries de petits pavés apparus un par un en ordre *"décroissant"*, *"croissant"* et *"aléatoire"* (Figure 4.5).

Concernant l'ordre *"décroissant"* d'apparition, les points d'intérêt les plus importants sont tout d'abord affichés sous forme de petits pavés, puis les points d'intérêt moins importants sont montrés progressivement. Contrairement à l'ordre *"décroissant"*, l'ordre *"croissant"* montre premièrement les points d'intérêt les moins importants puis, de manière ascendante, affiche ceux les plus importants. Quant à l'ordre *"aléatoire"*, les points d'intérêt sont affichés aléatoirement, certains points concernant des zones vides ou des informations contextuelles sont également choisis. Ces trois ordres de présentation différents renvoient donc à

trois types d'informations ayant un degré de pertinence différent. Autrement dit, ils construisent trois types de représentation de l'image ayant une saillance différente, à savoir un nombre différent de géons (les éléments géométriques de l'objet) selon l'ordre de présentation. L'ordre *"décroissant"* ayant le nombre de points d'intérêt le plus important, puis l'ordre *"croissant"* ayant un nombre de points d'intérêt modéré et enfin l'ordre *"aléatoire"* ayant le nombre de points d'intérêt le moins important.

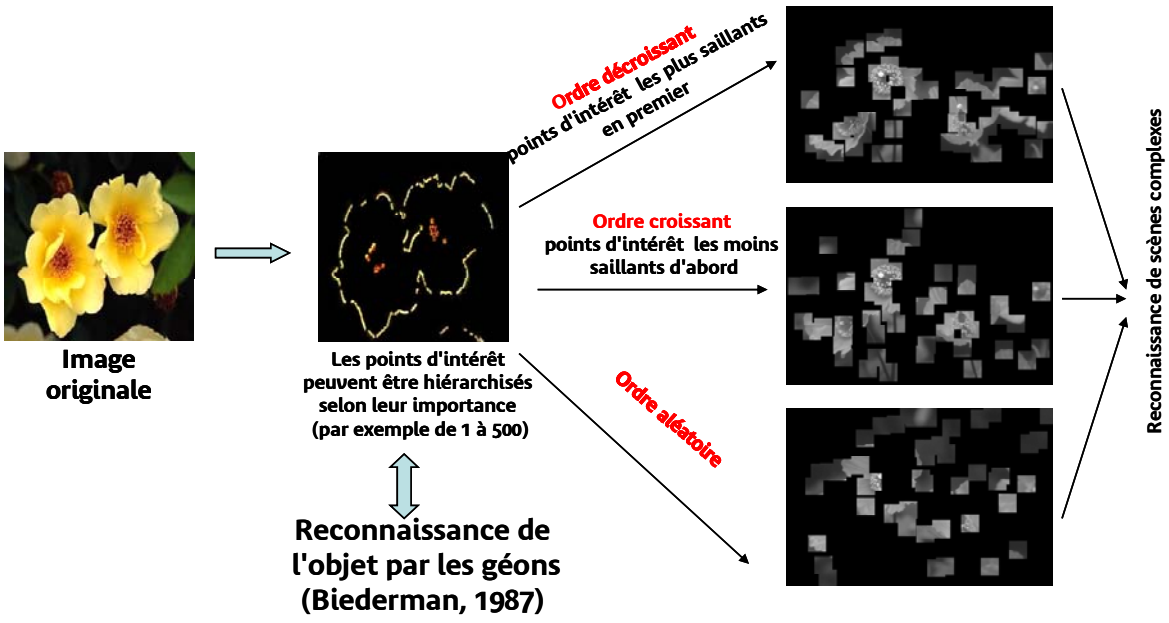


Figure 4.5. Illustration des trois ordres de présentations de points d'intérêt.

1.3.2. Scènes naturelles et scènes artéfactuelles

Afin d'étudier le rôle des points d'intérêt dans l'identification de scènes, nous introduisons une autre variable indépendante : la catégorie des scènes visuelles complexes. Les scènes visuelles sont divisées selon deux axes sémantiques (Rogowitz, Frese, Smith, Bouman, & Kalin, 1997), soit les scènes naturelles ne contenant pas ou peu d'objets conçus par l'homme et les scènes artéfactuelles dans lesquelles se trouvent principalement des objets conçus par l'homme. L'introduction de cette variable permet de comparer la performance d'identification de ces deux types de scène aux recherches récentes du domaine.

Selon Gaillard et Bourges (1999), les scènes de type "ville" (scènes artéfactuelles) laissent entrevoir une organisation géométrique des objets les constituant : des lignes nettes délimitent des formes non accidentelles. La perception de ces scènes n'impose pas de traitement perceptif des détails représentant des ornements. Quand aux scènes de type "campagne" (scènes naturelles), le contour des objets ne représente généralement pas une organisation géométrique mais délimite plutôt des formes accidentelles. Ces deux types de scènes possèdent donc des caractères différents pouvant impliquer des mécanismes de traitements différents.

1.3.3. Image en niveau de gris

Pour essayer d'aller plus loin dans la compréhension des mécanismes mis en œuvre dans le système visuel, il est essentiel de déterminer quels sont les attributs les plus importants dans une image pour effectuer une tâche donnée. On peut manipuler une partie du contenu des images et observer l'impact de cette modification sur les performances et l'activité cérébrale pour savoir si ces informations sont essentielles au fonctionnement du système. C'est ce type de manipulation qu'ont fait Delorme et ses collègues dans une étude parue en 2000 (Delorme, Richard, Fabre-Thorpe, 2000).

Testés sur des images présentées aléatoirement en couleurs ou en niveaux de gris, les sujets (hommes ou singes) se sont montrés aussi rapides (temps de réaction minimal) et pratiquement aussi précis dans les deux conditions de présentation. Ainsi, de manière surprenante, retirer une importante partie des informations de l'image qui peuvent sembler très utiles pour interpréter la scène ou lever des ambiguïtés, ne ralentit pas la vitesse du traitement.

2. Méthode

2.1. Participants

Quarante étudiants inscrits en licence à l'Université de Rennes 2 ont participé à cette recherche (27 hommes et 13 femmes). Tous les participants ont une vue normale ou corrigée. Aucun d'entre eux n'a acquis de connaissances en rapport avec l'objet cette recherche. Ils étaient tous ignorants du point de vue des hypothèses de l'expérience.

2.2. Matériel

Quatre-vingt seize photographies (48 couples de photographies) composées de niveaux de gris provenant d'une source de base de données COREL (voir chapitre 2, Point 1.1) sont utilisées. La répartition des images selon les variables indépendantes a été la suivante :

- vingt-quatre couples de scènes naturelles complexes composés de quatre catégories : fleurs, montagnes, mers et déserts. Chaque catégorie, elle-même, est représentée par six couples d'exemples ;

- vingt-quatre couples de scènes artéfactuelles complexes de quatre catégories de photographies : monuments, villes, scènes d'intérieur et voitures.

2.3. Equipement

L'expérience, pilotée par un ordinateur, se déroule dans une salle de cours, dans des conditions d'éclairage confortables. Les participants sont installés à une distance approximative de 80 cm de l'ordinateur. L'ordinateur utilisé pour la programmation et la passation de l'expérience, ainsi que pour l'enregistrement des données est un PC portable HP équipé d'un écran 15 pouces et un clavier supplémentaire. Ce PC est du type Pentium M 1.60 GHz équipé de 512 Mo de RAM. L'expérience est programmée par le logiciel MACROMEDIA Director 10.

2.4. Procédure expérimentale

Avant de commencer l'expérience, les participants sont informés qu'une première image (image-cible) est présentée pendant 10 secondes, qu'ils doivent l'observer attentivement et mémoriser le contenu de l'image. Ensuite, une deuxième image (image-test) apparaît, le contenu de cette deuxième image étant montré successivement par de petits pavés avec un temps d'affichage de 300 ms. Ces pavés ont une taille de 4 cm^2 (2 x 2), correspondant aux 2 degrés du champ fovéal pour un sujet qui est placé à environ 80 cm d'un écran d'ordinateur.

Les participants effectuent alors une tâche de reconnaissance : dès qu'ils jugent que l'image-test et l'image-cible sont identiques, ils doivent appuyer sur la touche « V » du clavier. Inversement, dès qu'ils jugent que ces deux images sont différentes, ils doivent appuyer sur la touche « N » du clavier (Figure 4.6).

Au début de cette expérience, les participants ont pu lire le texte suivant sur l'écran de l'ordinateur :

> *"Vous devez d'abord observer attentivement une image complète appelée image-cible. Ensuite cette image va disparaître, puis vous verrez apparaître une nouvelle image appelée image-test. Celle-ci sera révélée progressivement par petits pavés. L'image-cible et l'image-test sont soit identiques, soit différentes. Attention, certaines images se ressemblent mais sont différentes. Dès que vous pensez que l'image-test est exactement la même que l'image-cible, appuyez sur la touche "V" du clavier. Si au contraire, vous pensez qu'elle n'est pas la même que l'image-cible, appuyez sur la touche "N" du clavier."*

Après en avoir pris connaissance, les participants sont invités à réaliser une phase d'entrainement sur dix essais. Les images utilisées sont des images différentes de celles utilisées lors de l'expérience. Une fois cette phase terminée, le candidat peut soit reproduire cette phase d'entraînement, soit démarrer ladite expérience.

Au total, la passation de l'expérience s'étale sur vingt minutes environ.

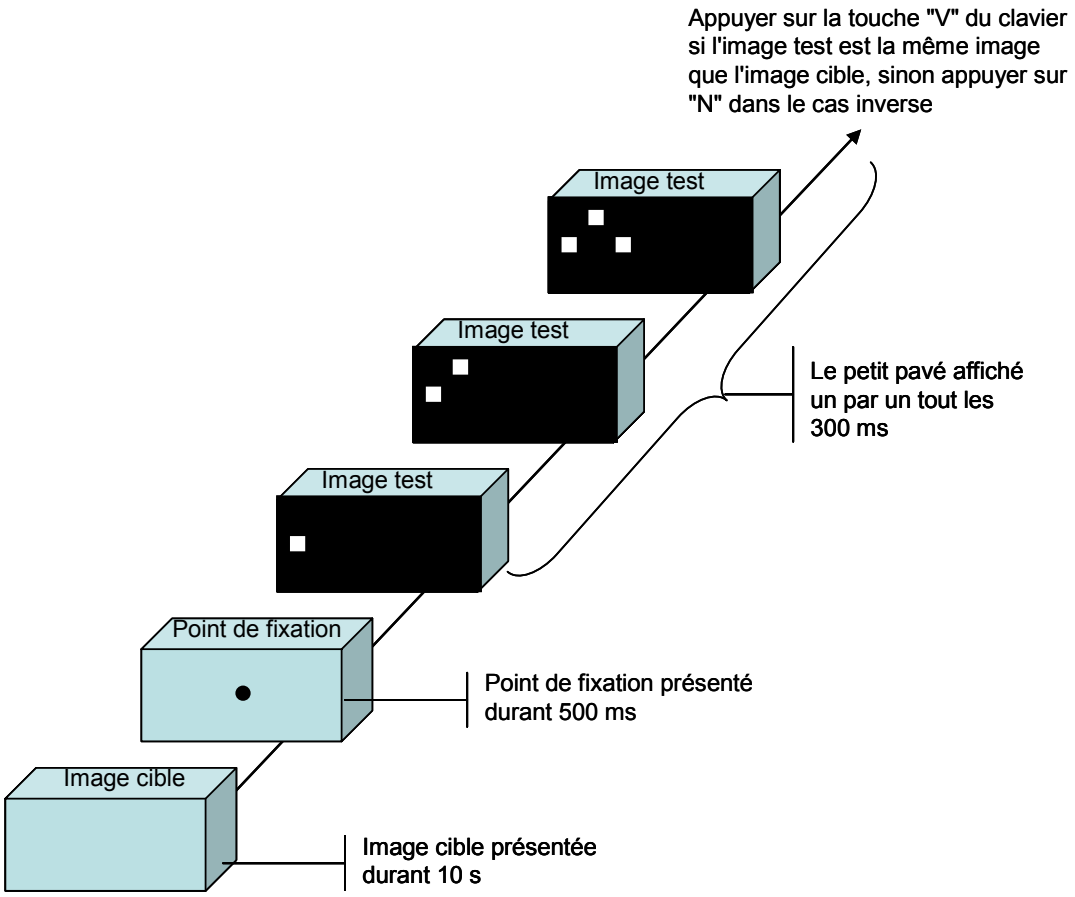


Figure 4.6. Paradigme de reconnaissance utilisé dans l'expérience 1.

2.5. Plan expérimental

Plan expérimental

S*[G2] *I2 *O3

Le facteur **S** correspond aux participants ; le facteur **G** correspond aux groupes (groupe avec les scènes naturelles *vs* groupe avec les scènes artéfactuelles) ; le facteur **I** correspond à la congruence d'images (l'image-cible et l'image-test sont identiques *vs*. l'image-cible et l'image-test sont différentes) ; le facteur **O** correspond à l'ordre de présentation des fenêtres (*"décroissant" vs*. *"croissant" vs*. *"aléatoire"*).

Variables dépendantes (VD)

Trois variables dépendantes sont observées : le temps de réaction, la surface affichée et le taux de réponses correctes.

Les données, recueillies par le logiciel MACROMEDIA Director 10, ont été analysées par le logiciel SPSS 10.0. et selon la Théorie de la Détection du Signal.

3. Analyse des résultats

Afin de nous permettre d'observer les effets de la pertinence des informations dans la reconnaissance d'images, les analyses que nous rapportons ici visent à présenter séparément le temps de réaction et le taux de réponses correctes, et ce, pour les deux types de scènes en fonction de la congruence entre l'image-cible et l'image-test :

- le premier cas est celui où les deux images du couple sont bien la même image, une bonne réponse consistant alors à répondre "**V**".

- le deuxième cas est celui où les deux images sont différentes, la bonne réponse étant alors "**N**" pour rejeter l'hypothèse que les deux images sont la même.

3.1. L'image-cible et l'image-test sont identiques

Les résultats obtenus pour les trois ordres de présentation sont résumés dans le Tableau 4.1.

Tableau 4.1

Temps de réaction (s) et taux de réponses correctes (%) en fonction de la catégorie et de l'ordre de présentation des points d'intérêt pour l'expérience 1. Erreurs type entre parenthèses.

	Ordre de présentation		
	Décroissant	Croissant	Aléatoire
Temps de réaction (s)			
Scène naturelle	11,1 (1,9)	12,9 (2,1)	13,1 (2,6)
Scène artéfactuelle	12,0 (1,7)	11,1 (1,1)	13,4 (1,7)
Taux de réponses correctes (%)			
Scène naturelle	92,5 (3,2)	88,3 (4,2)	89,6 (3,3)
Scène artéfactuelle	88,8 (3,7)	86,3 (4,6)	88,8 (4,4)

3.1.1. Le temps de réaction

L'analyse du temps de réaction ne montre pas d'effet significatif selon les deux catégories de scènes, F (1,19) = 1.618, p> .05. L'ordre de présentation n'influence pas le temps de réaction (F<1), ni pour l'interaction avec la catégorie de scènes (F<1).

Les participants répondent en moyenne avec un temps comparable pour chaque ordre pour ces deux types de scènes (11,1 s, 12,9 s et 13,1 s pour l'ordre *"décroissant"*, "*croissant*" et *"aléatoire"* des scènes naturelles ; 12 s, 11,1 s et 13,4 s pour l'ordre *"décroissant"*, *"croissant"* et *"aléatoire"* respectivement pour les scènes artéfactuelles) (Figure 4.7)

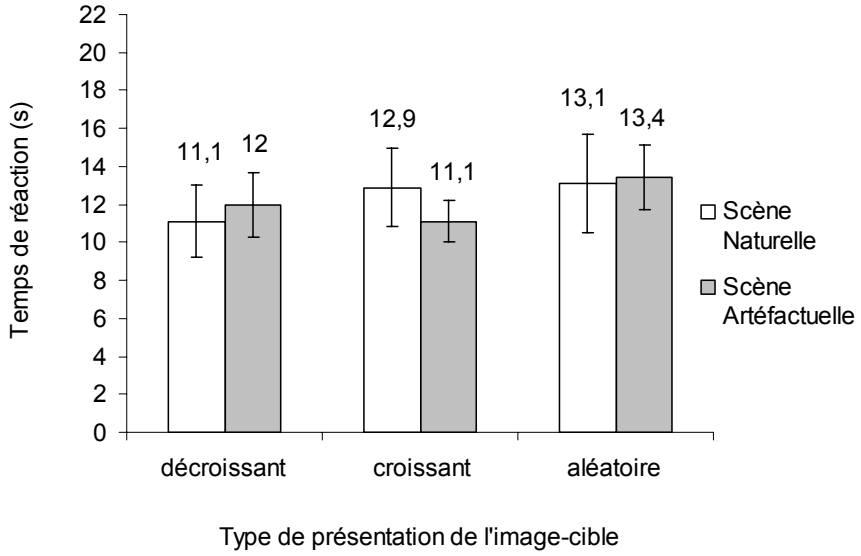


Figure 4.7. Temps de réaction pour les scènes naturelles et les scènes artéfactuelles en fonction de l'ordre de présentation.

3.1.2. Analyse du taux de réponses correctes

Aucun effet significatif n'est observé au niveau du taux de réponses correctes entre les trois ordres de présentation (F<1), que ce soit pour les scènes naturelles ou les scènes artéfactuelles (92,5%, 88,3% et 89,6% pour l'ordre *"décroissant"*, *"croissant"* et *"aléatoire"* des scènes naturelles ; 88,8%, 86,3% et

88,8% pour l'ordre *"décroissant"*, *"croissant"* et *"aléatoire"* respectivement pour les scènes artéfactuelles) (Figure 4.8).

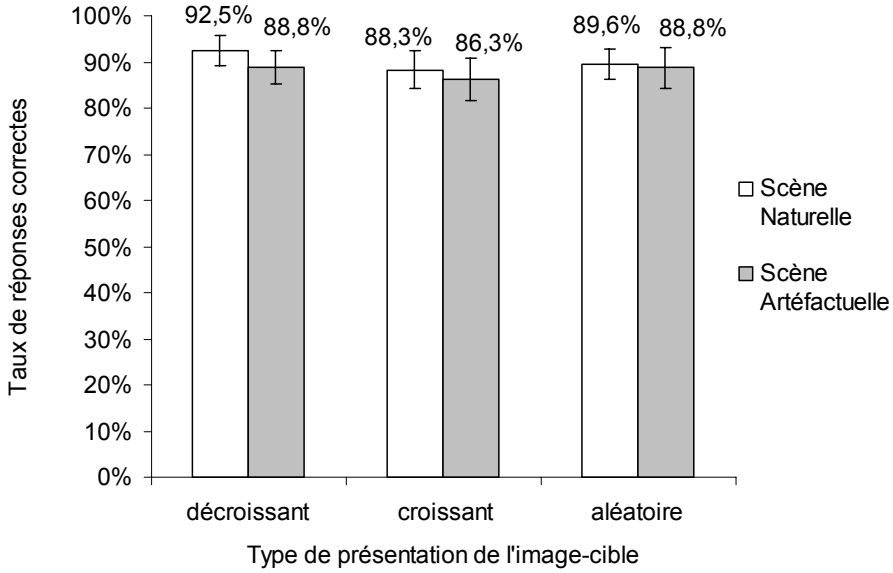


Figure 4.8. Taux de réponses correctes pour les scènes naturelles et les scènes artéfactuelles en fonction de l'ordre de présentation.

3.2. L'image-cible et l'image-test sont différentes

Les résultats obtenus pour les trois ordres de présentation sont résumés dans le Tableau 4.2.

Tableau 4.2

Temps de réaction de rejet (s) et taux de rejets corrects (%) en fonction de la catégorie et de l'ordre de présentation des points d'intérêt pour l'expérience 1. Erreurs type entre parenthèses.

	Ordre de présentation		
	Décroissant	Croissant	Aléatoire
Temps de réaction de rejet (s)			
Scène naturelle	9,0 (1,4)	8,8 (2,3)	10,6 (2,9)
Scène artéfactuelle	9,1 (1,9)	9,3 (2,1)	16,8 (3,7)
Taux de réponses correctes (%)			
Scène naturelle	78,8 (2,7)	86,0 (3,1)	85,0 (3,6)
Scène artéfactuelle	93,6 (3,1)	90,0 (2,4)	75,3 (3,1)

3.2.1. Le temps de réaction de rejet

L'analyse du temps de réaction de rejet montre une interaction entre l'ordre de présentation et la catégorie de scènes, F (2, 38) = 10.088, p< .001. Les participants répondent beaucoup plus rapidement pour les scènes naturelles en ordre *"aléatoire"* (10,6 s) que pour des scènes artéfactuelles (16,8 s), F (1, 19) = 32.552, p< .001. Pour l'ordre *"décroissant"* et *"croissant"*, le temps de réaction de rejet entre ces deux types de scène ne varie pas significativement (F<1). En ce qui concerne les scènes naturelles, les participants ont un temps de rejet assez proche (autour de 9,5 s). Quant aux scènes artéfactuelles, les participants répondent lentement tant pour l'ordre *"aléatoire"* (16,8 s), que pour l'ordre *"décroissant"* (9,1 s), F (1, 19) = 26.335, p< .001, que pour l'ordre *"croissant"* (9,3 s), F (1, 19) = 24.131, p< .001 (Figure 4.9).

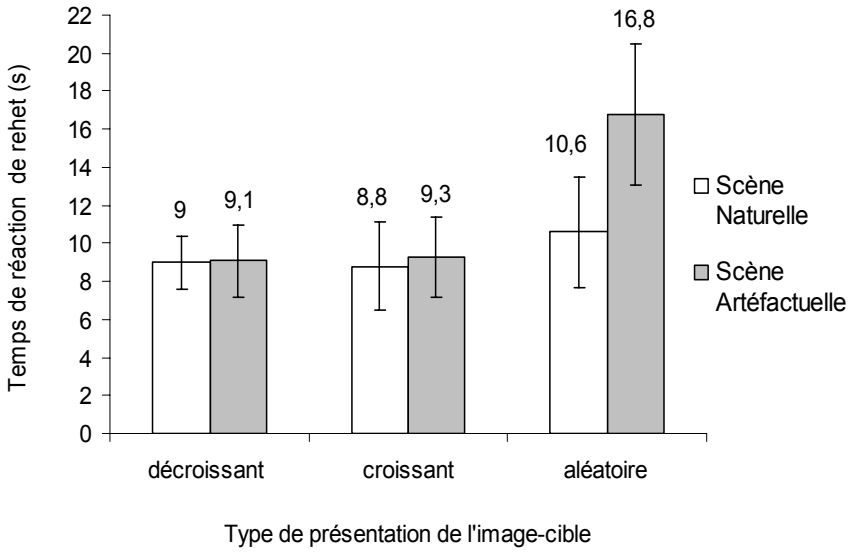


Figure 4.9. Temps de réaction de rejet pour les scènes naturelles et les scènes artéfactuelles en fonction de l'ordre de présentation.

3.2.2. Analyse du taux de rejets corrects

L'analyse du taux de rejets corrects indique une interaction entre l'ordre de présentation et la catégorie de scènes, F (2, 38) = 20.727, p< .001. En ordre *"décroissant"*, les scènes naturelles (78,8%) sont moins biens reconnues que les

scènes artéfactuelles (93,6%), *F* (1, 19) = 30.584, *p*< .001. En ordre *"aléatoire"*, les scènes naturelles (85%) sont mieux reconnues que les scènes artéfactuelles (75,3%), *F* (1,19) = 31.641, *p*< .001 (Figure 4.10).

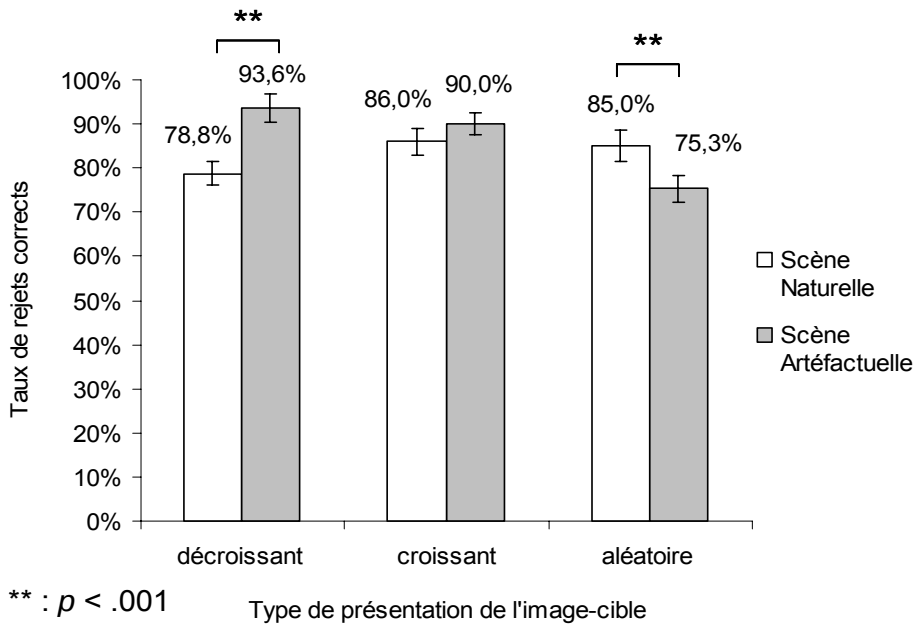


Figure 4.10. Taux de rejets corrects pour les scènes naturelles et les scènes artéfactuelles en fonction de l'ordre de présentation.

Quand on observe les résultats au sein de chaque catégorie de scènes, les scènes naturelles en ordre *"décroissant"* sont moins biens reconnues par rapport aux deux autres ordres de présentation, le taux de réponses correctes de ces derniers étant assez proche (85%). A l'inverse, les scènes artéfactuelles en ordre *"aléatoire"* sont moins bien reconnues qu'en ordre *"décroissant"* et *"croissant"*, dont le taux de réponses correctes ne varie pas beaucoup (F<1).

3.3. Analyse de la surface affichée

Le temps de réaction des participants correspond à un autre indicateur, à savoir la surface affichée de l'image-test au moment où la réponse des participants se concrétise. En effet, le temps de réponse est lié directement au nombre de pavés ouverts (chaque pavé étant affiché toutes les 300 ms par la programmation de l'expérience). La fonction entre le temps de réaction et le

nombre de pavés ouverts est linéaire. En revanche, la fonction entre le temps de réaction et la surface affichée n'est pas linéaire en raison de la contrainte méthodologique de l'affichage des points d'intérêt. En effet, lorsque certains points d'intérêt sont très proches au niveau de la saillance, la programmation de l'expérience les affiche de manière superposée. Il arrive que certains pavés se superposent partiellement pour les trois ordres de présentation. Par conséquent, il nous apparaît intéressant de rendre compte de la surface affichée dont les participants ont besoin pour reconnaître l'image. En outre, la surface couverte par les pavés ne présente pas le même type de pertinence informationnelle selon les trois ordres de présentation. Ainsi, nous pouvons considérer les informations disponibles pour la reconnaissance d'image (surface et pertinence) comme une mesure de l'information dont le participant a besoin.

Les participants répondent en moyenne en 10,9 s, ce qui correspond à 20,4% de surface affichée (de l'image-test) et avec un taux de réponses correctes de 86,7% pour les scènes naturelles. Quant aux scènes artéfactuelles, ils répondent en moyenne en 12 s, ce qui correspond à 20,3% de surface affichée, avec un taux de réponses correctes de 87,1%.

La surface affichée pour ces deux types de scènes est quasiment identique. De même, aucune différence significative n'apparaît entre la condition de la congruence d'images concernant le pourcentage de surface affichée (21,7% dans la condition où les deux images sont identiques, 22% dans la condition où les deux images sont différentes). Bien que le pourcentage de surface affichée soit équivalent entre ces deux conditions, il n'entraîne pas le même taux de réponses correctes (88,2% dans le cas où l'image-cible et l'image-test sont des images différentes, 84,8% dans le cas où elles sont identiques).

Le pourcentage de surface affichée varie en fonction de l'ordre de présentation, les participants répondant en moyenne à partir de 18,4% de la surface affichée concernant l'ordre *"décroissant"*, 18,2% concernant l'ordre *"croissant"* et 23,5% concernant l'ordre *"aléatoire"* (Figure 4.11).

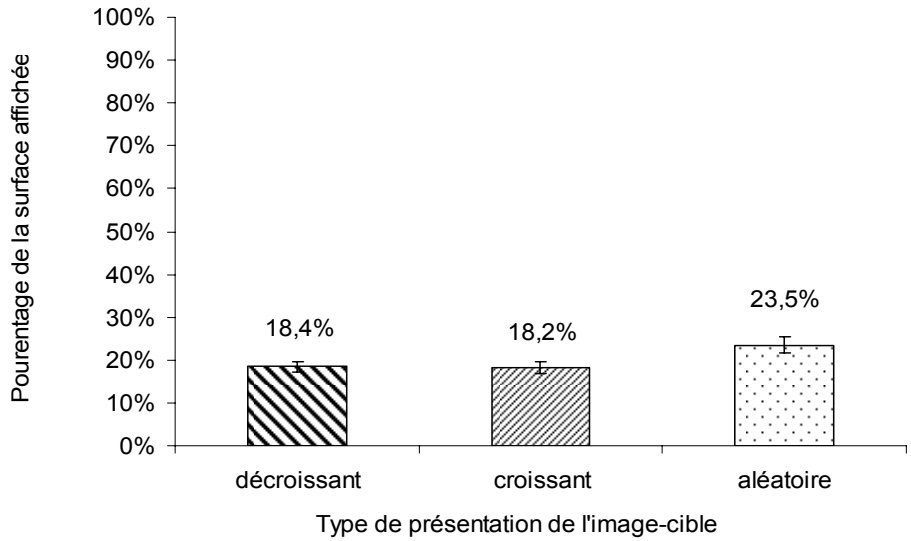


Figure 4.11. Pourcentage de surface affichée en fonction de l'ordre de présentation.

Les résultats montrent que les participants ont vu une plus grande surface de contenu de l'image-test avant de répondre en ordre *"aléatoire"*, qu'en ordre *"décroissant"* et *"croissant"*. Malgré une grande surface affichée, les participants rencontrent plus de difficultés à reconnaître l'image-test pour cet ordre. Deux graphiques de résumé sont présentés selon la congruence entre l'image-cible et l'image-test.

Lorsque l'image-cible et l'image-test sont identiques, les participants répondent en moyenne avec 23,8% de la surface affichée pour l'ordre *"aléatoire"*, et leur taux de réponses correctes est équivalent (83%) par rapport aux deux autres ordres de présentation dont les surfaces affichées sont très proches (autour de 21%) (Figure 4.12).

Lorsque l'image-cible et l'image-test sont différentes, les participants répondent en moyenne avec 23,5% de surface affichée pour l'ordre *"aléatoire"*, mais leur taux de réponses correctes (81%) est inférieur à celui des deux autres ordres (87% pour *"croissant"* et 92% pour *"décroissant"*). Or, la surface dont ces derniers ont besoin pour répondre est relativement faible, puisqu'elle est de 15,3% pour l'ordre *"croissant"* et de 16,4% pour l'ordre *"décroissant"* (Figure 4.13).

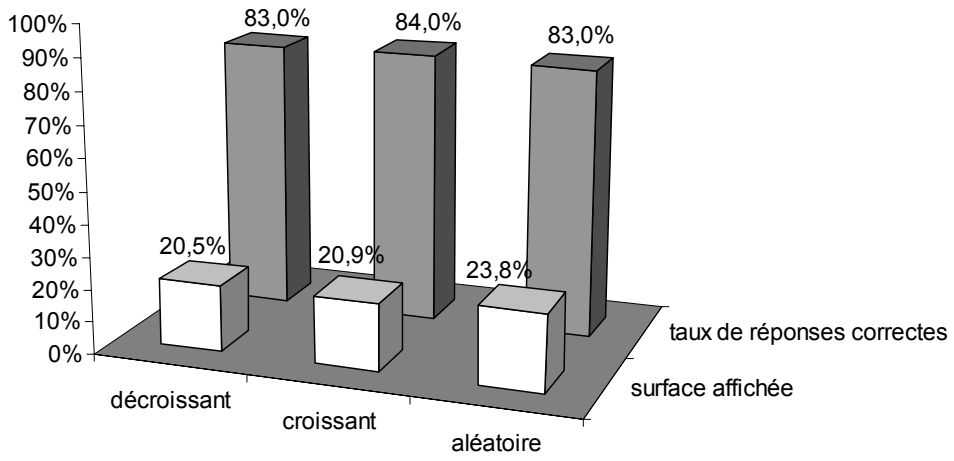


Figure 4.12. Surface affichée et taux de réponses correctes en fonction de l'ordre de présentation lorsque les deux images du couple sont identiques.

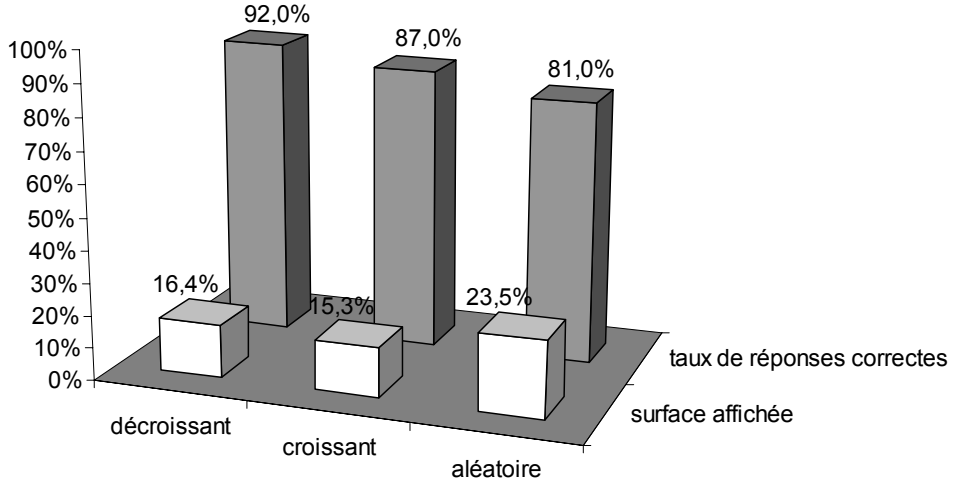


Figure 4.13. Surface affichée et taux de réponses correctes en fonction de l'ordre de présentation lorsque les deux images du couple sont différentes.

3.4. Stratégies utilisées par les participants

Les participants réagissent différemment pour les trois ordres de présentations en fonction de la catégorie de scènes visuelles. Nous pouvons supposer que la présence du stimulus représente pour le participant un certain coût. Il ne donnera sa réponse que s'il a un niveau de certitude suffisant. Dans

cette perspective, le fait que le participant n'ait pas donné de réponse positive ne signifie pas nécessairement qu'il n'a rien perçu du stimulus.

Nous sommes alors amenés à déterminer la part de la sensibilité sensorielle et celle de la stratégie des participants. Pour cela, il est nécessaire d'effectuer une analyse en fonction de la Théorie de la Détection du Signal (TDS) car elle nous permet d'étudier les stratégies employées par les participants dans cette tâche de reconnaissance. La TDS s'applique dans une tâche dont la réponse est binaire ("OUI" ou "NON"), elle regroupe les réponses du sujet selon quatre situations (Tableau 4.3) :

Tableau 4.3

Illustration des quatre situations des réponses selon la TDS

		Types d'essai	
		Signal + Bruit (S + B)	Bruit (B)
Réponses	**OUI**	Détection correcte (DC)	Fausse alarme (FA)
	NON	Omission (O)	Rejet Correct (RC)

- "Détection Correcte (DC)", c'est la réponse "oui" donnée par le sujet considérant le cas où le signal et le bruit sont présents ;
- "Omission (O)", c'est la réponse "non" donnée par le sujet considérant le cas où le signal et le bruit sont présents ;
- "Fausse Alarme (FA)", c'est la réponse "oui" donnée considérant le sujet dans le cas où seul le bruit est présent ;
- "Rejet Correct (RC)", c'est la réponse "non" donnée par le sujet considérant le cas où seul le bruit est présent ;

Le signal est le stimulus présenté au participant. Le bruit signifie un ensemble de phénomènes non pertinents qui demeurent en l'absence du signal. Ainsi, nous distinguons les essais où seul le bruit est présent (B) et les essais où le signal est présent (S+B).

La TDS permet de connaître la relation liant la sensibilité sensorielle et la stratégie de décision du sujet en réponse à un stimulus. Ces deux réponses intermédiaires sont décrites par deux paramètres notés d' et β respectivement. Voici, une présentation des significations pour ces deux paramètres (Figure 4.14).

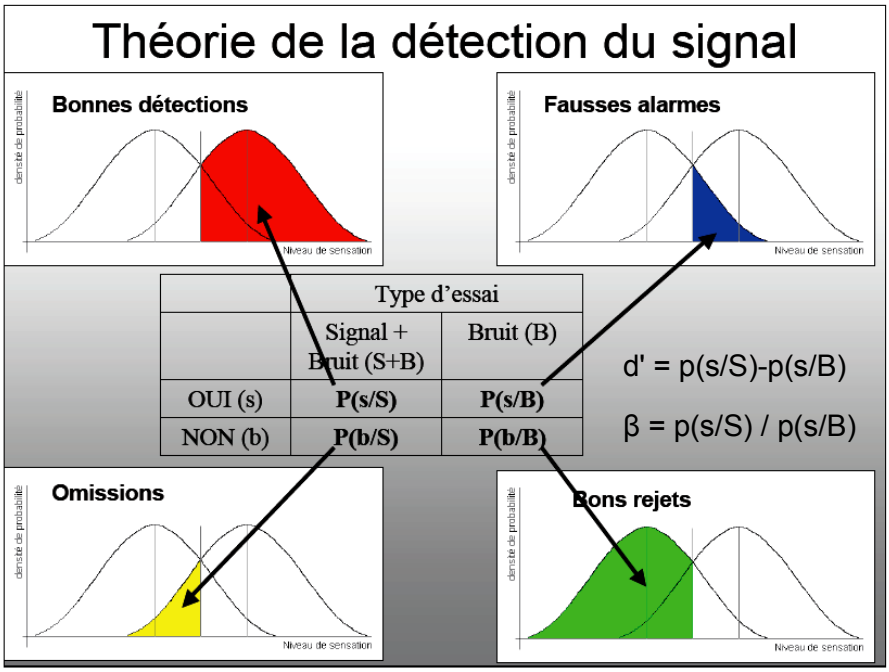


Figure 4.14. Illustration des paramètres d' et β de la TDS.
(http://unpc.univ-lyon2.fr/putois/psychophysiqueputoisTDS.pdf)

- Le paramètre **d'** (d' = p(s/S)-p(s/B)) : capacité sensorielle du sujet.
  - Si d' est élevé : tâche facile ou bonne capacité de discrimination.
  - Si d' est faible : tâche difficile ou mauvaise capacité de discrimination.
  - Si d' est nul : sujet répond au hasard.

- Le paramètre **β** (β = p(s/S) / p(s/B)) : critère décisionnel du sujet.
  - Si β est élevé : comportement le plus prudent (sujet réservé), le sujet dit avoir perçu quelque chose que s'il en est vraiment sûr.
  - Si β est faible : comportement le plus risqué (sujet téméraire ou libéral), le sujet dit avoir perçu quelque chose tout en n'étant pas certain (le sujet suit son intuition, il devine).
  - Si β est proche de 1 : comportement neutre.

3.4.1. Stratégies utilisées pour reconnaître les scènes naturelles

Nous avons récolté les réponses des participants selon les quatre situations de la théorie de la TDS pour les scènes naturelles (Tableau 4.3) :

- "Détection Correcte (DC)", c'est la réponse "V" donnée par les participants considérant le cas où les deux images du couple sont identiques ;
- "Omission (O)", c'est la réponse "N" donnée par les participants considérant le cas où les deux images du couple sont identiques ;
- "Fausse Alarme (FA)", c'est la réponse "V" donnée par les participants considérant le cas où les deux images du couple sont différentes ;
- "Rejet Correct (RC)", c'est la réponse "N" donnée par les participants considérant le cas où les deux images du couple sont différentes ;

Tableau 4.4

Réponses des participants selon l'ordre de présentation des pavés, en intégrant la théorie TDS, considérant les scènes naturelles

		Décroissant	
		même image	image différente
Réponse	oui	DC 92,5%	FA 21,2%
	non	O 7,5%	RC 78,8%
		1	1

d' = 2,24
β = 0,5554

		Croissant	
		même image	image différente
Réponse	oui	DC 88,3%	FA 14%
	non	O 11,7%	RC 86%
		1	1

d' = 2,27
β = 0,9077

		Aléatoire	
		même image	image différente
Réponse	oui	DC 89,6%	FA 15%
	non	O 10,4%	RC 85%
		1	1

d' = 2,3
β = 0,8283

DC = "Détection Correcte" ; O = "Omission" ; FA = "Fausse Alarme" ; RC = "Rejet Correct"
d' = critère de sensibilité ; β = critère décisionnel du sujet

La représentation graphique de ces résultats est la suivante (Figure 4.15).

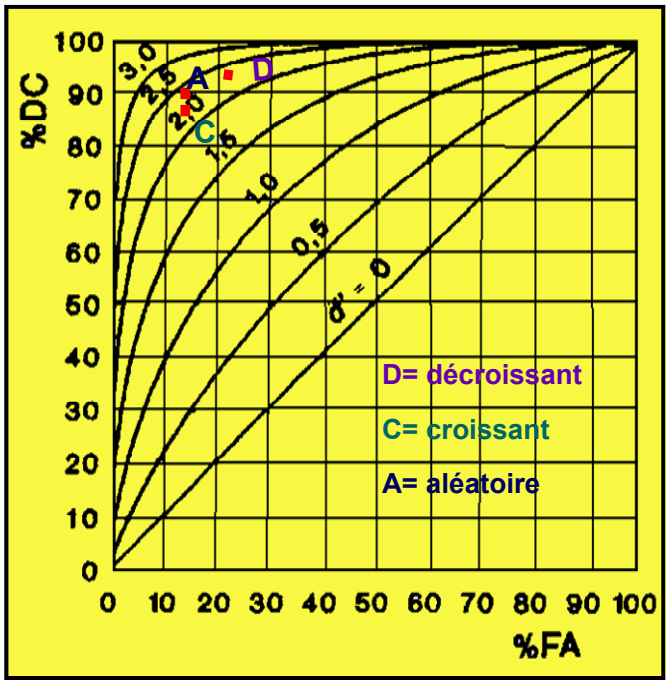


Figure 4.15. Stratégies utilisées par les participants pour reconnaître les scènes naturelles selon l'ordre de présentation des pavés.

Globalement, les trois ordres de présentation entraînent la même difficulté de la réalisation de la tâche. Cependant, les participants changent de stratégie de décision pour ces trois ordres. En ce qui concerne l'ordre de présentation *"décroissant"*, les participants ont un comportement plus risqué, ils font plus de fausses alarmes (21,2%). Lorsque les images ont présentées aléatoirement, ils utilisent une stratégie plus prudente. Quant à l'ordre *"croissant"*, leur stratégie est intermédiaire entre les deux situations précédentes.

3.4.2. Stratégies utilisées pour reconnaître les scènes artéfactuelles

Les réponses des participants en fonction des quatre situations de la théorie TDS pour les images artéfactuelles sont résumées dans le Tableau 4.5.

Tableau 4.5

Réponses des participants selon l'ordre de présentation des pavés, en intégrant la théorie TDS, considérant les scènes artéfactuelles

		Décroissant	
		même image	image différente
Réponse	oui	DC 88,8%	FA 6,4%
	non	O 11,2%	RC 93,6%
		1	1

d' = 2,738
β = 1,2517

		Croissant	
		même image	image différente
Réponse	oui	DC 86,3%	FA 10%
	non	O 13,7%	RC 90%
		1	1

d' = 2,376
β = 1,1715

		Aléatoire	
		même image	image différente
Réponse	oui	DC 88,8%	FA 24,7%
	non	O 11,2%	RC 75,3%
		1	1

d' = 1,88
β = 0,5625

DC = "Détection Correcte" ; O = "Omission" ; FA = "Fausse Alarme" ; RC = "Rejet Correct"
d' = critère de sensibilité ; β = critère décisionnel du sujet

La représentation graphique de ces résultats est la suivante (Figure 4.16)

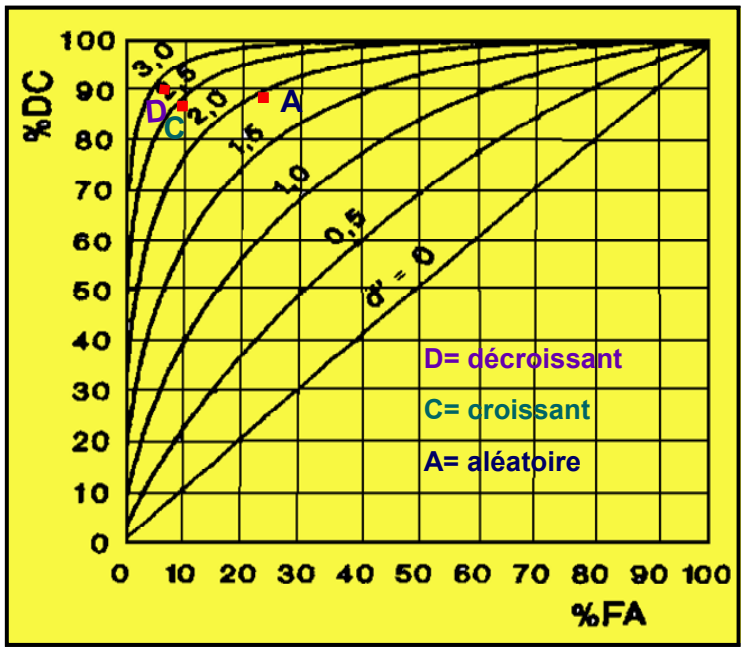


Figure 4.16. Stratégies utilisées par les participants pour reconnaître les scènes artéfactuelles selon l'ordre de présentation des pavés.

Les ordres *"décroissant"* et *"croissant"* sont jugés plus faciles à identifier par rapport à l'ordre *"aléatoire"*. Les participants utilisent une stratégie plus prudente pour les images présentées en ordre "*décroissant*" qu'en "*croissant*". Ils ont un comportement risqué pour l'ordre *"aléatoire"* et procèdent à beaucoup de fausses alarmes (24,7%).

4. Discussion

4.1. Dominance en ordre "*décroissant*" et "*croissant*"

Dans cette expérience, nous avons analysé les effets de la pertinence des informations locales dans la reconnaissance de scènes visuelles complexes.

Nous pouvons déjà remarquer que nos résultats confortent partiellement la théorie de Biederman (1987). En effet, selon lui, la reconnaissance des objets repose sur la perception d'éléments géométriques de base à partir desquels l'objet peut être construit.

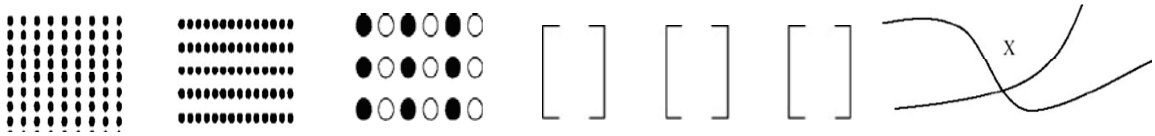
Dans notre expérience, le fait que les participants aient besoin de moins de surface, lorsque l'ordre de présentation des pavés est *"décroissant"* et *"croissant"*, comparé à une présentation *"aléatoire"*, est lié à la pertinence des informations exposées lors des démonstrations "*décroissante*" et "*croissante*" (Figure 4.12. et Figure 4.13). Comme nous le savons, ces deux ordres se caractérisent par des points d'intérêt beaucoup plus importants qu'en situation d'ordre *"aléatoire"*. Donc, la saillance des points d'intérêt a un effet sur la reconnaissance de l'image.

Les participants perçoivent mieux les informations structurales des scènes en ordre de présentation *"décroissant"* et *"croissant"* qu'en ordre de présentation *"aléatoire"*, ce qui s'explique, en premier lieu, par la méthodologie d'affichage des pavés selon l'ordre *"aléatoire"*. En effet, les pavés sont dispersés, parmi lesquels certains ne montrent aucune information de l'objet (par exemple, certains pavés concernent des régions vides). Bien qu'en situation d'ordre *"aléatoire"*, les

participants ont eu besoin d'une plus grande surface pour répondre, la surface affichée ne donnant pas plus d'informations pertinentes sur l'objet que les deux autres situations de présentation.

En second lieu, ce résultat peut être lié à la pertinence des informations affichées dans les pavés en condition d'ordre *"décroissant"* et *"croissant"*. Par la prédiction de l'algorithme, ces deux ordres montrent les informations les plus saillantes (les points d'intérêt les plus importants) de l'image. Après avoir examiné les propriétés des informations exposées au sein de ces pavés, nous trouvons alors une forte densité de bords, de contours et un fort contraste de luminance. Des chercheurs sont ont déjà montrés un lien entre ces caractéristiques et les mouvements oculaires (Mannan, Ruddock, & Wooding, 1996, 1997 ; Reinagel & Zador, 1999).

Selon nos résultats, celles-ci permettraient donc de créer une meilleure représentation de l'image à identifier grâce aux lois de clôture et de continuité de perception (Figure 4.17). Par conséquent, les participants pouvaient construire facilement une représentation partielle concernant un ou plusieurs objets (Friedman, 1979) ou la structure globale de la scène (Fei Fei, VanRullen, Koch, & Perona, 2002, 2005), ce qui permet d'aboutir à la reconnaissance de l'image. Cependant, ces conditions nécessaires à la reconnaissance ne se retrouvaient pas en situation de présentation *"aléatoire"*. C'est la raison pour laquelle le temps de réaction et le taux de réponses correctes obtenus sont moins bon selon cette présentation par rapport aux deux autres.



(a)
Les figures sont perçues comme juxtaposition de colonnes (à gauche) ou de lignes (à droite) grâce à la propriété de proximité

(b)
La figure est perçue comme une série de colonnes par similarité de couleurs

(c)
Bien que tous les segments soient espacés de la même distance, ils sont groupés 2 à 2 de manière à être perçus comme des objets fermés

(d)
La figure est perçue comme deux lignes continues se coupant en X plutôt que comme deux courbes en V se touchant en X (propriété continuité)

Figure 4.17. Illustration des propriétés de regroupement : proximité (a), similarité (b), fermeture (c) et continuité (d).

Une autre explication concernant la supériorité des présentations en ordre *"décroissant"* et *"croissant"* consiste à faire l'hypothèse que ces deux ordres offrent de meilleures conditions de construction d'une représentation spatiale-temporelle de l'image, ce qui faciliterait la coordination en mémoire de travail, quelles que soient les variantes théoriques (Baddeley, 1986, 1992, 2000; Lieury, 2005; Sweller, 2005). La présentation des pavés en ordre *"décroissant"* et *"croissant"* offrirait ainsi une meilleure visualisation des informations. En effet, l'affichage des pavés pour ces deux ordres est souvent en continu spatio-temporellement. Les informations ayant des propriétés perceptives similaires permettraient donc le maintien d'un même type de traitement cognitif. L'administrateur central ne serait en séquence pas en difficulté.

Par ailleurs, dans ces deux ordres de présentation, les informations possèdent des caractéristiques homogènes durant un certain temps d'affichage. Il y a donc peu d'informations hétérogènes, coûteuses en ressources attentionnelles des participants. Ainsi, ces deux ordres de présentation demanderaient des traitements cognitifs moins coûteux (Wolf, 1989). Inversement, en condition d'ordre *"aléatoire"*, les pavés sont non seulement très espacés (la trajectoire entre deux pavés peut être longue), mais aussi les informations affichées sont hétérogènes, ce qui solliciterait une charge mentale supplémentaire des participants. Par conséquent, ceux-ci auraient des difficultés à construire une représentation de l'image à identifier. La mémoire dans cette condition serait mise en difficulté.

Cependant, pour les taux de réponses correctes, nous n'avons pas trouvé de différence significative entre la présentation en ordre de "*décroissant*" et *"croissant"*. Bien que la saillance des informations affichées soit différente au démarrage de l'affichage, à mesure que des pavés s'ouvrent, l'ensemble des informations affichées construirait une carte spatio-temporelle de l'image. Les participants trouveraient alors presque le même type d'informations structurantes de la scène à mesure de l'affichage d'un certain nombre de pavés. Ainsi, les participants répondraient en moyenne sur la base de trente pavés affichés environ. Autrement dit, après cette présentation, soit près de 19% de la surface de l'image, les informations montrées par ces pavés pour ces deux ordres de présentations

couvrent une partie pertinente du contenu de la scène, ce qui n'entraîne pas de différence significative au niveau du type de réponses correctes.

4.2. Scènes naturelles et scènes artéfactuelles

Les trois ordres de présentation entraînent le même niveau de sensibilité décisionnelle des participants pour les scènes naturelles (Tableau 4.4). Ces trois ordres de présentation fournissent donc trois types d'informations jugées de même importance par les participants. Néanmoins, les stratégies utilisées par les participants ne sont pas les mêmes. Ils ont un comportement plus risqué concernant la situation d'ordre *"décroissant"*. Inversement, ils utilisent une stratégie plus prudente concernant la situation d'ordre "*croissant*" et "*aléatoire*".

Quant aux scènes artéfactuelles, les situations d'ordre *"décroissant"* et *"croissant"* ne provoquent pas de différence concernant le critère de sensibilité décisionnelle (Tableau 4.5). Cependant, les participants deviennent très prudents pour la situation d'ordre *"décroissant"* par rapport à la même situation pour les images naturelles. La situation d'ordre "*aléatoire*" devient plus difficile au regard du nombre de fausses alarmes, comparativement à celle des scènes naturelles.

Les participants réagissent différemment pour les scènes naturelles et les scènes artéfactuelles selon l'ordre de présentation.

En situation d'ordre *"décroissant"*, les participants ont un comportement plus risqué pour les scènes naturelles avec un nombre des fausses alarmes important (21,2%, β = 0,56). Cependant, ils sont prudents et sont sources de moins de fausses alarmes pour les scènes artéfactuelles (6,4%, β = 1,25). En revanche, en situation d'ordre *"croissant"*, le comportement des participants est similaire pour ces deux types de scènes (Tableau 4.4, Tableau 4.5).

Les différentes stratégies utilisées par les participants selon ces deux ordres de présentation peuvent être liées, à la caractéristique des deux types de

scène. En effet, d'après Gaillard et Bourges (1999), les scènes naturelles sont composées d'éléments ou de structures spatiales non réguliers, discontinus, avec beaucoup de hautes fréquences. A l'opposé, les scènes artéfactuelles sont composées d'éléments réguliers, continus, avec beaucoup de basses fréquences. Il y a donc plus de similitudes aux formes géométriques dans les scènes artéfactuelles que dans les scènes naturelles. En conséquence, les participants semblent formuler facilement des "géons" beaucoup plus structurés pour les scènes artéfactuelles que pour les scènes naturelles (Gaillard, Boulliou & Gautier, 1996; Gaillard & Bourges, 1999). En effet, la situation d'ordre *"décroissant"* fournit une densité d'éléments localisés concernant des bords, des contours, des contrastes de luminance plus forts, ce qui permettrait aux participants de mieux percevoir les invariants des objets tout en activant les "schémas de situation de base" ("Situations Cognitives de Référence", Rossi, 2005). Ces derniers faciliteront l'identification de la scène. Par conséquent, ils répondent rapidement avec peu de fausses alarmes pour les scènes artéfactuelles et inversement pour les scènes naturelles.

Quant à la supériorité de l'ordre *"décroissant"* par rapport à l'ordre *"croissant"*, celle-ci est probablement liée à la saillance des informations spatio-temporelles. En situation d'ordre *"décroissant"*, les points d'intérêt les plus saillants s'affichent en premier, puis au fur et à mesure, les participants visualisent des informations beaucoup plus structurales que dans l'ordre *"croissant"*. Ce mode de présentation permet de mieux percevoir un ou plusieurs objets typiques de la scène ainsi que leur relation spatiale. Cela permettrait d'activer l'identification de la scène (Friedman, 1979, De Graef, Christiaens, & d'Ydewalle, 1990).

En situation d'ordre aléatoire, les participants rencontrent plus de difficultés pour reconnaître les scènes artéfactuelles (d' = 1,88) que pour les scènes naturelles (d' = 2,3). Pour cette situation, les informations affichées sont caractérisées par des contours discontinus et une structuration moins lisible. Cependant, les informations affichées apparaissent moins pertinentes pour les scènes artéfactuelles. Ainsi, les participants ont un comportement plus risqué, avec beaucoup de fausses alarmes, pour les scènes artéfactuelles (24,7%, β = 0,56) comparé aux scènes naturelles (15%, β = 0,82%).

Cette expérience a mis en évidence l'effet de la pertinence des informations dans la reconnaissance de scènes complexes. Les participants ont une meilleure performance pour l'ordre *"décroissant"* et *"croissant"* que pour l'ordre *"aléatoire"*. En effet, les informations issues de ces deux ordres possèdent deux types de caractéristiques différentes : les contours et le contraste de luminance. Alors que les contours permettent de former les objets et de les identifier, le contraste de luminance permet de construire une structuration spatiale de l'image ainsi que de localiser les objets.

Cependant, nous pouvons nous interroger sur le mécanisme de traitement des informations utilisées par les participants.

Quels sont les rôles joués par les contours et par la structuration spatiale de luminance dans cette tâche de reconnaissance?

L'identification d'une scène passe-t-elle par l'aboutissement de la représentation sémantique (signification des objets qui sont formés par les contours) ou bien par l'aboutissement préalable de l'activation des propriétés globales de la scène (représentation perceptive, structuration spatiale de luminance)?

L'expérience 2 a pour objectif de répondre à ces questions.

Chapitre 5 Rôle des contours et de la structuration spatiale des différentes zones de luminance -- tâche de reconnaissance (expérience 2)

Cette expérience vise à étudier deux facteurs pouvant influencer la reconnaissance d'une scène visuelle complexe, à savoir les contours des objets qu'elle contient et la structuration spatiale des différentes zones de luminance la composant. Afin d'observer l'effet de ces deux facteurs, nous avons réalisé des modifications de scènes visuelles complexes sur des photographies de sorte que seuls restent, soit les contours, soit la structuration spatiale des différentes zones de luminance. Nous obtenons ainsi trois types d'images (Figure 5.1)

- "***Image-originale***" : Images en niveau de gris, transformées à partir de photographies couleur avec le logiciel IrfanView 4.10.

- "***Image-contour***" : Images en niveau de gris, transformées par ce même logiciel. Elles contiennent uniquement des bords et des contours des objets.

- **"*Image-luminance*" :** Images en niveau de gris, figurées par des différentes zones de luminance, sans aucune information précise concernant les objets. Les zones de luminance sont déterminées de manière suivante : l'image est découpée en petits carrés de taille 5x5 (ou 20x20, 35x35, 60x60) pixels, la valeur moyenne de l'intensité lumineuse dans la zone est ensuite calculée, puis cette valeur est affectée à tous les pixels de la zone.

(a) Image-originale (b) Image-contour (c) Image-luminance

Figure 5.1. Illustration des trois types de transformations de scènes visuelles complexes utilisées au cours de l'expérience 2 : (a) *"image-originale"* ; (b) *"image-contour"* ; (c) *"image-luminance"*.

1. Méthode

1.1. Participants

Trente-quatre étudiants (20 hommes et 14 femmes) inscrits en licence à l'Université de Rennes 2 ont participé à cette recherche. Tous les participants, naïfs du point de vue des objectifs de cette recherche, ont attesté d'une vue normale ou corrigée. Aucun d'entre eux n'a participé à l'expérience 1.

1.2. Matériel

Quatre-vingt dix couples d'images en niveau de gris ont été choisies dans la base de données COREL. Certaines de ces photographies avaient été utilisées dans l'expérience 1. Ce sont toutes des scènes naturelles appartenant aux catégories suivantes : mer, montagne, plage, désert et champ.

Le masque expérimental, du même format que les photographies, est constitué de différentes figures grisées (des croix).

1.3. Equipement

L'expérience, pilotée par ordinateur, se déroule dans une salle découpée en quatre box contenant chacun un ordinateur. Le matériel informatique est strictement identique pour les quatre postes avec un processeur du type Pentium III, cadencé à 1 Ghz équipé de 128 Mo de RAM, un écran de 17 pouces et un clavier. L'expérience fait appel au même logiciel et au même programme que l'expérience précédente.

1.4. Procédure expérimentale

La procédure utilisée dans l'expérience est assez proche de celle de l'expérience 1.

Les participants sont assis face à l'écran à une distance de 100 centimètres. Les images (cibles et tests) couvrent un angle visuel de 30° environ.

Avant de commencer l'expérience, une consigne informe les participants que deux images (image-cible et image-test) sont présentées successivement. La première image (image-cible) est montrée tout d'abord aux participants sous un des trois types de transformations (*"image-originale", "image-contour" et "image-luminance"*) pendant un des trois types de temps d'affichage (50 ms, 150 ms ou 300 ms). Après la présentation de l'image-cible, un masque est immédiatement affiché pendant 200 ms, puis l'image-test apparaît. Cette image est toujours en niveau de gris[6] dont le temps d'affichage est programmé à 10 secondes, au delà duquel, faute de réponse du participant, l'expérience se poursuit automatiquement. Une fois l'image-test apparue, le participant doit décider si cette images et l'image-cible sont identiques en appuyant sur la touche "V" si c'est le cas, en appuyant sur la touche "N" dans le cas contraire. (Figure 5.2)

Le texte consigne de cette expérience est le suivant :

"Vous devez d'abord observer attentivement une image de paysage appelée image-cible. Cette image-cible *sera présentée sous des formes différentes : normale ou dégradée. Ensuite, cette image va disparaître, puis vous verrez apparaître une nouvelle image de paysage appelée image-test.*

L'image-cible et l'image-test sont soit identiques soit différentes.

Dès que vous pensez que l'image-test est la même que l'image-cible, appuyez le plus vite possible sur la touche "V" du clavier.

Si au contraire, vous pensez que l'image-test n'est pas la même que l'image-cible, appuyez le plus vite possible sur la touche "N" du clavier".

[6] L'image-test est toujours une image originale en niveau de gris dans les expériences 2, 3, 4 et 5.

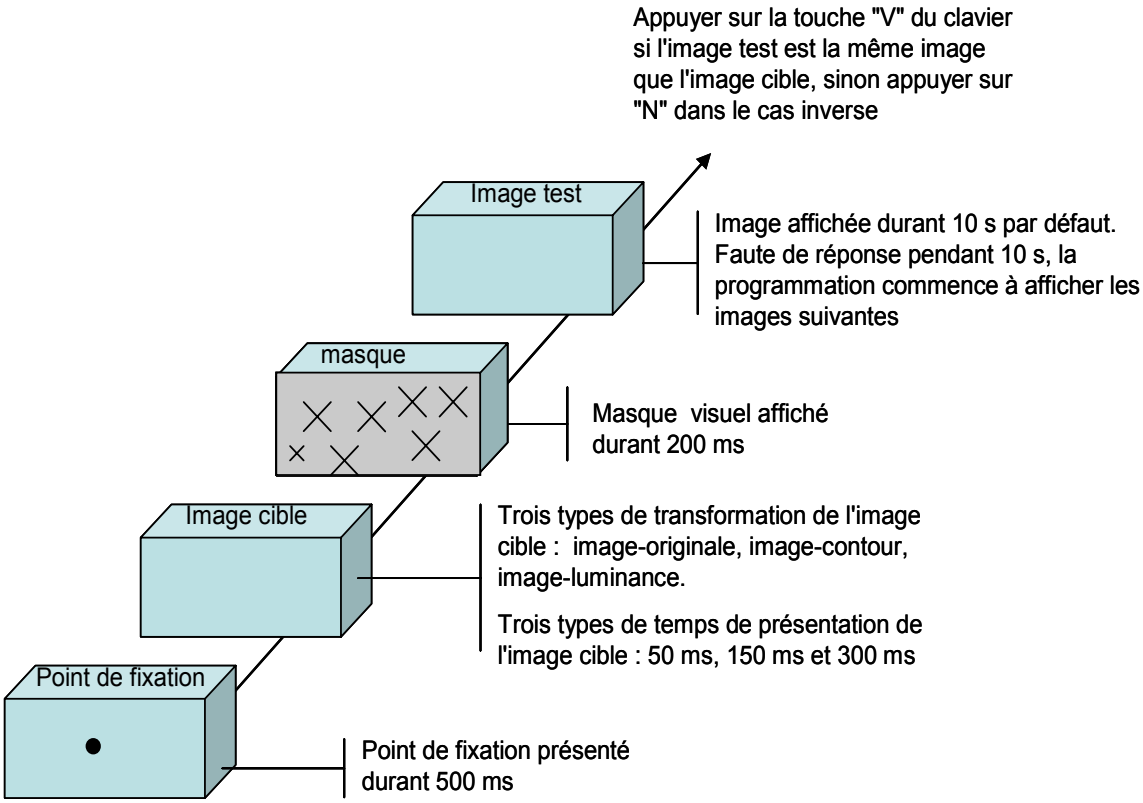


Figure 5.2. Procédure de l'expérience 2.

Plusieurs exemples sont montrés aux participants dans l'objectif d'expliquer la tâche qu'ils vont réaliser (Figure 5.3). Une précision importante est signalée : les participants doivent comparer le contenu de l'image-test et celui de l'image-cible.

- "Image-cible = **montagne-originale**" *vs* "image-test = **même montagne en niveau de gris**".

- "Image-cible = **plage-contour**" *vs* "image-cible = **une autre plage en niveau de gris**" (l'image-cible et l'image-test sont différentes, mais elles appartiennent à la même catégorie).

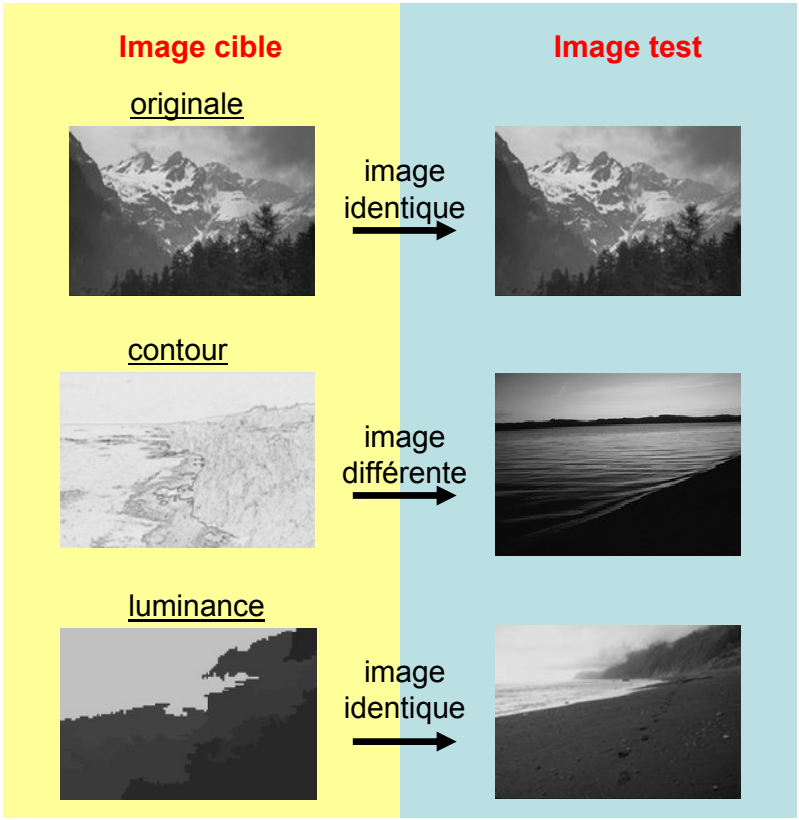


Figure 5.3. Illustration d'images-cibles et d'images tests pour l'expérience 2.

Comme dans l'expérience précédente, les participants passent d'abord par une phase d'entraînement en utilisant des images différentes de celles utilisées durant l'expérience.

L'expérience 2 dure approximativement quinze minutes.

1.5. Plan expérimental

Plan expérimental

S_{34}*T3*I3*C2

Le facteur **S** correspond aux participants ; le facteur **T** correspond au temps de présentation de l'image-cible (50 ms *vs*. 150 ms *vs.* 300 ms) ; le facteur **I** correspond au type de transformation de l'image-cible ("*image-originale*" *vs. "image-contour" vs. "image-luminance"*) ; le facteur **C** correspond à la congruence d'images (image identique *vs.* image différente).

Variables dépendantes (VD)

Le temps de réaction (ms) et le taux de réponses correctes (%) sont les deux variables indépendantes dans cette expérience.

2. Analyse des résultats

Comme dans l'expérience 1, les résultats portent sur le taux de réponses correctes et le temps de réaction. Une analyse de la variance (ANOVAs à mesures répétées) a été réalisée ainsi qu'une analyse dérivée de la Théorie de la Détection du Signal (TDS).

Dans cette expérience, il s'agit d'évaluer le rôle des contours et de la structuration spatiale de luminance dans la reconnaissance de scènes naturelles complexes. Comme dans l'expérience précédente, les analyses seront effectuées séparément selon la congruence entre l'image-cible et l'image-test, soit dans le cas où les deux images du couple sont bien deux images identiques (une bonne réponse consistant alors à répondre "V" pour vrai), soit dans le cas où les deux images du couple sont différentes (la bonne réponse est alors "N" pour rejeter l'hypothèse que les deux images correspondent à la même image).

2.1. Les deux images du couple sont identiques

Les résultats obtenus pour les trois types de transformation de l'image-cible sont résumés dans le Tableau 5.1.

Tableau 5.1

Temps de réaction (ms) et taux de réponses correctes (%) en fonction du temps de présentation et du type de transformation de l'image-cible pour l'expérience 2. Erreurs type entre parenthèses.

	Temps de présentation			Transformation de l'image-cible		
	50 ms	150 ms	300 ms	Originale	Contour	Luminance
Temps de réaction (ms)						
	1290	1061	1074	981	1144	1301
	(66)	(49)	(47)	(52)	(56)	(61)
Taux de réponses correctes (%)						
	68,2	85,3	86,3	90,6	79,2	70,0
	(2,4)	(2,1)	(1,6)	(1,5)	(2,7)	(2,2)

2.1.1. Analyse du temps de réaction

Le temps de présentation des images-cibles a un effet sur le temps de réaction, F (2, 66) = 27.247, p< .001. Les participants répondent plus lentement pour les images présentées en 50 ms (1290 ms) que les images présentées en 150 ms (1061 ms), F (1, 33) = 21.523, p< .001, et que les images présentées 300 ms (1074 ms), F (1, 33) = 18.115, p< .001. Parmi ces dernières, le temps de réaction ne varie pas significativement (F<1) (Figure 5.4).

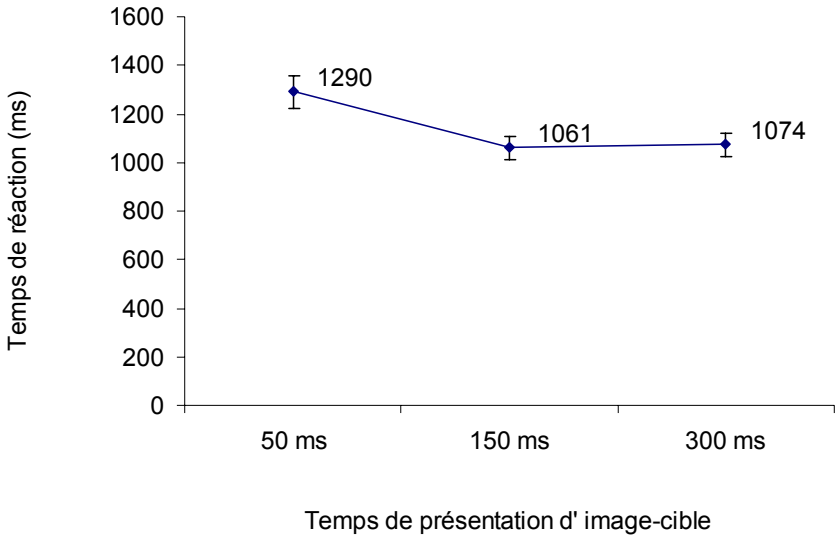


Figure 5.4. Temps de réaction en fonction du temps de présentation de l'image-cible.

Le temps de réaction varie significativement en fonction du type de transformation de l'image-cible, F (2, 66) = 27.171, p< .001. Les participants répondent le plus rapidement pour les *"images-originales"* (981 ms) que pour les *"images-contours"* (1144 ms), F (1, 33) = 25.227, p< .001, et que pour les *"images-luminance"* (1301 ms), F (1, 33) = 15.337, p< .001. Parmi ces dernières, la différence du temps de réaction est également significative F (1, 33) = 17. 253, p< .001 (Figure 5.5).

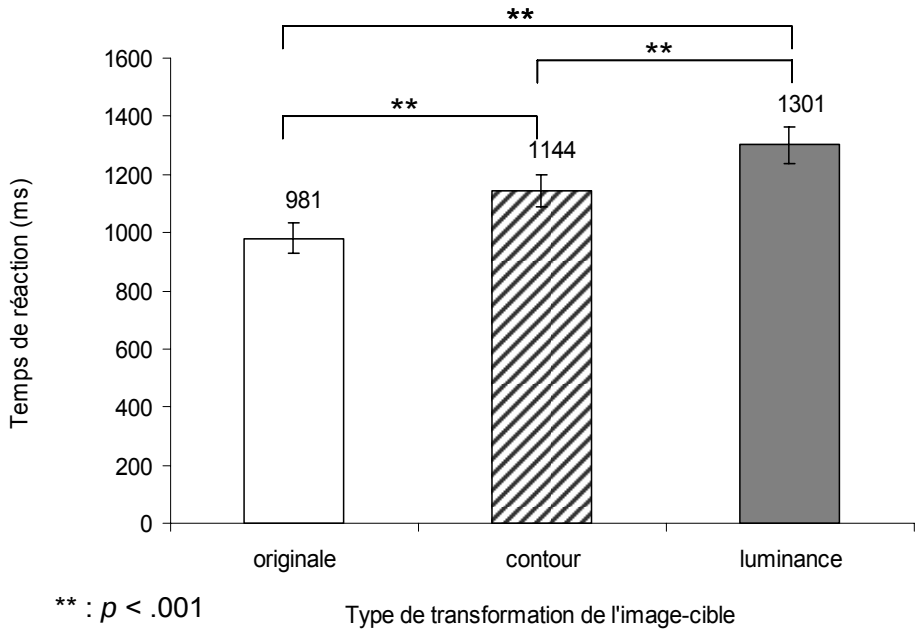


Figure 5.5. Temps de réaction en fonction du type de transformation de l'image-cible.

2.1.2. Analyse du taux de réponses correctes

Quant au taux de réponses correctes, les participants répondent différemment en fonction du temps de présentation de l'image-cible, F (2, 66) = 41.706, p< .001. Ils répondent moins bien dans le cas où les images sont présentées en 50 ms (68,2%) que dans le cas où elles sont présentées en 150 ms (85,3%), F (1,33) = 11.316, p< .05 et en 300 ms (86,3%), F (1, 33) = 7.313, p< .05. Cependant, elles restent stables à partir d'une durée de présentation de 150 ms (F<1) (Figure 5.6).

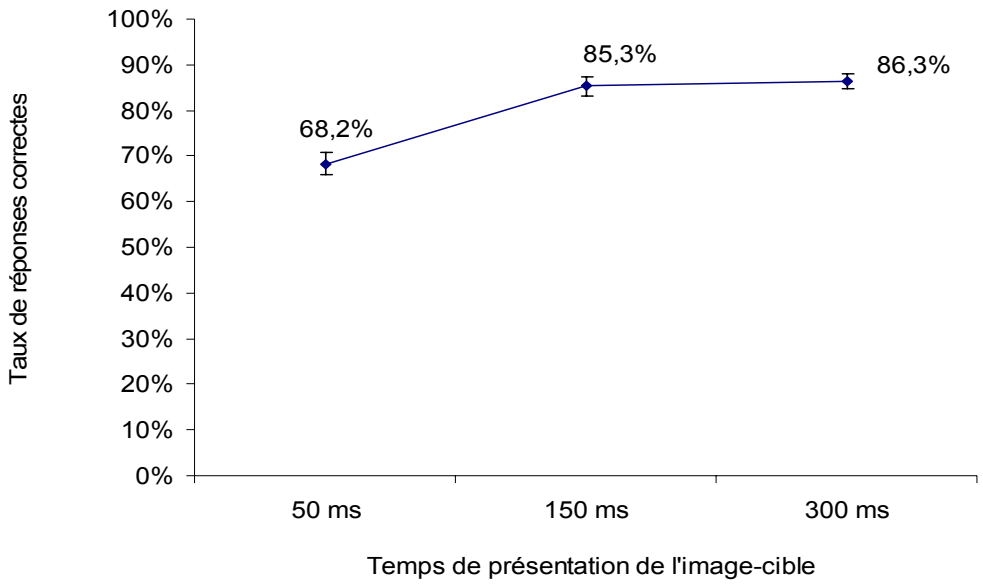


Figure 5.6. Taux de réponses correctes en fonction du temps de présentation de l'image-cible.

Les taux de réponses correctes varient significativement en fonction du type de transformation d'image, F (2, 66) = 31.353, p< .001.

Les participants reconnaissent préférentiellement les *"images-originales"* (90,6%) plutôt que les "*images-contours*" (79,2%), F (1,33) = 15.446, p< .05, alors que les "*images-luminance*" (70,0%) sont les moins reconnues, F (1,33) = 33.277, p< .001. Ils reconnaissent mieux les *"images-contours"* que les "*images-luminance*", F (1,33) = 13.336, p< .05 (Figure 5.7).

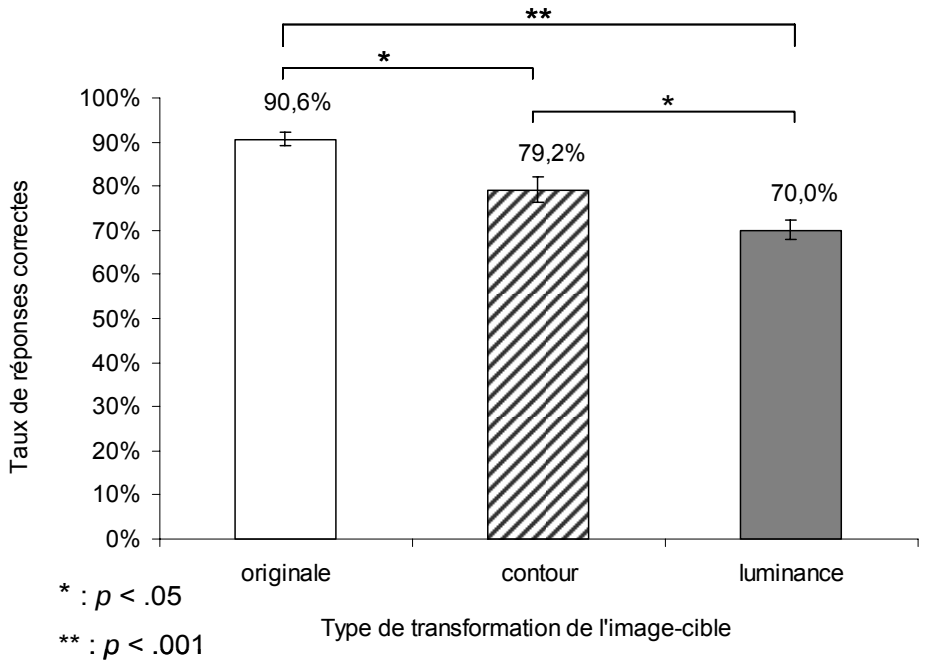


Figure 5.7. Taux de réponses correctes en fonction du type de transformation de l'image-cible.

Une interaction significative est observée entre le temps de présentation et le type de transformation de l'image, F (4, 132) = 4.644, p< .001 (Figure 5.8).

Les *"images-contours"* sont mieux reconnues que les *"images-luminance"* lorsque le temps de présentation de 150 ms (89% et 74%), F (1, 33) = 4.356, p< .05, et lorsque le temps de présentation est de 300 ms (88% et 74%), F (1, 33) = 4.113, p< .05.

En revanche, lorsque le temps de présentation est de 50 ms, la différence du taux de réponse correcte entre ces deux types de transformation d'images n'est pas significative (F<1). Une même tendance est trouvée concernant la performance pour les "*images-luminance*" en fonction du temps de présentation : elle s'améliore de 50 ms à 150 ms, F (1, 33) = 11.336, p< .001, puis devient stable entre 150 ms et 300 ms.

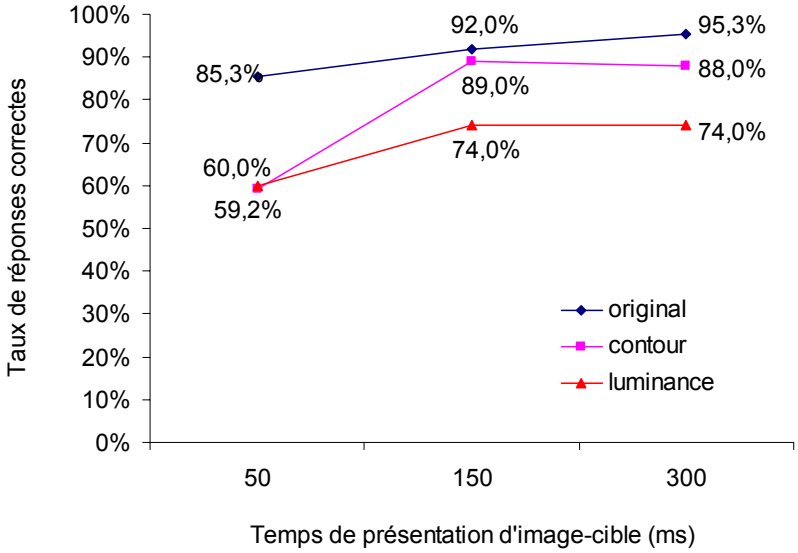


Figure 5.8. Taux de réponses correctes en fonction du temps de présentation de l'image-cible

2.2. Les deux images du couple sont différentes

Les résultats obtenus pour les trois types de transformation de l'image-cible sont résumés dans le Tableau 5.2.

Tableau 5.2

Temps de réaction de rejet (ms) et taux de rejets corrects (%) en fonction du temps de présentation et du type de transformation de l'image-cible pour l'expérience 2. Erreurs type entre parenthèses.

	Temps de présentation			Transformation de l'image-cible		
	50 ms	150 ms	300 ms	Originale	Contour	Luminance
Temps de réaction de rejet (ms)						
	1275	1253	1218	1045	1309	1392
	(53)	(64)	(57)	(43)	(64)	(73)
Taux de rejets corrects (%)						
	70,8	80,6	76,1	92,2	61,6	73,7
	(2,7)	(2,1)	(2,7)	(1,4)	(3,2)	(2,7)

2.2.1. Analyse du temps de réaction de rejet

Quand l'image-cible et l'image-test sont deux images différentes, les participants répondent avec un temps similaire quel que soit le temps de présentation des images-cibles (F<1) (Figure 5.9).

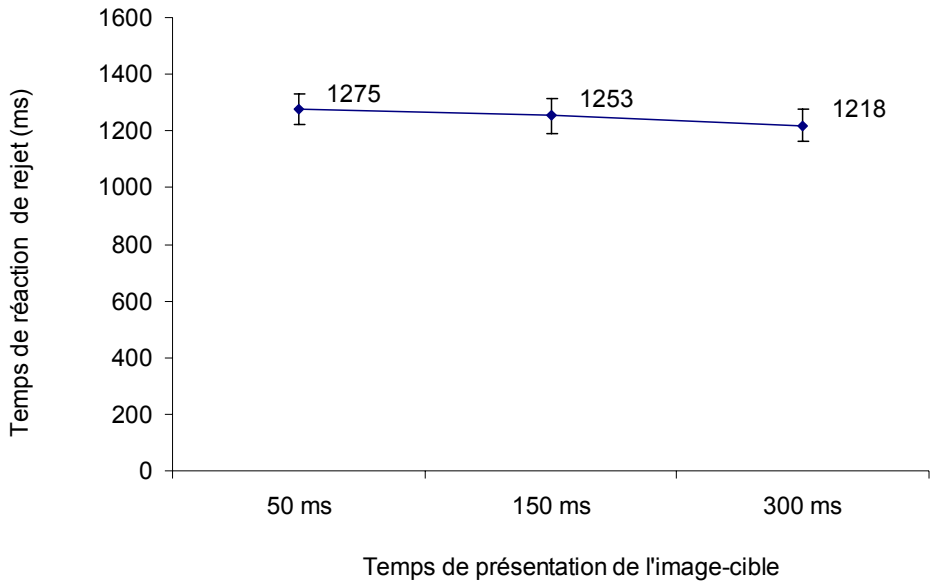


Figure 5.9. Temps de réaction de rejet en fonction du temps de présentation de l'image-cible

Les participants répondent moins rapidement pour les "*images-luminance*" (1392 ms) que pour les "*images-contours*" (1309 ms)", F (1, 33) = 24.513, p< .001, et d'autant plus pour les "*images-originales*" (1045 ms), F (1, 33) = 20.557, p< .001. Parmi ces dernières, la différence n'est pas significative (F<1) (Figure 5.10).

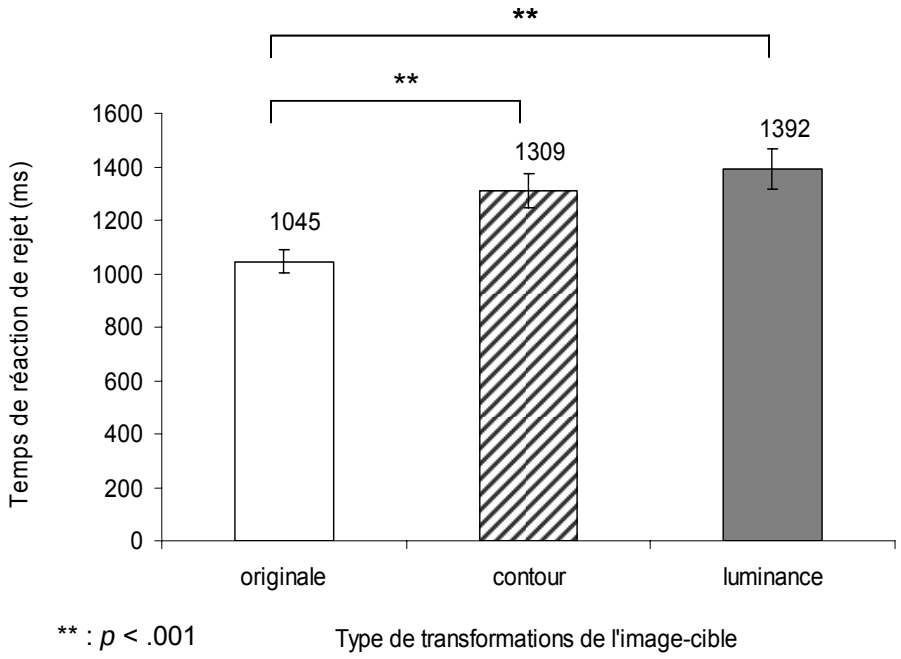


Figure 5.10. Temps de réaction de rejet en fonction du type de transformation de l'image-cible.

2.2.2. Analyse du taux de rejets corrects

Les rejets corrects varient significativement en fonction du temps de présentation, F (2, 66) = 6.590, p< .05.

Les images présentées en 50 ms (70,8%) sont moins bien rejetées que lorsqu'elles sont présentées en 150 ms (80,6%), F (1, 33) = 16.371, p< .05, et que lorsqu'elles sont présentées en 300 ms (76,1%), F (1, 33) = 12.441, p< .05, la différence entre ces dernières n'est pas significative (F<1) (Figure 5.11).

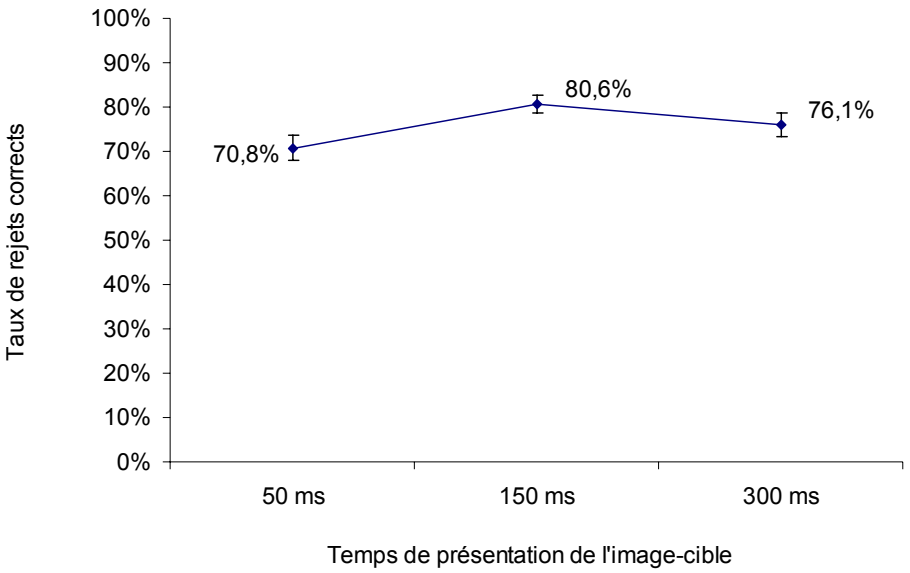


Figure 5.11. Taux de rejets corrects en fonction du temps de présentation de l'image-cible.

Quant aux rejets corrects en fonction du type de transformation de l'image, un effet significatif est observé, F (2, 66) = 61.594, p< .001. La performance est moins bonne pour les *"images-contours"* (61,6%) en comparaison des *"images-luminance"* (73,7%), F (1, 33) = 25.664, p< .001 (Figure 5.12).

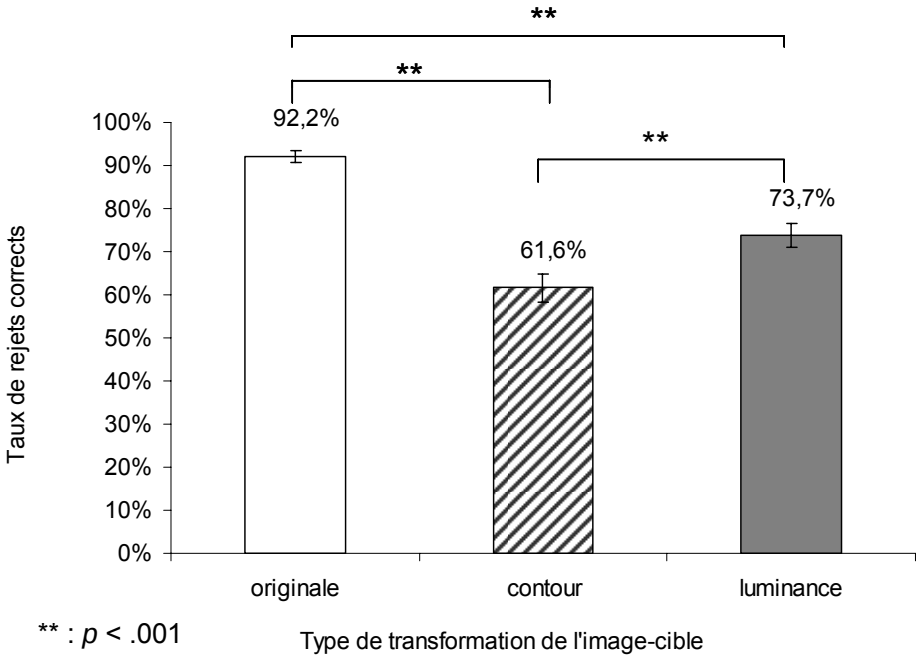


Figure 5.12. Taux de rejets corrects en fonction du temps de présentation de l'image-cible.

Dans le cas où l'image-cible et l'image-test sont différentes, les *"images-luminance"* sont mieux rejetées que les *"images-contours"* lorsque le temps de présentation est de 50 ms (69,4% pour les *"images-luminance"* et 57,2% pour les *"images-contours"*), F (1, 33) = 8.432, p< .05, et lorsque le temps de présentation est de 150 ms (85,0% et 62,1%), F (1, 33) = 23.441, p< .001. (Figure 5.13).

Les participants rejettent préférentiellement les "*images-luminance*" lorsqu'elles sont présentées en 150 ms que lorsqu'elles sont présentées en 50 ms, F (1, 33) = 15.117, p< .001, et en 300 ms, F (1, 33) = 11.537, p< .001. Le taux de rejets corrects pour les "*images-contours*" s'améliore entre 50 ms et 300 ms de temps de présentation, F (1, 33) = 7.226, p< .05.

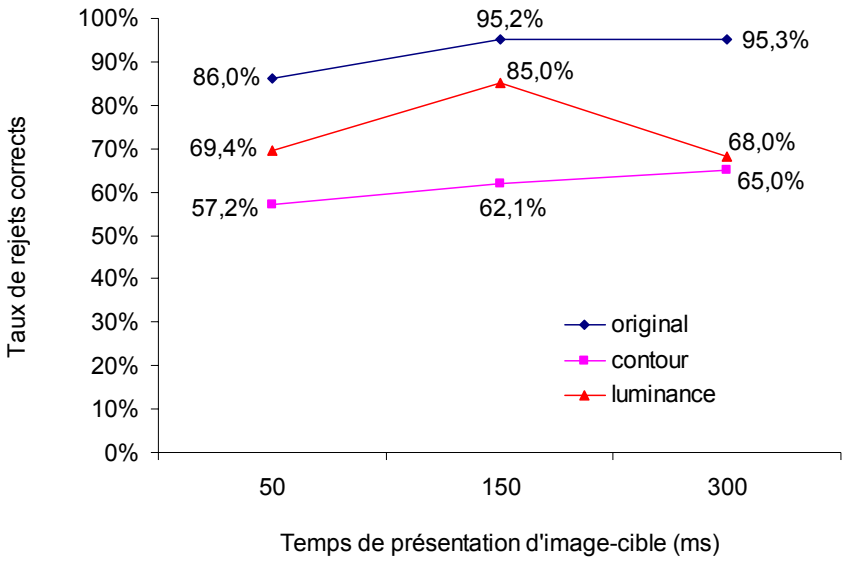


Figure 5.13. Taux de rejets corrects en fonction du temps de présentation de l'image-cible.

2.3. Stratégies utilisées par les participants

2.3.1. Le temps de présentation est de 50 ms

Les réponses des participants en fonction des quatre situations de la théorie TDS sont présentées dans le Tableau 5.3.

Tableau 5.3

Réponses des participants selon le type de transformation de l'image-cible pour une présentation de 50 ms en intégrant la théorie TDS

		Originale	
		même image	image différente
Réponse	oui	DC 84,7%	FA 13,5%
	non	O 15,3%	RC 86,5%
		1	1
		d' = 2,226	
		β = 1,0776	

		Contour	
		même image	image différente
Réponse	oui	DC 60%	FA 42,9%
	non	O 40%	RC 57,1%
		1	1
		d' = 0,432	
		β = 0,7062	

		Luminance	
		même image	image différente
Réponse	oui	DC 60%	FA 31,2%
	non	O 40%	RC 68,8%
		1	1
		d' = 0,743	
		β = 1,9349	

DC = "Détection Correcte" ; O = "Omission" ; FA = "Fausse Alarme" ; RC = "Rejet Correct"
D' = critère de sensibilité ; β = critère décisionnel du sujet

La représentation graphique suivante illustre les résultats du tableau ci-dessus (Figure 5.14).

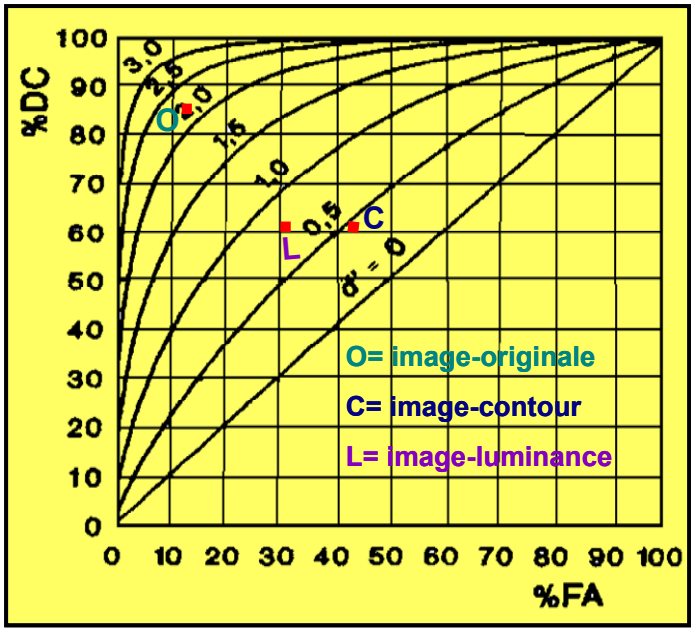


Figure 5.14. Stratégies utilisées par les participants pour identifier les images-cibles présentées en 50 ms.

Les participants rencontrent des difficultés à effectuer la tâche lorsque l'image-cible est une "*image-contour*" et une "*image-luminance*". Ils font plus de

fausses alarmes pour les *"images-contours"* (42,9%) que pour les *"images-luminance"* (31,2%). Les participants ont un comportement plus prudent avec les "*images-luminance*" qu'avec les "*images-contours*". Les *"images-originales"* apparaissent relativement faciles pour lesquelles, les participants ont une stratégie neutre.

2.3.2. Le temps de présentation est de 150 ms

Les réponses des participants en fonction des quatre situations de la théorie TDS pour les images-cibles présentées en 150 ms sont exposées ci-dessous (Tableau 5.4).

Tableau 5.4

Réponses des participants selon le type de transformation de l'image-cible pour une présentation de 150 ms en intégrant la théorie TDS

		Originale	
		même image	image différente
Réponse	oui	DC 91,8%	FA 5,3%
	non	O 8,2%	RC 94,7%
		1	1
		d' = 3,008	
		β = 1,1614	

		Contour	
		même image	image différente
Réponse	oui	DC 89,4%	FA 37,6%
	non	O 10,6%	RC 62,4%
		1	1
		d' = 1,564	
		β = 0,2532	

		Luminance	
		même image	image différente
Réponse	oui	DC 74,7%	FA 15,3%
	non	O 25,3%	RC 84,7%
		1	1
		d' = 1,688	
		β = 1,5391	

DC = "Détection Correcte" ; O = "Omission" ; FA = "Fausse Alarme" ; RC = "Rejet Correct"
D' = critère de sensibilité ; β = critère décisionnel du sujet

La représentation graphique suivante illustre les résultats du tableau ci-dessus (Figure 5.15).

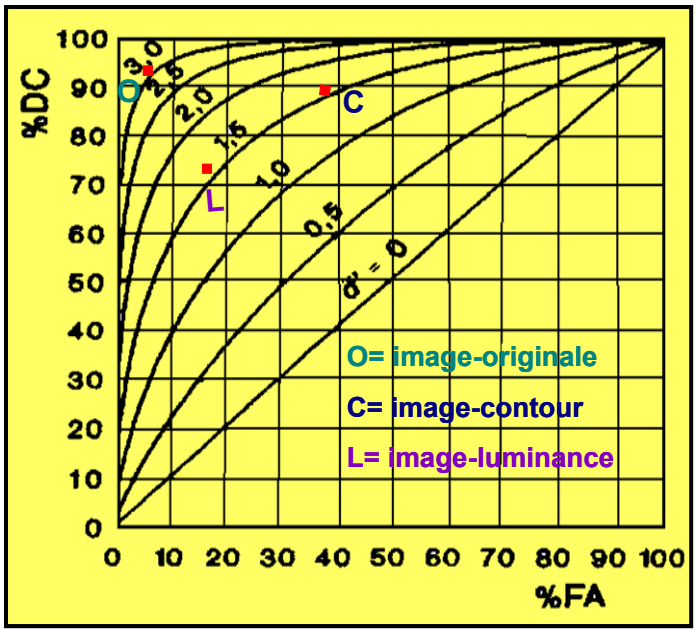


Figure 5.15. Stratégies utilisées par les participants pour identifier les images-cibles présentées en 150 ms.

Lorsque le temps de présentation est de 150 ms, la tâche devient relativement plus facile qu'en 50 ms. Les *"images-contours"* et les *"images-luminance"* induisent le même niveau de sensibilité. Cependant, les participants font plus de fausses alarmes avec les *"images-contours"* (37,6%) qu'avec les "*images-luminance*" (15,3%), ils utilisent une stratégie plus prudente avec les "*images-luminance*".

2.3.3. Le temps de présentation est de 300 ms

Le Tableau 5.5 résume les réponses des participants en fonction des quatre situations de la théorie TDS pour les images-cibles présentées en 300 ms.

Tableau 5.5

Réponses des participants selon le type de transformation de l'image-cible pour une présentation de 300 ms en intégrant la théorie TDS

		Originale	
		même image	image différente
Réponse	oui	DC 95,3%	FA 4,7%
	non	O 4,7%	RC 95,3%
		1	1

d' = 3,35
β = 1

		Contour	
		même image	image différente
Réponse	oui	DC 88,2%	FA 34,7%
	non	O 11,8%	RC 65,3%
		1	1

d' = 1,578
β = 0,332

		Luminance	
		même image	image différente
Réponse	oui	DC 75,3%	FA 32,4%
	non	O 24,7%	RC 67,6%
		1	1

d' = 1,042
β = 0,6675

DC = "Détection Correcte" ; O = "Omission" ; FA = "Fausse Alarme" ; RC = "Rejet Correct"
D' = critère de sensibilité ; β = critère décisionnel du sujet

La représentation graphique suivante illustre les résultats du tableau ci-dessus (Figure 5.16)

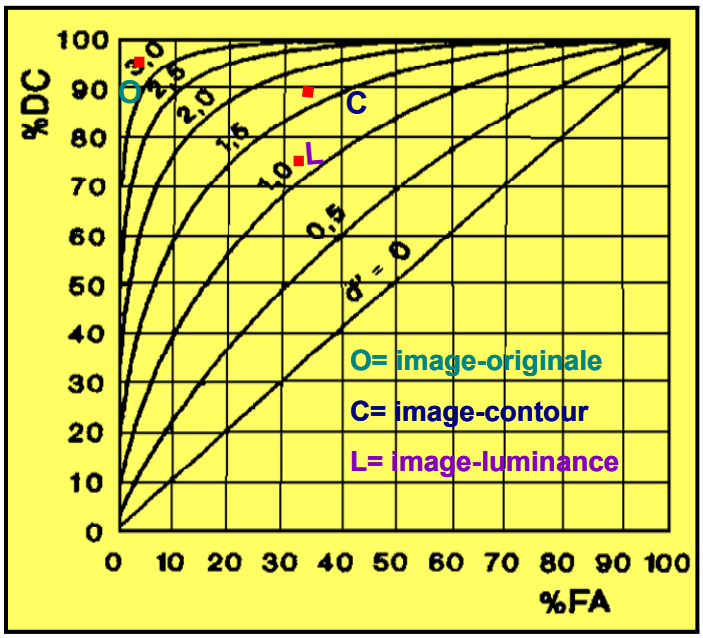


Figure 5.16. Stratégies utilisées par participants pour identifier les images-cibles présentées en 300 ms.

Lorsque le temps de présentation est de 300 ms, les "*images-originales*" et les "*images-contours*" ne présentent pas de différence comportementale par rapport à celles présentées durant 150 ms. Cependant, les *"images-luminance"*

paraissent plus difficiles à reconnaître à 150 ms, ce qui entraîne une stratégie plus risquée pour les participants qu'avec une durée de 300 ms.

3. Discussion

L'objectif de cette expérience était d'observer le rôle des contours et de la structuration spatiale de luminance dans la reconnaissance de scènes naturelles complexes.

Les trois types de transformation n'entraînent pas la même sensibilité lors de la réalisation de la tâche de reconnaissance.

Les *"images-originales"* sont les images-cibles les plus rapidement et correctement identifiées (Tableau 5.1, Tableau 5.2). Cette supériorité est due au fait qu'elles contiennent plus d'informations pertinentes par rapport aux deux autres types d'images ("*image-contour*" et "*image-luminance*"). En effet, les *"images-originales"* possèdent presque tous les éléments nécessaires qui permettent d'extraire une carte de saillance complète et pertinente (Itti & Koch, 2000; Treisman, 1998; Treisman & Gelade, 1980; Wolfe, 1994). Cependant, il reste la possibilité, en dehors de notre cadre théorique, d'un mécanisme de comparaison "pixel" par "pixel" de l'image-cible et de l'image-test sachant qu'elles sont identiques. Mais un tel mécanisme demanderait une charge de mémoire considérable. D'autre part, compte tenu du temps de stockage très court de l'image iconique (250 ms, Sperling, 1960, cité par Lieury, 2005, Rossi, 2005), il est peu vraisemblable qu'un tel mécanisme de comparaison puisse intervenir.

Les *"images-contours"* et les "*images-luminance*", quant à elles, possèdent uniquement un seul type de propriété informationnelle. Les informations contenues dans ces images sont donc relativement moins pertinentes, et par conséquent, la carte de saillance est plus difficilement extraite.

Différents types de représentation sont donc construits selon la pertinence des cartes de saillance extraites. Les *"images-originales"* permettent d'extraire une carte de saillance plus complète contenant un plus grand nombre d'éléments sélectionnés (Itti, Koch, & Niebur, 1998; Itti & Koch, 2000; Itti, Gold, & Koch, 2001; Itti & Koch, 2001; Niebur, Itti, & Koch, 2002). Ce qui permet au système de traitement d'avoir une meilleure représentation de l'image-cible et de faciliter la réalisation de la tâche. Quant aux *"images-contours"* et aux "*images-luminance*", faute d'autres types d'informations complémentaires, l'extraction de la carte de saillance est difficile, ce qui ne permet pas au mécanisme de traitement de construire une meilleure représentation de l'image-cible.

3.1. Supériorité des *"images-contours"* lorsque l'image-cible et l'image-test sont identiques

Lorsque les deux images du couple sont identiques, nous observons une supériorité de la reconnaissance des *"images-contours"* présentées durant 150 ms (89%) et 300 ms (88%) par rapport à la reconnaissance des *"images-luminance"* exposées durant ces mêmes durées (74% et 74%, Figure 5.8). Cette facilité de reconnaissance met en évidence la pertinence des informations portées par les contours comparativement aux informations portées par la structuration des différentes de zones de luminance.

En effet, les contours permettent de clore et mieux identifier des figures (cf. la loi de clôture des Gestaltistes) ; et éventuellement à aboutir à une signification ou à une identification de scènes. Puis, ces images contours sont représentées en mémoire soit par leur identité ("représentation qualitative", par exemple, l'image est représentée par sa catégorie), soit par leur propriété physique ("représentation quantitative", par exemple, l'image est représentée par une carte de contour). La représentation des *"images-contours"* est donc caractérisée par une "double propriété" ou double codage (Lieury, 1995, Lieury & Calvez, 1986). Ces "double propriétés" pourraient activer le schéma de scène tout en activant la mémoire sémantique (Rossi, 2005). Ainsi, lorsque l'image-cible est une *"image-contour"*,

ces deux types de représentations (quantitative et qualitative) faciliteraient leur identification.

Les "*images-luminance*", quant à elles, ne montrent aucune information concernant l'identité des objets, elles permettent uniquement au mécanisme de traitement d'avoir une représentation quantitative. Faute d'informations portant sur l'identité de l'image-cible, les participants semblent avoir une grande difficulté à réaliser la tâche de reconnaissance.

Néanmoins, les performances dans les situations *d'"images-contours"* et *d'"images-luminance"* restent stables entre 150 ms et 300 ms (Figure 5.8). Le fait d'augmenter le temps de présentation n'améliore pas la reconnaissance des scènes. Le critère de sensibilité lors de la réalisation de la tâche est peu changé (Tableau 5.4, Tableau 5.5). Cette observation suggère que les informations disponibles sont extraites avant 150 ms et au-delà de ce temps, elles ne permettent plus au système de traitement de mieux identifier les images-cibles. Ainsi, la représentation de *l'"image-contour"* et de *l'"image-luminance"* est bien réalisée avant 150 ms, puis elle n'est plus consolidée en mémoire selon l'augmentation du temps de présentation. En revanche, la performance des *"images-originales"* s'améliore de 150 ms (92%) à 300 ms (95,3%) de présentation. Cette amélioration est due au fait que ces images fournissent plus d'informations et permettent donc de mieux les identifier grâce à l'allongement du temps de présentation.

Le taux de réponses correctes entre les *"images-contours"* (60%) et les *"images-luminance"* (59,2%) ne diffère pas significativement lorsque le temps de présentation est de 50 ms. Ces deux types d'images-cibles apparaissent difficiles au regard des réponses sensorielles (d'= 0,43 pour les "*images-contours*" et d'= à 0,74% pour les "*images-luminance*"). Ils fournissent peu d'informations permettant d'avoir une représentation pertinente.

Lorsqu'une image-contour est affichée pendant 50 ms, environ 50% des parvocellulaires (Schmolesky *et al.*, 1998, Figure 1.11) sont activées tout en traitant les hautes fréquences. Cependant, ces activités neuronales ne peuvent

pas être maintenues en mémoire du fait d'un effet de masquage. Par conséquent, ce traitement trop court portant sur les *"images-contours"* ne permet pas d'extraire une carte de saillance pertinente. Ces images ne peuvent donc pas être représentées qualitativement (sémantiquement). Les participants ont ainsi un comportement risqué et produisent donc beaucoup de fausses alarmes.

Pour les "*images-luminance*", avec ce même temps de présentation, toutes les magnocellulaires sont activées très précocement par rapport aux parvocellulaires (avec environ 20 ms d'avance, Nowak, Munk, Girard, & Bullier, 1995; Nowak & Bullier, 1997). Bien que ces activités neuronales soient maintenues relativement longtemps en mémoire, elles permettent difficilement, en revanche, de construire une représentation de l'image-cible faute d'informations portant sur son identité. Par conséquent, ce type de transformation apparaît difficile pour les participants qui ont un comportent prudent mais produisent beaucoup de fausses alarmes.

3.2. Supériorité des *"images-luminance"* lorsque l'image-cible et l'image-test sont différentes

Le comportement des participants est différent dans le cas où les deux images du couple sont constituées de deux scènes différentes.

Les *"images-luminance"* permettent d'avoir une meilleure performance (rejet correct) que les *"images-contours"* lorsque le temps de présentation est de 50 ms et 150 ms (Figure 5.13). Ce résultat montre que les participants rejettent mieux l'image-test lorsque l'image-cible est une "*image-luminance*" que dans le cas où l'image-cible est une "*image-contour*". Cette supériorité des *"images-luminance"* suggère que le mécanisme de traitement serait basé sur la luminance spatiale, et ce, lorsqu'il s'agit d'identifier les différences entre les deux images du couple. Ainsi, il ne compare que la structure globale des deux images du couple afin de distinguer leur différence. Les traitements magnocellulaires jouent alors un rôle important.

Quant aux *"images-contours"*, le mécanisme rencontrent des difficultés lors de la comparaison de la représentation de l'image-cible à celle de l'image-test. En effet, le traitement des hautes fréquences portant sur l'image-cible ne peut être activé que tardivement, et ensuite affaibli en raison de l'intervention du masque. La représentation non pertinente de l'image-cible construite à la suite d'un traitement de faible mobilisation est ainsi oubliée rapidement en mémoire (Wolfe, 1999). Par conséquent, ce mécanisme se trouve en difficulté pour rejeter l'image-cible en se basant sur les *"images-contours"*.

La performance concernant la reconnaissance des *"images-luminance"* diminue significativement selon un le temps de présentation allant de 150 ms à 300 ms (Figure 5.13). Cette observation pourrait s'expliquer par la pertinence des informations fournies par l'"*image-luminance*" diminuant suivant le temps de présentation. Les stratégies utilisées par les mécanismes de traitement seraient changées également. Le traitement se focaliserait *à priori* sur les informations issues de la vision fovéale qui sont moins structurales, et ne permettraient pas d'avoir une vision globale de la scène. C'est pourquoi les participants reconnaissent moins bien les *"images-luminance"* au delà de 150 ms de temps de présentation.

Les résultats obtenus dans cette expérience montrent que les participants ont une performance assez proche pour la reconnaissance des *"images-contours"* et des *"images-luminance"* présentées pendant 50 ms (Figure 5.13). De manière générale, leurs performances en reconnaissance s'améliorent à partir de 150 ms et restent stables entre 150 ms et 300 ms.

Cette observation nous conduit à la réflexion suivante :

- la tâche de reconnaissance visuelle étudiée dans l'expérience 2 implique peut-être l'utilisation d'un mécanisme de catégorisation. Par exemple, dans le cas où l'image-cible et l'image-test sont identiques, non seulement elles appartiennent à la même catégorie, mais surtout l'image-test est considérée comme l'image la plus "typique" (représentative) de l'image-cible (la reconnaissance est considérée comme le cas extrême de la catégorisation). Tandis que dans le cas où les deux

images sont différentes, l'image-test n'est plus considérée comme l'image typique de l'image-cible, cependant elle est toujours une image de la même catégorie que l'image-cible. Il paraît alors intéressant d'observer le comportement des participants face à des *"images-contours"* et des *"images-luminance"* dans une tâche de catégorisation générale.

- comme nous avons observé que la performance des participants varie peu à partir de 150 ms, nous focaliserons donc nos observations sur des temps de présentations de 50 ms et de 150 ms.

L'objectif de l'expérience 3 est donc de tester ces hypothèses et d'identifier les rôles joués par les contours et la luminance spatiale dans la catégorisation d'images.

Chapitre 6 Rôle des contours et de la structuration spatiale -- tâche de catégorisation (expérience 3)

L'objectif de cette expérience est de comprendre le mécanisme de traitement des contours et de la structuration spatiale des différentes zones de luminance dans la catégorisation de scène. Comme l'expérience est centrée sur le rôle de la catégorisation sémantique, aucune image-test n'est perceptivement identique. Afin d'égaliser les conditions, toutes les images-tests sont soit congruentes sur le plan catégoriel (sémantique) soit différentes.

Pour ce faire, nous avons construits des couples d'images, soit de même catégorie[7] (par exemple, "**image-cible** = montagne originale " *vs* "**image-test**= une autre image de montagne en niveau de gris[8]"), soit de catégorie différente (par exemple, "**image-cible** = plage-luminance " *vs* "**image-test** =une autre image en niveau de gris d'une catégorie différente". Par exemple, une image de mer, ou une image de montagne, etc.) (Figure 6.1).

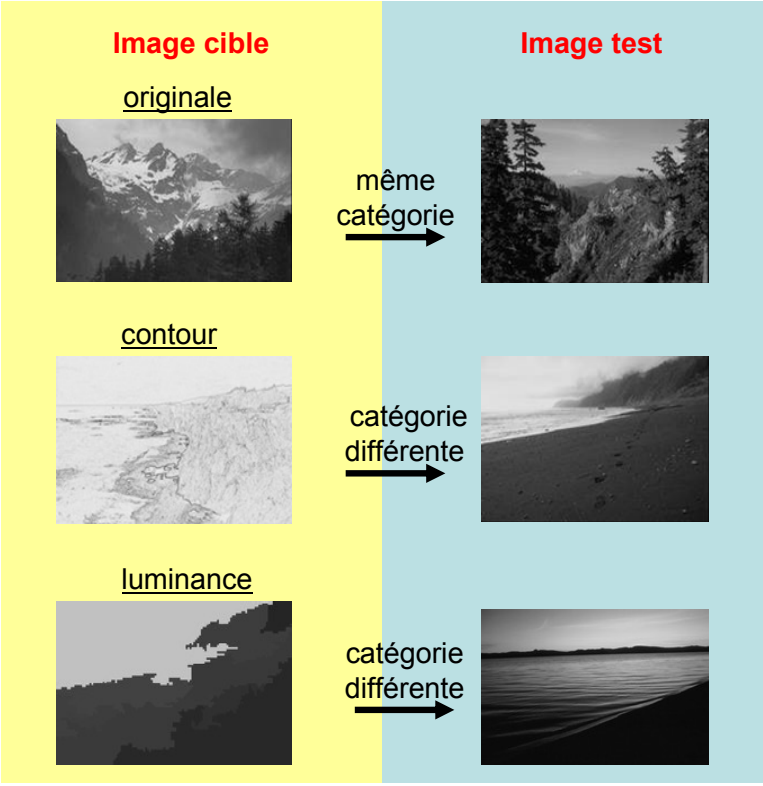


Figure 6.1. Illustration d'images-cibles et d'images-tests pour l'expérience 3.

[7] Comme dans l'expérience 2, ici, la catégorie est réservée uniquement au terme du contenu mais pas au type de transformation de l'image-cible, elle s'applique également pour les expériences 4 et 5.
[8] L'image-test est toujours une image originale en niveau de gris dans les expériences 2, 3, 4 et 5.

Dans cette expérience, nous nous sommes intéressés, au processus de traitement ascendant (*bottom-up*) portant sur les contours et la structuration spatiale des différentes zones de luminance dans une tâche de catégorisation de scènes visuelles. En d'autres termes, il s'agit de savoir, sans introduire préalablement les informations sémantiques concernant la catégorie de scène, si les participants sont capables de déterminer sa catégorie, et ce, en se basant sur des informations portant sur les objets qu'elle contient ou sur la structuration spatiale de luminance qu'elle possède. Par exemple, une image de plage est transformée en trois zones de luminance différentes (le sable, l'eau en bas, le ciel clair en haut ayant un degré de luminance différent). Cette image est présentée aux participants sans leur indiquer sa catégorie. Sont-ils capables de dire s'il s'agit d'une plage?

Afin de garantir le processus de traitement ascendant, il est nécessaire de minimiser le rôle d'un traitement sémantique des participants portant sur la catégorie de scènes. Il convenait alors de limiter au maximum l'activation de leurs connaissances préalables concernant les scènes présentées lors de cette expérience. Pour contrôler cet effet de pré-activation des connaissances sur la catégorie de scène, nous avons recouru à deux types de mesures.

Tout d'abord, nous ne présentions pas aux participants les noms des catégories de scènes utilisées dans l'expérience.

Ensuite, nous avons modifié la consigne en utilisant un terme neutre concernant la description de la tâche qu'ils devaient réaliser. Nous avons utilisé l'expression "*peut aller ensemble*" pour faire référence à l'idée de "catégorisation" d'images. D'après nous, cette expression renvoie à plusieurs caractéristiques, par exemple, le contenu, l'organisation des objets dans l'image, etc. tandis que la "catégorisation" risquerait d'activer plusieurs niveaux de représentation de haut niveau liée à la sémantique de l'image. Par exemple, les chiens et les chats intègrent la même super-catégorie "Animaux" (niveau supérieur), cependant ils appartiennent à deux catégories spécifiques différentes : catégorie "chien" et catégorie "chat" (niveau de base) (Rosch, 1977; Rossi, 2005).

1. Méthode

1.1. Participants

Trente étudiants (18 femmes et 12 hommes) inscrits en licence à l'Université de Rennes 2 ont participé à cette recherche. Tous les participants ont attesté d'une vue normale ou corrigée. Ils étaient tous ignorants des objectifs de cette recherche. Aucun d'entre eux n'a participé aux expériences 1 et 2.

1.2. Matériel

Cent quarante quatre couples d'images en niveaux de gris, sélectionnées dans la base de données COREL constituent des scènes naturelles appartenant aux catégories suivantes : mer, montagne, plage, désert et champ. Certaines photographies avaient été déjà utilisées dans les deux expériences précédentes.

1.3. Equipement

L'expérience se déroule dans la même salle que l'expérience 2 en utilisant les mêmes équipements.

1.4. Procédure expérimentale

La procédure utilisée est la même que dans l'expérience précédente. La seule différence est la modification de la consigne visant à préciser la tâche de la catégorisation.

> *"Vous devez d'abord observer attentivement une image de paysage appelée image-cible. Cette image-cible sera présentée sous des formes différentes : normale ou dégradée. Ensuite, cette image va disparaître, puis*

vous verrez apparaître une nouvelle image de paysage appelée image-test. L'image-cible et l'image-test sont toujours différentes. Dès que vous pensez que l'image-cible et l'image-test peuvent aller ensemble, appuyez le plus vite possible sur la touche "V" du clavier. Si au contraire, vous pensez que l'image-cible et l'image-test ne peuvent pas aller ensemble, appuyez le plus vite possible sur la touche "N" du clavier".

La durée de cette expérience était environ de quinze minutes.

1.5. Plan expérimental

Plan expérimental

S$_{30}$***T2*I3*C2**

Le facteur **S** correspond aux participants ; le facteur **T** correspond au temps de présentation de l'image-cible (50 ms *vs.*150 ms) ; le facteur **I** correspond au type de transformation de l'image-cible *("image-originale" vs. "image-contour" vs. "image-luminance"*) ; le facteur **C** correspond à la congruence d'images (même catégorie *vs.* catégorie différente).

Variables dépendantes

Les variables dépendantes sont les mêmes que dans l'expérience précédente.

2. Analyse des résultats

Des analyses statistiques similaires à celles de l'expérience 2 ont été réalisées.

Tout d'abord, nous pouvons remarquer que les participants ont un très bon taux global de réponse (79,5%) et un temps moyen de réaction de 1048 ms. Ils répondent en moyenne avec un temps similaire lorsque les images appartiennent à la même catégorie (1037 ms avec un taux de réponses correctes de 76,7%) et lorsqu'elles sont une catégorie différente (1059 ms avec un taux de réponses correctes de 82,3%).

Comme dans l'expérience précédente, nous commencerons par présenter les résultats correspondent au cas où les deux images du couple intègrent la même catégorie.

2.1. L'image-cible et l'image-test appartiennent à la même catégorie

Les résultats obtenus sont présentés dans le Tableau 6.1.

Tableau 6.1

Temps de réaction (ms) et taux de réponses correctes (%) en fonction du temps de présentation et du type de transformation de l'image-cible pour l'expérience 3. Erreurs type entre parenthèses.

	Temps de présentation		Transformation de l'image-cible		
	50 ms	150 ms	Originale	Contour	Luminance
Temps de réaction (ms)					
	1063 (43)	1012 (57)	934 (43)	1012 (54)	1166 (59)
Taux de réponses correctes (%)					
	77,1 (2,9)	76,3 (2,3)	88,2 (1,7)	83,4 (1,6)	59,1 (2,6)

2.1.1. Temps de présentation de l'image-cible

Nous pouvons remarquer que le temps de présentation de l'image-cible (50 ms et 150 ms) n'a pas d'effet significatif ni sur le nombre de bonnes réponses ($F<1$), ni sur le temps de réaction F (1,29) = 4.022, $p> .05$. En moyenne, les

participants répondent en 1063 ms avec un taux de bonnes réponses de 77,1% lorsque l'image-cible est présentée en 50 ms, et en 1012 ms avec 76,3% de bonnes réponses lorsqu'elle est présentée en 150 ms.

2.1.2. Type de transformation de l'image-cible

Les réponses varient significativement selon le type d'image-cible, *F* (2, 58) = 94.545, *P*< .001, de même pour le temps de réaction, *F* (2, 58) = 28.477, *p*< .001. Tout abord, les participants répondent significativement mieux pour les *"images-originales"* (88,2% de bonnes réponses) que pour les *"images-contours"* (83,4% de bonnes réponses), *F* (1, 29) = 38.918, *p*< .001, et que pour les "*images-luminance*" *F* (1, 29) = 54.331, *p*< .001 (Figure 6.2). Le taux de bonnes réponses des *"images-luminance"* est moins bon (59,1%) qu'avec les *"images-contours"*, *F* (1, 29) = 26.861, *P*< .001.

Les *"images-luminance"* sont moins rapidement reconnues (1166 ms) et que pour les *"images-contours"* (1012 ms), *F* (1, 29) = 11.761, *P*< .05, et que pour les *"images-originales", F* (1, 29) = 20.131, *P*< .001, (Figure 6.3).

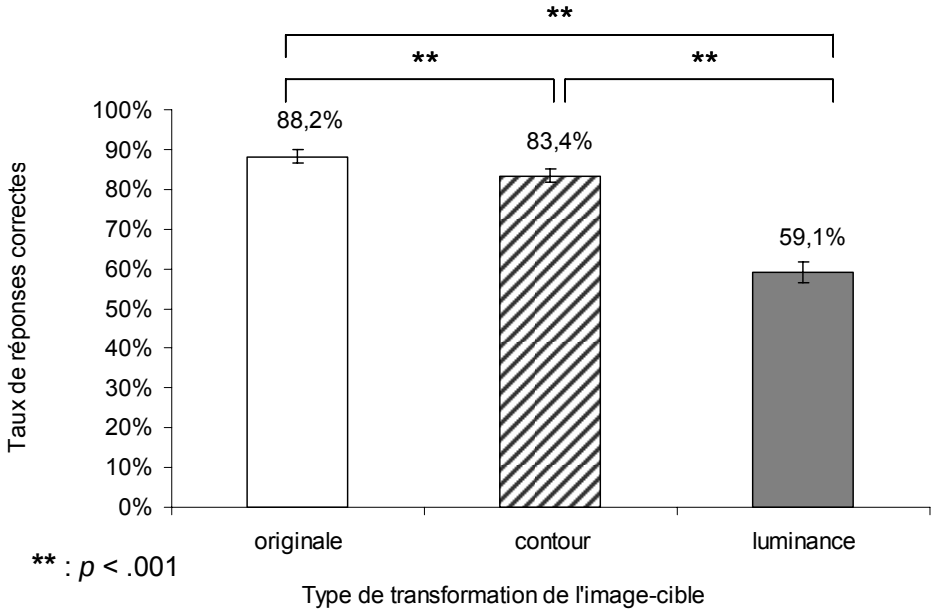


Figure 6.2. Taux de réponses correctes en fonction de type de transformation de l'image-cible.

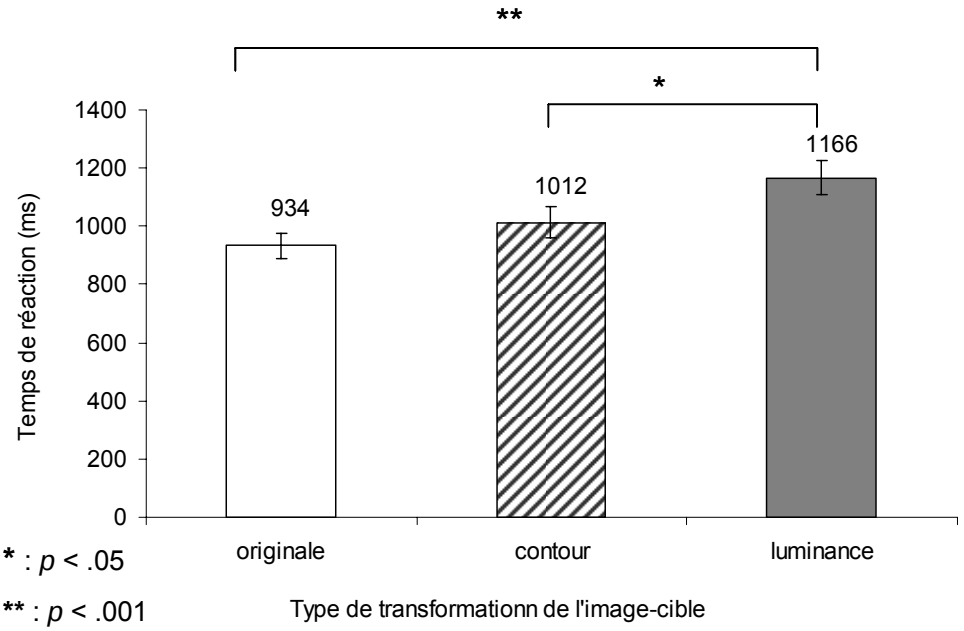


Figure 6.3. Temps de réaction en fonction du type de transformation de l'image-cible.

2.1.3. Catégorie d'images

Il est possible qu'il y ait des différences de réponses comportementales en fonction de la nature des catégories, par exemple entre montagne ou plage, etc. C'est pourquoi nous avons procédé à une analyse des catégories d'images à posteriori.

La catégorie de l'image influence les réponses, F (3, 87) = 13.165, p< .001. Nous avons ainsi les meilleures réponses pour les images de montagne (environ 85% de bonnes réponses). Nous avons ensuite des résultats semblables pour mer et plage et des résultats moins bons pour le désert (environ 70% de bonnes réponses). Les images de montagne obtiennent significativement plus de bonnes réponses que les images de désert, F (1, 29) = 167.196, P< .001.

Nous avons vu les effets directs des variables indépendantes. Nous allons maintenant étudier les interactions entre elles et quelles influences pourraient-elles exercer conjointement sur les réponses.

2.1.4. Interaction entre temps de présentation et type de transformation d'image

Aucune interaction significative n'est observée entre le temps de présentation et le type d'image-cible pour le temps de réaction (F<1) (Figure 6.4).

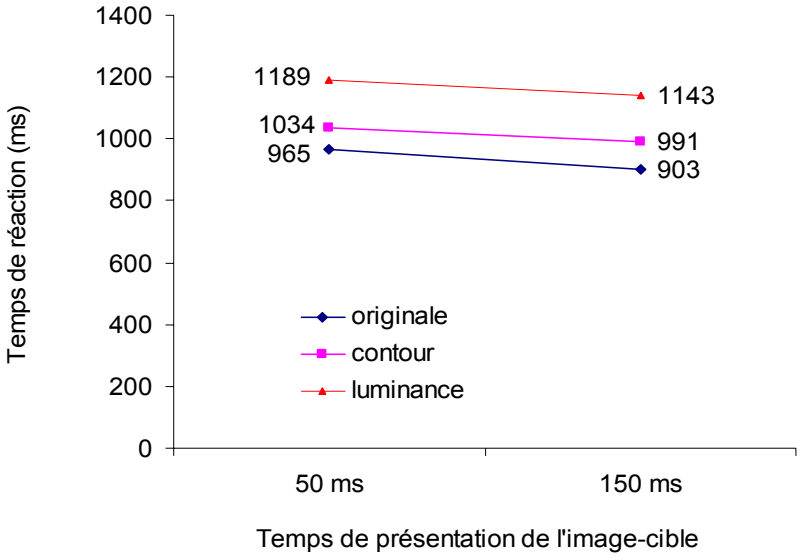


Figure 6.4. Temps de réaction en fonction du type de transformation de l'image-cible et du type de temps de présentation de cette image.

En revanche, nous obtenons une interaction significative entre ces deux variables pour les taux de réponses correctes, F (2, 58) = 8.931, p<.001 (Figure 6.5).

Nous remarquons une diminution significative, F (1, 29) = 20.535, P< .001 du nombre de bonnes réponses dans le cas d'une "*image-luminance*" avec un temps de présentation de 50 ms (63,9% de bonnes réponses). Cette diminution est encore plus accentuée dans le cas d'une "*image-luminance*" avec un temps de présentation de 150 ms (53,9% de bonnes réponses).

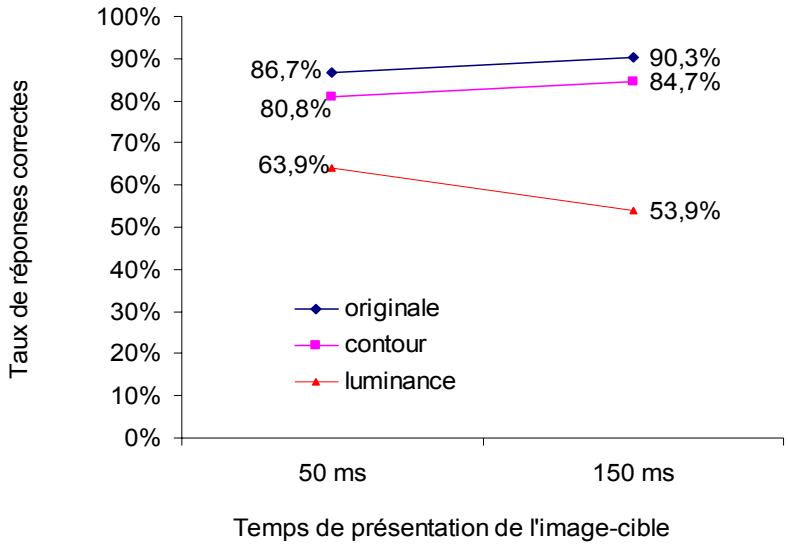


Figure 6.5. Taux de réponses correctes en fonction du type de transformation de l'image-cible et du type de temps de présentation de cette image.

2.2. L'image-cible et l'image-test sont de catégories différentes

Tableau 6.2

Temps de réaction de rejet (ms) et taux de rejets corrects (%) en fonction du temps de présentation et du type de transformation de l'image-cible pour l'expérience 3. Erreurs type entre parenthèses.

	Temps de présentation		Transformation de l'image-cible		
	50 ms	150 ms	Originale	Contour	Luminance
Temps de réaction de rejet (ms)					
	1103 (45)	1015 (51)	929 (43)	1079 (64)	1170 (71)
Taux de rejets corrects (%)					
	81,1 (2,6)	83,5 (1,9)	91,9 (1,1)	84,0 (1,8)	74,9 (2,3)

2.2.1. Temps de présentation

Lorsque le temps de présentation de l'image-cible augmente (passant de 50 ms à 150 ms), le nombre de bons rejets augmente légèrement de 81% à 83,5%, F (1,29) = 2.584, p> .05 et le temps de réaction de rejet diminue significativement de 1103 ms à 1015 ms, F (1, 29) = 19.292, P< .05.

2.2.2. Type de transformation de l'image-cible

Le type de transformation de l'image-cible a un effet significatif sur le temps de réaction de rejet, F (2, 58) = 18.148, p< .001. En effet, les *"images-originales"* induisent des réponses significativement plus rapides (929 ms) que les *"images-contours"* (1079 ms), F (1, 29) = 30.217, p< .001, et que les "*image-luminance*" (1170 ms), F (1, 29) = 43.833, p< .001.

Les participants mettent plus longtemps pour les "*images-contours*" que pour les "*image-luminance*", F (1, 29) = 39.241, P< .001 (Figure 6.6)

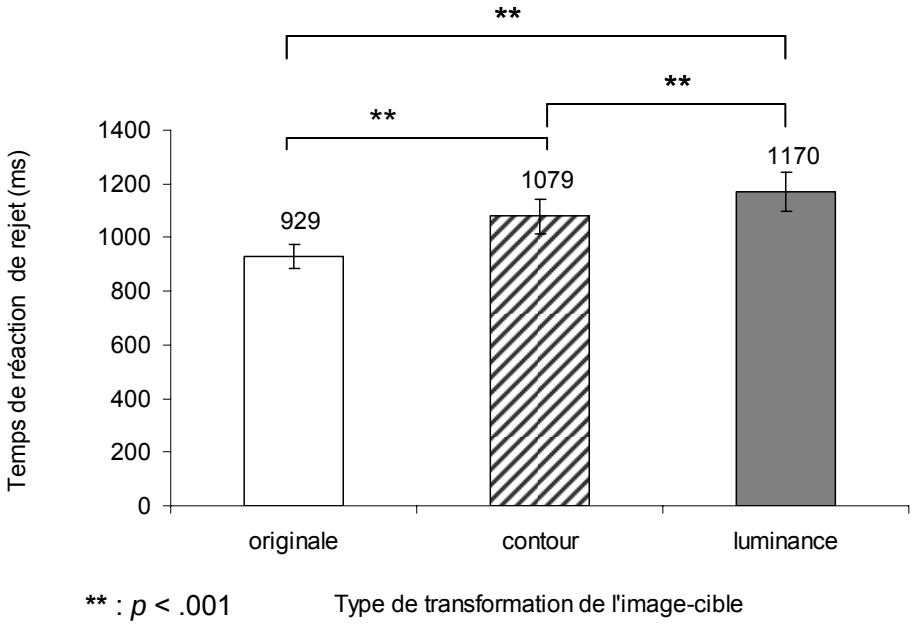


Figure 6.6. Temps de réaction de rejet en fonction du type de transformation de l'image-cible.

L'analyse indique aussi un effet significatif sur le taux de rejets corrects, F (2, 58) = 28.745, p< .001. La situation des *"images-originales"* a le meilleur taux de rejets corrects (92%) que la situation des *"images-contours"* (84%), F (1, 29) = 43,446, p< .001, et que la situation des "*images-luminance*" (74,9%), F (1, 29) = 61,334, p< .001.

Les "*images-luminance*" sont donc moins facilement rejetées que les *"images-contours"*, F (1, 29) = 26.254, P< .001 (Figure 6.7).

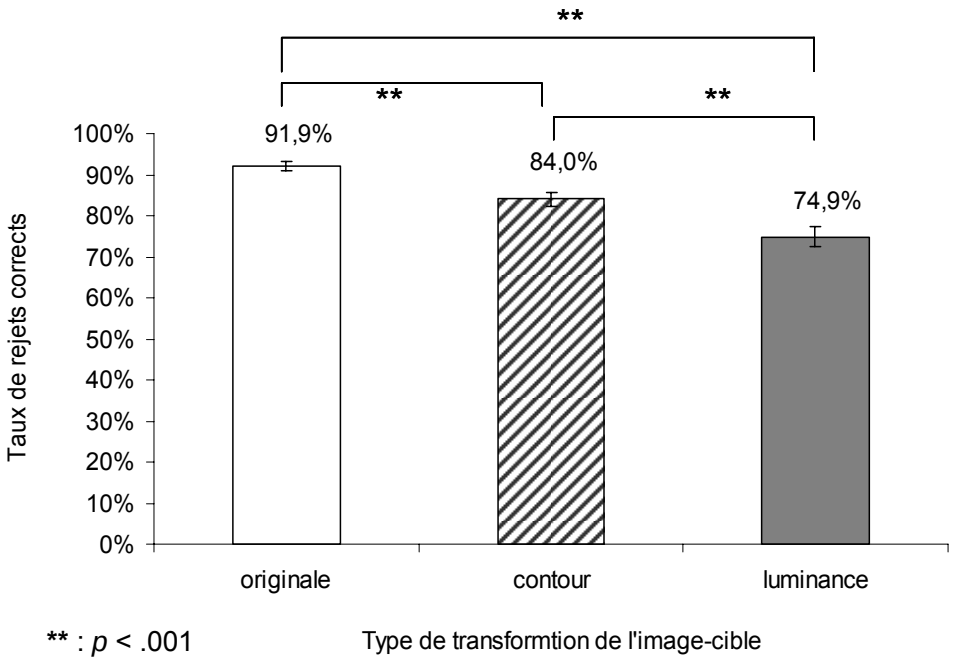


Figure 6.7. Taux de rejets corrects en fonction du type de transformation de l'image-cible.

2.2.3. Catégorie d'images

La catégorie d'images n'a pas d'effet significatif sur le temps de réaction de rejet, F (3, 87) = 1.826, p> .05. Cependant, elle influence le nombre de bons rejets, F (3, 87) = 6.601, p< .05.

2.2.4. Interaction entre temps de présentation et type de transformation d'image

L'interaction entre le temps de présentation et le type d'image-cible pour le temps de réaction de rejet n'est pas significative (F<1) (Figure 6.8).

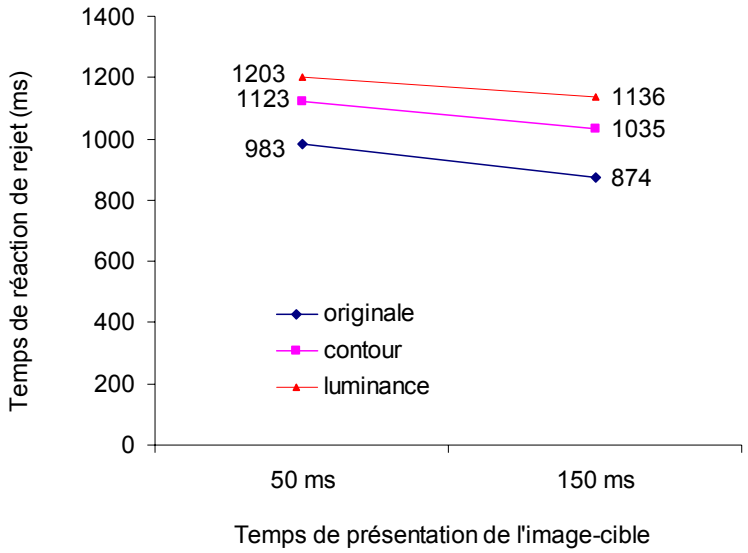


Figure 6.8. Temps de réaction de rejet en fonction du type de transformation de l'image-cible et du type de temps de présentation de cette image.

En revanche, nous obtenons une interaction significative entre ces deux variables pour les taux de rejets corrects, F (2, 58) = 5,916, p<.05. La performance des participants pour les *"images-contours"* s'améliore en fonction du temps de présentation F (1,29) = 7.353, p< .05 (Figure 6.9).

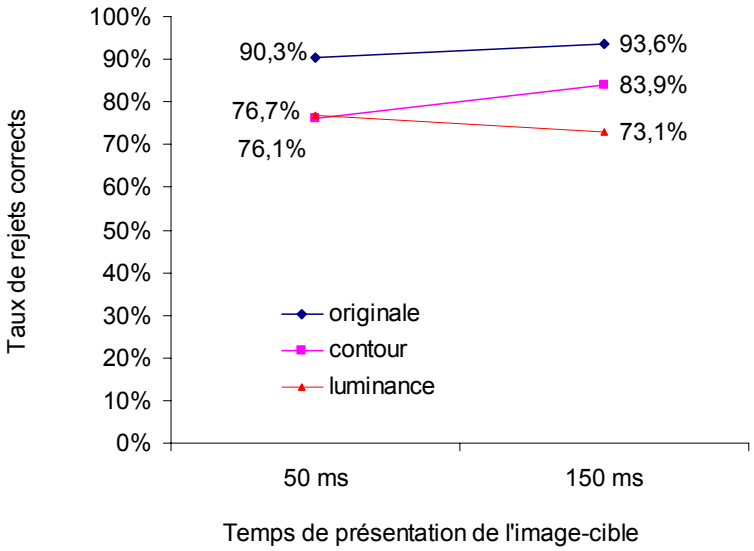


Figure 6.9. Taux de rejets corrects en fonction du type de transformation de l'image-cible et du type de temps de présentation de cette image.

2.3. Stratégies utilisées par les participants

2.3.1. Le temps de présentation est de 50 ms

Un résumé des réponses des participants en fonction des quatre situations de la théorie TDS est présenté par le tableau ci-dessous (Tableau 6.3).

Tableau 6.3

Réponses des participants selon le type de transformation de l'image-cible pour une présentation de 50 ms en intégrant la théorie TDS

		Originale	
		même image	image différente
Réponse	oui	DC 86,7%	FA 9,7%
	non	O 13,3%	RC 90,3%
		1	1
		d' = 2,441	
		β = 1,1677	

		Contour	
		même image	image différente
Réponse	oui	DC 80,8%	FA 23,9%
	non	O 19,2%	RC 76,1%
		1	1
		d' = 1,581	
		β = 0,815	

		Luminance	
		même image	image différente
Réponse	oui	DC 63,9%	FA 23,3%
	non	O 36,1%	RC 76,7%
		1	1
		d' = 1,085	
		β = 2,049	

DC = "Détection Correcte" ; O = "Omission" ; FA = "Fausse Alarme" ; RC = "Rejet Correct"
d' = critère de sensibilité ; β = critère décisionnel du sujet

La représentation graphique suivante illustre les résultats du tableau ci-dessus (Figure 6.10).

Concernant les *"images-contours"* et les "*images-luminance*", les participants font un nombre de fausses alarmes similaire (proche de 23,6%). Cependant, les *"images-luminance"* apparaissent plus difficiles à identifier (63,9% de DC) que les "*images-contours*" (80,8%). Les *"images-originales"* sont les plus faciles à identifier (d'=2,441, 86,7% de DC).

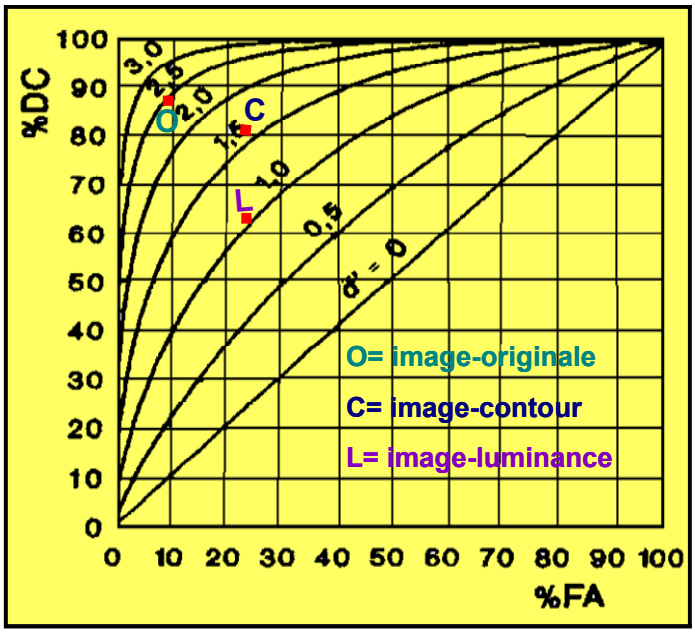


Figure 6.10. Stratégies utilisées par les participants pour identifier les images-cibles présentées en 50 ms.

2.3.2. Le temps de présentation est de 150 ms

Le tableau ci-dessous résume les réponses des participants en fonction des quatre situations de la théorie TDS (Tableau 6.4).

Tableau 6.4

Réponses des participants selon le type de transformation de l'image-cible pour une présentation de 150 ms en intégrant la théorie TDS

		Originale	
		même image	image différente
Réponse	oui	DC 90,3%	FA 6,4%
	non	O 9,7%	RC 93,6%
		1	1

d' = 2,821
β = 1,1718

		Contour	
		même image	image différente
Réponse	oui	DC 84,7%	FA 16,1%
	non	O 15,3%	RC 83,9%
		1	1

d' = 2,014
β = 0,9675

		Luminance	
		même image	image différente
Réponse	oui	DC 53,9%	FA 26,9%
	non	O 46,1%	RC 73,1%
		1	1

d' = 0,714
β = 6,2896

DC = "Détection Correcte" ; O = "Omission" ; FA = "Fausse Alarme" ; RC = "Rejet Correct"
d' = critère de sensibilité ; β = critère décisionnel du sujet

La représentation graphique suivante illustre les résultats du tableau ci-dessus (Figure 6.11).

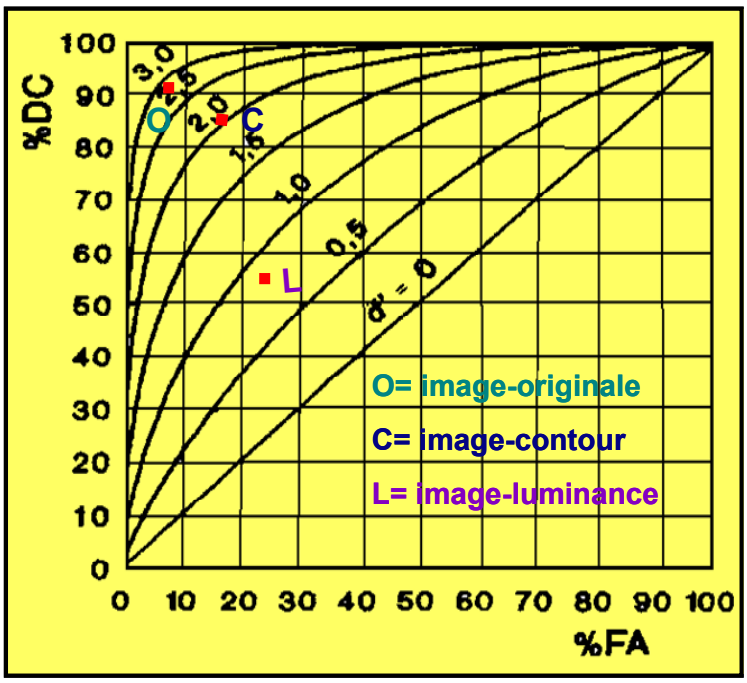


Figure 6.11. Stratégies utilisées par les participants pour identifier les images-cibles présentées en 150 ms.

Lorsque le temps de présentation augmente à 150 ms, l'identification des *"images-originales"* (d'=2,821) et des *"images-contours"* (d'=2,014) apparaît relativement plus facile. Les participants ont une stratégie plutôt neutre (β est proche de 1. Ils produisent moins de fausses alarmes (6,4% pour les "*images-originales*", 16,1% pour les "*images-contours*").

Quant aux "*images-luminance*", elles deviennent de plus en plus difficiles à identifier (d'=0,714, 53,9% de DC), les participants utilisant une stratégie prudente que nous pouvons apparenter au hasard.

3. Discussion

Cette expérience consistait à étudier le rôle des contours et de la structuration spatiale de luminance dans la catégorisation de scènes.

Les résultats concernant les taux de réponses correctes (DC) sont très nets. La performance pour les "*images-contours*" (83,4%) est très proche de celle des "*images-originales*" (88,2%) tandis que la performance liée aux "*images-luminance*" (59,1%) s'apparente au hasard. Les contours semblent donc absolument décisifs dans la catégorisation sémantique, ce qui rejoint une longue tradition dans les théories de la perception, de la loi de clôture des Gestaltistes jusqu'aux primitives de Biederman (1987) (voir aussi Gaillard, Boulliou & Gautier, 1996). De même, dans le domaine de la mémoire, celle des images est aussi performante lorsque les dessins sont des simples contours que lorsqu'ils sont en couleurs et détaillés. Les simples contours permettent donc une identification sémantique (Lieury, 2005).

En revanche, pour les réponses concernant les rejets, les "*images-luminance*" sont un peu plus performantes (74,9%). Ce qui nous laisse penser que la surface ou le gradient de luminance sont des indices qui, à des niveaux perceptifs, peuvent conduire à détecter des différences d'images, en l'occurrence à rejeter correctement les images-tests lorsqu'elles sont différentes des images-cibles.

3.1. Les deux images du couple sont de même catégorie

3.1.1. Supériorité des "images-contours"

Nous observons une supériorité des *"images-contours"* par rapport aux *"images-luminance"* quelque soit le temps de présentation (Tableau 6.1, Tableau 6.2). Ce résultat suggère que les contours sont des informations porteuses de la

catégorisation de l'image. Ils permettent d'avoir un accès rapide au sens général de la scène.

Les contours permettent de clore et de mieux identifier des figures (cf. la loi de clôture des Gestaltistes). L'identification d'un ou plusieurs objets typiques active la catégorisation de scènes (Friedman, 1979). Or, les *"images-contours"* sont stockées en mémoire par double codage (Paivio et Lieury, cité par Lieury, 2005; Lieury, 1995, Lieury & Calvez, 1986), ce qui faciliterait l'activation du schéma de scène en mémoire sémantique (Rossi, 2005).

D'un point de vue, neuropsychologique, l'activité neuronale du traitement pourrait s'effectuer par la manière suivante. En 50 ms de temps de présentation, le sens général de l'image semble être extrait en se basant sur des activités neuronales d'environ 50% de parvocellulaires (Schmolesky *et al*., 1998, Figure 1.11). Bien que ces activités soient faibles, les informations extraites par des contours apparaissent pertinentes pour identifier la catégorisation de cette scène.

Quand le temps augmente de 50 ms à 150 ms, le temps de traitement augmente tout en maintenant une activité neuronale relativement durable. Les parvocellulaires traitent donc plus d'informations portant sur les contours, ce qui permettrait une meilleure catégorisation.

3.1.2. Paradoxe des "images-luminance" en fonction du temps de présentation

Paradoxalement, lorsque le temps de présentation passe à 150 ms, les "*images-luminance*" sont moins performantes qu'à 50 ms. Comment expliquer ce paradoxe puisqu'en général, la performance est meilleure avec l'allongement du temps de présentation? L'explication ne semble pas pouvoir venir des stratégies d'exploration oculaire puisqu'une simple fixation demande en moyenne 250 ms.

L'explication que nous pourrions apporter suppose que les informations vision-fovéales joueraient un rôle central pour l'identification de la scène. En effet,

lorsque le temps de présentation est de 50 ms, les magnocellulaires sont sollicités pour traiter les informations extra-fovéales. Ce traitement fabriquerait une vision globale de l'image-cible, ce qui pourrait donner le sens général de scène. Ainsi, ce temps de présentation permettrait d'élaborer une représentation structurale suffisamment précise pour permettre soit une catégorisation sémantique, soit même une représentation perceptive suffisante pour conduire à une décision d'identification. En revanche, lorsque le temps de présentation est de 150 ms, les informations dans le champ fovéal deviennent plus saillantes que celles relatives au champ périphérique. Cependant, les informations dans le champ fovéal seraient beaucoup moins structurées comparées aux informations globales, ce qui ne permettrait pas d'avoir une meilleure vision structurale de cette image. En conséquence, la construction de la représentation visuospatiale en mémoire, basée sur les informations prélevées dans le champ fovéal, serait encore plus difficile pour identifier les "*images-luminance*" (Lieury, 2005, Lorant-Royer & Lieury, 2003).

3.2. Les deux images du couple sont de catégories différentes

Lorsque les deux images du couple sont de catégories différentes, le comportement des participants diffère en fonction du type de transformation de l'image.

En considérant le temps de présentation de 50 ms, les *"images-luminance"* (76,7%) ont une performance similaire à celles des *"images-contours"* (76,1%). Donc, nous pouvons en déduire que la représentation perceptive de la structuration spatiale semblerait être considérée comme un critère pertinent pour rejeter l'image-test, comparé aux informations issues des contours. Cependant, faute d'informations disponibles portant sur l'identité des "*images-luminance*", la performance des participants ne s'améliore pas avec l'allongement du temps de présentation (73,1% en 150 ms de présentation). En revanche, et contrairement, la performance des participants avec les *"images-contours"* s'améliore (83,9% en 150 ms de temps de présentation). Cet effet bénéfique dû au fait que

l'augmentation du temps permettrait de clore et de mieux identifier des figures (cf. la loi de clôture des Gestaltistes), et éventuellement à aboutir à une catégorisation de scènes.

Cette expérience a donnée des résultats similaires pour les *"images-contours"* et les *"images-luminance"* comparés à ceux de l'expérience 2. Ces résultats suggèrent que le mécanisme de perception utilisé est similaire dans ces deux expériences. Cependant, les *"images-contours"* sont mieux reconnues dans l'expérience 3 que dans l'expérience 2. La différence des performances est liée aux différents mécanismes de traitement correspondant aux caractéristiques de la tâche à effectuer.

En effet, dans l'expérience 2, le mécanisme de traitement devait identifier à la fois le sens général de l'image et la localisation de tous les éléments de détails qu'elle contenait, puisque c'était une tâche de reconnaissance. Donc, la voie ventrale et la voie dorsale étaient fortement mobilisées afin de traiter les informations disponibles.

Dans l'expérience 3, le mécanisme de traitement a besoin d'identifier le sens général de l'image. Il ne rendrait pas compte la localisation des éléments puisque c'est une tâche de catégorisation. Dans ce cas, la voie ventrale semble être plus fortement mobilisée que la voie dorsale, ces deux voies visuelles ne seraient donc pas en compétition. Par conséquent, une meilleure coordination des voies visuelles est offerte dans le cas de catégorisation par rapport au cas de reconnaissance. C'est la raison pour laquelle les participants répondent mieux dans l'expérience 3.

Nous venons de voir l'effet des contours et de la structuration spatiale de luminance dans la catégorisation d'images en 50 ms et 150 ms. Le fait que les *"images-luminance"* présentées en 50 ms (63,9%) soient mieux reconnues qu'en 150 ms (53,9%), tiendrait à ce que, dans le premier cas (50 ms), elles offriraient une meilleure représentation perceptive de l'ensemble de cette image. La représentation grossière de la scène visuelle construite à partir de ces informations magnocellulaires précoces pourrait être très souvent suffisante pour

déclencher une réponse correcte dans la tâche de catégorisation utilisée ici. Et ce, sans que l'analyse plus fine apportée par les informations parvocellulaires ne soit nécessaire.

Il nous faudra regarder dans une prochaine expérience en considérant une tâche de catégorisation dans laquelle le temps de présentation devient très bref, par exemple 20 ms et 35 ms, si les participants sont capables de catégoriser les images en se basant sur les informations de la structuration spatiale de la luminance.

Le niveau de catégorisation à partir duquel cette tâche deviendra impossible nous donnera des indications sur la finesse des traitements précoces basée notamment sur des informations magnocellulaires.

Chapitre 7 Rôle des contours et de la structuration -- tâche de catégorisation précoce (expérience 4)

L'objectif de l'expérience 4 consiste à étudier le mécanisme précoce de catégorisation de scènes basé sur les contours et la structuration spatiale de luminance. Nous cherchons à comprendre comment le système visuel extrait ces deux types d'information lorsque le temps de présentation devient ultra-court. Nous définissons lors de cette étude, deux types de temps de présentation de l'image-cible :

- soit, 20 ms de temps de présentation, les magnocellulaires viennent de commencer à être activées, mais pas les parvocellulaires (Schmolesky *et al*., 1998).

- soit, 35 ms de temps de présentation, environ 80% des magnocellulaires et 10% des parvocellulaires sont activées (Schmolesky *et al*., 1998).

A quel point les *"images-contours"* et les "*images-luminance*" permettent d'effectuer une tâche de catégorisation ?

1. Méthode

1.1. Participants

Trente-deux étudiants (20 femmes et 12 hommes) inscrits en licence à l'Université de Rennes 2 ont participé à cette recherche. Tous les participants ont une vue normale ou corrigée. Ils étaient tous ignorants des objectifs de cette recherche. Aucun d'entre eux n'a participé aux expériences 1, 2 et 3.

1.2. Matériel

Le matériel utilisé est le même que l'expérience 3.

1.3. Equipement

L'expérience se déroule dans la même salle que l'expérience précédente et avec les mêmes équipements.

1.4. Procédure et plan expérimental

La procédure est identique à celle de l'expérience précédente.

Plan expérimental

S$_{32}$*T2*I3*C2

Le facteur **S** correspond aux participants ; le facteur **T** correspond au temps de présentation de l'image-cible (20 ms *vs.* 35 ms) ; le facteur **I** correspond au type de transformation de l'image-cible (*"image-originale" vs. "image-contour" vs. "image-luminance"*) ; le facteur **C** correspond à la congruence d'images (même catégorie *vs.* catégorie différente).

Variables dépendantes

Les variables dépendantes sont les mêmes que dans l'expérience précédente.

2. Analyse des résultats

Les mêmes analyses que celles effectuées au cours de l'expérience 3 ont été utilisées.

Le taux global de réponse est assez bon (68,3%) et le temps de réaction moyen est de 1352 ms environ. Les participants répondent en moyenne en 1380 ms avec un taux de bonnes réponses de 60,9% quand les images du couple sont de la même catégorie, et en 1324 ms avec 75,8% de bonnes réponses lorsqu'elles sont de catégories différentes.

Comme dans l'expérience précédente, nous présentons les résultats en deux cas séparés. Le premier cas est celui où les deux images du couple sont de la même catégorie. L'autre cas est celui où les deux images du couple sont de catégories différentes.

2.1. Les images du couple sont de même catégorie

Les sujets répondent en moyenne en 1380 ms et identifient correctement dans 60.9% des cas. Nous allons maintenant étudier les facteurs qui peuvent expliquer ces réponses. Les résultats obtenus sont résumés dans le Tableau 7.1.

Tableau 7.1

Temps de réaction (ms) et taux de réponses correctes (%) en fonction du temps de présentation et du type de transformation de l'image-cible pour l'expérience 4. Erreurs type entre parenthèses.

	Temps de présentation		Transformation de l'image-cible		
	20 ms	35 ms	Originale	Contour	Luminance
Temps de réaction (ms)					
	1385 (63)	1374 (72)	1255 (45)	1376 (56)	1509 (62)
Taux de réponses correctes (%)					
	60,8 (3,3)	60,9 (2,7)	72,4 (3,6)	59,9 (3,7)	50.4 (3,4)

2.1.1. Temps de présentation de l'image-cible

Nous pouvons remarquer que le temps de présentation de l'image-cible n'a pas d'effet significatif sur le nombre de bonnes réponses, F (1, 31) = 1.172, p> .05 ou sur le temps de réaction (F<1). En effet, les résultats obtenus varient peu suivant le temps de présentation de l'image-cible (1385 ms de temps de réaction et 60,8% de bonnes réponses pour 20 ms de présentation ; 1374 ms et 60,9% de bonnes réponses pour 35 ms de temps de présentation)

2.1.2. Type de transformation de l'image-cible

Le type de transformation de l'image-cible influence significativement les temps de réaction, *F* (2, 62) = 10.421, *P*< .001. Les participants répondent plus rapidement pour les *"images-originales"* (1255 ms) que pour les *"images-contours"* (1376 ms), *F* (1, 31) = 14.273, *p*< .001, et que pour les *"images-luminance"* (1509 ms), *F* (1, 31) = 30.681, *p*< .001. Parmi ces dernières, il existe une différence significative au niveau du temps de réaction, *F* (1, 31) = 31.203, *P*< .05 (Figure 7.1).

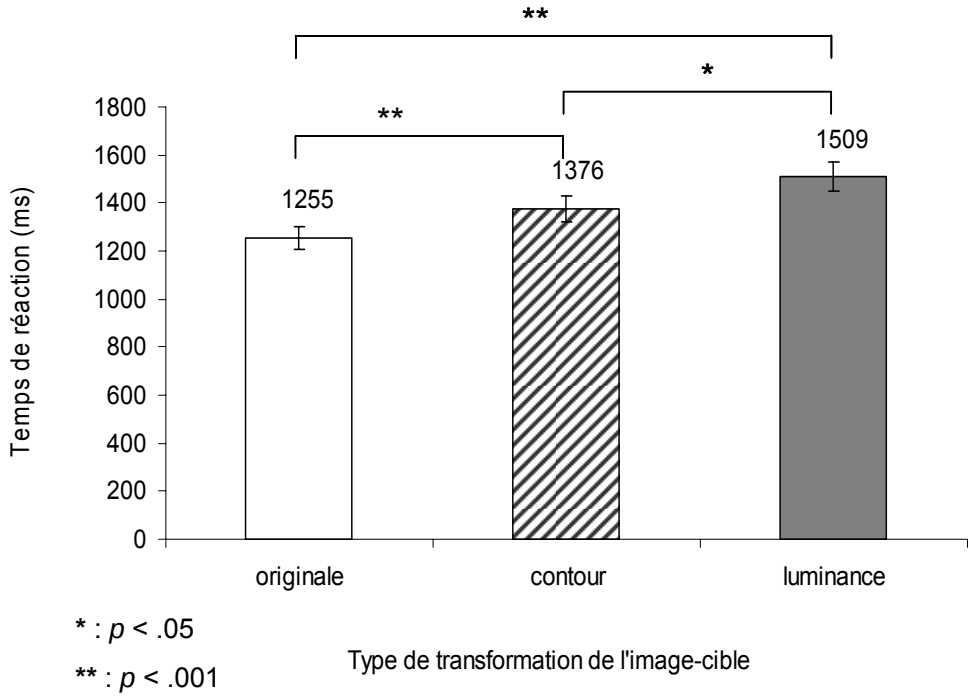


Figure 7.1. Temps de réaction en fonction du type de transformation de l'image-cible.

L'analyse indique également un effet significatif pour les taux de réponses correctes, *F* (2, 62) = 17.568, *P*< .001. Les participants répondent mieux pour les *"images-originales"* (72,4%) que pour les *"images-contours"* (59.9%), *F* (1, 31) = 100.977, *P*< .001, de même pour les "*images-luminance*" (50,3%) *F* (1, 31) = 137.331, *P*< .001. Les *"images-luminance"* sont moins bien reconnues que les *"images-contours"*, *F* (1, 31) = 23.771, *p*< .001 (Figure 7.2).

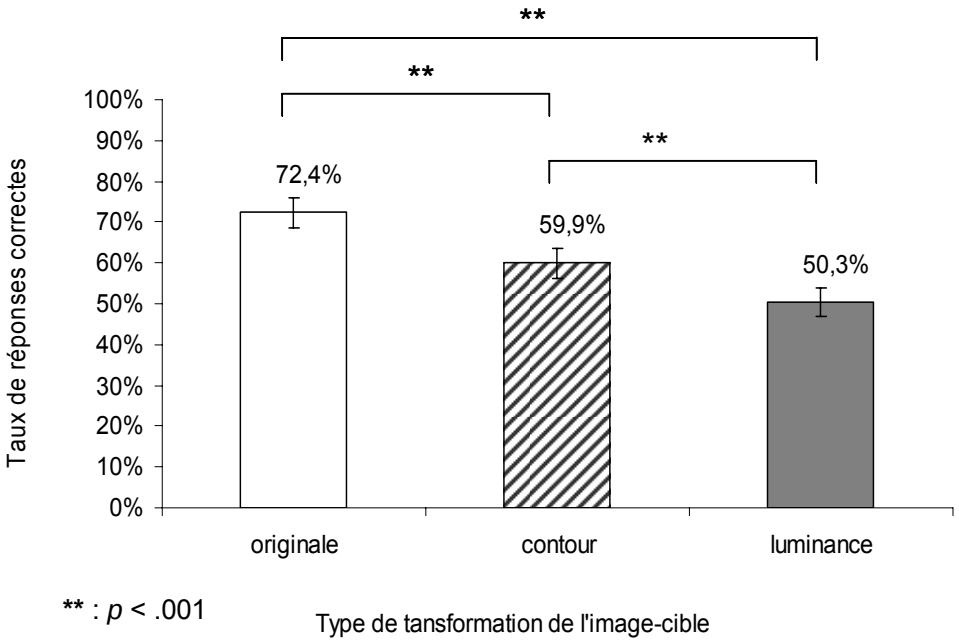


Figure 7.2. Taux de réponses correctes en fonction du type de transformation de l'image-cible.

2.1.3. Catégorie d'images

Comme dans l'expérience précédente, l'analyse effectuée pour la variable "catégorie d'images" est réalisée selon le plan séparé. La catégorie d'images influence significativement les réponses, F (3, 93) = 23.266, P< .001. Les images de montagne obtiennent plus de bonnes réponses que les images de mer, F (1, 31) = 41.650, P< .001), qui elles-mêmes obtiennent plus de bonnes réponses que les images de plage et de désert, F (1, 31) = 154.048, P< .001.

Le temps de réaction est influencé par la catégorie d'images, F (3, 93) = 7.970, p< .05. Les participants mettent significativement moins de temps à répondre pour les images de montagne que pour les images de plage ou de désert.

2.1.4. Interaction entre temps de présentation et type de transformation d'image

L'analyse indique une interaction significative des facteurs "temps de présentation" et "type de transformation d'image-cible" sur le temps de réaction, *F* (2, 62) = 5.546, *p*< .05. Les participants répondent plus rapidement pour les "*images-contours*" en fonction du temps de présentation (1436 ms et 1316 ms pour un temps de présentation de 20 et 35 ms respectivement), *F* (1,31) = 10.447, *P*< .05. Une tendance inverse est observée pour les "*images-luminance*". Les participants répondent moins rapidement (1443 ms et 1575 ms pour un temps de présentation de 20 et 35 ms respectivement), *F* (1,31) = 8.407, *P*< .05 (Figure 7.3).

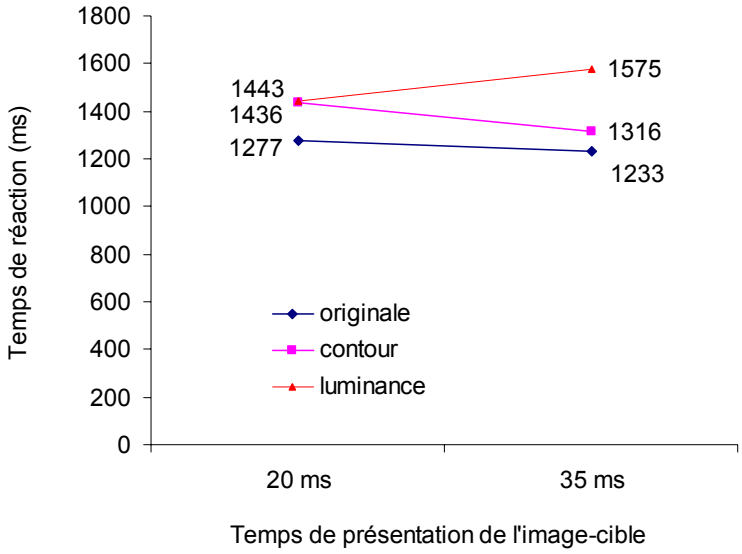


Figure 7.3. Temps de réaction en fonction du type de transformation de l'image-cible et du temps de présentation de cette image

L'analyse indique une interaction significative de ces deux variables sur le taux de réponses correctes, *F* (2, 62) = 13.275, *p*< .001. En 35 ms de temps de présentation, les taux de bonnes réponses pour les *"images-originales"* et les *"images-contours"* ont été respectivement de 75,3% et 63,8%. En 20 ms de temps de présentation, ces taux ont été de 69,5% et 56% (Figure 7.4). En revanche, nous remarquons une diminution significative, *F* (1,31) = 16.353, *P*< .05 du

nombre de bonnes réponses dans le cas d'une "*image-luminance*" lorsque le temps de présentation passe de 20 ms (56.8%) à 35 ms (43.8%).

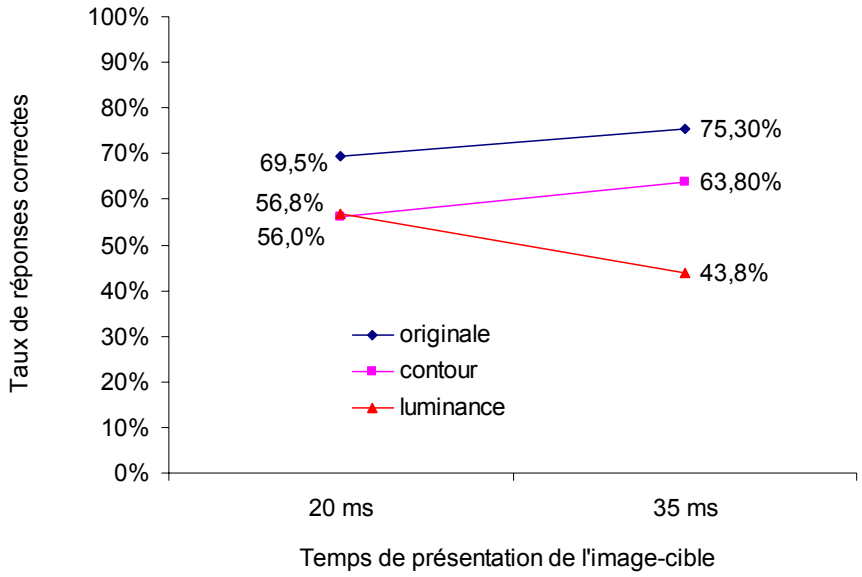


Figure 7.4. Taux de réponses correctes en fonction du type de transformation de l'image-cible et du temps de présentation de cette image

2.2. Les images du couple sont de catégories différentes

Il n'y a pas de différence significative pour le temps de réaction de rejet que ce soit des images du couple de même catégorie ou de catégories différentes ($F<1$). Le temps de réaction de rejet, en moyenne, a été respectivement de 1380 ms et 1324 ms.

En revanche, les participants donnent plus souvent de bons rejets pour les images de catégories différentes (75,8%) que pour les images de même catégorie (60,9%). Ils sont donc meilleurs pour rejeter les couples d'images de catégories différentes que pour reconnaître les couples d'images de même catégorie (détection correcte). Les résultats obtenus sont résumés dans le Tableau 7.2.

Tableau 7.2

Temps de réaction de rejet (ms) et taux de rejets corrects (%) en fonction du temps de présentation et du type de transformation de l'image-cible pour l'expérience 4. Erreurs type entre parenthèses.

	Temps de présentation		Transformation de l'image-cible		
	20 ms	35 ms	Originale	Contour	Luminance
Temps de réaction de rejet (ms)					
	1325 (58)	1323 (53)	1242 (67)	1308 (63)	1422 (93)
Taux de rejets corrects (%)					
	74,7 (3,9)	76,9 (3,3)	82,4 (3,8)	73,3 (2,3)	71,7 (2,1)

2.2.1. Temps de présentation

Le temps de présentation de l'image-cible n'a pas d'effet significatif sur les réponses que ce soit pour le temps de réaction de rejet (F<1) ou pour le taux de rejets corrects, F (1, 31) = 2.709, p> .05. En effet, les résultats obtenus ne varient pas significativement suivant le temps de présentation (1325 ms de temps de réaction de rejet et 74,7% de bons rejets en 20 ms de temps de présentation, 1323 ms et 76,9% en 35 ms de temps de présentation).

2.2.2. Type de transformation de l'image-cible

Le type de transformation de l'image-cible a un effet significatif sur le temps de réaction de rejet, F (2, 62) = 5.712, p< .05. Les participants mettent significativement plus de temps à répondre pour les *"images-luminance"* (1422 ms) que pour les *"images-contours"* (1308 ms), F (1, 31) = 8.408, p< .001, et que pour les "*images-originales"* (1242 ms)*,* F (1, 31) = 14.953, p< .001 (Figure 7.5).

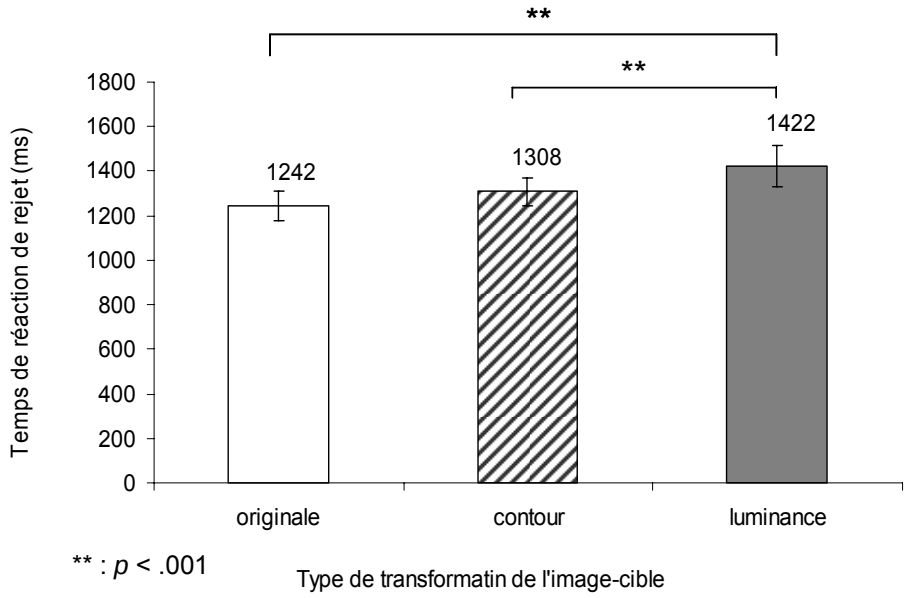


Figure 7.5. Temps de réaction de rejet en fonction du type de transformation de l'image-cible.

Le type de transformation de l'image-cible a un effet sur les taux de rejets corrects, F (2, 62) = 3.991, p< .05. Les *"images-originales"* obtiennent plus de bons rejets (82,4%) que les *"images-contours"* (73,3%), F (1, 31) = 21.331, P< .001, et que les *"images-luminance"* (71,7%), F (1, 31) = 6.532, P< .05 (Figure 7.6).

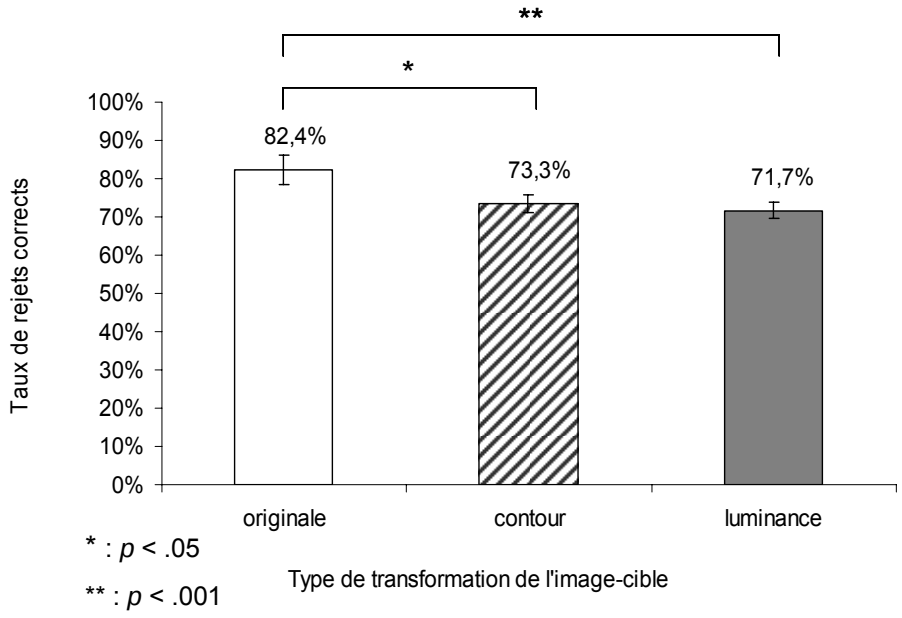


Figure 7.6. Taux de rejets corrects en fonction du type de transformation de l'image-cible.

2.2.3. Catégorie d'images

La catégorie d'images n'a pas d'effet significatif sur le temps de réaction de rejet (F <1). Les participants répondent en moyenne avec un temps proche de 1350 ms quelle que soit la catégorie d'images.

La catégorie de l'image influence significativement le nombre de bonnes réponses (F (3, 93) = 18.671, P< .001). Les images de montagne ont 84,2% de taux de rejets corrects, 79,2% pour les images de mer, 73,4% pour les images de plage et 66,5% pour les images de désert.

2.2.4. Interaction entre temps de présentation et type de transformation de l'image

L'analyse n'indique pas d'interaction significative entre le facteur "temps de présentation" et le "type de transformation d'image-cible" sur le temps de rejet (F <1). (Figure 7.7).

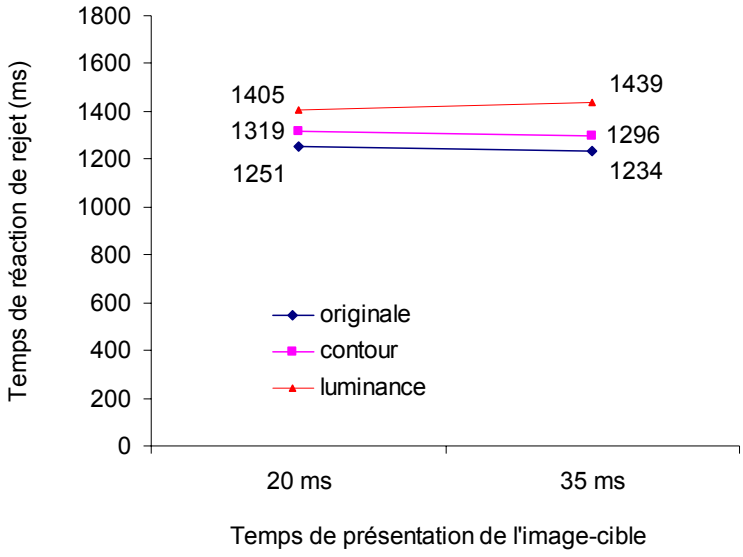


Figure 7.7. Temps de réaction de rejet en fonction du type de transformation de l'image-cible et du temps de présentation de cette image

L'analyse indique une interaction significative de ces deux variables sur le taux de rejets corrects, F (2, 62) = 3.175, p< .05. Le taux de rejets corrects s'améliore en fonction du temps de présentation pour les *"images-contours"* (69,8% et 76,8% pour un temps de présentation de 20 et 35 respectivement) F (1,31) = 6.551, P< .05 (Figure 7.8).

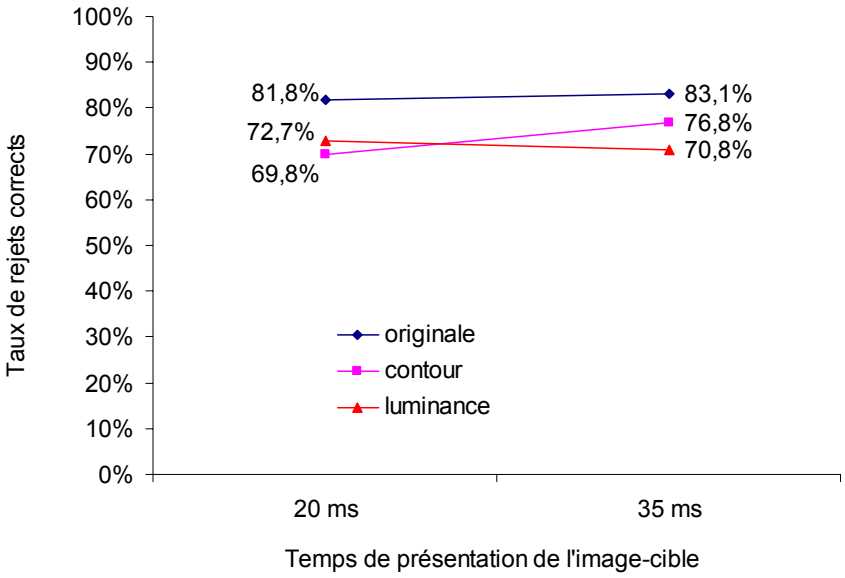


Figure 7.8. Taux de rejets corrects en fonction du type de transformation de l'image-cible et du temps de présentation de cette image

2.3. Stratégies utilisées par les participants

2.3.1. Le temps de présentation est de 20 ms

Les réponses des participants en fonction des quatre situations de la théorie TDS pour les image-cibles présentées en 20 ms sont résumées dans le Tableau 7.3.

Tableau 7.3

Réponses des participants selon le type de transformation de l'image-cible pour une présentation de 20 ms en intégrant la théorie TDS

		Originale	
		même image	image différente
Réponse	oui	DC 69,5%	FA 18,2%
	non	O 30,5%	RC 81,8%
		1	1
		d' = 1,418	
		β = 1,7797	

		Contour	
		même image	image différente
Réponse	oui	DC 56%	FA 30,2%
	non	O 44%	RC 69,8%
		1	1
		d' = 0,671	
		β = 3,4355	

		Luminance	
		même image	image différente
Réponse	oui	DC 56,8%	FA 27,3%
	non	O 43,2%	RC 72,7%
		1	1
		d' = 0,775	
		β = 3,5429	

DC = "Détection Correcte" ; O = "Omission" ; FA = "Fausse Alarme" ; RC = "Rejet Correct"
D' = critère de sensibilité ; β = critère décisionnel du sujet

La représentation graphique suivante illustre les résultats du tableau ci-dessus (Figure 7.9).

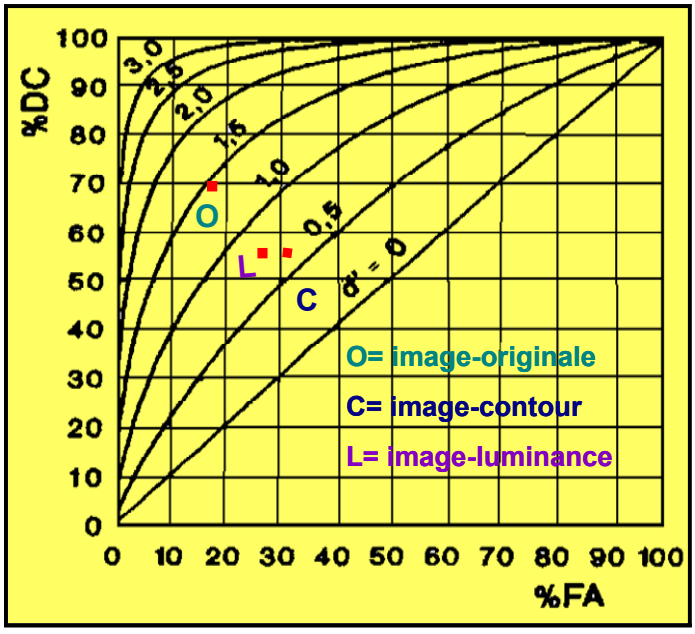


Figure 7.9. Stratégies utilisées par les participants pour identifier les images-cibles présentées en 20 ms.

Les *"images-contours"* (d' = 0,671, β = 3,4355) et les *"images-luminance"* (d' = 0, 0,775, β = 3,5429) déterminent les mêmes comportements, elles sont donc toutes plus difficiles à identifier que les "*images-originales*" (d' = 1,418, β = 1,7797). Les participants restent prudents avec beaucoup de fausses alarmes, qui sont respectivement de 30,2% et 27,3%. Quant aux *"images-originales"*, elles sont donc relativement faciles à identifier. Les participants font moins de fausses alarmes (18,2%).

2.3.2. Le temps de présentation est de 35 ms

Les réponses des participants en fonction des quatre situations de la théorie TDS pour les image-cibles présentées en 35 ms sont résumées dans le Tableau 7.4.

Tableau 7.4

Réponses des participants selon le type de transformation de l'image-cible pour une présentation de 35 ms en intégrant la théorie TDS

		Originale	
		même image	image différente
Réponse	oui	DC 75,3%	FA 16,9%
	non	O 24,7%	RC 83,1%
		1	1

d' = 1,642
β = 1,4008

		Contour	
		même image	image différente
Réponse	oui	DC 63,8%	FA 23,2%
	non	O 36,2%	RC 76,8%
		1	1

d' = 1,085
β = 2,0737

		Luminance	
		même image	image différente
Réponse	oui	DC 43,8%	FA 29,2%
	non	O 56,2%	RC 70,8%
		1	1

d' = 0,392
β = -3,509

DC = "Détection Correcte" ; O = "Omission" ; FA = "Fausse Alarme" ; RC = "Rejet Correct"
D' = critère de sensibilité ; β = critère décisionnel du sujet

La représentation graphique suivante illustre les résultats du tableau ci-dessus (Figure 7.10).

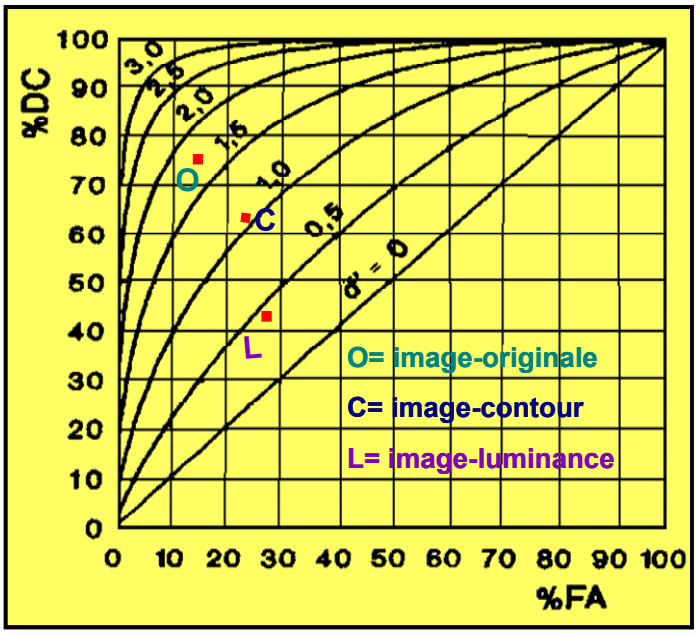


Figure 7.10. Stratégies utilisées par les participants pour identifier les images-cibles présentées en 50 ms.

Comparées aux images présentées en 20 ms, les participants réagissent de la même façon pour les *"images-originales"* lorsque le temps de présentation est de 35 ms. Les *"images-contours"* deviennent en revanche relativement plus faciles à identifier (d' = 1,085, β = 2,0737). Les participants ont un comportement moins prudent et produisent moins de fausses alarmes (23,2%) qu'en 20 ms (30,2%). Quant aux "*images-luminance*", elles sont les plus difficiles à identifier, les participants utilisent la stratégie la plus risquée avec beaucoup de fausses alarmes (29,2%) et beaucoup d'omissions (56,2%).

3. Comparaison avec l'expérience 3

L'expérience précédente était presque identique à celle-ci, la seule différence étant les temps de présentation. En effet, pour l'expérience 4 que nous venons de voir, les temps de présentation étaient de 20 ou 35 ms alors que précédemment, pour l'expérience 3, les temps étaient de 50 ou 150 ms. Nous verrons quelles modifications cela a apporté aux résultats. Nous comparons les résultats concernant les *"images-contours"* et les *"images-luminance"*.

3.1. Moins bons résultats pour l'expérience 4

Ce qui ressort le plus entre les expériences 3 et 4 est le fait que les participants répondent moins bien lors de l'expérience 4 qu'ils ne l'avaient fait lors de l'expérience 3.

En effet, ils mettent dans l'expérience 4 en moyenne 1352 ms pour répondre alors qu'avant (dans l'expérience 3), ils en mettaient 1048 ms (Figure 7.11).

De même, les participants donnent moins souvent la bonne réponse (dans 68,3% des cas) que dans l'expérience 3 (79,5% de bonnes réponses lors de l'expérience précédente) (Figure 7.12). Ce phénomène se retrouve globalement, pour les images de même catégorie et pour les images de catégories différentes.

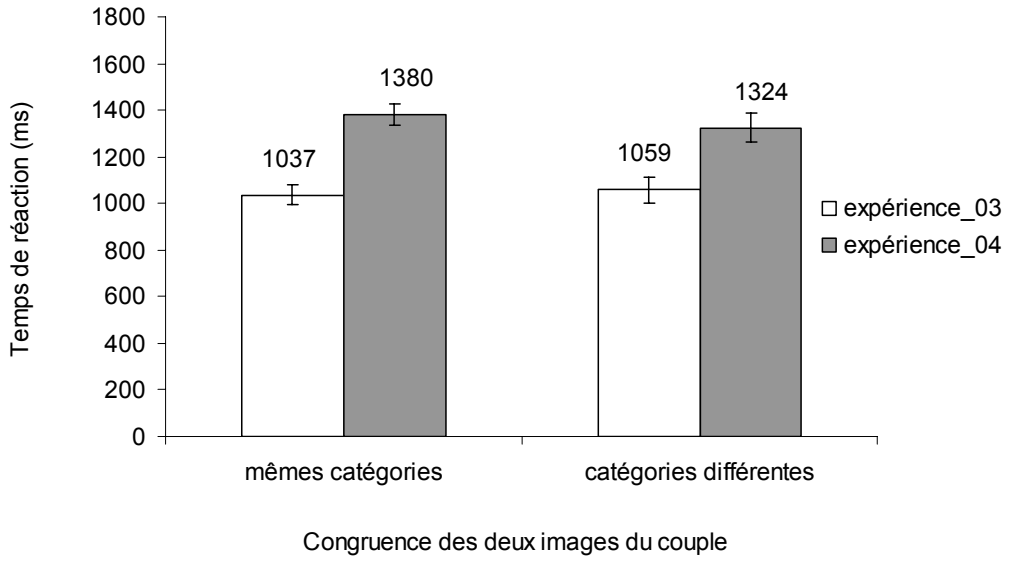


Figure 7.11. Temps de réaction pour les "*images-contours*" et les "*images-luminance"* selon la congruence entre l'image-cible et l'image-test.

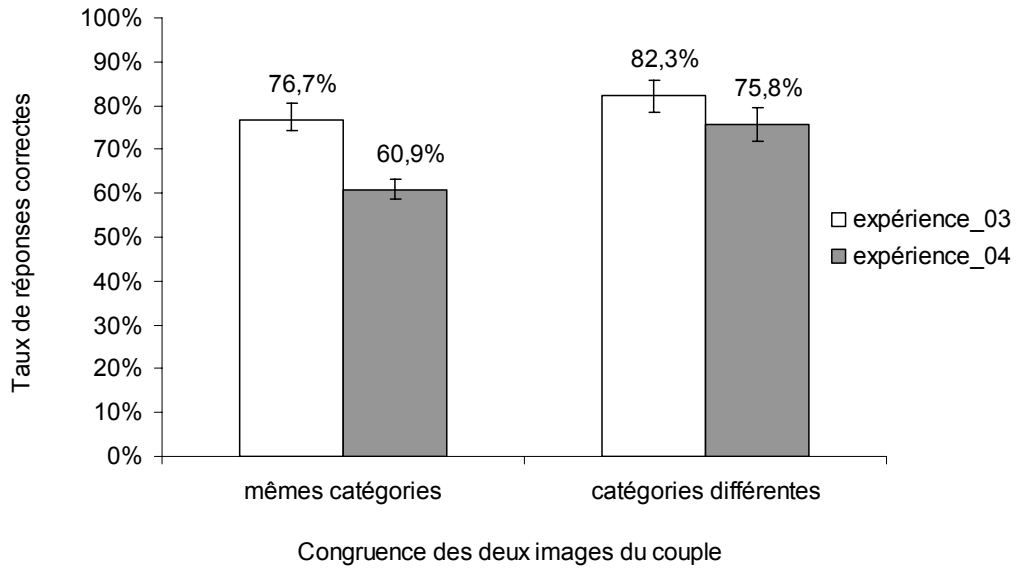


Figure 7.12. Taux de réponses correctes pour les "*images-contours*" et les "*images-luminance*" selon la congruence entre l'image-cible et l'image-test.

La première hypothèse explicative de ces résultats est que le temps de traitement de l'image-cible devient plus court. Cependant, le temps de réaction et le taux de bonnes réponses à 50 ms devraient être plus proches de ceux de 35 ms que de ceux de 150 ms.

Or, le graphique ci-dessous nous montre que ce n'est pas le cas (Figure 7.13, Figure 7.14). Nous constatons sur les deux figures une différence importante pour des taux de réponses correctes entre les expériences 3 et 4, puis pour chaque expérience, une petite variation des résultats due au temps de présentation.

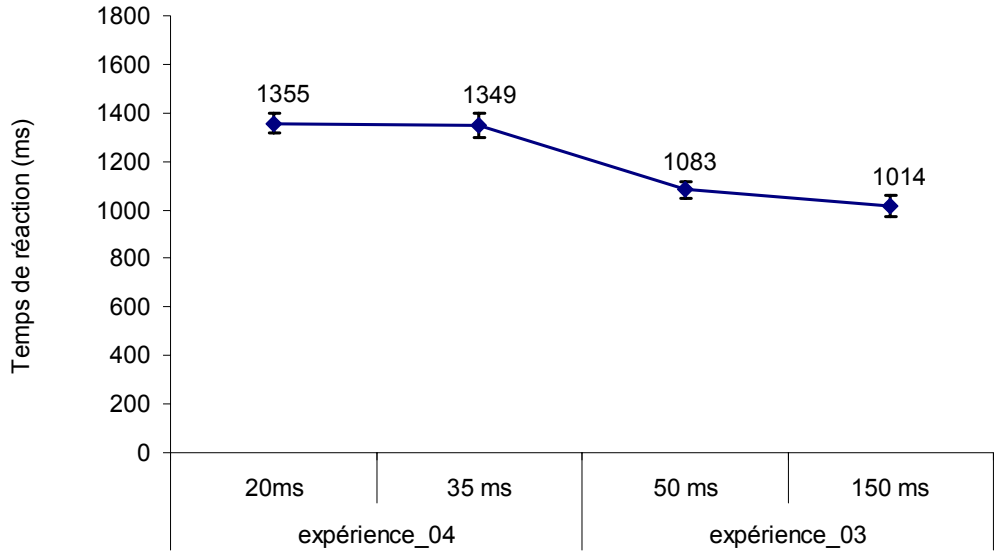


Figure 7.13. Evolution du temps de réaction pour les *"images-contours"* et les *"images-luminance"* dans les expériences 4 et 3.

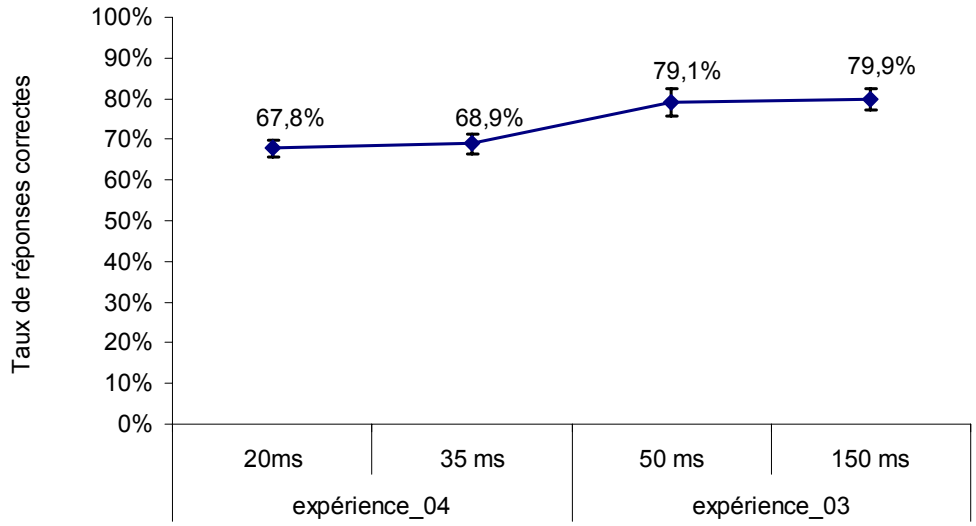


Figure 7.14. Evolution du temps de réaction pour les *"images-contours"* et les *"images-luminance"* dans les expériences 4 et 3.

Les participants répondent plus lentement dans l'expérience 4 que dans l'expérience 3. Ceci suggère que le mécanisme d'identification mis en jeu dans ces deux expériences est différent.

3.2. Confirmation de l'influence des facteurs

Dans la plupart des cas, les facteurs ou les interactions produisent un effet significatif sur les variables dépendantes. Ces résultats confirment les effets observés dans l'expérience précédente. Pour les deux dernières expériences, nous observons, par exemple, de meilleurs résultats pour les *"images-originales"*, puis pour les *"images-contours"* et pour les "*images-luminance*". L'effet du type d'image est donc le même. Les interactions ne sont pas toutes significatives pour les couples d'images de même catégorie et de catégories différentes.

Nous remarquons un effet "intéressant" : dans le cas de couples d'images de même catégorie, le nombre de bonnes réponses des *"images-luminance"* diminue significativement quand le temps de présentation des images-cibles augmente. Cet effet ressortait déjà lors de l'expérience précédente. Dans le cas d'"*images-luminance*", le mécanisme de traitement se trouve en difficulté selon l'allongement du temps de présentation. Le taux de bonnes réponses passe de 56,8% pour le temps de présentation de 20 ms, à 43,8% pour le temps de présentation de 35 ms.

Nous pouvons aussi noter que dans le cas d'images de même catégorie, comme dans le cas d'images de catégories différentes, nous observons un effet significatif des variables, aussi bien sur les taux de bonnes réponses que sur les temps de réaction.Enfin, la catégorie influence aussi les réponses. Nous avons de meilleurs résultats pour les images de montagne que pour les images de mer, de plage et de désert.

Nous pouvons supposer que les images de montagne sont plus reconnaissables, que les images de désert, de plage et de mer peuvent se ressembler. Que le temps de présentation soit de 20 ms ou de 35 ms ne semble pas modifier cet effet de catégorie.

4. Discussion

Cette expérience visait à étudier le mécanisme précoce portant sur les contours et la structuration spatiale de luminance dans la catégorisation précoce de scènes complexes.

Le premier constat contribue à une confirmation. Les *"images-originales"* sont mieux reconnues que les *"images-contours"* et les "*images-luminance*" (Tableau 7.1, Tableau 7.2). Bien que le temps de présentation soit extrêmement court (20 ms et 35 ms), les informations disponibles portées par les *"images-originales"* sont jugées "suffisantes" pour identifier leur catégorie. Le mécanisme de catégorisation semblerait basé sur les traitements préattentifs. Ces traitements seraient alloués à la structure spatiale ou à la localisation des traits les plus saillants de la scène. De plus, la vision fovéale devrait être extrêmement importante pour identifier la catégorie de l'image. La vision extra-fovéale, quant à elle activée très faiblement, donnerait des informations supplémentaires.

Les *"images-contours"* et les *"images-luminance"* ont une performance similaire que ce soit dans le cas de deux images de même catégorie (56% et 56,8%) ou de catégories différentes (69,8% et 72,7%). Cette observation suggère que ces deux types d'images fournissent un niveau de pertinence des informations équivalent. Chacun d'entre eux est jugé moins pertinent pour identifier leur catégorie. Lorsque les *"images-contours"* sont présentées en 20 ms, l'intensité du stimulus est tellement faible qu'un pourcentage quasi nul des parvocellulaires est sollicité (Schmolesky *et al.*, 1998). Ces traitements ne permettraient donc pas de percevoir les contours, de les clore ainsi que de les former en objet (loi de clôture des Gestaltistes). Ainsi, les participants ont des difficultés à identifier ces images et font beaucoup de fausses alarmes (30,2%).

Pour les "*images-luminance*", les magnocellulaires sont activées lorsque le temps de présentation est de 20 ms. Ces traitements, sembleraient porter sur la structuration spatiale de l'image. Ils seraient insuffisants pour identifier leur catégorie, et ce, faute d'informations supplémentaires. Ainsi, ces images ne

pourraient pas être représentées en mémoire soit visuo-spatialement soit sémantiquement (Lieury, 1995, Lieury & Calvez, 1986; Rossi, 2005). Par conséquent, les participants font également beaucoup de fausses alarmes lors de la réalisation de la tâche (27,3%).

Lorsque les deux images sont de même catégorie, nous n'observons pas d'amélioration significative pour les *"images-originales"* et les *"images-contours"* en fonction du temps de présentation. Cela suggère que l'intervalle de 15 ms ne constituerait pas un écart suffisant pour identifier une scène complexe. Cependant, Concernant les "*images-luminance*", une diminution de la performance est observée en fonction du temps de présentation (de 56,8% à 43,8%). Cette tendance a déjà été présentée dans l'expérience 3 (de 63,9% à 53,9% pour le temps de présentation respectivement de 50 ms et 150 ms). Cette diminution des réponses correctes pourrait être liée aux facteurs provoquant la diminution de la pertinence des informations. Cet effet serait lié au matériel utilisé durant l'expérience. Comme les *"images-luminance"* possèdent des contours irréguliers de transitions entre les différentes zones de luminance, ces contours irréguliers pourraient engendrer une perturbation sémantique car ce ne sont pas des contours d'objets tels que nous les rencontrons quotidiennement. Ainsi, certains traitements se focaliseraient sur ces contours, ce qui provoquerait une perturbation décisionnelle de la catégorisation de scène, même si le temps de présentation est très court.

Il serait intéressant de tester l'effet de ces contours irréguliers dans la catégorisation d'images. D'une part, ce travail permettrait de comparer les résultats de la nouvelle expérience avec ceux des expériences précédentes. D'autre part, il permettrait de savoir à quel point l'intervention sémantique modifie le mécanisme de catégorisation.

L'objectif de l'expérience 5 est de tester ces deux types de questionnement.

Chapitre 8 Effet du lissage des contours des zones de luminance -- tâche de catégorisation (expérience 5)

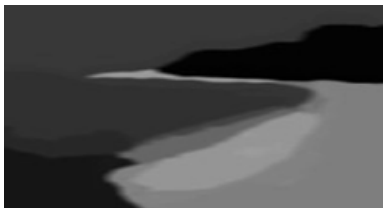
Dans les expériences 3 et 4, lorsque l'image-cible est une "*image-luminanc*e", la performance des participants diminue significativement en fonction du temps de présentation. Cette diminution de performance serait liée aux perturbations décisionnelles issues des traitements sur les contours irréguliers entre les différentes zones de luminance.

Le premier objectif de cette expérience consiste à tester cette hypothèse. Pour ce faire, nous avons lissé ces contours avec l'aide du logiciel Photoshop 7.0 afin que les transitions entre différentes zones de luminance soient progressives (Figure 8.1).

(a)
Image luminance sans lissage pour une montagne

(b)
Image luminance avec lissage pour la montagne de l'image (a)

(c)
Image luminance sans lissage pour une plage



(d)
Image luminance avec lissage pour la plage de l'image (c)

Figure 8.1. Illustration des *"images-luminance sans lissage "* et des "*images-luminance avec lissage*".

Le deuxième objectif est de tenter de répliquer des résultats concernant les *"images-luminance sans lissage"* utilisées dans les expériences 3 et 4. Ainsi, nous définissons trois types de temps de présentation de l'image-cible en 35 ms (utilisé dans l'expérience 4), 50 ms et 150 ms (utilisés dans l'expérience 3).

Après avoir réalisé une comparaison des résultats pour les trois expériences précédentes, nous décidons d'utiliser simplement deux types de catégories de scènes comme image-cible : la catégorie "montagne" et la catégorie "plage". En effet, les scènes de montagne et les scènes de plage ont obtenu des performances stables dans les trois expériences précédentes. Par conséquent, nous les considérons comme les images les plus représentatives pour réaliser l'expérience 5, les deux autres catégories d'images (mer et désert) ne sont pas prises en compte.

1. Méthode

1.1. Participants

Soixante étudiants (47 femmes et 13 hommes) inscrits en licence à l'Université de Rennes 2 ont participé à cette recherche. Tous les participants ont une vue normale ou corrigée, et étaient ignorants des objectifs de cette recherche. Aucun d'entre eux n'a participé aux expériences 1, 2, 3 et 4.

1.2. Matériel

Soixante-quatre couples d'images sont sélectionnés à partir du matériel utilisé dans l'expérience 4. Elles sont composées de deux catégories différentes (32 couples d'image de plage et 32 couples d'image de montagne).

1.3. Equipement

L'expérience se déroule dans la même salle expérimentale que l'expérience précédente et avec les mêmes équipements.

1.4. Procédure et plan expérimental

La procédure est identique à celle de l'expérience précédente.

Plan expérimental

S$_{20}$[T3]*I4*C2

Le facteur **S** correspond aux participants ; le facteur **T** correspond au temps de présentation de l'image-cible (35 ms *vs.* 50 ms *vs.* 150 ms) ; le facteur **I** correspond au type de transformation de l'image-cible (*"image-originale" vs. "image-contour" vs. "image-luminance sans lissage" vs. "image-luminance avec lissage"*) ; le facteur **C** correspond à la congruence d'images (même catégorie *vs.* catégorie différente).

Variables dépendantes

Les variables dépendantes sont les mêmes que dans l'expérience précédente.

2. Analyse des résultats

2.1. Les deux images du couple sont de même catégorie

Tableau 8.1

Temps de réaction (ms) et taux de réponses correctes (%) en fonction du temps de présentation et du type de transformation de l'image-cible pour l'expérience 5. Erreurs type entre parenthèses.

		Transformation de l'image-cible			
				Luminance	
		Originale	Contour	Avec lissage	Sans lissage
Temps de réaction (ms)					
	35 ms	1514 (129)	1705 (121)	1665 (160)	1493 (143)
	50 ms	1449 (157)	1573 (201)	1497 (147)	1524 (200)
	150 ms	1180 (55)	1229 (83)	1176 (55)	1337 (72)
Taux de réponses correctes (%)					
	35 ms	66,3 (3,9)	65,0 (6,2)	63,8 (4,3)	60,0 (4,8)
	50 ms	66,3 (4,1)	61,3 (4,0)	73,1 (7,0)	64,4 (3,1)
	150 ms	80,0 (3,4)	76,3 (4,5)	76,3 (2,9)	46,3 (3,3)

2.1.1. Temps de présentation de l'image-cible

L'analyse du temps de réaction montre un effet significatif pour le temps de présentation de l'image-cible, F (2, 38) = 12.307, p< .05. Les participants répondent plus rapidement lorsque les images-cibles sont présentées en 150 ms (1231 ms) que celles présentées en 50 ms (1511 ms), F (1, 19) = 5.327, p< .05, et que celles présentée en 35 ms (1594 ms), F (1, 19) = 3.424, p< .05 (Figure 8.2).

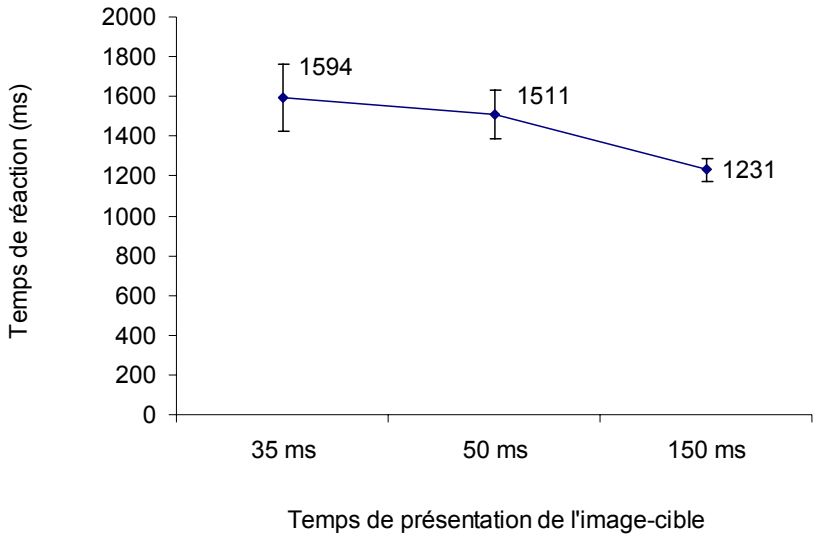


Figure 8.2. Temps de réaction en fonction du temps de présentation de l'image-cible.

L'analyse indique un effet significatif du temps de présentation de l'image-cible pour le taux de réponses correctes, F (2, 38) = 10.292, p< .001. Les participants catégorisent significativement mieux lorsque les images sont présentées en 150 ms (74,1%) que celles présentées en 50 ms (66,3%), F (1, 19) = 4.221, p< .05, et que celles sont présentées en 35 ms (60,3%), F (1, 19) = 8.663, p< .001 (Figure 8.3).

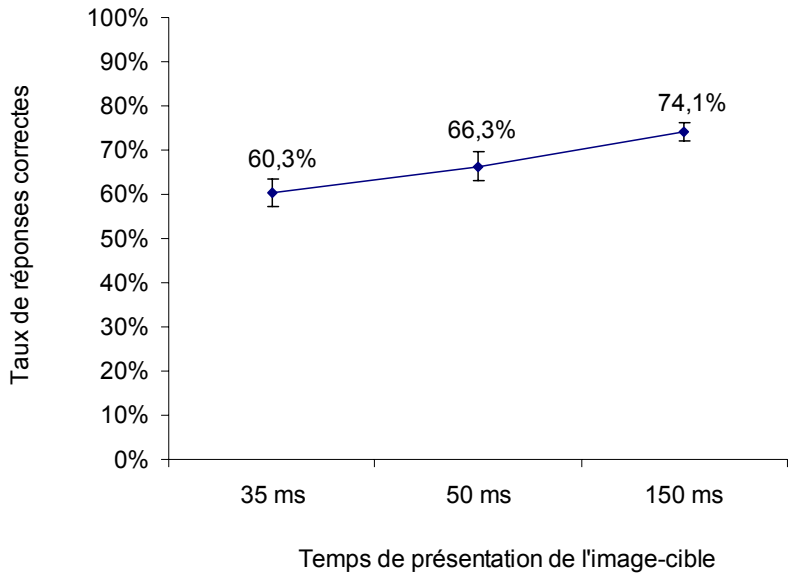


Figure 8.3. Taux de réponses correctes en fonction du temps de présentation de l'image-cible.

2.1.2. Type de transformation de l'image-cible

Dans cette partie, notre analyse porte essentiellement sur l'effet du lissage réservé aux "*images-luminance*". Les analyses portant sur les "*images-contours*" et les "*images-originales*" ne seront pas présentées en détail.

Le type de transformation de l'image-cible n'influence pas significativement le temps de réaction, F (3, 57) = 1.206, P> .05. Les participants répondent avec un temps équivalent pour les quatre types de transformations d'images-cibles (Figure 8.4).

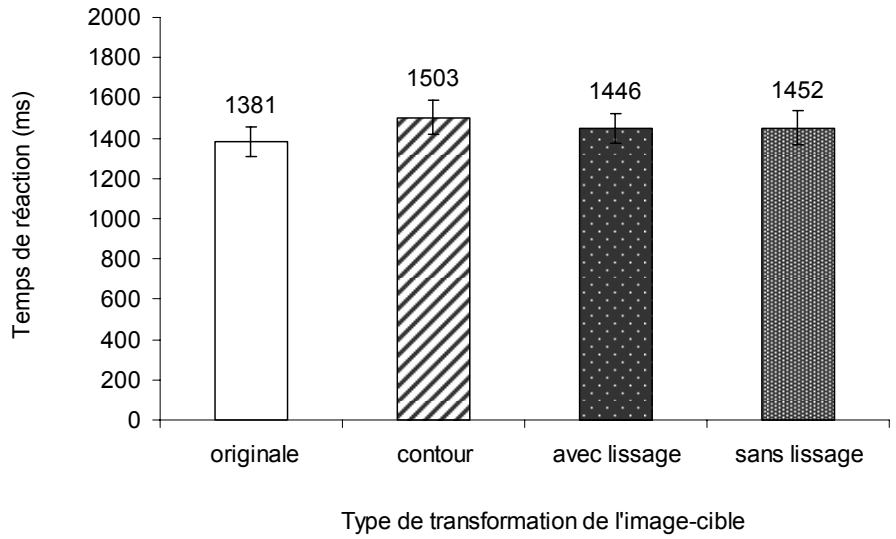


Figure 8.4. Temps de réaction selon le type de transformation de l'image-cible.

Le type de transformation de l'image-cible influence significativement le taux de bonnes réponses, *F* (3, 57) = 3.267, *P*< .05. Les *"images-luminance sans lissage"* sont moins bien reconnues (57,9% de bonnes réponses) que les "*images-luminance avec lissage*" (71,0%), *F* (1, 19) = 3.447, *p*< .05. Le taux de réponses correctes est équivalent entre les deux autres types d'images (70,8% pour les *"images-originales"*, 67,5% pour les *"images-contours"*) (F<1) (Figure 8.5).

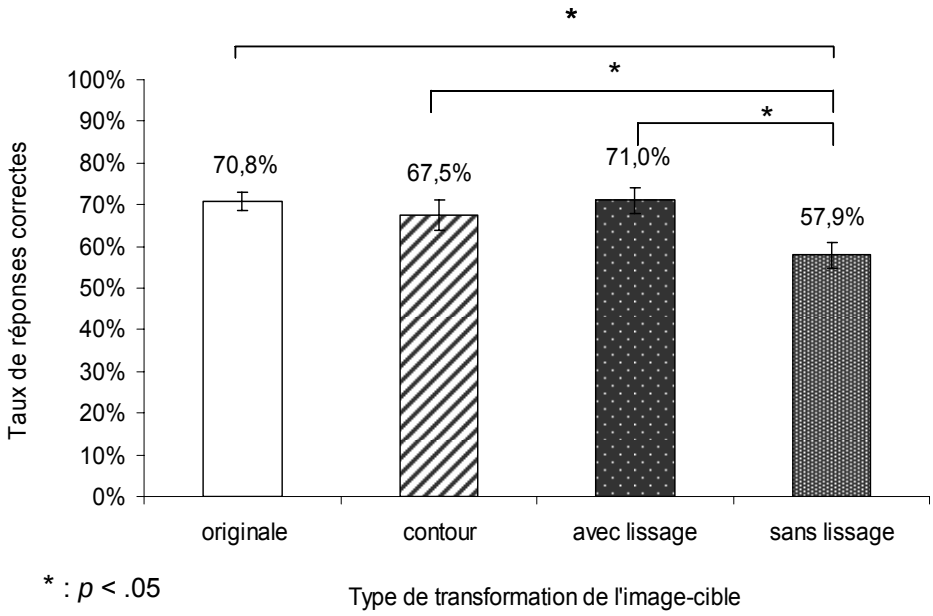


Figure 8.5. Taux de réponses correctes selon le type de transformation de l'image-cible.

Maintenant, nous allons voir les effets du type de transformation selon le temps de présentation d'images-cibles

2.1.2.1. Le temps de présentation est de 35 ms

Le type de transformation de l'image-cible n'influence pas significativement le temps de réaction, F (3, 57) = 1.919, p> .05. De même, pour les taux de bonnes réponses (F<1). Quel que soit le type de transformation de l'image-cible, les participants répondent en moyenne en 1594 ms, environ 63,8% de bonnes réponses (Figure 8.6, Figure 8.7).

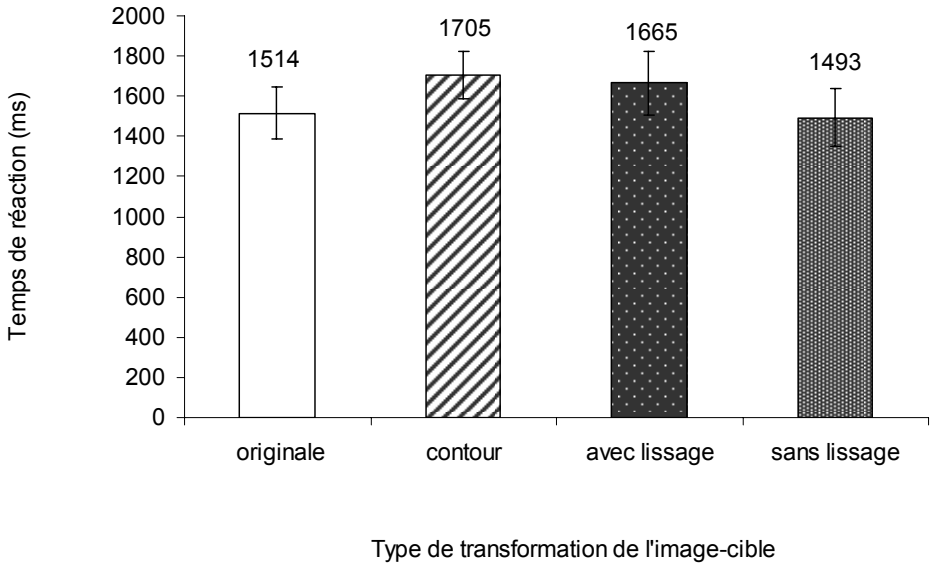


Figure 8.6. Temps de réaction pour les images-cibles présentées en 35 ms selon le type de transformation de ces images.

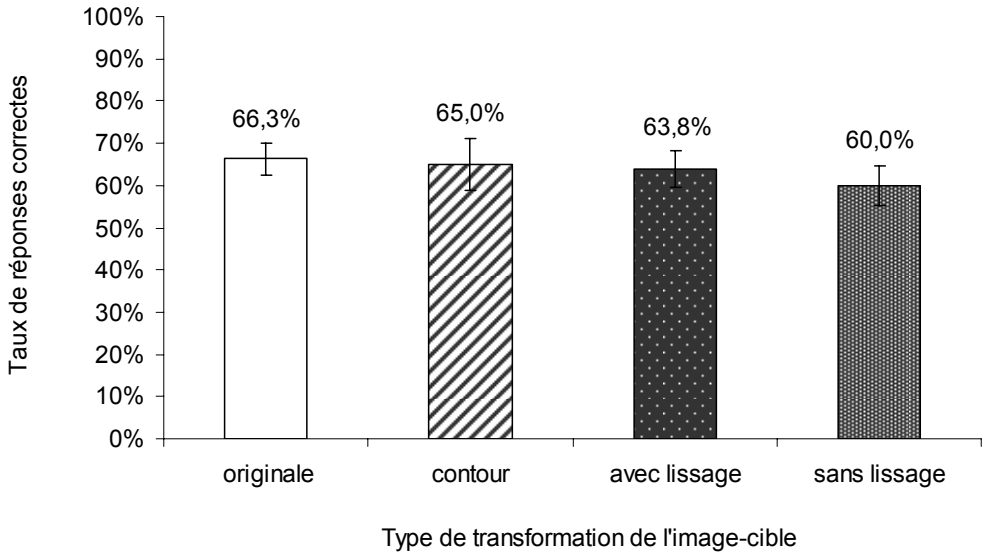


Figure 8.7. Taux de réponses correctes pour les images-cibles présentées en 35 ms selon le type de transformation de ces images.

2.1.2.2. Le temps de présentation est de 50 ms

Des résultats analogues sont observés que le temps de présentation soit de 50 ms ou de 35 ms. Le type de transformation de l'image-cible ne produit pas d'effet significatif, que ce soit sur le temps de réaction ($F<1$) ou sur le taux de bonnes réponses, F (3, 57) = 1.210, $P> .05$.

Les participants répondent avec un temps en moyenne de 1511 ms pour les quatre types d'images. Quant au taux de réponses correctes, il varie peu selon les *"images-contours"* (61,3%), les *"images-luminance avec lissage"* (73,1%) et les *"images-luminance sans lissage"* (64,4%) (Figure 8.8, Figure 8.9).

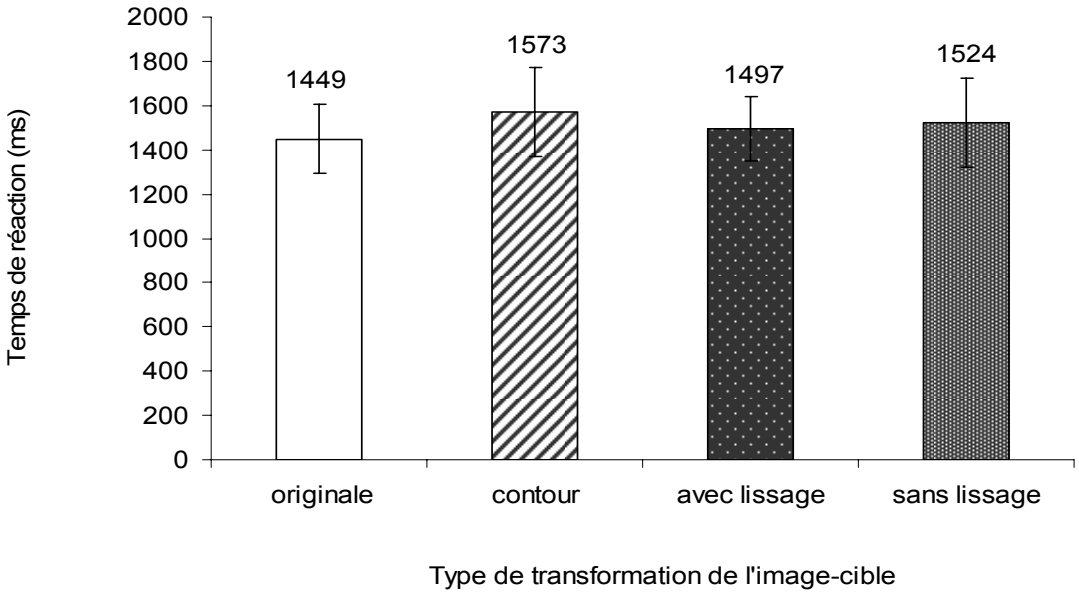


Figure 8.8. Temps de réaction pour les images-cibles présentées en 50 ms selon le type de transformation de ces images.

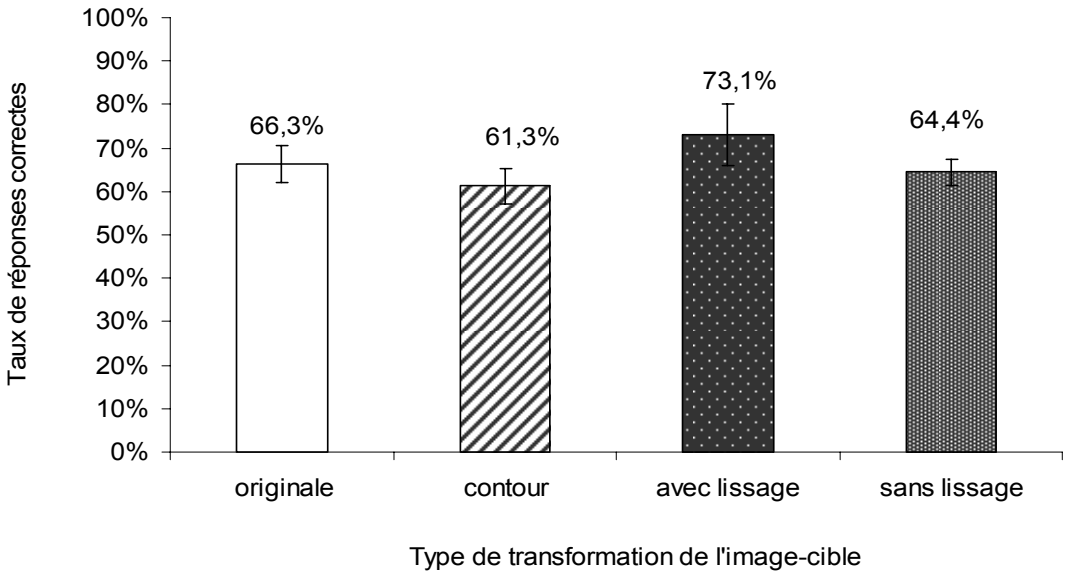


Figure 8.9. Taux de réponses correctes pour les images-cibles présentées en 50 ms selon le type de transformation de ces images.

2.1.2.3. Le temps de présentation est de 150 ms

L'analyse montre un effet significatif pour le temps de réaction, F (3, 57) = 3.556, p< .05.

Les *"images-luminance sans lissage"* sont moins rapidement reconnues (1337 ms) que les "*images-luminance avec lissage*" (1176 ms), F (1,19) = 15.533, p< .001. Ces dernières ont un temps de réaction comparables aux *"images-originales"* (1180 ms) et aux "*images-contours*" (1229 ms) (Figure 8.10).

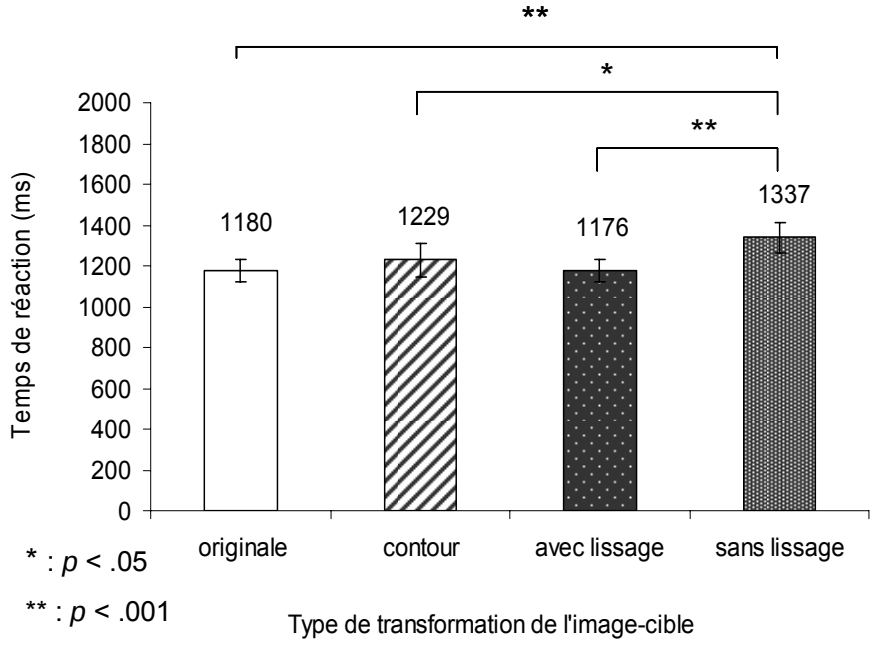


Figure 8.10. Temps de réaction selon le type de transformation de l'image-cible présentée en 150 ms.

L'analyse montre également un effet significatif pour les taux de bonnes réponses, F (3, 57) = 21.722, p< .001.

Les *"images-luminance sans lissage"* (46,3%) sont moins bien reconnues que *"images-luminance avec lissage"* (76,3%), F (1, 19) = 84.610, p< .001. Ces derniers ont des taux de réponses correctes comparables avec les deux autres types d'images, F (1, 19) = 2.632, p> .05 (Figure 8.11).

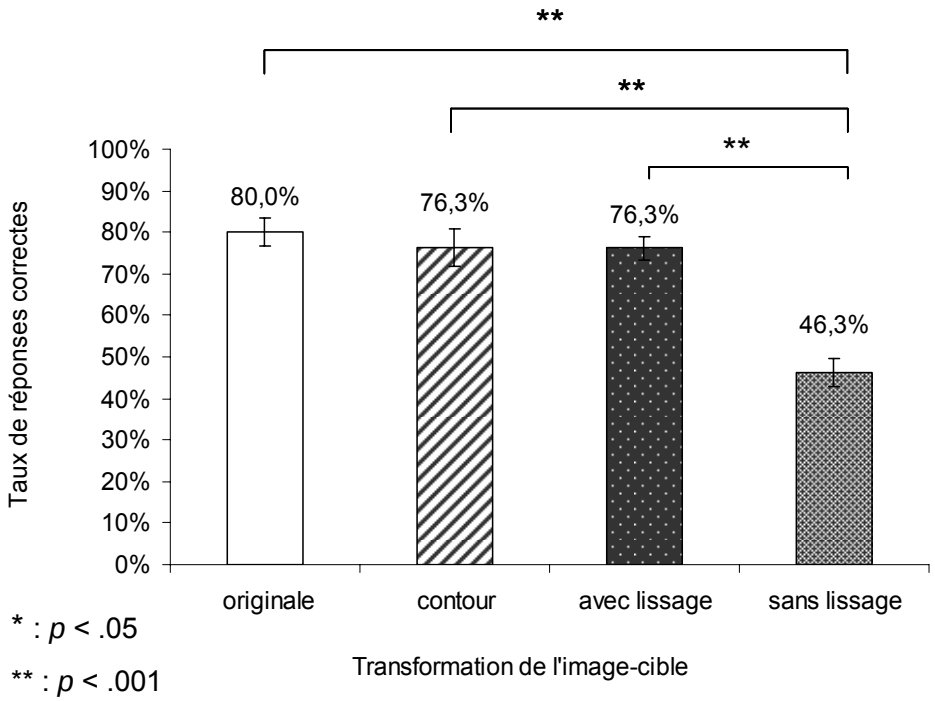


Figure 8.11. Taux de réponses correctes selon le type de transformation de l'image-cible présentées en 150 ms.

2.1.3. Catégorie d'images

La catégorie d'images influence significativement le temps de réaction, F (1, 19) = 29.684, p< .001 et le taux de réponses correctes, F (1, 19) = 28.682, P< .001.

Les images de montagne obtiennent significativement plus de bonnes réponses (75,4%) que les images de plage (58,2%), F (1, 19) = 41.650, P< .001. De plus, elles sont reconnues plus rapidement (1366 ms) que les images de plage (1524 ms), F (1,19) = 87.618, p< .001.

2.1.4. Interaction entre "le temps de présentation de l'image-cible" et "le type de transformation de l'image-cible"

Nous n'avons pas observé d'interaction entre "le temps de présentation de l'image-cible" et "le type de transformation de l'image-cible" pour le temps de réaction, F (6, 114) = 1,348, p> .05 (Figure 8.12).

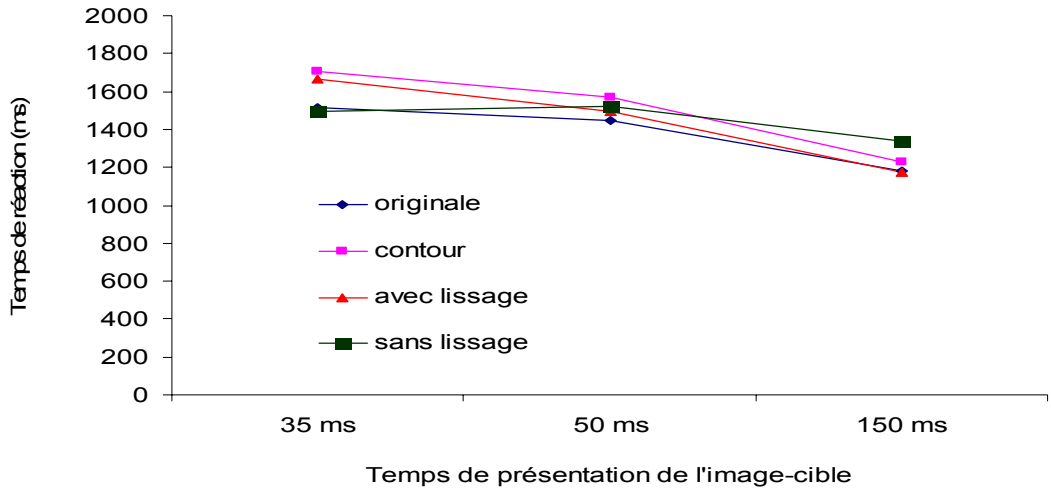


Figure 8.12. Temps de réaction en fonction du temps de présentation et du type de transformation de l'image-cible.

Nous observons une interaction significative entre ces deux facteurs pour les taux de réponses correctes, F (6, 114) = 5,696, p< .05.

Les participants catégorisent moins bien les "*images-luminance avec lissage*" de 35 ms (63,8%) que celles à 50 ms de temps de présentation (73,1%), F (1, 19) = 27.338, P< .001. Leurs taux de réponses correctes restent comparables lorsque le temps de présentation est de 50 ms et de 150 ms (F<1). Quant aux "*images-luminance sans lissage*", les participants n'améliorent pas leurs performances entre le temps de présentation de 35 ms (60,3%) et de 50 ms (64,4%) (F<1). De même, ils catégorisent moins bien lorsque le temps de présentation augmente à 150 ms (46,3%), F (1, 19) = 41.179, P< .001 (Figure 8.13).

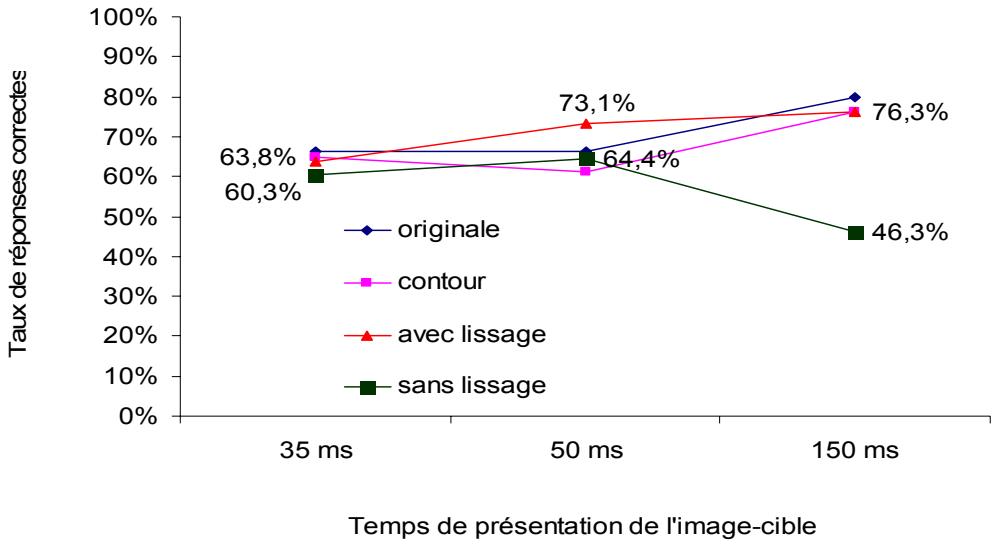


Figure 8.13. Taux de réponses correctes en fonction du temps de présentation et du type de transformation de l'image-cible.

2.2. Les deux images du couple sont de catégories différentes

Les résultats obtenus sont résumés dans le Tableau 8.2.

Tableau 8.2

Temps de réaction de rejet (ms) et taux de rejets corrects (%) en fonction du temps de présentation et du type de transformation de l'image-cible pour l'expérience 5. Erreurs type entre parenthèses.

		Transformation de l'image-cible			
				Luminance	
		Originale	Contour	Avec lissage	Sans lissage
Temps de réaction de rejet (ms)					
	35 ms	1537 (189)	1795 (205)	1699 (168)	1771 (194)
	50 ms	1413 (130)	1770 (159)	1814 (159)	1867 (161)
	150 ms	1047 (55)	1349 (65)	1403 (81)	1310 (54)
Taux de rejets corrects (%)					
	35 ms	85,0 (3,0)	64,4 (4,9)	55,0 (3,7)	51,9 (4,0)
	50 ms	82,5 (3,7)	63,8 (4,3)	63,1 (3,5)	60,0 (4,8)
	150 ms	95,6 (2,1)	66,3 (3,5)	70,0 (4,1)	63,8 (4,1)

2.2.1. Temps de présentation de l'image-cible

L'analyse du temps de réaction montre un effet significatif du temps de présentation de l'image-cible, F (2, 38) = 3.961, p< .05. Les images présentées en 150 ms (1277 ms) sont rejetées plus rapidement que celles présentées en 50 ms (1716 ms), F (1, 19) = 11.344, p< .001, et que celles présentées en 35 ms (1700 ms), F (1, 19) = 8.641, p< .001 (Figure 8.14). Cependant, le taux de rejets corrects ne varie pas significativement (F<1) suivant le temps de présentation (64,1%, 67,3% et 71,4% pour 35 ms, en 50 ms et 150 ms respectivement) (Figure 8.15)

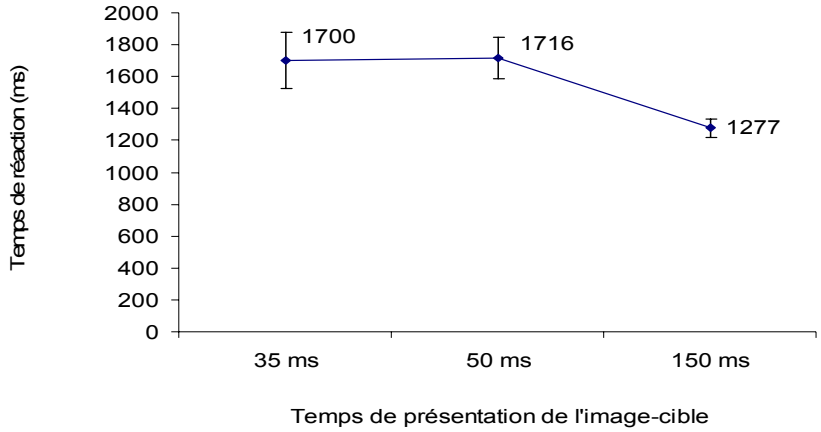


Figure 8.14. Temps de réaction de rejet en fonction du temps de présentation de l'image-cible.

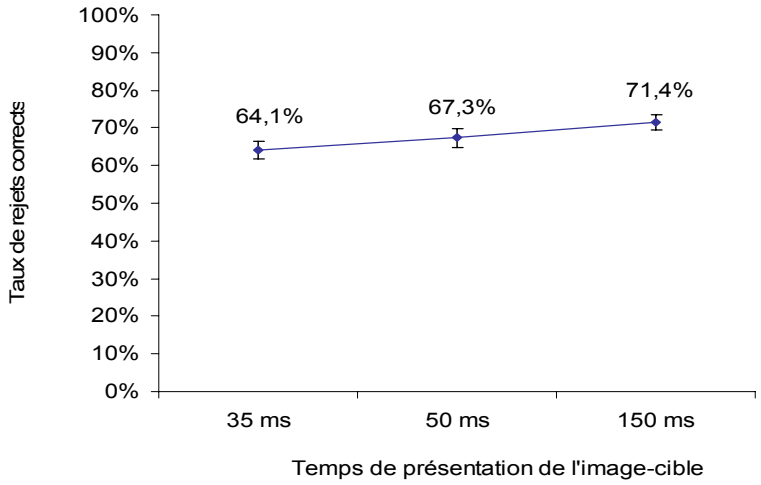


Figure 8.15. Taux de rejets corrects en fonction du temps de présentation de l'image-cible.

2.2.2. Type de transformation de l'image-cible

L'analyse montre un effet significatif du type de transformation de l'image-cible sur le temps de réaction de rejet, F (3, 57) = 10.968, p< .001.

La situation des *"images-originales"* est rejetée plus rapidement (1333 ms) qu'en situation des "*images-contours*" (1638 ms), F (1, 19) = 21.506, p< .001. Ces dernières ont un temps de rejet comparable aux deux autres types d'images (F<1). (Figure 8.16).

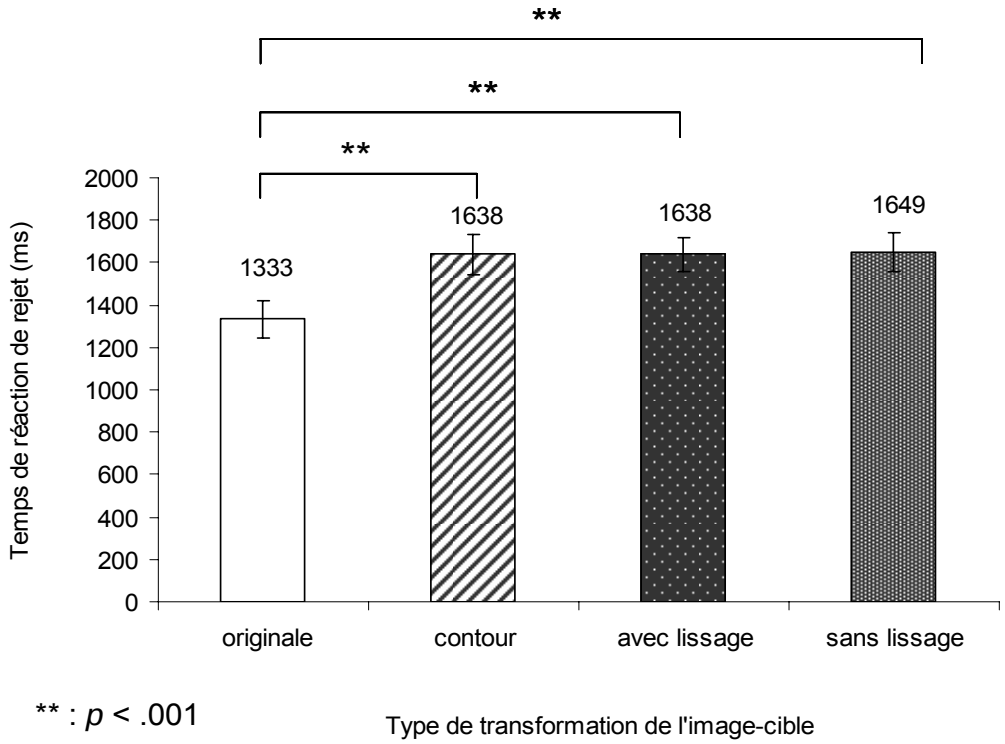


Figure 8.16. Temps de réaction de rejet selon le type de transformation de l'image-cible.

Le type de transformation de l'image-cible influence significativement le taux de rejets corrects, F (3, 57) = 52,754, p< .001. La situation des *"images-luminance avec lissage"* est mieux rejetée (62,7%) qu'en situation des *"images-luminance sans lissage"* (58,5%), F (1, 19) = 5.354, p< .05 (Figure 8.17).

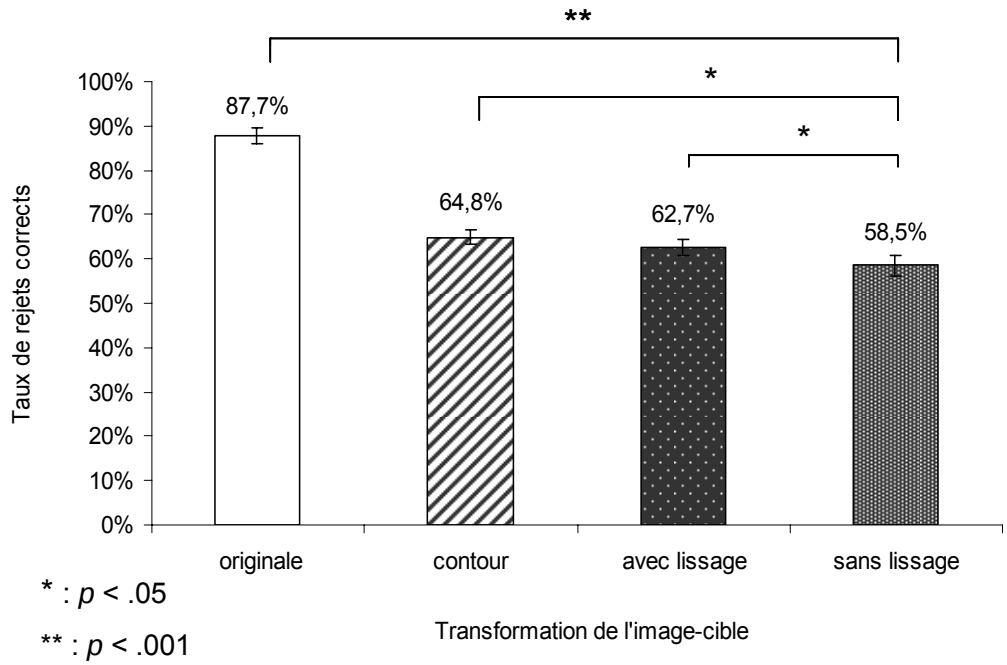


Figure 8.17. Taux de rejets corrects selon le type de transformation de l'image-cible.

Maintenant, nous allons nous intéresser aux effets du type de transformation selon le temps de présentation de l'image-cible.

2.2.2.1. Le temps de présentation est de 35 ms

Le type de transformation de l'image-cible n'entraîne pas significativement le temps de réaction de rejet, F (3, 57) = 1.966, p> .05. Il influence les taux de rejets corrects, F (3, 57) = 6.447, p< .05.

Les participants ont un meilleur taux de performances pour les *"images-originales"* (85% de bonnes réponses), que pour les *"images-contours"* (64%) et que pour les *"images-luminance avec lissage"* et les *"images-luminance sans lissage*" (55% et 51,9%) (Figure 8.18, Figure 8.19).

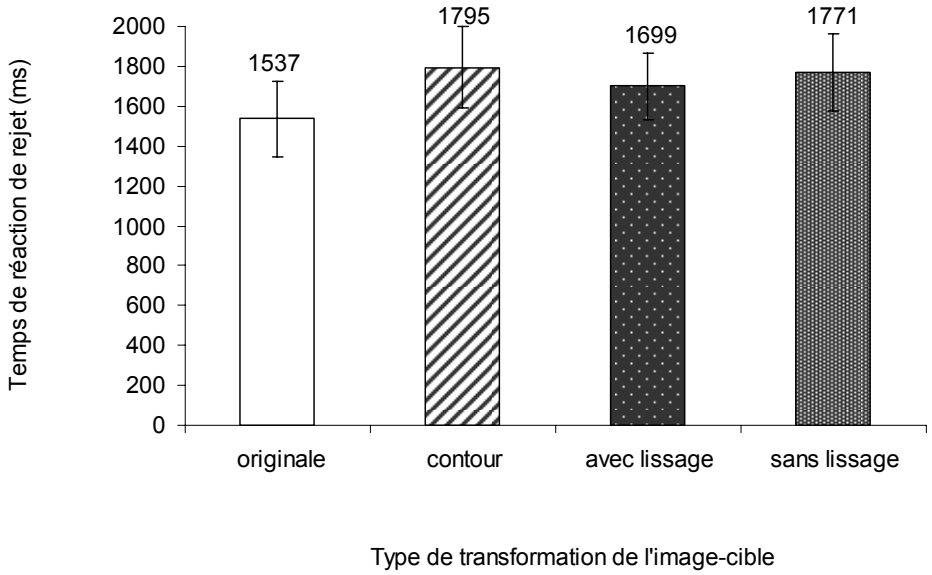


Figure 8.18. Temps de réaction de rejet pour les images-cibles présentées en 35 ms selon le type de transformation de ces images.

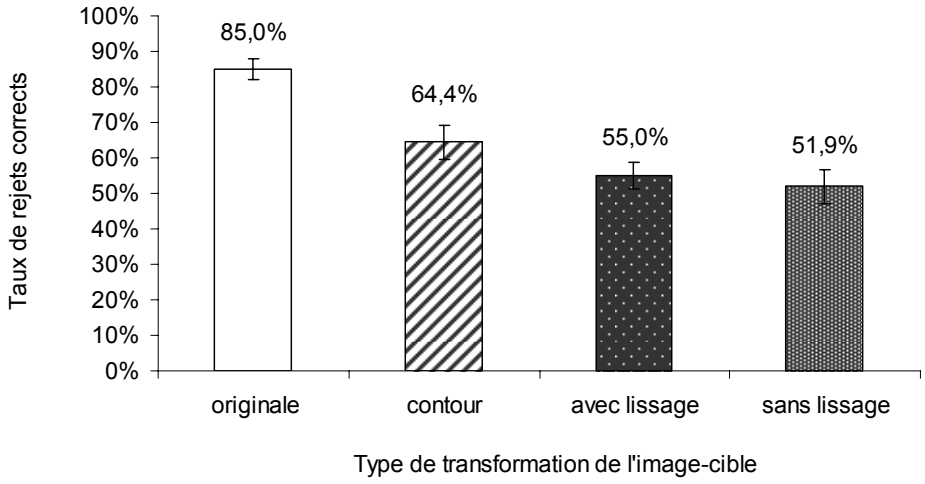


Figure 8.19. Taux de rejets corrects pour les images-cibles présentées en 35 ms selon le type de transformation de ces images.

2.2.2.2. Le temps de présentation est de 50 ms

Le type de transformation de l'image-cible n'entraîne pas significativement pour le temps de réaction de rejet, F (3, 57) = 3.007, p> .05. Cependant, il influence les taux de rejets corrects, F (3,57) = 11.427, p< .05. Les participants ont

une meilleure performance pour les *"images-originales"* (82,5% de bonnes réponses), pour les *"images-contours"* (63,8%) et pour les *"images-luminance avec lissage"* et les "*images-luminance sans lissage*" (63,1% et 60%) (Figure 8.20, Figure 8.21).

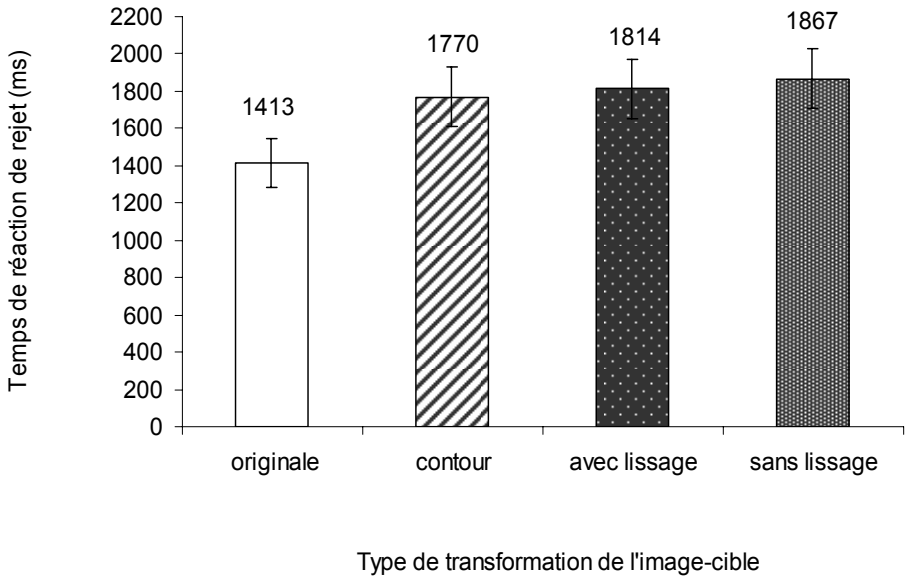


Figure 8.20. Temps de réaction de rejet pour les images-cibles présentées en 50 ms selon le type de transformation de ces images.

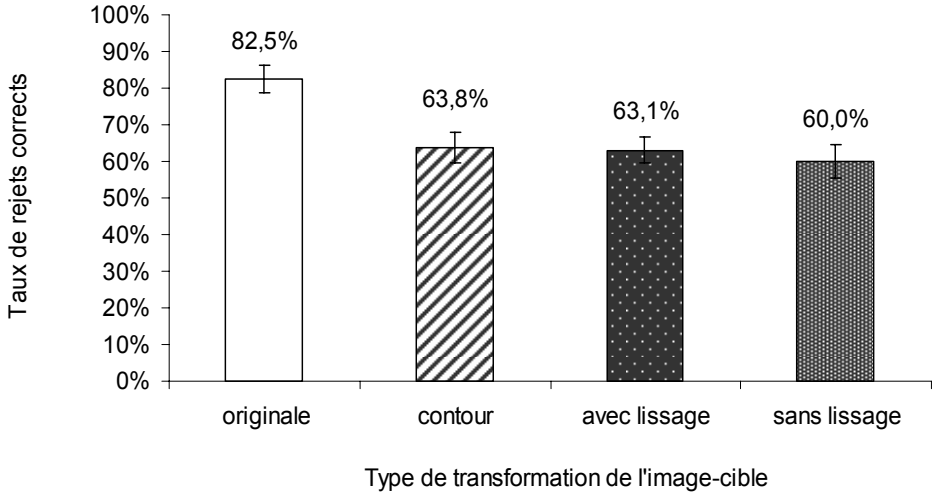


Figure 8.21. Taux de rejets corrects pour les images-cibles présentées en 50 ms selon le type de transformation de ces images.

2.2.2.3. Le temps de présentation est de 150 ms

L'analyse indique un effet significatif du type de transformation de l'image-cible sur le temps de réaction de rejet, F (3, 57) = 16.416, p< .001. Les *"images-originales"* sont reconnues plus rapidement (1047 ms) que les "*images-contours*" (1349 ms), F (1, 19) = 21.305, p< .001. Ces dernières ont un temps de rejet comparable aux deux autres types d'images (F<1) (Figure 8.22).

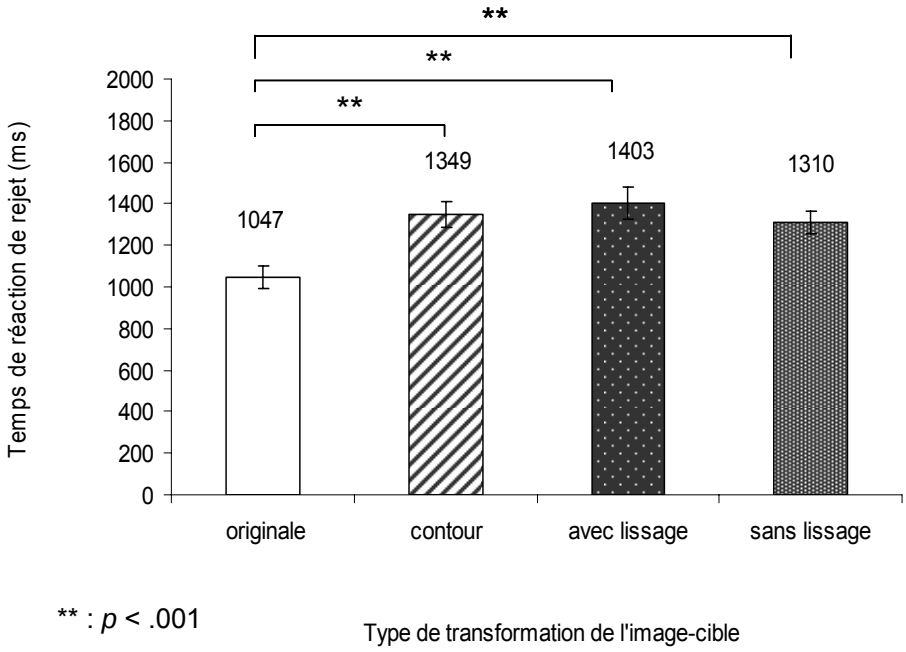


Figure 8.22. Temps de réaction de rejet pour les images-cibles présentées en 150 ms selon le type de transformation de ces images.

L'analyse montre également un effet significatif du taux de bonnes réponses, F (3,57) = 23.185, p< .001. Les participants rejettent mieux en situation des *"images-originales"* (95,6%) qu'en situation des *"images-contours"* (66,3%), F (1, 19) = 25.083, p< .001. Ces dernières ont des taux de rejets corrects comparables aux deux autres types d'images (F<1) (Figure 8.23).

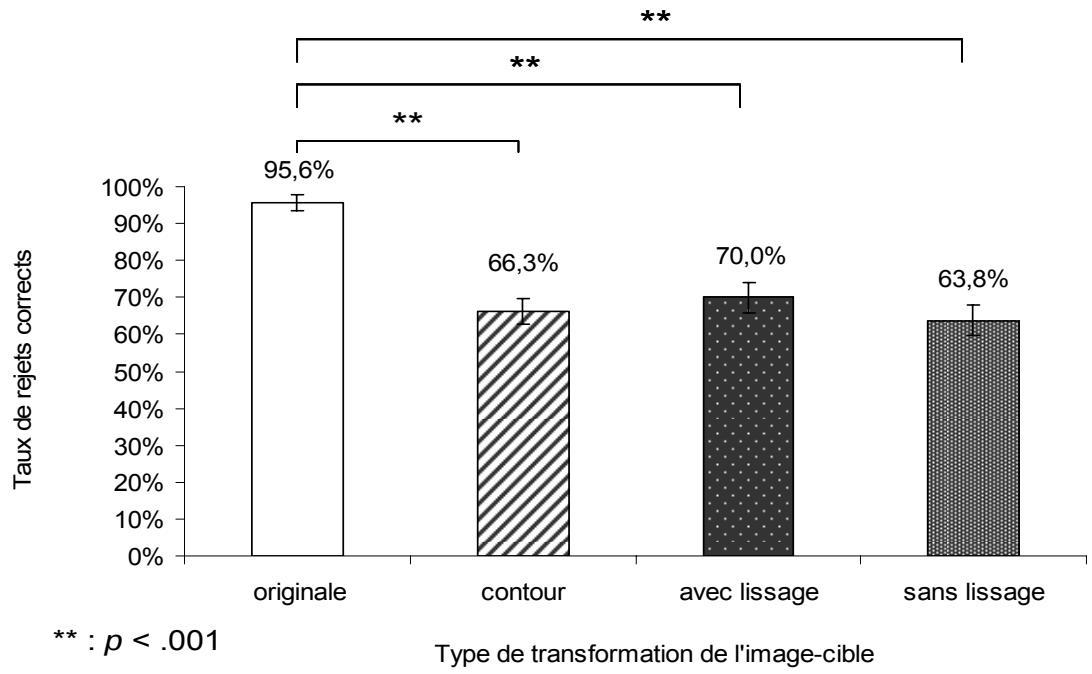


Figure 8.23. Taux de rejets corrects pour les images-cibles présentées en 150 ms selon le type de transformation de ces images

2.2.3. Catégorie d'images

La catégorie d'images n'influence pas significativement le temps de réaction de rejet (F <1).

Les participants répondent en moyenne en 1559 ms pour les images de montagne et en 1570 ms pour les images de plage. L'analyse ne montre pas d'effet significatif pour le taux de rejets corrects (F<1). Les images de montagne obtiennent 68,1% de bonnes réponses contre 67,1% pour les images de plage.

2.2.4. Interaction entre "le temps de présentation d'image-cible" et "le type de transformation de cette image"

L'analyse ne montre pas d'interaction entre le temps de présentation des images-cibles et le type de transformation des images-cibles sur le temps de réaction (F<1). De même, l'analyse n'indique pas d'effet significatif de l'interaction

entre le temps de présentation des images-cibles et la catégorie d'images (F<1) au niveau du temps de réaction de rejet. Quant au taux de rejets corrects, l'analyse ne montre pas non plus d'effet de l'interaction entre ces deux facteurs (F<1).

2.3. Stratégies utilisées par les participants

2.3.1. Le temps de présentation est de 35 ms

Ci-dessous se trouvent les réponses des participants selon les quatre situations de la théorie TDS (Tableau 8.3).

Tableau 8.3

Réponses des participants selon le type de transformation de l'image-cible pour une présentation de 35 ms en intégrant la théorie TDS

		Originale	
		même catégorie	catégorie différente
Réponse	oui	DC 62,5%	FA 15%
	non	O 37,5%	RC 85%
		1	1

d' = 1,355
β = 3,2527

		Contour	
		même catégorie	catégorie différente
Réponse	oui	DC 61,3%	FA 35,6%
	non	O 38,7%	RC 64,4%
		1	1

d' = 0,656
β = 1,2857

		Luminance-avec-lissage	
		même catégorie	catégorie différente
Réponse	oui	DC 65,6%	FA 45%
	non	O 34,4%	RC 55%
		1	1

d' = 0,528
β = 0,3129

		Luminance-sans-lissage	
		même catégorie	catégorie différente
Réponse	oui	DC 65,6%	FA 48,1%
	non	O 34,4%	RC 51,9%
		1	1

d' = 0,45
β = 0,1186

DC = "Détection Correcte" ; O = "Omission" ; FA = "Fausse Alarme" ; RC = "Rejet Correct"

d' = critère de sensibilité ; β = critère décisionnel du sujet

La représentation graphique suivante illustre les résultats du tableau ci-dessus (Figure 8.24).

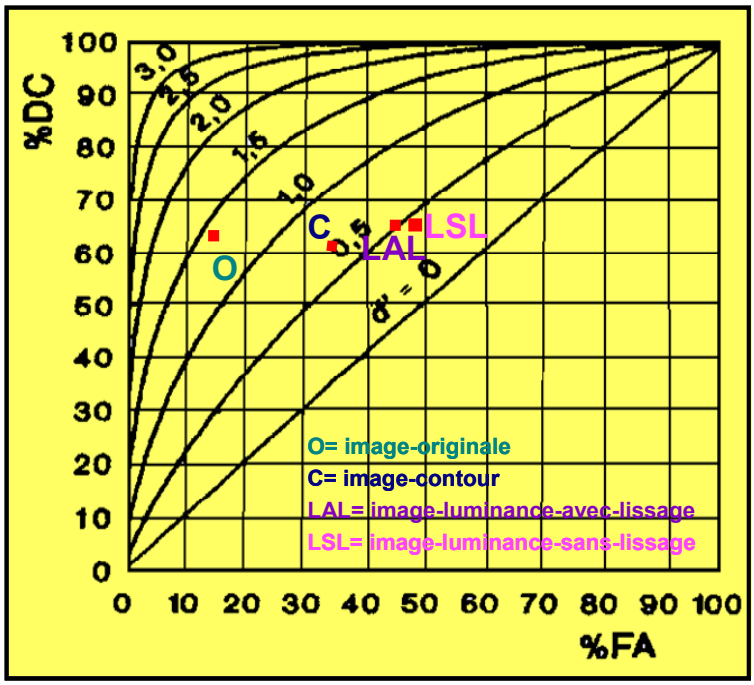


Figure 8.24. Stratégies utilisées par les participants pour identifier les images-cibles présentées en 35 ms.

Les *"images-originales"* sont les images les plus faciles à identifier. Les trois autres types d'images-cibles entraînent presque les mêmes comportements. Les participants produisent moins de fausses alarmes et restent un peu plus prudents pour les *"images-contours"* (35,6%) que pour les *"images-luminance avec lissage"* (45%) et les *"images-luminance sans lissage"* (48,1%). Pour ces dernières, les participants ont un comportement plus risqué et font beaucoup de fausses alarmes.

2.3.2. Le temps de présentation est de 50 ms

Les réponses des participants selon les quatre situations de la théorie TDS pour les images-cibles présentées en 50 ms sont montrées dans le Tableau 8.4.

Tableau 8.4

Réponses des participants selon le type de transformation de l'image-cible pour une présentation de 50 ms en intégrant la théorie TDS

		Originale	
		même catégorie	catégorie différente
Réponse	oui	DC 66,3%	FA 17,5%
	non	O 33,7%	RC 82,5%
		1	1

d' = 1,357
β = 2,2217

		Contour	
		même catégorie	catégorie différente
Réponse	oui	DC 61,3%	FA 36,2%
	non	O 38,7%	RC 63,8%
		1	1

d' = 0,64
β = 1,2297

		Luminance-avec-lissage	
		même catégorie	catégorie différente
Réponse	oui	DC 73,1%	FA 29,2%
	non	O 26,9%	RC 70,8%
		1	1

d' = 1,164
β = 0,8891

		Luminance-sans-lissage	
		même catégorie	catégorie différente
Réponse	oui	DC 64,4%	FA 40%
	non	O 35,6%	RC 60%
		1	1

d' = 0,622
β = 0,6863

DC = "Détection Correcte" ; O = "Omission" ; FA = "Fausse Alarme" ; RC = "Rejet Correct"

d' = critère de sensibilité ; β = critère décisionnel du sujet

La représentation graphique ci-dessous résume les stratégies utilisées par les participants pour quatre types différents de présentation de l'image-cible (Figure 8.25).

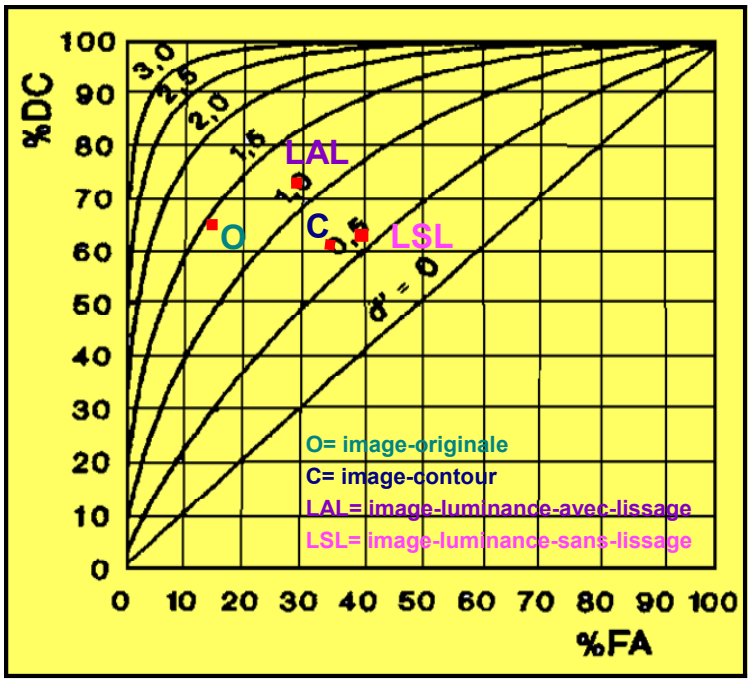


Figure 8.25. Stratégies utilisées par les participants pour identifier les images-cibles présentées en 50 ms.

Les *"images-originales"* et les *"images-luminance avec lissage"* sont jugées relativement plus faciles à identifier que les deux autres types d'images. Les participants utilisent une stratégie prudente pour les "*images-originales*". Leurs stratégies sont plutôt neutres pour les "*images-luminance avec lissage*" (β = 0,8891). Quant aux *"images-contours"* (d'=0,64) et aux *"images-luminance sans lissage"* (d'=0,622), elles entraînent le même niveau de difficulté lors de la réalisation de la tâche de catégorisation. Les participants ont un comportement plus risqué avec plus de fausses alarmes pour les *"images-luminance sans lissage"* (40%) que pour les *"images-contours"* (36,2%).

2.3.3. Le temps de présentation est de 150 ms

Ci-dessous se trouve un résumé des réponses des participants les quatre situations de la théorie TDS pour les images-cibles présentées en 150 ms (Tableau 8.5.)

Tableau 8.5

Réponses des participants selon le type de transformation de l'image-cible pour une présentation de 150 ms en intégrant la théorie TDS

		Originale	
		même catégorie	catégorie différente
Réponse	oui	DC 80%	FA 4,4%
	non	O 20%	RC 95,6%
		1	1

d' = 2,548
β = 2,0271

		Contour	
		même catégorie	catégorie différente
Réponse	oui	DC 76,3%	FA 33,7%
	non	O 23,7%	RC 66,3%
		1	1

d' = 1,137
β = 0,5875

		Luminance-avec-lissage	
		même catégorie	catégorie différente
Réponse	oui	DC 76,3%	FA 30%
	non	O 23,7%	RC 70%
		1	1

d' = 1,24
β = 0,7324

		Luminance-sans-lissage	
		même catégorie	catégorie différente
Réponse	oui	DC 46,3%	FA 36,2%
	non	O 53,7%	RC 63,8%
		1	1

d' = 0,26
β = -3,802

DC = "Détection Correcte" ; O = "Omission" ; FA = "Fausse Alarme" ; RC = "Rejet Correct"

d' = critère de sensibilité ; β = critère décisionnel du sujet

La représentation graphique ci-dessous montre les stratégies utilisées par les participants pour les quatre types de présentation de l'image-cible (Figure 8.26).

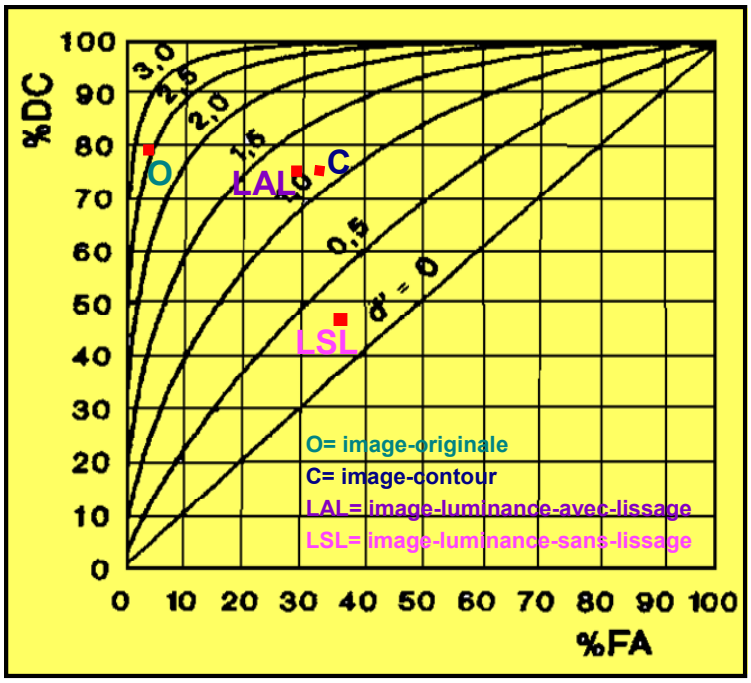


Figure 8.26. Stratégies utilisées par les participants pour identifier les images-cibles présentées en 150 ms.

Les *"images-originales"* sont les images les plus faciles à identifier. Les *"images-contours"* et les *"images-luminance avec lissage"* sont jugées relativement faciles à identifier. Elles entraînent le même type de comportement. Les "*images-luminance sans lissage*" sont jugées difficiles à identifier (d'=0,26), les participants ont un comportement plus risqué en les considérant et font beaucoup de fausses alarmes (36,2%).

3. Discussion

Dans cette expérience, nous tentions de révéler l'effet du lissage pour les *"images-luminance"* lors d'une tâche de catégorisation de scènes visuelles complexes.

Le résultat principal est qu'à la vitesse de 150 ms, les *"images-luminance avec lissage"* permettent un taux excellent de détections correctes (76,3%), aussi

bon que pour les "*images-originales*" (80%). Cependant, cette performance ne s'observe qu'avec un temps de présentation de l'image cible le plus long (150 ms).

Lorsque les temps de présentation sont de 35 ms et de 50 ms, nous n'avons pas trouvé de différence significative pour les taux des réponses correctes concernant les *"images-luminance avec lissage"* (63,8% et 60% pour 35 ms et 50 ms respectivement), de même, concernant les *"images-luminance sans lissage"* (73,1% et 64,4% pour 35 ms et 50 ms respectivement) (Figure 8.7, Figure 8.9).

Pour ces deux types de présentation, le mécanisme de traitement semblerait se baser sur la structuration spatiale des différentes zones de luminance. Ainsi, les contours irréguliers dans les "*images-luminance sans lissage*" devaient être très peu perçus. Par conséquent, la voie dorsale semble être principalement sollicitée.

Passant d'un temps de 50 ms à 150 ms de présentation, deux types de performances différentes sont observés selon le type de lissage. Nous observons une performance comparable pour les *"images-luminance avec lissage"* (73,1% et 76,3% pour 50 ms et 150 ms respectivement). Cependant, la performance des *"images-luminance sans lissage"* diminue significativement (64,4% et 46,3% pour 50 ms et 150 ms respectivement) (Figure 8.13). Cette différence serait liée au mécanisme de traitement impliqué selon ces deux types d'images :

- les *"images-luminance avec lissage"* sont relativement faciles à identifier car elles permettraient de construire une structure spatiale, aboutissant à une catégorisation.

- au contraire, les *"images-luminance sans lissage"* entraînent des difficultés pour leur catégorisation. Les contours irréguliers des *"images-luminance"* sont donc perturbants et pourraient entraîner des difficultés d'identification de la scène.

Plusieurs hypothèses sont discutées afin de rendre compte de cet effet de perturbation qui est issue des traitements portant sur les contours irréguliers lorsque le temps de présentation est de 150 ms.

En effet, le fait d'augmenter le temps de présentation augmente l'activité parvocellulaire (Schmolesky *et al.,* 1998). Ainsi, à partir de 50 ms, la voie ventrale prendrait un rôle important tout en traitant les contours (hautes fréquences) de scène. Selon le principe de traitement "*l'analyse des images depuis les grandes formes jusqu'aux détails*" (*coarse to fine*) (Schyns & Oliva, 1994; Hérault, Beaudot, & Oliva, 1995; Schyns & Oliva, 1997), le système évolue d'un mode purement perceptif vers un mode sémantique. Ainsi, d'une saillance perceptive le système évoluera vers la sélection des régions à forte saillance cognitive (Figure 8.27).

Au cours du temps d'exploration de la scène, le système de traitement porterait sur les éléments informatifs, à savoir les contours irréguliers des *"images-luminance sans lissage"*. N'ayant aucune signification, ces contours seraient donc considérés comme des éléments prégnants tout en sollicitant davantage des mécanismes (pré-)attentionnels (De Graef, Christiaens, & d'Ydewalle, 1990; Friedman, 1979; Henderson, Weeks, & Hollingworth, 1999; Hollingworth, 2003). Ne pouvant pas construire de représentation des *"images-luminance sans lissage"* correspondantes en mémoire à long terme, le mécanisme de traitement serait perturbé en orientant sur des fausses pistes de catégorisation sémantique.

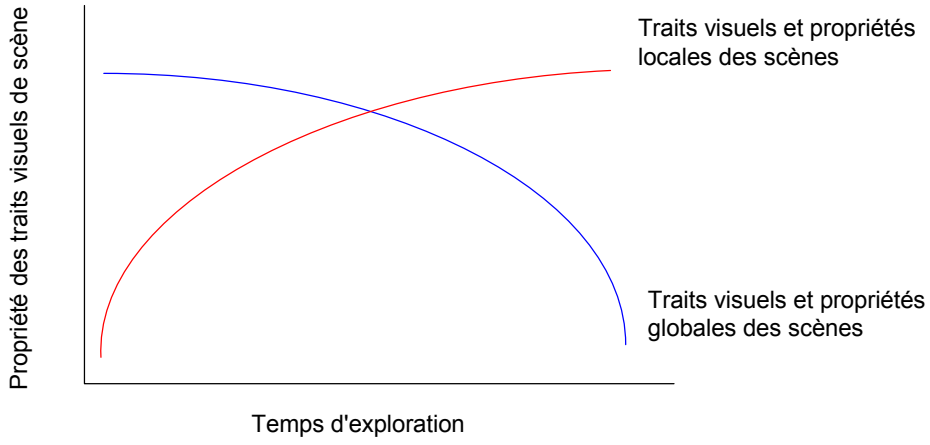


Figure 8.27. Illustration de l'évolution d'un traitement purement perceptif vers un traitement purement cognitif (Henderson & Hollingworth, 1999).

Selon une autre hypothèse appelée "amnésie d'inattention" (*"inattentional amnesia"*, Wolf, 1999), la représentation visuelle serait instantanée et ne resterait pas en mémoire. Ainsi, l'attention étant indispensable à la mémorisation, il s'ensuit que les stimuli, n'étant pas des cibles de l'attention, pourraient être perçus mais seraient instantanément oubliés. Lorsque le temps de présentation augmente, faute d'attention focalisée portant sur la structuration spatiale de l'image, la performance des *"images-luminance sans lissage"* diminuerait. C'est le cas des *"images-luminance sans lissage",* pour lesquelles les cellules parvocellulaires seraient sollicitées tout en traitant les contours irréguliers. Ne pouvant pas construire de lien entre ce que signifient ces contours irréguliers et les représentations sémantiques stockées en mémoire à long terme, le mécanisme d'identification se trouverait en difficulté en raison du manque d'activation sémantique.

Au contraire, pour les *"images-luminance avec lissage"*, la transition entre les zones de luminance reste floue. Elle est caractérisée par de basses fréquences. L'allongement du temps de présentation maintiendrait relativement longtemps l'intensité des stimuli, ce qui permettrait au mécanisme de traitement de mieux percevoir la structuration de l'ensemble de l'image. Ces images seraient alors jugées relativement faciles à identifier et la performance des participants s'améliorant d'autant.

En ce qui concerne les *"images-originales"*, elles n'entraînent pas de différence comportementale lorsque le temps de présentation est de 35 ms ou de 50 ms. La différence de temps n'est pas suffisante pour provoquer une différence perceptible des informations disponibles.

Quant aux *"images-contours"*, elles sont jugées relativement difficiles à identifier suivant le temps de présentation par rapport aux *"images-originales"*. Perceptivement, les *"images-contours"* fournissent moins d'informations disponibles que les *"images-originales"*. Le mécanisme d'identification se trouve en difficulté faute de disposer d'autres informations pertinentes. Comme dans le cas des *"images-originales"*, les participants réagissent de la même façon pour les *"images-contours"* et ce, pour un temps de présentation de 35 ms et de 50 ms.

Discussion finale

Nous avons montré, dans les chapitres 2 et 3, que les scènes visuelles complexes peuvent être identifiées très rapidement dès les premières fixations oculaires sur la scène (Intraub, 1981; McCauley, Parmelee, Sperber, & Carr, 1980; Potter, Staub, Rado, & O'Connor, 2002; Thorpe, Fize, & Marlot, 1996; VanRullen & Thorpe, 2001). Néanmoins, de nombreuses études sur la cécité au changement ("*change blindness*") montre que notre capacité à détecter une modification dans une image est relativement faible (Henderson & Hollingworth, 1999 ; O'Regan, 2001; O'Regan, Rensink, & Clark, 1999; Rensink, 2002 ; Shinoda, Hayhoe, & Shrivastava, 2001; Simons & levin, 1997).

Ces deux capacités humaines liées à la perception de scènes semblent contradictoires. En effet, elles impliquent deux mécanismes de traitements, l'un concerne l'identification de scènes, l'autre sa représentation en mémoire.

Le fait qu'une scène puisse être identifiée très rapidement dépend en premier lieu des différents types d'informations utilisés, que ce soient un ou plusieurs objets typiques (Friedman, 1979), leurs relations spatiales (De Graef, Christiaens, & d'Ydewalle, 1990), les fréquences spatiales (Schyns & Oliva, 1994), la couleur (Oliva & Schyns, 2000), les informations contextuelles et la connaissance de la tâche (Rossion & Gauthier, 2002; Schyns, 1998; Ullman, Vidal-Naquet, & Sali, 2002), que ce soient d'autres types d'informations caractérisées par leur propriétés de "scène-niveau" (Biederman, 1985, 1995; Oliva, Torralba, Guérin-Dugué, & Hérault, 1999; Guérin-Dugué & Oliva, 2000).

En second lieu, cette identification rapide dépend des différents types de traitements du système visuel pour ces informations. La voie ventrale est impliquée notamment dans la reconnaissance des objets, la voie dorsale l'est surtout dans la détermination de la position et du mouvement des objets (Ungerleider & Haxby, 1994; Watanabee, 2003).

Cependant, le mécanisme de représentation d'une scène en mémoire paraît très compliqué.

Premièrement, nous n'avons sans doute pas besoin de reconstruire une représentation complexe d'une scène que nous venons de voir. Selon la théorie de la mémoire conceptuelle à court terme (Potter, 1999,), nous percevons beaucoup d'objets lorsque nous regardons une scène mais seulement une petite partie de ces informations sera encore disponible lorsque nous regarderons une nouvelle scène. Ainsi, la nouvelle scène remplacerait la scène précédente dont seul le sens général serait mémorisé.

Deuxièmement, il est possible que notre système visuel soit en grande partie amnésique, comme le suggère Wolfe (1999) et que sans attention focalisée, nous ayons recours à la richesse du monde qui nous entoure plutôt qu'à une hypothétique représentation interne, comme le suggèrent O'Regan (1992) et Rensink (2000a, 2000b). Pour que le contenu d'une scène soit représenté en mémoire, il est nécessaire que les différentes régions des scènes soient la cible de l'attention focalisée (Henderson & Hollingworth, 1999 ; Henderson, Williams, Castelhano, Falk, 2003; Intraub, 1999).

Enfin, 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 aboutir à une représentation explicite.

La relation entre l'identification et la représentation d'une scène a été peu étudiée, à notre connaissance, en regard de l'intérêt sémantique et de la pertinence d'éléments visuelles par rapport à la tâche.

Etant donné les limites de traitement dans une tâche d'identification rapide de scène (Irwin, 1992; Luck & Vogel, 1997; Melcher, 2001; Vogel, Woodman, & luck, 2001), il semble évident que tous les éléments d'une scène ne sont pas traités en même temps, et que seules des informations traitées en attention focalisée soient représentées en mémoire. Il est donc important de déterminer quelles informations visuelles sont extraites de la scène et suffisamment représentées en mémoire.

L'objectif de cette thèse était donc, dans un premier temps, de déterminer à quel point la pertinence des informations locales d'une scène permet d'aboutir à son identification.

Dans un second temps, d'évaluer plus précisément comment ces informations sont représentées dans une tâche de reconnaissance et de catégorisation de scène.

1. Synthèse et interprétation des résultats obtenus

L'expérience 1 avait pour objectif d'étudier dans quelles mesures la prédiction des points d'intérêt définis par l'algorithme de traitement d'images, développé par France Telecom R&D, permet aux participants d'identifier une scène visuelle complexe.

Les résultats montrent que ces points d'intérêt jouent un rôle différent selon leur saillance. Lorsque les scènes sont affichées avec les points d'intérêt les plus saillants, elles ont une meilleure performance. Inversement, lorsque les scènes sont affichées avec les points d'intérêt les moins saillants, leur performance sont les plus faibles. Les points d'intérêt étant caractérisés à la fois par leurs contours et leur structuration spatiale de la luminance, ces deux types d'informations sont donc jugés comme informations porteuses de l'identification de scène.

Cependant, l'expérience 1 ne permettait pas de savoir quels rôles ont joué ces deux types d'informations dans l'identification de scène.

L'expérience 2 avait pour objectif d'observer l'effet des contours et de la structuration spatiale des luminances dans la reconnaissance d'une scène. Différents types d'images-cibles ont été utilisées dans cette expérience : *"image-originale"*, "*image-contour*" et "*image-luminance*". Les "*images originales*" sont mieux reconnues que les *"images-contours"* et les "*images-luminance*". Parmi ces dernières, les résultats varient en fonction de la congruence entre l'image-cible et

l'image-test. Lorsque les deux images du couple sont identiques, les *"images-contours"* sont mieux reconnues que les "*images-luminance*". Dans le cas inverse, ces deux types d'images ont une performance équivalente. Ces résultats suggèrent que le mécanisme de traitement soit basé sur les contours dans le cas où les deux images sont identiques, et, soit basé sur la structuration spatiale de luminance dans le cas où ces deux images sont différentes.

Les trois dernières expériences (3, 4 et 5) permettent d'étudier plus finement l'effet de ces deux facteurs dans une tâche de catégorisation.

Les expériences 3 et 4 consistaient à observer l'évolution de l'effet de ces deux facteurs selon le temps de présentation (50 ms et 150 ms pour l'expérience 3, et 20 ms et 35 ms pour l'expérience 4). Deux effets similaires sur le comportement sont observés dans ces deux expériences. Ainsi, les *"images-contours"* sont mieux reconnues que les *"images-luminance"* en fonction du temps de présentation. Le fait que les *"images-contours"* entraînent une meilleure performance est dû au mécanisme de traitement possédant de plus en plus d'informations disponibles lui permettant de mieux les identifier en fonction du temps de présentation. Cependant, la performance des *"images-luminance"* diminue significativement en fonction du temps de présentation. Cette diminution serait liée à un mécanisme perturbant impliqué pour ce type d'image.

Les résultats des expériences 3 et 4 conduisent à la réalisation de l'expérience 5 dans laquelle l'effet du lissage des contours entre les zones de luminance était observé. Nous avons crée deux conditions différentes : soit des *"images-luminance"* ne contenant pas d'informations portant sur les hautes fréquences (il n'existe plus de contours non réguliers, la transition entre les différentes zones de luminances est floue), soit les *"images-luminance"* avec des contours non réguliers (même type d'"*image-luminance*" que dans les expériences 2, 3, et 4).

Les résultats observés varient en fonction du temps de présentation de l'image-cible. La structuration spatiale peut être très rapidement extraite et peut être représentée en mémoire. Les *"images-luminance avec lissage*" sont mieux

reconnues que les "*images-luminance sans lissage"* notamment à partir de 50 ms de temps de présentation. Ces résultats montrent que le système de traitement visuel évolue d'un mode purement perceptif vers un mode sémantique (*coarse to fine*). Ainsi, d'une saillance perceptive, le système évoluera vers la sélection des régions à forte saillance cognitive (Schyns & Oliva, 1994; Hérault, Beaudot, & Oliva, 1995; Schyns & Oliva, 1997).

2. *Un modèle de traitement des contours et la structuration de luminance dans la perception d'une scène visuelle complexe*

Les résultats obtenus à l'issue de ce travail de thèse peuvent permettre de proposer un modèle de traitement des contours et de structuration des luminances dans la perception d'une scène complexe.

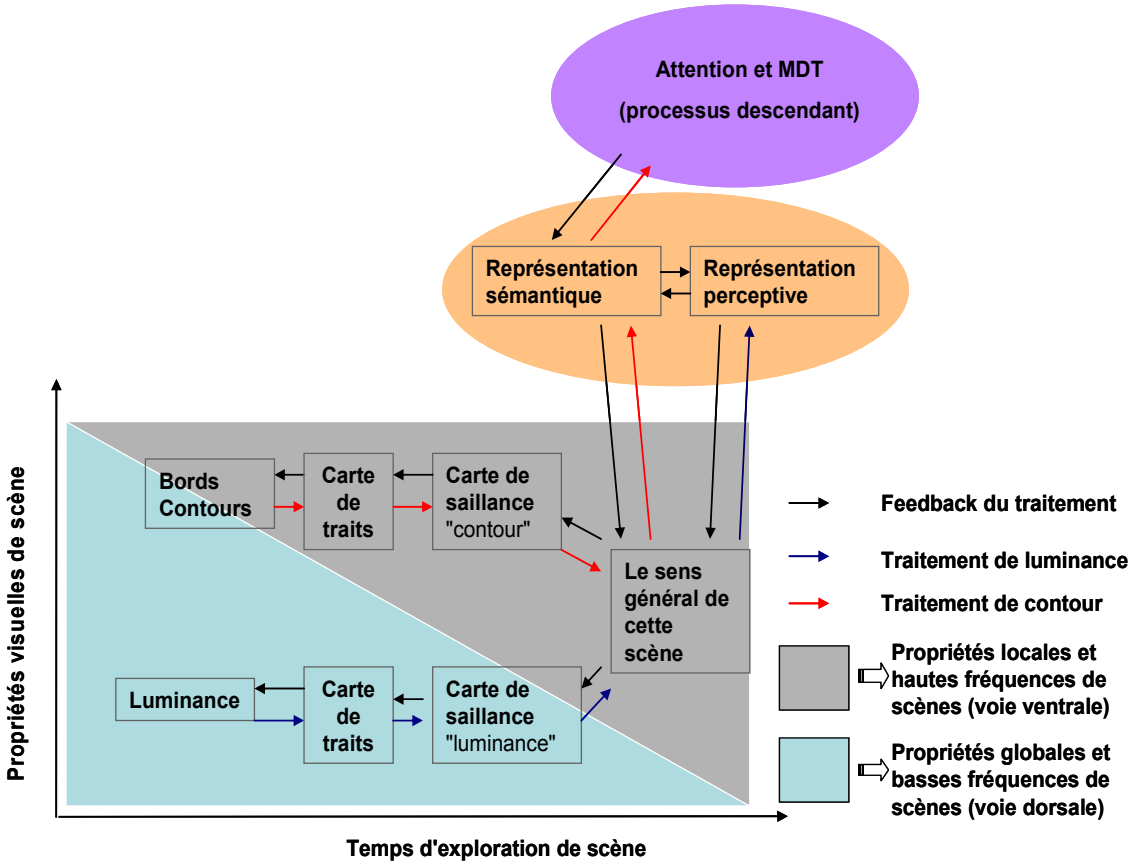


Figure 9.1. Modèle de traitement des contours et de la structuration de luminance dans la perception d'une scène complexe.

Dans ce modèle (Figure 9.1), les premières étapes de la perception d'une scène visuelle sont dévolues aux traitements préattentifs de différentes propriétés visuelles (dimensions de traits) basiques de scène. Ces traitements préattentifs pourraient être considérés comme des traitements massivement parallèles. Cependant, la luminance serait traitée avec environ 20 ms d'avance par rapport aux contours.

A l'issue de ces traitements préattentifs se matérialise, par l'émergence de cartes de traits, qui sont ensuite fusionnées, une carte de saillance (Itti & Koch, 2000; Treisman, 1998; Treisman & Gelade, 1980; Wolfe, 1994). La carte de saillance ainsi produite, fournit des informations concernant le sens général de la scène qui constitue soit une représentation sémantique, soit une représentation perceptive. Ces deux représentations sont ensuite stockées en mémoire.

La représentation sémantique du sens général d'une scène active très rapidement la connaissance de cette scène en mémoire, à savoir la catégorisation, le contexte, la structuration spatiale et le schéma de scène (Biederman, 1972; Biederman, Mezzanotte, & Rabinowitz, 1982; Chun & Nakayama, 2000; De Graef, 1992; Friedman, 1979; Intraub, Bender, & Mangels, 1992; Oliva, 2005). Ensuite, la connaissance sur cette scène guiderait, en retour, l'attention sur les objets d'intérêt dans les scènes visuelles afin d'appréhender leurs significations. Lorsque, la signification de ces objets est cohérente avec leur représentation en mémoire à long terme, cet appariement pourrait permettre d'identifier la scène. En cas de besoin d'autres informations supplémentaires pour l'identification de cette scène, l'attention se déplacera sur d'autres objets d'intérêt. Dans le cas où la signification de ces objets n'est pas cohérente avec leur représentation en mémoire à long terme, l'identification de scène n'aura pas lieu. Par la suite, ces objets deviennent de plus en plus informatifs et seront focalisés par l'attention.

Lorsque nous réalisons une tâche d'identification rapide d'une scène visuelle complexe en se basant sur deux types d'informations disponibles, l'un concerne les contours, l'autre concerne la structuration spatiale de luminance. Le traitement des contours permet d'identifier le sens général de la scène, qui est ensuite représenté sémantiquement en mémoire. Cette représentation sémantique

coderait non seulement le détail des objets d'intérêt central de la scène, mais serait également alimentée par les informations concernant la catégorisation, le sens général, le contexte, la structuration spatiale et le schéma de la scène (Hollingworth, 2004). Au contraire, le traitement des zones de luminance différentes ne permet que de construire une seule représentation perceptive, qui coderait simplement la structure grossière de la scène.

Ces deux représentations jouent un rôle différent selon le type de tâche à réaliser.

Dans une tâche de reconnaissance rapide d'une scène visuelle, les deux types de représentations sont utilisés dans l'appariement entre la représentation de l'image-cible et celle de l'image-test. En effet, la représentation sémantique permet de faire l'appariement entre l'identité des deux images du couple. La représentation perceptive permet, en revanche, d'identifier les localisations des objets de celles-ci. Les deux voies visuelles sont donc sollicitées dans la reconnaissance rapide d'une scène.

Dans une tâche de catégorisation rapide, la représentation sémantique semble être dominante par rapport à la représentation perceptive. Lorsque la représentation sémantique est extraite, la tâche de catégorisation s'effectue rapidement car le mécanisme de traitement ne compare que l'identité de ces deux images du couple. En revanche, une représentation perceptive permet également de catégoriser les scènes visuelles complexes à condition qu'elle ne soit caractérisée seulement que par les basses fréquences. Dans ce cas, seul le traitement réalisé par la voie dorsale permet d'effectuer une tâche de catégorisation. Etant donné que le système de traitement fonctionne selon la typologie d'"*analyse des images depuis les grandes formes jusqu'aux détail*s", seule la représentation perceptive est extraite. Le système visuel continue à chercher d'autres informations pertinentes afin de construire une représentation sémantique. L'attention serait donc focalisée sur les éléments jugés informatifs. Ainsi, dans une tâche de catégorisation, la voie dorsale et la voie ventrale sont toutes deux en compétition. Cependant, la voie ventrale est dominante. Elle

établie le lien entre le traitement en cours portant sur les contours et leurs significations stockées en mémoire à long terme.

3. Perspectives de recherche

Nos travaux de recherche montrent que, dès les premiers regards sur une scène visuelle, les traitements visuels portant sur les contours et la structuration spatiale des zones de luminance peuvent permettre d'engendrer le sens général de la scène. Ce dernier est ensuite représenté sémantiquement ou visuellement en mémoire. La représentation sémantique d'une scène déclenche des connaissances concernant la catégorisation, le contexte, la structuration spatiale et le schéma de scène en mémoire à long terme, qui en retour, guideront spécifiquement l'attention visuelle sur les éléments d'intérêt pour la compréhension et l'interprétation de la scène.

Le fait que la structuration spatiale des différentes zones de luminance permet de construire une représentation grossière de scène constitue la première étape de son identification. Cette représentation perceptive ne serait pas maintenue sans intervention sémantique portant sur cette scène.

Il serait alors intéressant d'étudier les facteurs pouvant faciliter la construction d'une représentation sémantique de scène en se basant sur les différentes zones de luminance.

Cette facilité pourrait-elle être introduite pour fournir des informations supplémentaires ayant des caractéristiques différentes ?

- Par exemple, par la couleur, en ajoutant cette propriété visuelle pour toutes les zones de luminance différente. Une "*image-luminance-couleur*" peut-elle faciliter son identification?

- Par l'amorçage de la signification (par exemple, présentation d'un objet typique) pour une des différentes zones de luminance. Cet amorçage faciliterait-t-il la construction d'une représentation sémantique de cette "*image-luminance"*?

Les futures recherches pourraient étudier cet amorçage partiel de la sémantique pour les différentes zones de luminance dans une tâche de catégorisation de scènes. De plus, toutes ces études pourraient utiliser l'enregistrement des mouvements oculaires afin d'avoir une analyse plus fine.

Une autre réflexion porterait sur le caractère excentré de l'image-cible par rapport à la zone de vision fovéale. Le fait de présenter l'image cible avec différents angles visuels (dans le champ périfovéal ou dans le centre vision-fovéale, par exemple) ne fournit pas la même type d'informations pour l'observateur. La mesure de l'écart entre la position du champ fovéal et la position de l'image à l'écran permettrait de finaliser le fonctionnement du mécanisme de traitement portant sur la structuration spatiale de la luminance dans une tâche d'identification d'images.

Bibliographies

Aguirre, G. K., & D'Esposito, M. (1999). Topographical disorientation: a synthesis and taxonomy. *Brain, 122* (9), 1613-1628.

Alleyson, D. (1999). *Le Traitement Chromatique dans la Rétine : un Modèle de Base pour la perception Humaine des couleurs*. Thèse de Doctorat, université Joseph Fourier, Grenoble.

Altmann, G. T. M., & Kamide, Y. (2004). Now you see it, now you don't: Mediating the mapping between language and the visual world. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology, 103,* 62-70.

Antes, J. R., & Penland J. G. (1981). Picture context effect on eye movement patterns. In D. F. Fisher, R. A. Monty, & J. W. Senders (Eds.), *Eye movements: Cognition and visual perception*. Hillsdale, NJ: Erlbaum.

Bacon-Macé, N., Macé, M. J.-M., Fabre-Thorpe, M., & Thorpe, S. (2005). The time course of visual processing: Backward masking and natural scene categorisation*. Vision Research, 45,* 1459-1469*.*

Baddeley, A. D. (1986). *Working memory.* New York: Oxford University Press.

Baddeley, A. D. (1992). Working memory. *Science, 255*, 556-559.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*, 417-423.

Bar, M. (2004). Visual objects in context. *Nature Review Neuroscience, 5,* 617-629.

Bar, M, & Ullman, S. (1996). Spatial context in recognition. *Perception* 25:343-52.

Barlow, H. B. (1953). Summation and inhibition in the frog's retina. *Journal of Physiology,* 119, pp. 69-88*.*

Beaudot, W. (1994). *Le traitement neuronal de l'information dans la rétine des vertébrés : Un creuset d'idées pour la vision artificielle*. Thèse de doctorat, Institut National Polytechnique, Grenoble.

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* (pp. 213-253). Hillsdale, NJ: Erlbaum.

Biederman, I. (1985). Human image understanding: recent research and a theory. *Computer Vision, GRaphics, and Image Processing, 32*, 29-73.

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. (1995). Visual object recognition. In S. M. Kosslyn & D. N. Osherson (Eds.), *An invitation to cognitive science: Visual cognition*. Cambridge, MA: MIT Press.

Biederman, I., Blickle, T. W., Teitelbaum, R. C., & Klatsky, G. J. (1988). Object search in non-scene arrays. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 456-467.

Biederman, I., Glass, A. L., & Stacy, E. W. Jr. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology, 97*, 22-27.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene Perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*, 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*, 597-600.

Bock, K., Irwin, D. E., & Davidson, D. J. (2004). Putting first things first. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Boyce, S. J, Pollatsek A, Rayner K. (1989). Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception and Performance,15*: 556-66.

Bridgeman, B., Hendry, D., & Stark, L. (1975). Failure to detect displacements of the visual world during saccadic eye movements.*Vision Research, 15*, 719-722.

Bridgeman, B., & Stark, L. (1979). Omnidirectional increase in threshold for image shifts during saccadic eye movements. *Perception & Psychophysics, 25*, 241-243.

Bullier, J. (1997). Organisation anatomique et fonctionnelle des voies visuelles. *Cours, Ecole de Printemps des Neurosciences et Sciences de l'Ingénieur sur la perception visuelle.*

Bullier, J. (1998). Architecture fonctionnelle du système visuel. *In : Boucart, Hennaff & Belin (eds) La vision : aspects perceptifs et cognitifs. Edition SOLAL Neuropsychologie.*

Bülthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex, 5*(3), 247-260.

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.

Buser, P. & Imbert, M. (1987). *Vision*, Hermann Edition, Paris.

Buswell, G. T. (1935). *How people look at pictures*. Chicago: University of Chicago Press.Carpenter, R. H. S. (1988). *Movements of the eyes*. London: Pion.

Castelhano, M. S., & Henderson, J. M. (2005). Incidental Visual Memory for Objects in Scenes. *Visual Cognition, 12*, 1017-1040.

Chauvin, A. (2003). *Perception des scènes naturelles : étude et simulation du rôle de l'amplitude, de la phase et de la saillance dans la catégorisation et l'exploration de scènes naturelles.* Thèse de doctorat, Université Joseph Fourier, Grenoble.

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences, 4*, 170-178.

Chun, M. M. & Nakayama, K. (2000). On the functional role of implicit Visual memory for the adaptative deployment of attention cross scenes. *Visual Cognition, 7*, 65-81.

Currie, C., McConkie, G., Carlson-Radvansky, L. A., & Irwin, D. E. (2000). The role of the saccade target object in the perception of a visually stable world. *Perception & Psychophysics, 62*, 673-683.

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.

De Graef, P. (1992). Scene-context effects and models of real-word perception. In K. Rayner (Eds.), *Eye movements and visual cognition: Scene perception and reading* (pp. 243-259): Springer-Verlag.

De Graef, P, Christiaens D, & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research, 52*, 317-329.

Delorme, A., Richard, G., & Fabre-Thorpe, M. (2000). Ultra-rapid categorization of natural scenes does not reply on color cues: A study in monkeys and humans. *Vision Research, 40*, 2187-2200.

Delorme, A., Flückiger, M. (2003). *Perception et réalité*. Belgique : De Boeck.

Dougherty, R. F., Kocj, V. M., Brewer, A. A., Fischer, B., Modersitzki, J., & Wandell, B. A. (2003). Visual field representations and locations of visual areas V1/2/3 in human cortex. *Journal of Vision, 3*(10), 568-598.

Duncan, J., Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review, 96*(3), 433-458.

Duncan, J., Humphreys, G., & Ward, R. (1997). Competitive brain activity in visual attention. *Current opinion in Neurobiology*, 7(2), 255-261.

Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a sliency map for overt visual attention ? *European Journal of Neuroscience, 17*, 1089-1097.

Engel, S. A., Glover, G. H., & Wandell, B. A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cerebral Cortex, 7*, 181- 192.

Epstein, R. (2005). The cortical basis of visual scene processing. In J. M. Henderson (Eds.), *Real-Word Scene Perception* (pp. 954-978). New York: Psychology Press.

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*, 865-876.

Epstein, R., Harris, A., Stanley, D., & Kanwisher, N. (1999). The parahippocampal place area: Recognition, navigation, or encoding. *Neuron, 23*, 115-125.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*, 598-601.

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*, 171-180.

Farah, J. M. (1990). *Visual agnosia: disorders of object recognition and what they tell us about normal vision*. Cambridge: MIT Press.

Fei-Fei, Li, Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *IEEE C VPR 2004, Workshop on Generative-Model Based Vision*. 2004.

Fei Fei, Li., VanRullen, R., Koch, C. & Perona, P. (2002). Natural scene categorization in the near absence of attention. In*: Proceedings of the National Academy Sciences, 99*(14), pp 9596-9601.

Fei Fei, Li., VanRullen, R., Koch, C. & Perona, P. (2005). Why does natural scene categorization require little attention? In J. M. Henderson (Eds.), *Real-Word Scene Perception* (pp. 893-924). New York: Psychology Press.

Felleman, D. (1981). A comparison of the receptive field properties of neurons in the middle temporal visual area (MT) and striate cortex of the owl monkey, Aotus trivirgatus. *PhD. Dissertation, department of psychology, Vanderbilt University, Nashville TN.*

Fodor, J. A. (1983). *Modularity of Mind. Cambridge*, MA: MIT Press.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General, 108*, 316-355.

Friedman, A., & Liebelt, L. S. (1981). On the time course of viewing pictures with a view towards remembering. In D. F. Fisher, R. A. Monty, & J. W. Senders (Eds.), *Eye movements: Cognition and visual perception* (pp. 137-155). Hillsdale, NJ: Erlbaum.

Gaillard, J. P., & Bourges, P (1999). Effet du filtrage spatial sur la reconnaissance de paysages. *L'année psychologique*, *99,* 415-445.

Gaillard, J. P., Boulliou, R., & Gautier, C. (1996). Théorie des "géons" et interprétation quantitative dans une tâche d'indentification d'objets. *L'année psychologique, 96*, 561-586.

Greene, M. R., & Oliva, A. (2006). Natural scene categorization from the conjunction of ecological global properties. *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society,* 291-296

Grimes, J. (1996). On the failure to detect changes in scenes across saccades. In K. Akins (Ed.), *Perception: Vancouver studies in cognitive science* (pp. 89-110). Oxford: Oxford University Press.

Guérin-Dugué, A., & Oliva, A. (2000). Classifications of scene photographs from local orientations features. *Pattern Recognition Letters, 21*, 1135-1140.

Guyader, N. (2004). *Scènes visuelles: Catégorisation basée sur des modèles de perception. Approches (neuro) computationnelle et psychophysique.* Thèse de doctorat, Université Joseph Fourier, Grenoble.

Hartline, H. K. (1938). The response of single optic nerve fibers of the vertebrate eye to illumination of the retina, *American Journal of Physiology,* 121, pp. 400–415.

Hayhoe, M. M., Bensinger, D. G., & Ballard, D. H. (1998). Task constraints in visual working memory. *Vision Research, 38*, 125-137.

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision, 3*, 49-63.

Henderson, J. M. (2003). Human gaze control in real-world scene perception. *Trends in Cognitive Sciences, 7*, 498-504.

Henderson, J. M. & F. Ferreira. (2004). Scene perception for psycholinguists. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Henderson, J. M., & Hollingworth, A. (1998). Eye movements during scene viewing: An overview. In G. Underwood (Ed.), *Eye Guidance in Reading and Scene Perception* (pp. 269-283). Oxford: Elsevier.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology. 50*, 243-271.

Henderson, J. M., & Hollingworth, A. (1999a). The role of fixation position in detecting scene changes across saccades. *Psychological Science, 10*, 438-443.

Henderson, J. M., & Hollingworth, A. (1999b). 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*, 58-71.

Henderson, J. M., Weeks, P. A. Jr., & Hollingworth, A. (1999). Effects of semantic consistency on eye movements during scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 210-228.

Henderson, J. M., Williams, C. C., Castelhano, M. S., & Falk, R. J. (2003). Eye movements and picture processing during recognition. *Perception & Psychophysics, 65*, 725-734.

Hérault, J. (2001). De la rétine biologique aux circuits neuromorphiques. In : J. M. Jolion (Eds), *Les Systèmes de Vision. Edition Hermès.*

Hérault, J., Beaudot, W., & Oliva, A. (1995). *Perception Coarse-to-fine par un modèle de rétine.* Paper presented at the GRETSI, France.

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. (2004). Constructing visual representations of natural scenes: The roles of short- and long-term visual memory. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 519-537.

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General, 127*, 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 (Special Issue on Object Perception and Memory)*, 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*, 113-136.

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*, 761-768.

Hubel, D. & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, *Journal of Physiology*, 160, pp. 106–154.

Hubel, D. & Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology,* 195, pp. 215-243.

Hubel, D. H. & Wiesel, T.N. (1974). Sequence regularity and geometry of orientation columns in monkey striate cortex. *Journal of Computational Neurology,* 158, pp. 267-293.

Imbert, M. (1983).La neurobiologie de l'image, *La recherche*, vol. 14, pp. 600-13.

Intraub, H. (1979). Presentation rate and the representation of briefly glimpsed pictures in memory. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 78-87.

Intraub, H. (1980). The role of implicit naming in pictorial encoding. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 1-12.

Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 604-610.

Intraub, H. (1997). The representation of Visual Scenes. *Trends in the Cognitive Sciences, 1,* 217-221.

Intraub, H. (1999). Understanding and remembering briefly glimpsed pictures: Implications for visual scanning and memory. In Coltheart, V. (Ed.), *Fleeting Memories: Cognition of Brief Visual Stimuli*, (pp. 47-70), Cambridge, Massachusetts: MIT Press.

Intraub, H., Bender, R. S., & Mangels, J. A. (1992). Looking at pictures but remembering scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 180-191.

Irwin, D. E. (1992). Visual memory within and across fixations. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 146-165). New York: Springer-Verlag.

Irwin, D. E. (2004). Fixation location and fixation duration as indices of cognitive processing. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Irwin, D. E., & Yeomans, J. M. (1986). Sensory registration and informational persistence. *Journal of Experimental Psychology: Human Perception and Performance, 12,* 343-360.

Irwin, D. E., & Zelinsky, G. J. (2002). Eye movements and scene perception: Memory for things observed. *Perception & Psychophysics, 64*, 882-895.

Itti, L. (2000). *Models of Bottom-up and Top-down visual attention.* California Institute of Technology, Pasadena.

Itti, L., Gold, C., & Koch, C. (2001). Visual attention and target detection in cluttered natural scenes. *Optical Engineering, 40*, 1784-1793.

Itti., L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*, 1489-1506.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews: Neuroscience, 2*, 194-203.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern Analysis and Machine Intelligence. 20*, 1254-1259.

Jung, R. & Spillmann, L. (1970). Receptive-field estimation and perceptual integration in human vision. In F.A. Young and D.B. Lindsley (Eds)*, Early experience and visual information processing in perceptual and reading disorders.* Washington D.C.: National Academie of Sciences.

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

Koch, C., & Ullman, S. (1984). *Selecting one among the many: a simple network implementing shifts in selective visual attention* (No. A.I. Memo 770, C. B. I P Paper 003): Massachusetts Institute of technology.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiolog, 4*, 219-227.

Kosslyn, S. M. (1994). *Image and Brain*. Cambridge, MA: MIT Press.

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetsche, C. (2000). Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial Vision, 13*, 201-214.

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology,* 16, pp. 37-68.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research, 41*, 3559-3565.

Land, M. F., & Lee, D. N. (1994). Where we look when we steer. *Nature, 369*, 742-744.

Land, M. F., Mennie, N., & Rusted, J. (1999). Eye movements and the roles of vision in activities of daily living: Making a cup of tea. *Perception, 28*, 1311-1328.

Levy, I., Hasson, U., Avidan, G., Hendler, T., & Malach, R. (2001). Center-periphery organization of human object areas. *Nature Neuroscience, 4*, 533-539.

Lieury, A. (1995). Mémoire de images et double codage. *L'année Psychologique, 95*, 661-673.

Lieury, A. (2005). *Psychologie de la mémoire : Histoire, théories, expérience*. Paris : Dunod.

Lieury, A., & Calvez, F. (1986). Le double codage des dessins en fonction du temps de présentation et de leur ambigüité. *L'année Psychologique, 86*, 45-61.

Liu, A. (1998). What the driver's eye tells the car's brain. In G. Underwood (Ed.), *Eye Guidance in reading and scene perception* (pp. 431-452). Oxford: Elsevier.

Livingstone, M.S. & Hubel, D.H. (1987). Psychophysical evidence for separate channels for the percetion of form, color, movement, and depth. *J of Neuroscience*, 7, 3416-3468.

Livingstone, M.S. & Hubel, D.H. (1988). Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science*, 240, 740-749.

Loftus, G. R. (1985). Picture perception: Effects of luminance on available information and information-extraction rate. *Journal of Experimental Psychology: General, 114*, 342-356.

Loftus, G. R., Kaufman, L., Nishimoto, T., & Ruthruff, E. (1992). Effects of visual degradation on eye-fixation durations, perceptual processing, and long-term visual memory. In K. Rayner (Ed). *Eye movements and visual cognition: Scene perception and reading* (pp. 203-226). New York: Springer.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 565-572.

Loftus, G. R., & Nelson, W. W., & Kallman, H. J. (1983). Differential acquisition rates for different types of information from pictures. *Quarterly Journal of Experimental Psychology, 35A*, 187-198.

Lorant-Royer, S., & Lieury, A. (2003). La mémoire visuospatiale est-elle tridimensionnelle ? *Bulletin de Psychologie, 56*, 357-365.

Luck, S. J., Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390, 279-281.

Macé, M. J.-M. (2006). *Représentations visuelles précoces dans la catégorisation rapide de scènes naturelles chez l'homme et le singe.* Thèse de doctorat, Université Paul Sabatier, Toulouse.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics, 2*, 547-552.

Mannan, S., Ruddock, K. H., & Wooding, D. S. (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images. *Spatial Vision, 9*, 363-386.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision, 10*, 165-188.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997a). Fixation sequences made during visual examination of briefly presented 2D images. *Spatial Vision, 11*, 157-178.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997b). Fixation patterns made during brief examination of two-dimensional images. *Perception, 26*, 1059-1072.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society, London, B, 200*, 269-291.

Maunsell, J. & Newsome, W. (1987). Visual processing in monkey extra striate cortex, *Annual Review of Neuroscience*? 10, pp .363-401.

McCauley, C., Parmelee, C. M., Sperber, R. D., & Carr, T. H. (1980). Early extraction of meaning from pictures and its relation to conscious identification. J*ournal of Experimental Psychology: human perception and Performance, 6*, 265-276.

McConkie, G. W. (1990). *Where vision and cognition meet*. Paper presented at the Human Frontier Science Program Workshop on Object and Scene Perception, Leuven, Belgium.

McConkie, G. W. (1991). Perceiving a stable visual world. In J. Van Resnbergen, M. Devijver, & G. d'Ydewalle (Eds.), *Proceedings of the sixth European conference on eye movements* (pp. 5-7). Leuven, Belgium: Laboratory of Experimental Psychology.

McConkie, G. W., & Currie, C. B. (1996). Visual stability while viewing complex pictures. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 563-581.

McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics, 17*, 578-586.

Melcher, D. (2001). Persistence of visual memory for scenes. *Nature*, 412, 401-402.

Mendez, M. F., & Cherrier, M. M. (2003). Agnosia for scenes in topographagnosia. *Neuropsychologia, 41,* 1387-1395.

Movshon, J. A. (1978). The Hypercomplexities in the visual cortex. *Nature, 272*(23), 305-306.

Movshon, J. A., Thompson, I. D. & Tolhurst, D. J. (1978). Receptive field organization of complex cells in the cat's striate cortex, *Journal of Physiology,* 283, pp. 79-99.

Murphy, G. L., & Wisniewski, E. J. (1989). Categorizing objects in isolation and in scenes: What a superordinate is good for. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 4*, 572-586.

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 ET study. *Brain, 123* (9), 1903-1912.

Niebur, E., Itti, L., & Koch, C. (2002). Controlling the focus of visual selective visual attention. *In: Models of Neural Networks IV*, Springer Verlag.

Nowak, L. G., Munk, M. H., Girard, P. & Bullier, J. (1995). Visual latencies in areas V1 and V2 of the macaque monkey. *Vision Neuroscience,* 12, pp. 371-84.

Nowak, L. G, & Builler, 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.

O'Craven, K. M., & Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of Cognitive Neuroscience, 12*, 1013-1023.

Oliva, A. (2005). Gist of a scene. In L. Itti, G. Rees & J. K. Tsotsos (Eds.), *Neurobiology of Attention*. San Diego, CA : Elsevier.

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*, 72-107.

Oliva, A., & Schyns, P. G. (2000). Colored diagnostic blobs mediate scene recognition. *Cognitive Psychology, 41*,176-210.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal in Computer Vision, 42*, 145-175.

Oliva, A., & Torralba, A. (2003). Scenecentered description from spatial envelope properties. In H. H. Bulthoff et al. (Eds.), *Lecture notes in computer science: Biologically motivated computer vision*. New York: Springer-Verlag.

Oliva, A., & Torralba, A., Castelhano, M. S., & Henderson, J. M. (2003). Top-down control of visual attention in object detection. *IEEE Proceedings of the International Conference on Image Processing, 1*, 253-256.

Oliva, A., Torralba, A. B., Guérin-Dugué, A., & Hérault, J. (1999). *Super-Ordinate Representation of Scenes from Power Spectrum Shapes.* Paper presented at the Challenge of Image Retrieval (CIR99), Newcastle.

O'Regan, J. K. (2001). Thoughts on change blindness. In L. Harris & M. Jenkin (Eds.), *Vision and Attention* (pp. 281-302). New York: Springer Verlag.

O'Regan, J. K., Deubel, H., Clark, J. J., & Rensink, R. A. (2000). Picture change during blinks: Looking without seeing and seeing without looking. *Visual Cognition, 7,* 191-211.

O'Regan, J. K., Rensink, R. A., & Clark, J. J. (1999). Change-blindness as a result of 'mudsplashes'. *Nature*, 398, 34-35.

Palmer, S. E. (1975). Visual perception and world knowledge: notes on a model of sensory-cognitive interaction. In *Explorations in Cognition*, ed. DA Norman, DE Rumelhart, LNR Res. Group, pp. 279-307. San Francisco: Freeman.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*, 107-123.

Pelz, J. B., & Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research, 41*, 3587-3596.

Potter, M. C. (1975). Meaning in visual search. *Science, 187,* 965-966.

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

Potter, M. C. (1993). Very short-term conceptual memory. *Memory & Cognition, 21*, 156-161.

Potter, M. C. (1999). Understanding sentences and scenes: The role of conceptual short-term memory. In V. Coltheart (Ed.), *Fleeting memories* (pp. 13-46). Cambridge, MA: MIT Press.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology, 81*, 10-15.

Potter, M. C., Staub, A., & O'Connor, D. H. (2004). Pictorial and conceptual representation of glimpsed pictures. J*ournal of Experimental Psychology: Human Perception and Performance, 30,* 478-489.

Potter, M. C., Staub, A., Rado, J., & O'Connor, D. H. (2002). Recognition memory for briefly-presented pictures: The time course of rapid forgetting. *Journal of Experimental Psychology: Human Perception and Performance*, 28, 1163-1175.

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences, 3*, 111-169.

Pylyshyn, Z. W. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral & Brain Sciences*, 22, 341-423.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network: Computer and neural systems, 10*, 1-10.

Rensink, R. A. (2000a). The dynamic representation of scenes. *Visual Cognition*, 7, 17-42.

Rensink, R. A. (2000b). Seeing, sensing and scrutinizing. *Vision Research*, *40*, 1469-1487.

Rensink, R. A. (2002). Change detection. *Annual Review of Psychologye, 53*, 245-277.

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.

Rogowitz, B. E., Frese, T., Smith, J., Bouman, C. A., & Kalin, E. (1997). Perceptual image similarity experiments. In SPIE *Conference on Human Vision and Electronic Imaging*, pp. 576-590.

Rolls, E. T., & Deco, G. (2002). *Computational neurosciences of vision*. Oxford: Oxford university Press.

Rosch, E. (1977). Human categorization. In *Studies in cross-cultural psychology,* N. Warren, Ed. Academic Press, London.

Rossi, J. P. (2005). *Psychologie de la mémoire : De la mémoire épisodique à la mémoire sémantique.* Belgique : De Boeck.

Rossion, B., & Gauthier, I. (2002). How does the brain process upright and inverted faces? *Behavrioral and Cognitive Neuroscience Review, I(1),* 62-74.

Rousselet, G. A., Joubert, O. R., & Fabre-Thorpe. M. (2005). How long to get to the "gist" of real-word natural scenes? In J. M. Henderson (Eds.), *Real-Word Scene Perception* (pp. 852-877). New York: Psychology Press.

Ryan, J. D., Althoff, R. R., Whitlow, S., & Cohen, N. J. (2000). Amensia is a deficit in relational memory. *Psychological Science, 11*, 454-461.

Sanocki, T., & Epstein, W. (1997). Priming spatial layout of scenes. *Psychological Science, 8*, 374-378.

Sato, N., Nakamura, K., 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.

Schiller, P. H. (1997). Past and present ideas about how the visual scene is analyzed by the brain. In K. a. Rockland. (Ed.), *Cerebral Cortex* (Vol. 12): Plenum.

Shinoda, H., Hayhoe, M. M., & Shrivastava, A. (2001). What controls attention in natural environments? *Vision Research, 41*, 3535-3545.

Schmolesky, M.T., Wang, Y., Hanes, D.P., Thomson, K.G., Leutgeb, S., Schall, J.D. & Leventhal, A.G. (1998). Signal timing across the macaque visual system. *Journal of Neurophysiology,* 79(6), pp. 3272-3178.

Schnapf, J. L., Kraft, T.W., Nunn, B.J. & Baylor D.A. (1988). Spectral sensitivity of primate photoreceptors. *Visual neuroscience,* 1, pp. 255-261.

Schyns, P. G. (1998). Diagnostic recognition: task constraints, object information, and their interactions. *Cognition, 67*(1-2), 147-179.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science, 5*, 195-200.

Simons, D. J., & Levin, D.T. (1997). Change blindness. *Trends in Cognitive Sciences, 1*, 261-267.

Spivey, M. J., Richardson, D. C., & Fitneva, S. A. (2004). Thinking outside the brain: Spatial indices to visual and linguistic information. In J. M. Henderson, and F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Stark, L., & Ellis, S. R. (1981). Scanpaths revisited: Cognitive models direct active looking. In D. F. Fisher, R. A., Monty, & J. W. Senders (Eds.), *Eye movements: Cognition and visual perception* (pp. 193-226). Hillsdale, NJ: Erlbaum.

Steeves, J. K. E., Humphrey, G. K., Culham, J. C., Menon, R. S., & Goodale, M. A. (2004). Behavioral and neuroimaging evidence for a contribution of color information to scene recognition in a patient with impaired form recognition. *Journal of Cognitive Neuroscience, 16, 955-965.*

Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 19-30). New York: Cambridge University Press.

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

Torralba, A. (2003). Modelling global scene factors in attention. *Journal of the Optical Society of America, 20*, 1407-1418.

Torralba, A.,& Oliva., A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems, 14,* 391-412.

Treisman, A. (1998). The perception of features and objects. In R. D. Wright (Eds.), *Visual Attention* (PP. 26-54). New York Oxford: Oxford University Press.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*, 97-136.

Trueswell, J., & Gleitman, L. (2004). Children's eye movements during listening: Developmental evidence for a constraint based theory of sentence

processing. In J.M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Tversky, B., & Hemenway, K. (1983). Categories of environmental scenes. *Cognitive Psychology, 15*, 121-149.

Ullman, S. (1996). *High-Level Vision: Object Recognition and Visual Cognition*. Cambridge, MA: MIT Press.

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., & Haxby, J. (1994). What and where in the human brain. *Current Opinion in Neurobiology. 4*, 157-165.

van Diepen, P. M. J., De Graef, P., & d'Ydewalle, G. (1995). Chronometry of foveal information extraction during scene perception. In J. M. Findlay, R. Walker, & R. W.Kentridge (Eds.), *Eye movement research: Mechanisms, processes and applications* (pp. 349-362). Amsterdam: Elsevier.

van Diepen, P. M. J., Wampers, M., & d'Ydewalle, G. (1998). Functional division of the visual field: moving masks and moving windows. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 337-355). Oxford: Elsevier.

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*, 655-668.

Vogel, E. K., Woodman, G. F., & luck, S. J. (2001). Storage of features, conjunctions and objects in visual working memory. *Journal of Experimental Psychology: human Perception and Performance, 27*, 92-114.

Watanabee, K. (2003). Differential effect of distractor timing on localizing versus identifying visual changes. *Cognition, 88*, 243-257.

Wolfe, J. M. (1998). What do you know about what you saw? *Current Biology*, 8, R303-R304.

Wolfe, J. M. (1999). Inattentional amnesia. In V. Coltheart (Ed.), *Fleeting Memories* (pp. 71-94). Cambridge, MA: 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., & Bennett, S. C. (1997). Preattentive object files: shapeless bundles of basic features. *Vision Research,* 37(1), 25-43.

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.

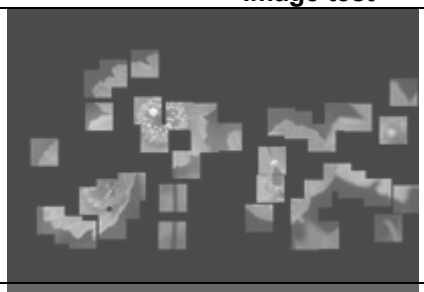
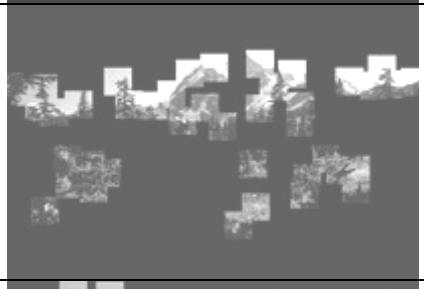
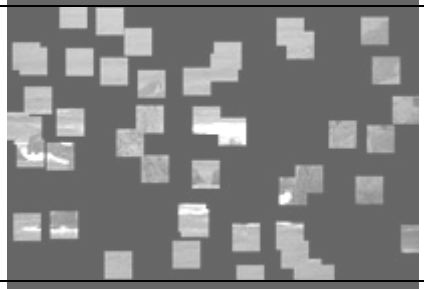
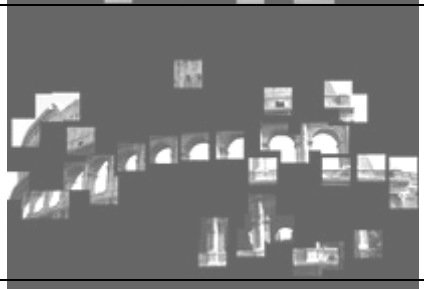
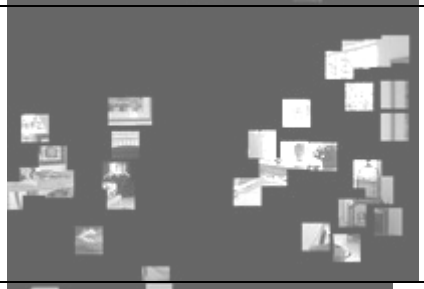
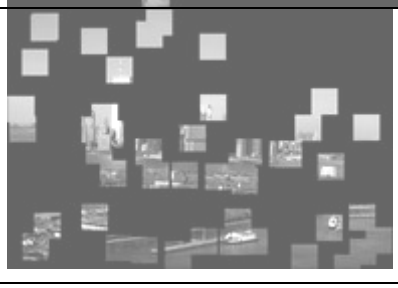
Wolfe, J. M., & Gancarz, G. (1996). Guided Search 3.0: A model of visual search catches up with Jay Enoch 40 years later. In V. Lakshminarayanan (Ed.), *Basic and clinical applications of vision science* (pp. 189-192). Dordrecht, Netherlands: Kluwer Academic.

Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum Press.

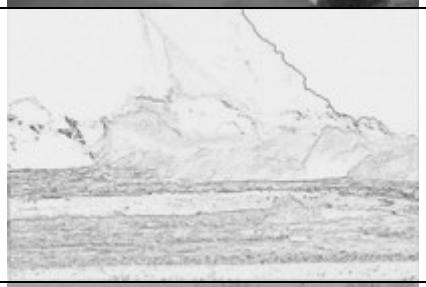
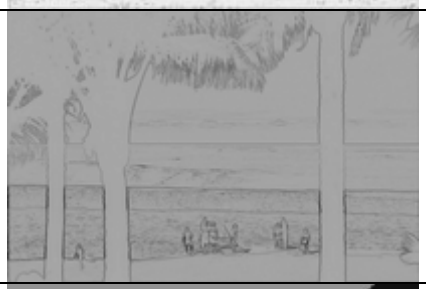
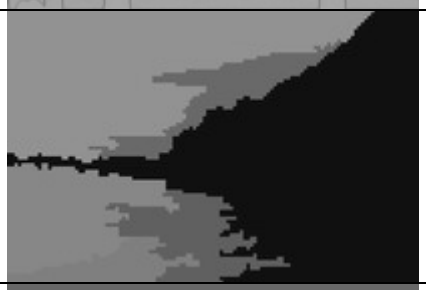
ANNEXES[9]

[9] N'ayant pas les droits de diffusion des images utilisées dans les cinq expérimentations, voici quelques extraits de ces images

ANNEXE 1 : CONSTRUCTION DU MATERIEL DE L'EXPERIENCE 1
Tâche de reconnaissance

Images naturelles			
	Image cible	Image test	
1			Même image point d'intérêt décroissant
2			Image différente point d'intérêt croissant
3			Même image point d'intérêt aléatoire
4			Image différente point d'intérêt décroissant
5			Même image point d'intérêt croissant
6			Image différente point d'intérêt aléatoire

ANNEXE 1 : CONSTRUCTION DU MATERIEL DE L'EXPERIENCE 2
Tâche de reconnaissance

Images naturelles			
Image cible		Image test	
1			Même image originale
2			Image différente contour
3			Même image contour
4			Image différente luminance
5			Même différente luminance

ANNEXE 3 : CONSTRUCTION DU MATERIEL DE L'EXPERIENCE 3 et 4
Tâche de catégorisation

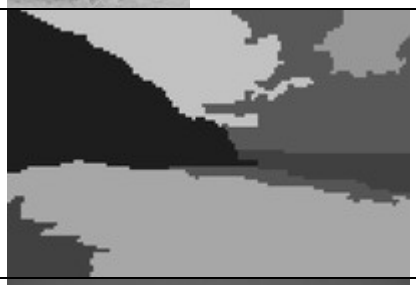
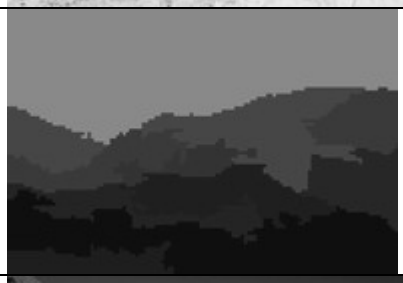
	Image cible	Image test	
1			Même catégorie originale 50 ms (expérience 3) 20 ms (expérience4)
2			Même catégorie contour 50 ms (expérience 3) 20 ms (expérience4)
3			Même catégorie luminance 50 ms (expérience 3) 20 ms (expérience4)
4			Même catégorie originale 150 ms (expérience 3) 35 ms (expérience 4)
5			Même catégorie contour 150 ms (expérience 3) 35 ms (expérience 4)
6			Même catégorie originale 150 ms (expérience 3) 35 ms (expérience 4)

	Image cible	Image test	
7			Catégorie différente originale 50 ms (expérience 3) 20 ms (expérience4)
8			Catégorie différente contour 50 ms (expérience 3) 20 ms (expérience4)
9			Catégorie différente luminance 50 ms (expérience 3) 20 ms (expérience4)
10			Catégorie différente originale 150 ms (expérience 3) 35 ms (expérience 4)
11			Catégorie différente contour 150 ms (expérience 3) 35 ms (expérience 4)
12			Catégorie différente originale 150 ms (expérience 3) 35 ms (expérience 4)

ANNEXE 4 : CONSTRUCTION DU MATERIEL DE L'EXPERIENCE 5
Tâche de catégorisation

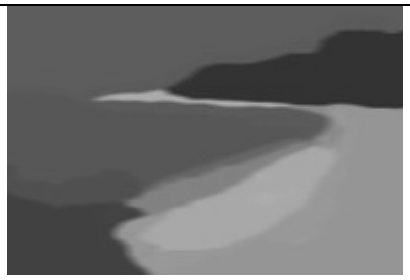
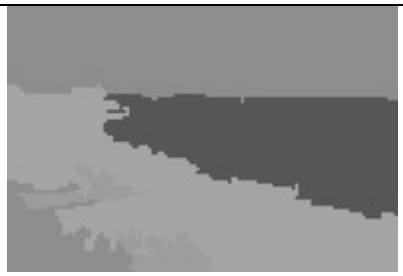
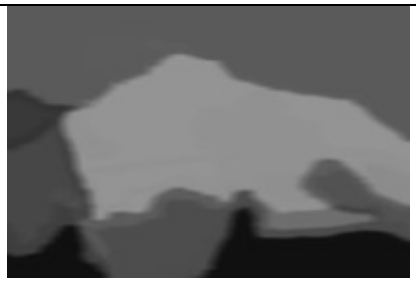
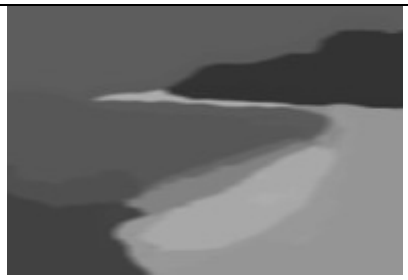
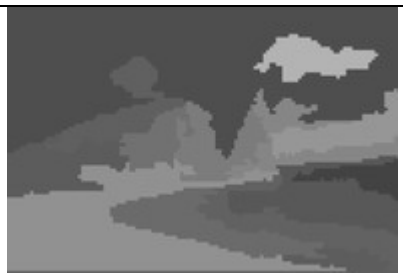
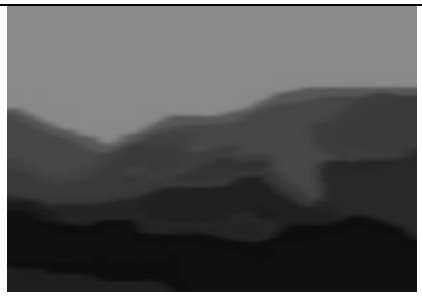
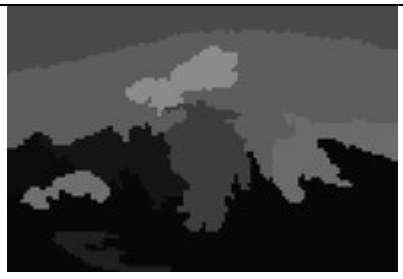
	Image cible	Image test	
1			Même catégorie Luminance avec lissage (35 ms ou 50 ms ou 150 ms)
2			Même catégorie Luminance sans lissage (35 ms ou 50 ms ou 150 ms)
3			Même catégorie Luminance avec lissage (35 ms ou 50 ms ou 150 ms)
4			Même catégorie Luminance sans lissage (35 ms ou 50 ms ou 150 ms)

	Image cible	Image test	
5			Catégorie différente Luminance avec lissage (35 ms ou 50 ms ou 150 ms)
6			Catégorie différente Luminance sans lissage (35 ms ou 50 ms ou 150 ms)
7			Catégorie différente Luminance avec lissage (35 ms ou 50 ms ou 150 ms)
8			Catégorie différente Luminance sans lissage (35 ms ou 50 ms ou 150 ms)

Index des figures

Index des tableaux

Traitements cognitifs mis en jeu dans la perception visuelle de scènes complexes et conséquences sur l'indexation automatique d'images

Résumé : Chez les sujets humains, la capacité d'identifier une scène visuelle complexe est remarquable. Avec une seule fixation d'une scène, de nombreuse d'informations sont disponibles : son contenu, son identité, sa structure spatiale et la catégorisation de cette scène (Potter, 1975; Schyns & Oliva, 1994; Thorpe, Fize, & Marlot, 1996). Plusieurs hypothèses sont développées : Tout abord, le sens général d'une scène peut être acquis grâce a l'identification d'un ou plusieurs objets typiques (Friedman, 1979) et leurs relations (De Graef, Christiaens, & d'Ydewalle, 1990). Alternativement à ce point de vue traditionnel considérant que l'identification d'une scène s'effectue par l'identification des objets qu'elle contient, une autre hypothèse suppose que l'identification d'une scène peut être acquise au moyen d'informations globales ayant des propriétés de "scène-niveau" donc sans savoir besoins d'informations portant sur les objets (Greene & Oliva, 2006; Schyns & Oliva, 1994; Oliva & Schyns, 2000). Ces propriétés de "scène-niveau" se caractériseraient par des grandes surfaces structurales ou d'autres types d'informations similaires (Biederman, 1995).

En outre, l'étude des mouvements oculaires montrent que les fixations précoces sont influencées par la densité de contours, le contraste local (Mannan, Ruddock, & Wooding, 1996, 1997; Reinagel & Zador, 1999) ainsi que la structure de la scène (Sanocki & Epstein, 1997; Castelhano & Henderson, 2003; Oliva & Torralba., 2003). Nous supposons donc que ces deux types d'informations fusionneraient en une représentation des différentes zones de luminance ne contenant pas d'informations précises sur les objets. Le but de cette thèse est d'étudier le rôle de la structuration spatiale des différentes zones de luminance dans l'identification rapide de scènes complexes. Les résultats suggèrent que les participants sont capables d'identifier une scène visuelle en se basant sur cette propriété.

Most clés : scène visuelle complexe, sens générale d'une scène, propriétés de "scène-niveau", contour, structuration spatiale des différentes de zones de luminance, identification de scène

Involving cognitive treatment in Visual perception of complex scenes and impacts on the automatic Indexation of images

Abstract: Human scene understanding is remarkable because, with only a brief glance at an image, an abundance of information is available: image content and meaning, spatial layout and semantic label (Potter, 1975; Schyns & Oliva, 1994; Thorpe, Fize, & Marlot, 1996) etc. Currently, several hypotheses have been advanced to explain how scenes are recognized so quickly. First, it could be that a diagnostic object is rapidly identified, and that the scene gist is inferred with from this object (Friedman, 1979) or a few objects and their spatial relationships (De Graef, Christiaens, & d'Ydewalle, 1990). Contrary to the traditional ideas of research in scene understanding that treat objects as the atoms of recognition, the real world scenes can be recognized without necessarily identifying the objects they contain (Greene & Oliva, 2006; Schyns & Oliva, 1994; Oliva & Schyns, 2000). There are some scene-level features that directly suggest identity and gist without requiring identification of any of the specific objects or any specific spatial relationships among them. Past suggestions for these features include large volumetric shapes or other similar large-scale image features (Biederman, 1995).

Studies of eye movement in scene recognition have shown that two kinds of information can be coded and stored during the early stages of low-level cognitive processing of complex scenes. These are contour density, local contrast (Mannan, Ruddock & Wooding, 1996, 1997; Reinagel & Zador, 1999), and global layout information (Sanocki & Epstein, 1997; Castelhano & Henderson, 2003; Oliva & Torralba., 2003). These two types of information are manipulated to transform the image into a "structural luminance image". The purpose of this work was to investigate how the "structural luminance image" is used by humans to process information in a real-world scene. The finding from extensive experiments demonstrates that subjects are able to identify natural scenes based on large structural regions of different luminance.

Keywords: real world scenes, scene gist, scene-level features, contour, large structural regions of different luminance, scene identification

Additif à la thèse de Jingqiang LI

Traitements cognitifs mis en jeu dans la perception visuelle de scènes complexes et conséquences sur l'indexation automatique d'images

Jingqiang LI

Li, Jingqiang. Traitements cognitifs mis en jeu dans la perception visuelle de scènes complexes et conséquences sur l'indexation automatique d'images, 2008

Sommaire

1) Saccades précoces, sémantique de scène et type de tâche à réaliser

Kirchner et Thorpe (2006) montrent que les sujets peuvent détecter et piloter leurs saccades vers l'animal présenté dans une scène avec un temps d'environ 120 ms, avec un taux de bonnes réponses de 90,1%. Dans leur expérience, deux images (une cible et un distracteur) sont présentées simultanément en 20 ms (sans masque) sur la partie gauche et droite de l'écran d'un ordinateur, chaque image est présentée avec un angle visuel de 6° par rapport à un sujet se plaçant devant l'écran d'un ordinateur à environ 80 cm. On a demandé aux sujets de fixer le plus vite possible où se trouve l'image contenant un animal (paradigme de l'orientation de saccades à choix forcé).

Afin de tester si les saccades sont déterminées par les attributs physiques de l'image cible et du distracteur (une autre image présentée en même temps que l'image cible), Kirchner et Thorpe (2006) ont effectué une analyse en employant la statistique des propriétés physiques entre les deux images. Ces analyses ne montrent pas que la performance des sujets est déterminée par les propriétés physiques des deux images (Johnson & Olshausen, 2005). Ces résultats conduisent Kirchner et Thorpe (2006) à conclure que le mécanisme de traitement serait basé sur des éléments locaux diagnostiques portant sur les animaux.

Des résultats similaires sont observés dans deux autres d'études ultérieures réalisées par Guyonneau, Kirchner, & Thorpe, (2006) et par Bacon-Macé, Kirchner, Fabre-Thorpe & Thorpe (2007). Dans l'étude de Guyonneau, Kirchner, & Thorpe (2006), le paradigme utilisé était presque identique que l'étude de Kirchner et Thorpe (2006). Ces chercheurs font varier la rotation (16 rotations) des images présentées. Une autre étude supplémentaire a été menée par Bacon-Macé, Kirchner, Fabre-Thorpe & Thorpe (2007), dans laquelle ces derniers introduisent des masques dans leur expérience. Les résultats montrent que les saccades précoces ne seraient pas du aux propriétés physiques entre l'image cible et le distracteur. Le mécanisme de traitement semble utiliser

plusieurs stratégies pour trouver les informations pertinentes. Les saccades seraient pré-activées par un processus sémantique ou par le but de la tâche.

2) Analyse du temps de traitement

Dans la plupart des études portant sur le temps de traitement d'image, le temps de réaction mesuré comprend plusieurs étapes de traitement, à savoir le temps de traitement d'informations visuelles et le temps d'exécution manuel afin de donner une réponse (appuyer un bouton, par exemple). Kirchner et Thorpe (2006) dans leur étude montrent que les saccades peuvent se déplacer vers l'image contenant un animal avec un temps minimum de 120 ms. Ce dernier inclut aussi le temps de préparation pour les saccades, qui est estimé à environ 20 ms. Cela dit, le temps effectif alloué au traitement de l'information visuelle est approximativement de 100 ms. Un temps aussi court de traitement semble en faveur d'un traitement qui se baserait sur des attributs locaux diagnostiques lié à la présence d'animaux, par exemple, les parties spécifiques d'animal tel que l'œil ou le bec, etc. Par ailleurs, cette rapidité suggère que le processus de traitement serait aussi très simple et rapide. D'après ces auteurs, il pourrait exister un parcours rapide de traitement de V4 (60-80 ms) vers SC ("*superior colliculus*" 40-60 ms) tout en pilotant la saccade vers l'image où se trouve un animal.

Un temps de traitement visuel d'environ 100 ms ne signifie pas que la catégorisation implique ce temps (Rousselet, Macé, Thorpe, & Fabre-Thorpe, 2007). En effet, le mécanisme de traitement enchainé fonctionnerait avec une stratégie différente. C'est-à-dire que l'activité neuronale serait mobilisée dans le but de former une catégorisation grossière de l'objet, ou de créer une carte de saillance pilotant par la suite le processus de traitement (Macé, Thorpe, & Fabre-Thorpe, 2005). Ces mêmes auteurs font l'hypothèse qu'une tâche de catégorisation de haut-niveau (par exemple, une catégorisation des animaux) n'aurait pas besoin d'une représentation de "haut-niveau". Un niveau de représentation "modéré" suffirait pour aboutir à la réalisation de cette tâche (par exemple, la détection de la face d'un animal permettant son identification). Cette

idée est largement développée par certains chercheurs essayant de modéliser la catégorisation de scènes naturelles avec une représentation de bas-niveau du sens général de la scène (Torralba, Oliva, Castelhano, & Henderson, 2006; Wang, Zhang & Fei-Fei, 2006).

Les études mentionnées plus haut, menés par l'équipe du CerCo, révèlent plusieurs points d'intérêt. Premièrement, elles permettent de distinguer, et d'affiner le temps alloué au traitement de l'image et à celui alloué à l'exécution motrice de la réponse. Deuxièmement, ces études sont les premières recherches intégrant la scène visuelle complexe à la mesure des mouvements oculaires afin de déterminer le processus de traitement dans une tâche de catégorisation d'images. Troisièmement, elles montrent que les mouvements oculaires précoces sont indissociables des informations descendantes (par exemple, la sémantique de scène, la demande de la tâche à réaliser). Enfin, la performance étonnante des sujets dans cette série d'étude implique que le système de traitement semblerait être un processus parallèle ne sollicitant pas l'attention focalisée, et il serait en interaction entre la voie dorsale et la voie ventrale (Evans & Treisman 2005).

3) Effet contextuel et mouvements oculaires

Quand on cherche un objet dans une scène naturelle, notre attention est pilotée par les informations contextuelles. L'observateur préfère fixer les zones les plus probablement associées à la présence d'objets spécifiques : par exemple, des voiture si l'hypothèse est que la scène est une rue (Najemnik & Geisler, 2005). Autrement dit, après un seul coup d'œil, le sens général d'une scène peut être identifié. Donc, le contenu de la scène va influencer immédiatement les mouvements oculaires pour les diriger vers des régions saillantes (Brockmole, & Henderson, 2006, 2008; Eckstein, Drescher, & Shimozaki, 2006; Neider & Zelinski, 2006; Over, Hooge, Vlaskamp, & Erkelens, 2007; Torralba, Oliva, Castelhano, & Henderson, 2006).

Depuis ces dernières années, les modèles computationnels (notamment des modèles de carte de saillance) se sont développés. L'idée originale était de prédire les régions qui attirent les fixations oculaires. Ces modèles deviennent de plus en plus complets en intégrant plusieurs facteurs descendants parmi lesquels le rôle du contexte et la demande de la tâche à réaliser (Elazary & Itti, 2008; Navalpakkam & Itti, 2006). Le rôle du contexte sur la reconnaissance d'un objet a fait l'objet de nombreuses études dans le champ des sciences de l'ingénieur afin de le prendre en compte par un système de traitement d'image (Wang, Zhang, & Fei-Fei, 2006; Vogel & Schiele, 2006).

4) Reconnaissance d'objets et contexte de scènes

Dans une scène naturelle, les objets ne sont pas isolés les uns des autres. Ces derniers ont un lien indissociable avec le contexte de la scène. Une revue de la littérature dans différents domaines permet de montrer cette évidence (Aminoff, Gronau, &. Bar ,2007; Auckland, Cave, & Donnelly, 2007; Gronau, Neta, & Bar ,2008). Le rôle du contexte fournit plusieurs types d'informations telles que la sémantique de scène, la configuration spatiale et la relation spatiale entre les objets. Des études récentes montrent que le contexte donne également des informations concernant l'objet, sur la probabilité de sa présence, sa position et sa taille (Eckstein, Drescher, & Shimozaki, 2006; Oliva & Torralba, 2006, 2007; Peters & Itti, 2008; Torralba, Oliva, Castelhano, & Henderson, 2006).

Les informations contextuelles interviennent à deux niveaux. Tout d'abord, sur l'étape de préactivation des objets stockés en mémoire. Par exemple, lorsqu'un sujet regarde une scène, la reconnaissance d'une ferme faciliterait celle d'un tracteur par rapport à un calamar se trouvant dans la même scène. Ensuite, les informations contextuelles interviennent sur le niveau perceptuel de l'analyse de l'objet. En effet, l'effet du contexte interviendrait très précocement dans le processus de traitement visuel d'un objet. Par exemple, lors de l'intégration des traits locaux pour former un objet (Auckland, Cave, & Donnelly, 2007; Davenport, 2007). Cet effet de contexte pourrait initier des processus de

traitement d'image divers. Cette hypothèse est étayée par des données neuro-anatomiques montrant une activation préférentielle de certaines zones corticales impliquées dans l'analyse de la structuration spatiale de scène, ou bien l'analyse de l'objet et de la relation spatiale d'objets dans une scène (Gronau, Neta, & Bar ,2008).

Si on retient l'hypothèse d'un effet de contexte, une scène contenant des éléments homogènes, telles que la taille, la position ou la texture par exemple, aurait une représentation contextuelle qui, en retour, influencerait le mécanisme de traitement ultérieur. En effet, ce mécanisme de traitement semblerait effectuer une sorte de statistique portant sur les éléments ayant les mêmes types de propriétés visuelles, pour former une représentation en mémoire et ainsi influencer l'exploration de la scène. Des études récentes sont en faveur de cette hypothèse. Ariely (2001) présente séparément soit une série de points ayant des tailles différentes soit un seul point aux sujets. Les points sont affichés toutes les 500 ms. Le résultat montre que les sujets estiment mieux la taille des points lors qu'ils sont affichés collectivement que lorsqu'ils sont affichés individuellement. D'autres études similaires sont proposées par Chong et Treisman (2005a, 2005b). Dans cette série d'études, les chercheurs montrent la supériorité de la performance des éléments présentés collectivement comparés à ceux présentés individuellement. La performance reste inchangé quel que soit la densité des éléments présentés. D'après ces auteurs, cette sorte de statistique des éléments homogènes semblerait être un mécanisme automatique, il pourrait être opérationnel très rapidement pendant l'exploration de la scène. Comme le confirment Alvarez & Oliva (2008), ce mécanisme ne solliciterait pas l'attention focalisée.

5) Représentation de la scène en mémoire

La question de la représentation en mémoire des scènes visuelles complexes est largement étudiée avec le paradigme de "la cécité au changement" (Lleras, Rensink, & Enns, 2007; Simons & Rensink, 2005). La

question principale est de savoir comment est représentée une scène en mémoire, de façon grossière ou détaillée, de façon perceptive ou sémantique, d'un seul niveau ou plusieurs niveaux, et ce que cette représentation contient ? Jusqu'à aujourd'hui, ces questions restent ouvertes.

Pendant l'exploration d'une scène, il y aurait un mécanisme statistique d'apprentissage "*statistical learning mechanisms*" (Brady & Oliva, sous-presse) qui interviendrait afin de minimiser le coût cognitif de traitement : par exemple, lier les traits physiques à la sémantique de la scène, puis stocker. En un seul coup d'œil, une représentation de scène avec une invariance de la taille serait extraite et pourrait être maintenue en mémoire afin de piloter les saccades oculaires lors de l'exploration ultérieure de la scène (Castelhano & Henderson, 2007). Ce pilotage impliquerait donc une représentation plus détaillée de la scène que le suggèrent les travaux antérieurs (Simons, 2000, Wolf, 1999). Dans cette hypothèse, la représentation de la scène fournit des informations concernant l'endroit le plus probable où se trouveraient les objets (Torralba, Oliva, Castelhano, & Henderson, 2006). De plus, cette représentation resterait disponible et permettrait de diriger les mouvements oculaires vers des zones n'étant explorées qu'en cas de besoin (Gronau, Neta, & Bar ,2008).

6) Sens général de scène et catégorisation perceptive de scène

Le sens général d'une scène est une notion peu précise dans la littérature, qui fait l'objet d'étude dans des domaines différents : la compréhension de la scène et la modélisation du système de catégorisation de la scène (Torralba, Oliva, Castelhano, & Henderson, 2006; Wang, Zhang & Fei-Fei, 2006). Ces références confirment l'hypothèse générale développée dans notre thèse, selon laquelle des propriétés de scène-niveau peuvent permettre de catégoriser une scène naturelle. En tenant compte de la spécificité de la scène par rapport aux objets, une scène ne peut pas être considérée simplement comme la somme des objets qu'elle contient. En effet, une scène a ses propres propriétés (le contexte, la structuration spatiale et les relations

spatiales entre les objets). Par ailleurs, de plus en plus de données neuro-anatomiques supposent qu'une scène soit traitée différemment par rapport aux autres objets (Yue, Vessel, & Biederman, 2006).

La rapidité de la compréhension de scènes ne pourrait s'expliquer que par des processus parallèles, mais les éléments locaux diagnostiques joueraient un rôle décisif en fonction du but de l'exploration visuelle. (Bacon-Macé, Kirchner, Fabre-Thorpe & Thorpe, 2007; Guyonneau, Kirchner, & Thorpe, 2006; Kirchner, & Thorpe, 2006).

7) Additif à la discussion

En ce qui concerne la première expérience, la performance des sujets est globalement meilleure pour l'ordre de présentation "*décroissante*" que les deux autres ordres de présentations ("*croissant*" et "*aléatoire*"). Plusieurs explications alternatives sont possibles à l'explication donnée antérieurement :

1) Cet ordre de présentation fournit aussi une meilleure condition de visualisation des éléments diagnostiques (Guyonneau, Kirchner, & Thorpe, 2006; Kirchner, & Thorpe, 2006), facilitant ainsi l'identification de l'image.

2) Cet ordre faciliterait la préactivation de l'effet contextuel permettant ainsi aux sujets d'améliorer leur performance (Brockmole, & Henderson, 2006, 2008; Eckstein, Drescher, & Shimozaki, 2006; Neider & Zelinski, 2006; Over, Hooge, Vlaskamp, & Erkelens, 2007; Torralba, Oliva, Castelhano, & Henderson, 2006).

3) Cet ordre induit une structuration de la scène facilitant l'activation des représentations de scène en mémoire (Alvarez & Oliva, 2008; Castelhano & Henderson, 2005; Hollingworth, 2006a, 2006b, 2007; Johnson, Hollingworth, & Luck, 2008).

Pour les expériences 2 à 4 (identification et catégorisation de scène), les performances (temps de réaction et le taux de réponses correctes) sont meilleures pour les "*images-originales"* que les "*images-contours"* et les "*images-luminance"*. Les publications récentes confortent l'hypothèse selon laquelle les contours permettant de construire le contexte et de créer la sémantique de scène (Brockmole, & Henderson, 2006, 2008; Castelhano & Henderson, 2005; Eckstein, Drescher, & Shimozaki, 2006; Hollingworth, 2006a, 2006b, 2007; Johnson, Hollingworth, & Luck, 2008; Neider & Zelinski, 2006; Over, Hooge, Vlaskamp, & Erkelens, 2007).

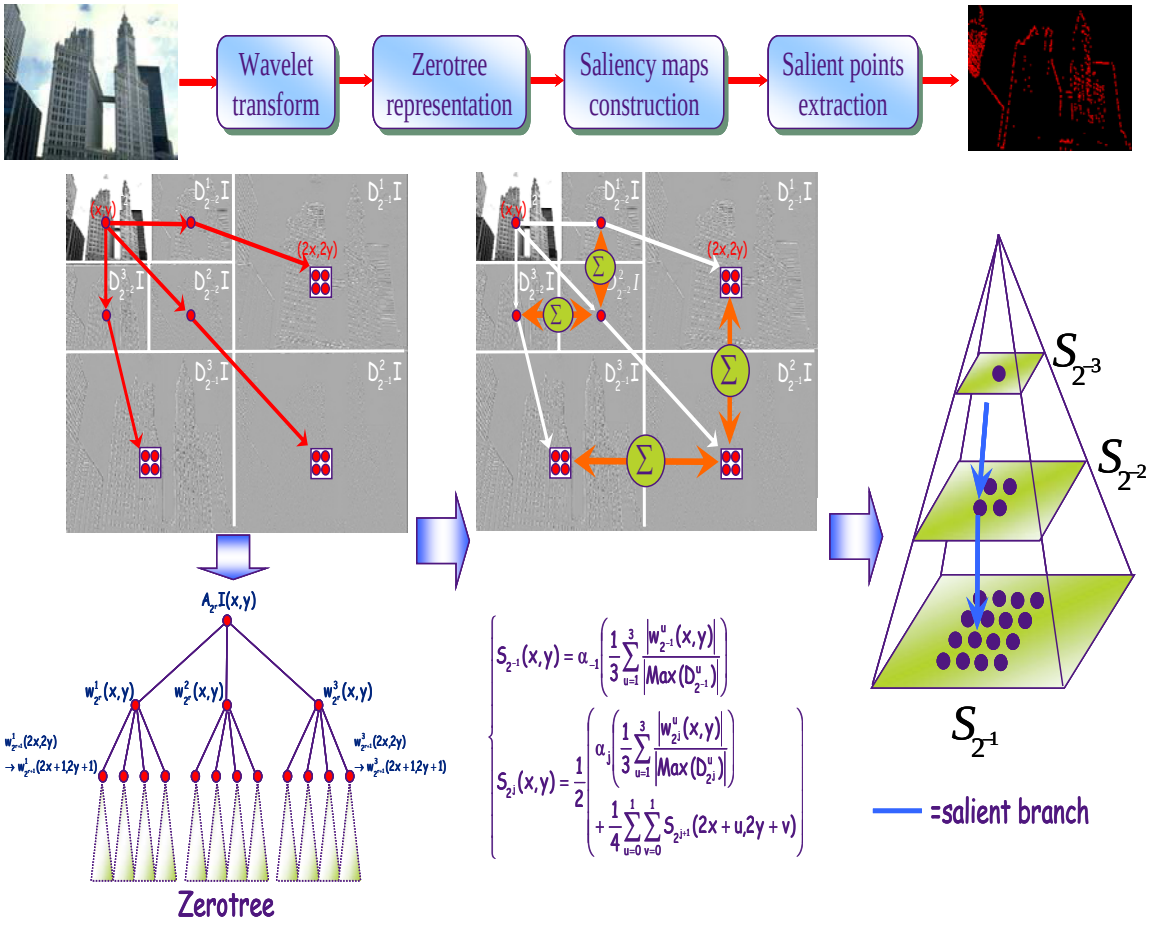
L'expérience 5 avait pour objectif d'étudier l'effet du lissage pour les "*images-luminance*" dans une tâche de catégorisation de scène. Les résultats montrent que le traitement de l'image débute par un traitement des propriétés globales de scène puis ce traitement s'oriente vers les propriétés locales (Oliva & Torralba, 2006). Bacon-Macé, Kirchner, Fabre-Thorpe et Thorpe (2007) et Guyonneau, Kirchner et Thorpe (2006) font l'hypothèse que dans une tâche de catégorisation des animaux, les deux voies visuelles sont en interaction. Comme le signalent ces auteurs, les résultats de cette expérience confirment cette hypothèse.

8) Présentation du principe de l'algorithme de traitement d'image de France Telecom Recherche et Développement

Le principe de l'algorithme de traitement d'image est présenté de manière descriptive en deux parties : Comment décrire une image? Comment représenter une image?

➢ Description d'image par l'algorithme

1) une image peut-être représentée par des points d'intérêts :



Un exemple :

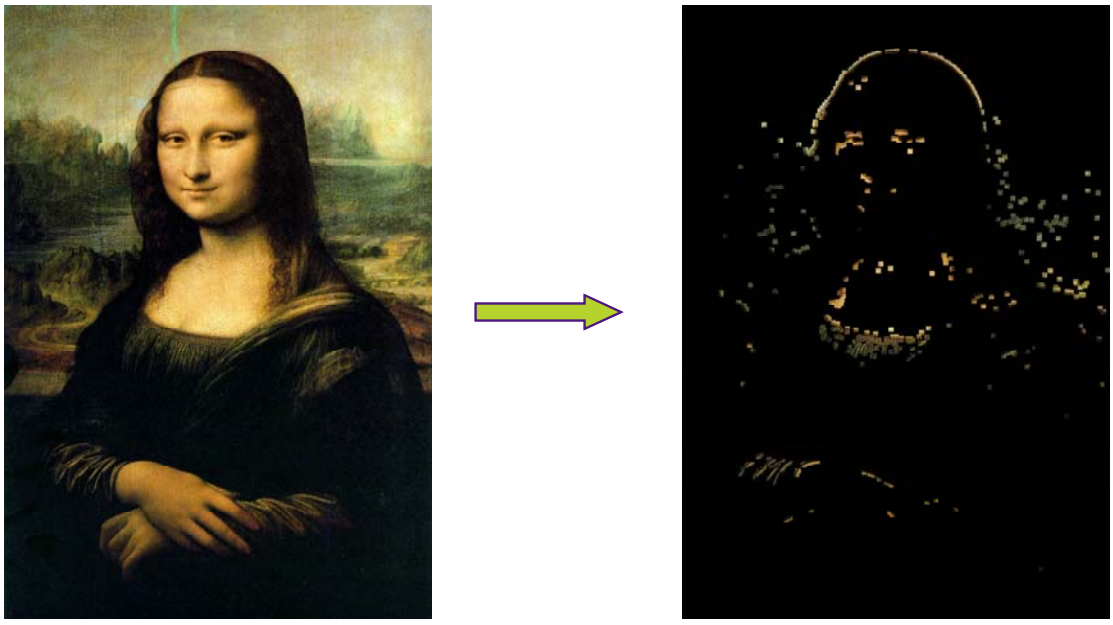
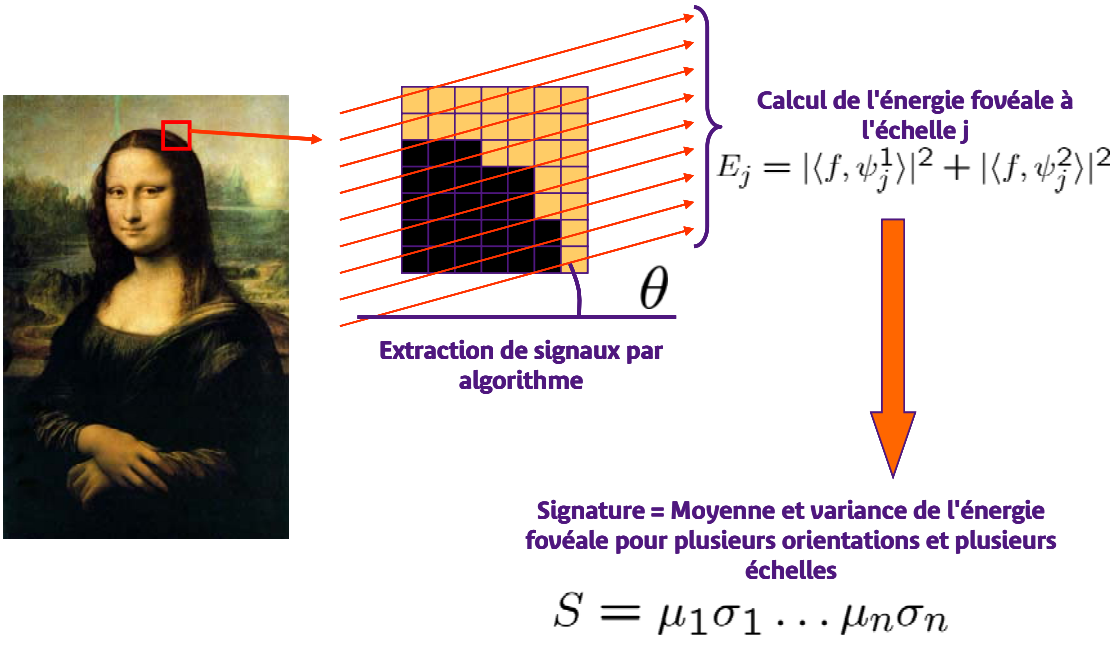


Image originale

Détecteur ondelettes

2) signature fovéale



➢ Représentation par des chaines

- ➢ Une image = Une chaîne de points (de signatures)
- ➢ Comparaison d'images = comparaison de chaînes

➢ Comparaison de chaines pour comparer les images

- ➢ Distance d'édition calculée grâce à la relation de récurrence suivante

$$D(i,j) = min \begin{cases} D(i-1,j-1) + \gamma(x_i, y_j) \\ D(i-1,j) + \gamma(x_i, \epsilon) \\ D(i,j-1) + \gamma(\epsilon, y_j)) \end{cases} \quad (1)$$

- ➢ Avec g le coût associé aux opérations d'édition
- ➢ Utilisation d'un algorithme de programmation dynamique
- ➢ Distance d'édition normalisée (Normalisation de la distance d'édition par le nombre d'opération d'édition)
- ➢ Distance entre histogrammes avec prise en compte de l'ordre (Une image est représentée par un histogramme des signatures prenant en compte leur place dans la chaîne)

Bibliographies

Alvarez, G. A., & Oliva, A. (2008). The Representation of Simple Ensemble Visual Features Outside the Focus of Attention. *Psychological Science, 19(4),* 392-398.

Aminoff, E., Gronau, N., & Bar, M (2007). The parahippocampal cortex mediates spatial and non-spatial associations. *Cerebral Cortex*, 27, 1493-1503.

Ariely, D (2001). “Seeing Sets: Representation by Statistical Properties” *Psychological Science*, *12 (2)*, 157- 162.

Auckland, M. E., Cave, K.R., & Donnelly, N. (2007). Non-target objects can influence perceptual processes during object recognition. *Psychonomic Bulletin and Review, 14*, 332-337.

Bacon-Mace, N., Kirchner, H., Fabre-Thorpe, M, &. Thorpe, S (2007). Effects of task requirements on rapid natural scene processing: From common sensory encoding to distinct decisional mechanisms. *Journal of Experimental Psychology: Human Perception and Performance 33(5)*: 1013-26.

Brady, T. F., & Oliva, A. (sous-presse). Statistical learning using real-world scenes: extracting categorical regularities without conscious intent. *Psychological Science.*

Brockmole, J. R., & Henderson, J. M. (2006). Recognition and attention guidance during contextual cueing in real-world scenes: Evidence from eye movements. *Quarterly Journal of Experimental Psychology, 59*, 1177-1187.

Brockmole, J. R., & Henderson, J. M. (2008). Prioritizing new objects for eye fixation in real-world scenes: Effects of object-scene consistency. *Visual Cognition, 16*, 375-390.

Castelhano, M.S., & Henderson, J.M. (2007). Initial Scene Representations Facilitate Eye Movement Guidance in Visual Search. *Journal of Experimental Psychology: Human Perception and Performance, 33(4)*, 753-763.

Chong, S. C., & Treisman, A. (2005 a). Attentional spread in the statistical processing of visual displays. *Perception & psychophysics, 67*, 1-13.

Chong, S. C., & Treisman, A. (2005 b). Statistical processing: computing the average size in perceptual groups. *Vision Research, 45*, 891-900.

Davenport, J.L. (2007). Consistency effects between objects in scenes. *Memory & Cognition, 35 (3)*, 393-401.

Eckstein, M. P., Drescher, B., & Shimozaki, S. S (2006). Attentional cues in real scenes, saccadic targeting and Bayesian priors. *Psychological Science, 17,* 973-80.

Elazary, L., & Itti, L (2008). Interesting objects are visually salient, *Journal of Vision,* Vol. 8, No. 3:3, pp. 1-15.

Gronau, N., M. Neta, M., & Bar, M (2008). Integrated contextual representation for objects' identities and their locations. *Journal of Cognitive Neuroscience*, 20(3), 371-388.

Guyonneau, R., Kirchner, H., & Thorpe, S. J. (2006). Animals roll around the clock: The rotation invariance of ultrarapid visual processing. *Journal of Vision*, *6*(10):1, 1008-1017.

Henderson, J.M., Brockmole, J.R., Castelhano, M.S., & Mack, M. (2007). Image salience versus cognitive control of eye movements in real-world scenes: Evidence from visual search. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movement research: Insights into mind and brain* (pp. 537–562). Oxford, England: Elsevier.

Hollingworth, A. (2006a). Visual memory for natural scenes: Evidence from change detection and visual search. *Visual Cognition, 14,* 781-807.

Hollingworth, A. (2006b). Scene and position specificity in visual memory for objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 58-69.

Hollingworth, A. (2007). Object-position binding in visual memory for natural scenes and object arrays. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 31-47.

Johnson, J. S., & Olshausen, B. A. (2005). The earliest EEG signatures of object recognition in a cued-target task are postsensory. *Journal of Vision, 5*(4), 299–312.

Johnson, J. S., Hollingworth, A., & Luck, S. J. (2008). The role of attention in the maintenance of feature bindings in visual short-term memory. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 41-55.

Kayeart, G., Biederman, I, & Vogels, R. (2005). Representation of regular and irregular shapes in macaque inferotemporal cortex. *Cerbral Cortex, 15*, 1308-1321.

Kirchner, H., & Thorpe, S. (2006) Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research, 46,* 1762-1776.

Lleras A, Rensink, R A, & Enns, J T (2007). Consequences of display changes during interrupted visual search: Rapid resumption is target-specific. *Perception & Psychophysics*, *69*: 980-993.

Macé, M. J-M., Thorpe, S, & Fabre-Thorpe, M (2005). Rapid categorization of achromatic natural scenes: how robust at very low contrasts? *European Journal of Neuroscience 21(7)*: 2007-2018.

Mermillod M., Guyader N. & Chauvin A**.** (2005). The Coarse-to-fine Hypothesis Revisited: Evidence from Neuro-Computational Modeling. *Brain & Cognition. 57(2)*, 151-157.
Najemnik, J, & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature* 434, 387–391.

Navalpakkam, V., & Itti, L (2006) Top-down attention selection is fine-grained, *Journal of Vision,* Vol. 6, No. 11, pp. 1180-1193.

Neider, M.B., & Zelinski, G.J. (2006). Scene context guides eye movements during visual search. *Vision Research. 46*, 614–621.

Rousselet, G. A, Mace, M. J.. Thorpe, S, & Fabre-Thorpe, M (2007). Limits of Event-related Potential Differences in Tracking Object Processing Speed. *The Journal of Cognitive Neuroscience 19(8)*: 1241-58.

Oliva, A. & Torralba, A. (2006). Building the Gist of a Scene: The Role of Global Image Features in Recognition. *Progress in Brain Research: Visual perception, 155,* 23-36.

Oliva, A. & Torralba, A. (2007). The Role of Context in Object Recognition. *Trends in Cognitive Sciences, 11(12),* 520-527*.*

Over, E.A.B., Hooge, I.T.C., Vlaskamp, B.N.S., & Erkelens C.J. (2007). Coarse-to-fine eye movement strategy in visual search. *Vision Research 47,* 2272-2280.

Peters, R. J., & Itti, L (2008). Congruence between model and human attention reveals unique signatures of critical visual events. Advances in Neural Information Processing Systems, Vol. 21.

Simons, D. J. (2000). Current approaches to change blindness. *Visual Cognition: Special Issue on Change Detection and Visual Memory, 7*, 1-16.

Simon, D, J, & Rensink, R, A (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences, 9*: 16-20.

Torralba, A., Oliva, A., Castelhano, & Henderson, M (2006) Contextual Guidance of Attention in Natural scenes: *The role of Global features on object search Psychological Review, 113(4)* 766-786.

Vogel, J., & Schiele, B. (2006). Semantic Modeling of Natural Scenes for Content-Based Image Retrieval. *International Journal of Computer Vision 72*, 133–157.

Wang, G., Zhang, Y., & Fei-Fei, L (2006). Using dependent regions for object categorization in a generative framework**.** *IEEE Comp. Vis. Patt. Recog.*

Wolfe, J. M. (1999). Inattentional amnesia. In V. Coltheart (Ed.), *Fleeting Memories* (pp. 71-94). Cambridge, MA: MIT Press.
Yue, X M., Vessel, E A., & Biederman, I. (2007). The neural basis of scene preferences. *NeuroReport. 18(6)*, 525-529.