HAL
open science

Le système visuel humain au secours de la vision par ordinateur

Alexandre Benoit

To cite this version:

Alexandre Benoit. Le système visuel humain au secours de la vision par ordinateur. Traitement du signal et de l'image [eess.SP]. Institut National Polytechnique de Grenoble - INPG, 2007. Français. NNT : . tel-00193715

HAL Id: tel-00193715

https://theses.hal.science/tel-00193715

Submitted on 4 Dec 2007

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

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

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque
|__|__|__|__|__|__|__|__|__|__|

T H E S E

pour obtenir le grade de

DOCTEUR DE L'INP Grenoble

Spécialité : SIPT : *Signal, Image, Parole, Telecoms*

préparée au Laboratoire Gipsa-Lab DIS

dans le cadre de **l'Ecole Doctorale** *EEATS: Electronique, Electrotechnique, Automatique et Traitement du signal*

présentée et soutenue publiquement

par

Alexandre BENOIT

le 14 février 2007

TITRE:

Le système visuel humain au secours de la vision par ordinateur

DIRECTRICE DE THESE : Alice CAPLIER

JURY

M. Pierre-Yves COULON	, Président
M. Patrick BOUTHEMY	, Rapporteur
M. Dominique BARBA	, Rapporteur
M. Alice CAPLIER	, Directeur de thèse
M. Jean-Philippe THIRAN	, Examinateur

Remerciements

Une fois la thèse rédigée et soutenue, vient le temps de la prise de recul et la vision de toutes les personnes qui ont contribuées à son bon déroulement. La thèse ne permet pas seulement de travailler sur un sujet tout seul dans son bureau, elle est également l'occasion de rencontrer tout un groupuscule sympathique de personnes travaillant sur des thématiques plus ou moins proche de la votre : les non moins fameux "collègues du laboratoire" avec qui l'on peu beaucoup échanger. Bien plus que de simples collaborateurs, ils deviennent complices de sa recherche et amis dans la vie de tous les jours.

Je remercie tout d'abord Alice Caplier, qui à l'origine, ma directrice de thèse s'est révélée être une personne de confiance, pleine d'énergie et source de motivation pour mon travail tant par les discussion et les différents projets qu'elle sait concevoir avec talent que par nos discussions skating et autres sujets sportifs et montagneux. Merci Alice pour tes encouragements et tes conseils, ce fut un plaisir de réaliser cette thèse avec toi.

Merci également à Jeanny Herault avec qui j'ai eu le plaisir d'engager d'infinie conversations passionnantes sur cette fameuse rétine et tout ce que l'on pourrai en faire, sans oublier tous ces bons moments à plaisanter. Merci pour tous ces moments d'échanges qui ont pu me permettre d'apprendre toujours plus et d'éclaircir mes idées le tout dans une ambiance des plus conviviales et constructive.

Un grand merci à Pierre-Yves Coulon qui a accepté de présider mon jury mais aussi avec qui j'ai eu l'occasion de partager de bons moments tant en discutant recherche durant ma thèse que lors des enseignements que nous donnions ensemble, une expérience enrichissante et bien sympathique.

Merci également à Jean Marc Chassery et Christian Jutten qui m'ont permis de réaliser ma thèse au sein du laboratoire et avec qui j'ai pu échanger tant sur le travail que pour tout autre sujet intéressant.

Merci à toutes les personnes que j'ai pu côtoyer au laboratoire en somme, les pages pourraient être nombreuses, alors, pour n'en citer que quelques un, Marie-No, Denis, Patricia, Gégé, Brice, Barth, Nico, Rosi, Manu, Jeam-Marc, Hervé et tous les autres avec qui j'ai eu le plaisir d'échanger des mots, du travail, bref, des moments de la vie.

Merci à tous mes proches qui ont su me soutenir durant cette thèse, non pas que j'ai souffert, bien au contraire, la thèse était passionnante et pleine de vie, mais merci pour tous ces moments de discussion, de soutient sur ce travail de thèse obscur. "Tu vas être docteur? tu vas soigner qui et quoi?" "Heu... je vais aider les ordinateurs à voir...." "Ah bon..... et sinon, ca se passe bien?". Merci aussi à miss K pour son soutient et à tous les amis proches.

Merci enfin à mes rapporteurs Dominique Barba et Patrick Bouthemy qui ont eu la patience de relire et rapporter sur mon manuscrit. Merci à Jean-Phillipe Thiran qui a accepté d'être mon examinateur pour ma soutenance. Nous avons pu discuter de façon ouverte et passionnante sur mon travail tout en partageant nos expérience, cela a été très enrichissant.

Enfin, un grand merci à ceux que je n'ai pas cité ici mais qui ont croisé mon sillage et avec qui j'ai pu passer de bons moments tant pour le travail que pour tout moment agréable ;o).

Sommaire

Préambule 12

Introduction 15

Partie I: Système visuel humain: modélisation et traitement d'images bas niveau 17

Chapitre I : Système visuel humain et modélisation 19

I.1. Introduction 19

I.2. Présentation générale du système visuel humain 20

I.3. La rétine 21

I.3.1. Les Photorécepteurs 22

I.3.2. La Couche Plexiforme Externe (PLE) 29

I.3.3. La Couche Plexiforme Interne (PLI) 44

I.4. En direction du cerveau 52

I.4.1. Transmission de l'information visuelle vers le cerveau 52

I.4.2. Le cortex visuel 53

I.5. Fonctionnement et modélisation de l'aire V1 55

I.5.1. Le cortex V1 55

I.5.2. Modélisation 56

I.6. Conclusion 62

Chapitre II: Système visuel humain et traitement d'images 65

II.1. Introduction 65

II.2. Extraction de contours 65

II.2.1. Extraction de tous les contours: filtre Parvo contours 65

II.2.2. Extraction de contours en mouvement: filtre MagnoY contours mobiles 69

II.3. Analyse fréquentielle 72

II.3.1. Analyse des orientations dominantes de l'image à partir du spectre log polaire 72

II.3.2. Détection des changements temporels: détecteur d'événements 75

II.4. Segmentation de mouvement 84

II.4.1. Principe 84

II.4.2. Limitations initiales de la méthode 85

II.4.3. Correction du système 87

II.4.4. Performances 89

II.5. Estimation de vitesse 93
II.5.1. Introduction 93
II.5.2. Filtres de vitesse large bande monodimensionnels 95
II.5.3. Estimation de vitesse 2D 96
III Conclusion 102

Partie II: Applications bio-inspirées, les modélisations du système visuel humain au service de la vision par ordinateur 104

Chapitre III: Analyse du visage 106
III.1. Introduction 106
III.2. Localisation des yeux 107
III.2.2. Principe 108
III.2.3. Résultats 109
III.3 Mouvements de tête globaux versus mouvements locaux 111
III.3.1. Principe 111
III.3.2. Classification du type de mouvement 114
III.3.3. Performances 115
III.4 Détection de l'état ouvert ou fermé de la bouche et des yeux 116
III.4.1. Principe 116
III.4.2. Méthode de détection 118
III.5. Interprétation des mouvements locaux du visage 125
III.5.1 Description des clignements d'yeux 125
III.5.2. Interprétation des mouvements de bouche 126
III.6 Extraction de l'information de mouvement global de la tête 128
III.6.1. Système proposé 128
III.6.2. Principe 130
III.6.3. Détection des hochements d'approbation et de négation 133
III.7. Intégration dans un système de détection de stress et d'hypovigilance chez les conducteurs 139
III.7.1. Présentation 139
II.7.2. Analyse visuelle de l'état de fatigue du conducteur 141
III.7.3. Extensions à envisager 144
III.8. Intégration dans un système d'apprentissage de la langue des signes 144
III.8.1. Description de l'application développée 145
III.8.2. Logiciel proposé 150
III.8.3. Extensions à envisager 151
VII. Conclusion du chapitre 152

Chapitre IV: Suivi et identification d'objets 153
IV.1. Suivi d'objet 153
IV.1.1. Algorithme proposé 153
IV.1.2. Suivi court terme avec attribution et conservation de l'identifiant 154
IV.1.3. Filtrage des zones segmentées dues au bruit 160
IV.1.4. Performances de l'algorithme de suivi avec identification des zones de bruit 165
IV.1.5. Conclusion sur le système de suivi 167
IV.2. Reconnaissance d'objets 168
IV.2.1. Algorithme proposé 168
IV.2.2. Analyse de performances 172
IV.2.3. Conclusion de la méthode d'identification et de classification 186
VI.3. Perspectives: système de suivi long terme d'objets 186
Chapitre V: Conclusion et perspectives 188
I. Conclusion 188
II. Perspectives 189

Annexe: synthèse des paramètres 191
1. Paramètres associés aux filtres Parvo et MagnoY. 192
1.1. Filtre Parvo contours 192
1.2. Filtre MagnoY contours mobiles 193
2. Paramètres de l'analyseur spectral 194
3. Paramètres associés au détecteur d'évènements 194
3.1. Réglage de la réponse temporelle 194
3.2. Réglage des seuils de déclenchement pour la segmentation d'évènements 194
3.3. Seuil de déclenchement du détecteur binaire d'alerte de mouvement 194
4. Paramètres de l'algorithme de segmentation de mouvement 195
4.1. Estimation très localisée de la quantité de mouvement: filtrage Seg_p 195
4.2. Estimation de la quantité de mouvement moyenne locale: filtrage Seg_h 195
4.3. Ajustement de la pondération de Seg_h pour la segmentation 195
5. Paramètres du système d'estimation de vitesse 196
5.1. Paramètre des filtres de vitesse 196
5.2. Paramètre des filtres passes bas spatiaux placés en aval des filtres de vitesse 196
BIBLIOGRAPHIE 197
Publications 202

Index des figures

Figure I.1: schéma global du système visuel humain [McGillSite] 20
Figure I.2: schéma général de l'oeil [Kolb96] 21
Figure I.3: organisation des couches cellulaires de la rétine [GsbmeSite] 22
Figure I.4: réponse d'un photorécepteur en fonction de la lumière reçue dans son voisinage 23
Figure I.5: sensibilité des différents types de photorécepteurs à la longueur d'onde de la lumière 24
Figure I.6: a, concentration des photorécepteurs en fonction de l'excentricité 25
Figure I.7: comparaison entre le capteur naturel , l'oeil et un capteur numérique standard. 25
Figure I.8: ajustement de la sensibilité des photorécepteurs par R_0 lié à la luminance locale. 26
Figure I.9: exemples d'effets de la compression logarithmique 28
Figure I.10: symbole associé au module de compression logarithmique 29
Figure I.11: triade synaptique au sein d'un cône 29
Figure I.12: notion de champ récepteur pour une cellule bipolaire 31
Figure I.13: modélisation des connexions entres cellules de même type au sein de la PLE 32
Figure I.14: fonction de transfert du filtre passe bas spatio-temporel 33
Figure I.15: illustration de l'effet du filtre passe-bas 34
Figure I.16: diagramme pour la mise en oeuvre du filtre passe-bas spatio-temporel 35
Figure I.17: symbole associé au filtre passe-bas spatio-temporel 36
Figure I.18 :modèle de PLE 37
Figure I.19: Modélisation et illustration avec un signal spatial 1D de la PLE à l'aide des modules élémentaires 38
Figure I.20: fonction de transfert de la couche PLE de la rétine. 39
Figure I.21: effet de l'enchaînement des traitements entre les différentes cellules au sein de la PLE 40
Figure I.22: extraction des contours dans une image 41
Figure I.23: blanchiment spectral de la rétine 42
Figure I.24: réponse impulsionnelle du filtre PLE 43
Figure I.25: effet spatio-temporel de la PLE. 43
Figure I.26: réponse impulsionnelle du filtre passe haut temporel des cellules amacrines 45
Figure I.27: la perception de la luminance du cercle est différente suivant celle du fond. 46
Figure I.28: modélisation du traitement de la voie parvocellulaire au niveau de la PLI 48
Figure I.29: schéma du traitement de la voie Magnocellulaire de la PLI 49
Figure I.30: signaux de sortie de la voie Magnocellulaire 49
Figure I.31: modélisation globale du traitement de l'information visuelle par la rétine 51
Figure I.32: cheminement de l'information de la rétine au CGL [McGillSite] 52
Figure I.33: les aires corticales du système visuel humain [McGillSite]. 53
Figure I.34: communication entre les 2 voies de la rétine avec le CGL et le cortex 54
Figure I.35: échantillonnage du spectre d'amplitude à l'aide d'une rosace de filtres GloP 57

Figure I.36: banc de filtres GloP 1D 58
Figure I.37: spectre échantillonné par une rosace de filtres GloP 59
Figure I.38: fenêtrage de Hanning de l'image entrante 60
Figure I.39: comportement du spectre log polaire 61
Figure I.40: symbole associé à l'analyseur de spectre "log polaire" 62
Figure I.41: résumé des principaux "éléments" du système visuel humain 64
Figure II.1: architecture filtre Parvo contours (a) et symbole associé (b) 66
Figure II.2: effet du filtrage Parvo contours pour l'extraction de détails 67
Figure II.3: a. extrait d'une séquence vidéo avec 2 objets l'un mobile, l'autre immobile 68
Figure II.4: filtre MagnoY contours mobiles 70
Figure II.5: effet du filtrage MagnoY contours mobiles 71
Figure II.6: exemple de calcul de courbe d'énergie cumulée par orientation à partir de différentes images spectrales. 73
Figure II.7: courbe d'énergie par orientation et effet zoom. 74
Figure II.8: symbole associé au détecteur d'orientations 75
Figure II.9: schéma du détecteur d'événements 75
Figure II.10: analyse de l'évolution de l'énergie totale du spectre log polaire de l'image des contours mobiles 76
Figure II.11: schéma électrique équivalent permettant de calculer l'énergie $E_1(t)$ 77
Figure II.12: détection d'événements de mouvement 78
Figure II.13: décomposition image par image d'un clignement d'oeil filmé à 30 images par seconde 79
Figure II.14: a, extraits d'une séquence de gestes de main. b, énergies E(t) et $E_1(t)$. c, indicateur de mouvement α(t) 79
Figure II.15: extrait de la base de test pour l'évaluation du détecteur d'événements de mouvement 81
Figure II.16: comparaison de méthodes de détection de mouvement 82
Figure II.17: symbole du détecteur d'alertes de mouvements (changements spatio-temporels) 84
Figure II.18: segmentation de zones en mouvement 85
Figure II.19: réponse des filtres F_1 et F_2 à l'énergie de la voie MagnoY contours mobiles 86
Figure II.20: résultat de la segmentation par comparaison de Seg_p et Seg_h 87
Figure II.21: segmentation en différentiant les pondération des sorties Seg_p et Seg_h 88
Figure II.22: exemple de segmentation avec la méthode proposée 90
Figure II.23: comparaison de différentes méthodes de segmentation 91
Figure II.24: symbole associé l'algorithme de segmentation de mouvement 93
Figure II.25: particularité du spectre d'un objet variant temporellement 94
Figure II.26: estimation de la vitesse par utilisation de filtres placés dans le domaine spatio-temporel 95
Figure II.27: fonction de transfert du filtre de vitesse unidimensionnel 96
Figure II.28: chaîne complète du calcul de flot optique 97
Figure II.29: exemple de calcul de flot optique avec la méthode proposée 98
Figure II.30: illustration du problème d'ouverture 99
Figure II.31: symbole associé l'estimateur de flot optique 100
Figure II.32: exemples de calcul de flot optique moyen sur des objets en mouvement 101
Figure III.1: Informations de base pour l'analyse d'un visage dans une scène vidéo 107
Figure III.2: boite englobante du visage (en pointillés), zone de recherche de chaque oeil 108
Figure III.3: schéma de détection de la position d'un oeil 109
Figure III.4: détection de la position de l'oeil 110

Figure III.5: classification des mouvements de tête 113
Figure III.6: définition des zones d'analyse permettant la classification des mouvements détectés sur le visage 114
Figure III.7: système d'analyse de l'état ouvert ou fermé 117
Figure III.8: système de détection des ouvertures et fermetures d'un oeil ou de la bouche 119
Figure III.9: système de détection de l'état ouvert/ fermé de la bouche ou d'un oeil 120
Figure III.10: algorithme de détection de l'état ouvert/fermé lors d'alerte de mouvement 121
Figure III.11: exemple d'évolution temporelle de l'énergie E(t) en sortie du filtre Parvo contours 122
Figure III.12 exemple d'évolution de l'énergie E(t) en sortie du filtre Parvo lors de mouvements de bouche 124
Figure III.13: exemple d'évolution de l'énergie en sortie du filtre Parvo contours 125
Figure III.14: exemple d'évolution de l'indicateur 127
Figure III.15: schéma global du processus d'analyse des mouvements globaux de la tête 129
Figure III.16: mouvements globaux simples et composés de la tête dans son ensemble 130
Figure III.17: répartition de l'énergie des contours sur le spectre log polaire d'un visage immobile 131
Figure III.18: exemples de modèles utilisés pour l'estimation de la pose et des mouvements de la tête 133
Figure III.19: séquence de hochements de tête d'approbation et de négation 135
Figure III.20: algorithme de reconnaissance des hochements de tête 136
Figure III.21: séquence de hochements de tête d'approbation et de négation 137
Figure III.22: extraits de la base de tests pour l'évaluation du détecteur de hochements de tête 138
Figure III.23: interface graphique de l'application de détection d'approbation et de négation. 139
Figure III.24: illustration du système de détection d'hypovigilance chez le conducteur 141
Figure III.25: architecture de l'analyseur basé vidéo de l'état de fatigue du conducteur 143
Figure III.26: exemple de signes 145
Figure III.27: exemple de séquence de langage des signes 147
Figure III.28: système d'analyse des mouvements de tête 148
Figure III.29: exemple d'évolution temporelle des indicateurs de mouvement 149
Figure III.30: interface du logiciel d'apprentissage de la langue des signes 151
Figure IV.1: système de suivi 154
Figure IV.2: méthode d'extraction d'informations sur les zones segmentées de chaque image 155
Figure IV.3: exemples de suivi de 3 balles avec la méthode proposée 159
Figure IV.4: comparaison entre les sorties des filtres Parvo ON-OFF et MagnoY contours mobiles 161
Figure IV.5: système de différenciation entre bruit et mouvement 162
Figure IV.6: illustration de la méthode de différenciation bruit et mouvement 163
Figure IV.7: extrait de la séquence de test pour l'identification des zones de mouvement et de bruit. 164
Figure IV.8: test du système de suivi temporel d'objets en mouvement 166
Figure IV.9: principe de fonctionnement de l'analyseur/identificateur d'objets 168
Figure IV.10: comportement du spectre log polaire lors d'une transformation de l'image 170
Figure IV.11: autocorrélation du spectre log polaire lors d'une translation 171
Figure IV.12: images extraites de la base de scènes naturelle 173
Figure IV.13: évolution des spectres log polaires et des autocorrélations 174
Figure IV.14: exemples de résultats de classification avec la méthode proposée 176
Figure IV.15: autocorrélations moyennes de trois classes d'images 177
Figure IV.16: images extraites de la base COIL100 178

Figure IV.17: évolution du spectre log polaire et de l'autocorrélation d'un même objet 180
Figure IV.18: évolution de la distance 181
Figure IV.19: exemples de résultats d'identification avec la méthode proposée 182
Figure IV.20: deux objets de forme proche et d'orientation identique dans la base de test 183
Figure IV.21: exemples de résultats de classification 184
Figure IV.22: exemples de résultats de classification par comparaison entre prise de vue identique 185
Figure IV.23: exemple d'objets appartenant à la même classe mais présentant des autocorrélations différentes 186
Figure IV.24: ajout de l'analyseur/identificateur d'objets au système de suivi d'objets 187

Index des tables

Table I.1: synthèse des modules associés au système visuel 63
Table II.1: comparaison de la réponse au bruit avec ou sans filtrage Parvo contours ou Sobel 67
Table II.2: comparaison de la réponse du filtrage MagnoY contours mobiles et de la différentielle temporelle 71
Table II.3: performances du détecteur d'événements 82
Table II.4: synthèse des modules d'analyse d'image bio-inspirés 102
Table III.1: performance de l'algorithme de détection des yeux 111
Table III.2: performances de l'identification des mouvements du visage 116
Table III.3: relation entre niveau d'énergie dans les zones d'analyse et état d'ouverture 117
Table III.4: performances de l'analyseur d'état de l'oeil 123
Table III.5: performances de la détection des états ouvert et fermé de la bouche 124
Table III.6: performances de la détection des clignements 126
Table III.7: performances de la détection des bâillements 127
Table III.8: performances du détecteur de périodes d'absence de parole 128
Table III.9: performances du système de détection de mouvement rigides du visage 132
Table III.10 : performance du détecteur de hochement de tête 138
Table III.11: présentation des différents signes de la base de test 146
Table IV.1: évaluation des performances du système de reconnaissance du bruit 164
Table IV.2: évaluation des performances du calcul de trajectoire des zones en mouvement. 167
Table IV.3: classification manuelle des objets de la base COIL100 179

Préambule

En commençant cette thèse en 2003, je ne savais pas encore que j'allais répondre à l'examen du Baccalauréat de l'académie de Grenoble posé le 20 juin 1972 pour la série D...celui de ma chère maman. "Maman, je vais essayer de t'aider...", en voici un extrait de l'énoncé :

72-53-1.

Le candidat traitera l'un des deux sujets suivants, au choix :

1er SUJET

Question I : (6 points)

En vous basant sur les observations faites en travaux pratiques au cours de la dissection d'un oeil de Mammifère, quelle légende mettriez-vous sur le croquis ci-joint, dû à Vésale, anatomiste du 16ème siècle ?

- Que pensez-vous de cette représentation de la structure de l'oeil ?
- Quel croquis proposeriez-vous à la place de celui de Vésale ?

Question II : (14 points)

Dans un ouvrage intitulé "Le cerveau vivant", le physiologiste Grey Walter parlant de l'oeil écrit : " ... l'analogie avec la photographie s'écroule quand nous venons à regarder le mécanisme de la vision ... Il est vrai que la lentille de l'oeil ressemble à celle d'un appareil photographique à bon marché, mais à un point de vue plus important, la rétine, sur laquelle l'image est projetée, n'est pas du tout semblable à un film photographique. La surface entière du film photographique possède un grain uniforme : les particules de matière chimiquement sensible qui la composent, sont réparties à une distance égale les unes des autres et sont d'une sensibilité égale. La rétine ne possède pas cette uniformité. Seule, une toute petite surface d'environ un tiers de millimètre, au centre de la rétine, a un grain suffisamment fin pour recevoir et transmettre une image très détaillée . ... ".

- A partir de l'étude de la rétine (schémas à l'appui), des observations et des expériences que vous avez pu réaliser, pourriez-vous justifier les affirmations de ce texte ?

Plus loin, le même auteur écrit encore : " La phase suivante du processus de la vision est le transfert de l'image rétinienne vers la zone de projection du cerveau ... " .

- Expliquez sous quelle forme se fait le "transfert de l'image rétinienne" et quel est le rôle du cerveau dans la vision.

Coïncidence amusante, j'ai travaillé sur une thématique de recherche et d'ingénierie, l'analyse d'image, qui m'en a fait approfondir une autre, la vision humaine. Et, étonnamment, me voici à même de répondre de manière précise à ce sujet d'examen d'il y a 35 ans...

Dans cette thèse, il est proposé une plongée dans la vision humaine, sa modélisation et son utilisation en traitement d'images. Bien qu'un grand nombre de points reste encore obscur, la recherche a, depuis cet examen de 1972, avancé et nos connaissances se sont enrichies.

Néanmoins, nous verrons que bien que l'on connaisse beaucoup de choses en vision humaine, et ce, depuis fort longtemps, nous commençons tout juste à utiliser ces connaissances pour nos besoins. En traitement d'images notamment, les connaissances du système visuel humain commencent tout juste à être exploitées dans des méthodes algorithmiques ou électroniques.

Dans cette thèse sont abordés deux domaines : la vision par ordinateur et les sciences cognitives. Les sciences cognitives permettent d'analyser les processus cognitifs qui permettent aux êtres vivants de communiquer et comprendre leur environnement. Les études dans ce domaine ont montré que les algorithmes de traitement qui en découlent peuvent parfaitement s'insérer dans les problématiques dites de "vision par ordinateur" pour lesquelles elles pourraient apporter des gains d'efficacité, de fiabilité et surtout d'adaptabilité. La vision par ordinateur couvre un large domaine d'applications dédiées à l'analyse de l'environnement par des méthodes mathématiques et algorithmiques. Les domaines tels que le suivi (d'objets, de personnes, etc.), l'identification ou la reconnaissance sont depuis longtemps des sujets à importants efforts de recherche pour lesquels de nombreuses approches sont développées.

Néanmoins, sciences cognitives et vision par ordinateur ne sont pas encore suffisamment associées et les connaissances en sciences cognitives n'apportent pas encore beaucoup d'aide à la vision par ordinateur. Nous ne sommes qu'au début d'une véritable période d'échanges pour laquelle chaque domaine pourrait apporter à l'autre.

Dans cette thèse est décrit un ensemble d'outils issus de la modélisation du système visuel humain (rétine et premières aires corticales) dédiées au traitement de l'information visuelle. Nous verrons que les applications qui en découlent suivent la structure du modèle biologique et démontrent d'intéressantes capacités en terme d'analyse de scènes de tous types, comme nous le faisons nous-mêmes sans nous en apercevoir.

Introduction

Dans notre environnement mêlant à la fois informatique et information visuelle, nous ne voyons que la finalité : télédiffusion, films, jeux vidéo, enregistrements vidéo, montages vidéo, etc. Ce monde est constitué de solides fondations de recherche. Celles-ci permettent d'accéder à des fonctionnalités et usages d'outils (appareils photographiques, caméra, application de traitement ou d'analyse d'image, etc.) créés à partir de principes et de méthodes performantes de traitement d'images. Ces sciences du traitement de l'information et de la vision permettent de concevoir des outils toujours plus performants et permettent à l'utilisateur final d'obtenir toujours plus de leurs images : des couleurs plus belles, des images plus rayonnantes, une meilleure précision, etc.

Néanmoins, les algorithmes associés utilisent principalement des approches de traitement d'images totalement déconnectées de notre propre façon de voir. Qui n'a jamais pris une photographie qui s'est révélée inutilisable à cause d'un contre-jour ? Etrange, le photographe voyait pourtant de ses yeux les choses différemment... Or, "comment voit-on?", "est-ce que ce que je vois en ce moment est vraiment la réalité ou bien n'est-ce qu'une représentation personnelle ?". Ces questions sont éloignées de la vision par ordinateur, elles touchent plutôt la connaissance de notre système visuel. Quel est donc le lien entre ce que nous voyons avec notre système visuel et la vision par ordinateur qui nous présente par ses propres méthodes un contenu visuel ?

Le rapprochement entre ces deux mondes est limité par la connaissance que l'on a du fonctionnement de notre système visuel. Les efforts de recherche sont nombreux depuis bien longtemps, mais on commence tout juste à utiliser notre compréhension du vivant pour des applications utiles à nos besoins actuels.

L'objectif de cette thèse est de montrer comment il est possible de développer des algorithmes de traitement d'images "bio-inspirés " pour réaliser par exemple une segmentation et un suivi d'objets mobiles. Partant du postulat que l'ensemble (oeil+cerveau) humain est capable de réaliser avec fiabilité une analyse de scènes (ce qui a été et est encore l'une des garanties de la survie de l'espèce humaine), l'idée est de montrer qu'il est avantageux de s'intéresser au fonctionnement du système visuel humain pour développer en vision par ordinateur des algorithmes d'analyse de scène.

Dans ce document, nous nous limitons à la compréhension et à la modélisation de la rétine et du cortex visuel primaire en travaillant exclusivement sur des images en niveau de gris. Le but est de montrer quels types d'analyse de scène il est possible de réaliser à partir de la modélisation de ces deux éléments du système visuel humain.

Ce document est organisé de la manière suivante : dans la première partie, nous proposons de voir comment la compréhension et la modélisation de la rétine et du cortex visuel primaire (cf. chapitre I) permettent de créer des outils bas niveau d'analyse d'image tels qu'un module d'extraction de contours, un module d'extraction de contours mobiles, un système d'analyse des orientations présentes dans une image, un détecteur d'alertes de mouvements, une méthode de segmentation de zones en mouvement et un système de calcul de vitesse (cf. chapitre II).

Dans la seconde partie, nous montrerons comment l'enchaînement spécifique de certains de ces outils bas niveau permet de résoudre des problèmes plus complexes tels que l'interprétation des mouvements du visage (cf. chapitre III), le suivi d'objets en mouvement et leur reconnaissance automatique (cf. chapitre IV).

Partie I: Système visuel humain: modélisation et traitement d'images bas niveau

Introduction

Dans cette partie, nous proposons une description du fonctionnement du système visuel humain ainsi que des modélisations qui en découlent en nous focalisant avant tout sur la rétine et le cortex visuel primaire. Pour chaque élément du système visuel, nous présentons successivement son fonctionnement et son effet en terme de traitement de l'information visuelle ce qui conduit à la définition de modules d'analyse "bas niveau" d'une image. A chacun de ces modules, nous associons un symbole caractéristique qui sera réutilisé dans toute la suite du manuscrit. Ces modules sont par la suite enchaînés les uns aux autres pour construire des algorithmes de traitement de plus "haut niveau".

Chapitre I : Système visuel humain et modélisation

I.1. Introduction

Dans cette partie sont décrites les connaissances acquises sur le système visuel humain. La progression de notre présentation suit le trajet de l'information visuelle, depuis les photorécepteurs jusqu'au cerveau.

La compréhension du corps humain et de son fonctionnement a depuis toujours passionné la recherche. Le but est de comprendre l'architecture et les interactions entre les différents éléments qui le composent. Le système visuel n'échappe pas à cette règle et son exploration a progressivement amené à la découverte des cellules et neurones qui le compose, mais nous sommes encore bien loin de connaître toutes les règles de fonctionnement. Néanmoins, ces connaissances permettent une meilleure compréhension du système dans son ensemble et autorisent déjà à mieux appréhender certaines pathologies.

Comprendre le corps humain ne se résume pas seulement à être capable de soigner tous ses troubles. A mesure que les découvertes avancent, on peut y voir toutes les ingéniosités et trésors de performance qu'offre notre corps, notamment en vision. Notre faculté à pouvoir observer et comprendre notre environnement proche ou éloigné qu'il soit connu ou inconnu, dans toutes les conditions d'observation de la vie courante montre que notre système visuel est de loin bien plus efficace que tous les algorithmes de vision par ordinateur.

Nous proposons donc de présenter dans ce chapitre une synthèse des travaux de recherche sur l'analyse du système visuel humain. L'objet de ce chapitre est de présenter l'ensemble des prés requis nécessaires sur le système visuel humain pour pouvoir développer des algorithmes "bio-inspirés" d'analyse d'images.

Chaque étape de traitement effectué par le système visuel humain est un processus optimisé de génération en génération. Il a pour but de permettre à l'individu de traiter au mieux les données visuelles dans l'objectif de survivre et de s'adapter dans son milieu. On peut donc partir du postulat que le système visuel humain est, à l'échelle de nos besoins, le système "idéal" capable de s'adapter à toutes les situations courantes. Le système visuel constituera dans ce travail un système de référence par rapport aux systèmes de vision par ordinateur. Le but est alors de tendre vers un degré de performance équivalent.

Comme il est montré dans cette partie, le système visuel optimise dès le début les données visuelles entrantes afin de simplifier les traitements ultérieurs. Chaque étage a une fonction précise pour extraire une donnée d'un type particulier et tire partie des traitements effectués en amont. Cette stratégie entre souvent en opposition avec celles qui consistent habituellement à introduire des post-traitements complexes pour compenser les défauts des données initiales.

I.2. Présentation générale du système visuel humain

L'information visuelle est reçue par l'oeil puis transmise au cerveau après différents traitements préliminaires réalisés par la rétine. Les traitements de plus haut niveau sont réalisés dans le cerveau au niveau des aires visuelles que l'on trouve dans la partie arrière des deux hémisphères cérébraux (cf. fig I.1). D'après les études physiologiques [Bullier98], les zones corticales dédiées au traitement de l'information visuelle occupent une place importante. De plus, les connexions avec les autres aires du cerveau font du système visuel un ensemble très complexe. La figure I.1 montre les principaux traitements qui vont faire l'objet de notre étude: les traitements rétiniens puis après transmission de l'information via le nerf optique, les traitements effectués par la première aire du cortex visuel primaire (aire V1).

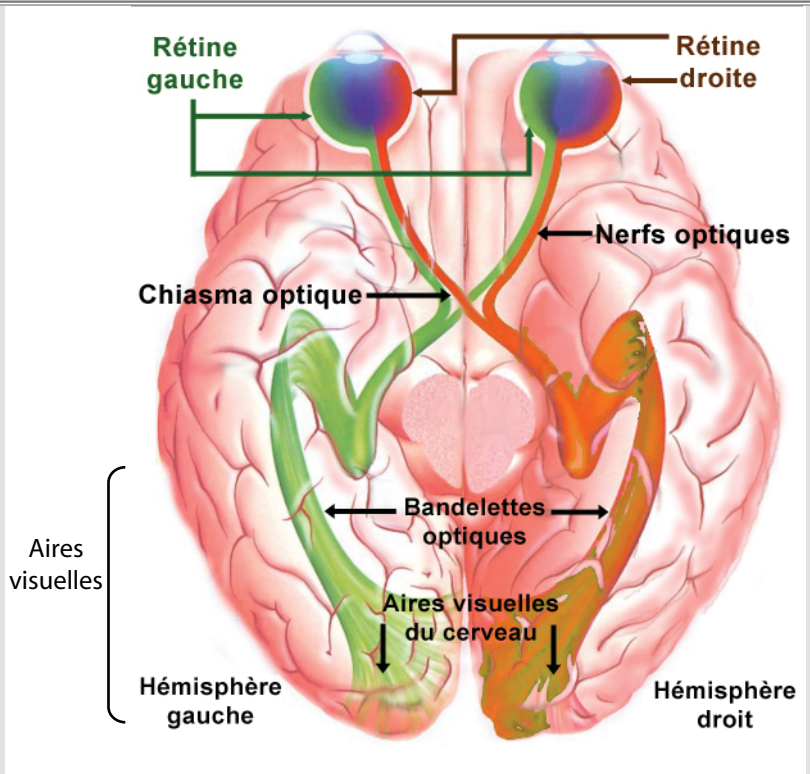


Figure I.1: schéma global du système visuel humain [McGillSite]

I.3. La rétine

L'oeil est le capteur du système visuel. Il est composé d'un système optique adaptatif constitué des éléments suivants (cf. fig. I.2):

→ la cornée, barrière protectrice transparente entre le milieu extérieur et l'oeil à proprement parler. Elle constitue l'un des premiers systèmes de focalisation.

→ l'iris et le cristallin qui permettent respectivement la modulation de la quantité de lumière entrante et l'ajustement de la focale du système optique.

→ la rétine tapissée de photorécepteurs, siège de réception de l'information lumineuse.

Notre étude sur le traitement de l'information visuelle commence au niveau de la rétine. Cornée, iris et cristallin ne sont pas considérés dans la suite.

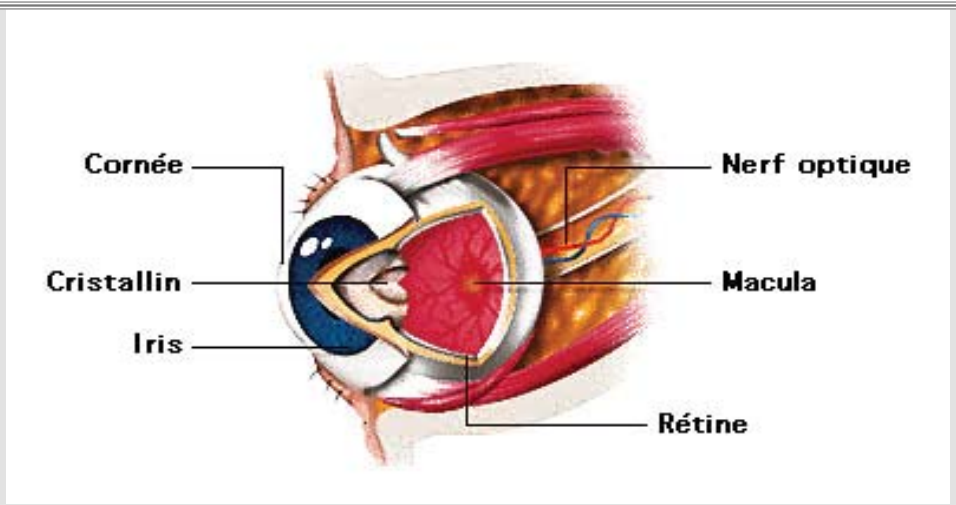


Figure I.2: schéma général de l'oeil [Kolb96]

La rétine est située sur le fond de l'oeil. Elle est constituée d'un assemblage de cellules de différents types, organisées en couches et ayant des propriétés particulières. La figure I.3 montre l'organisation de ces couches cellulaires. On y trouve: la couche des photorécepteurs (cônes et bâtonnets) qui captent l'information lumineuse, la couche des cellules horizontales, des cellules bipolaires et des cellules amacrines et enfin la couche des cellules ganglionnaires. Ces couches sont séparées par des zones de connexions synaptiques : la couche plexiforme externe (PLE) et la couche plexiforme interne (PLI).

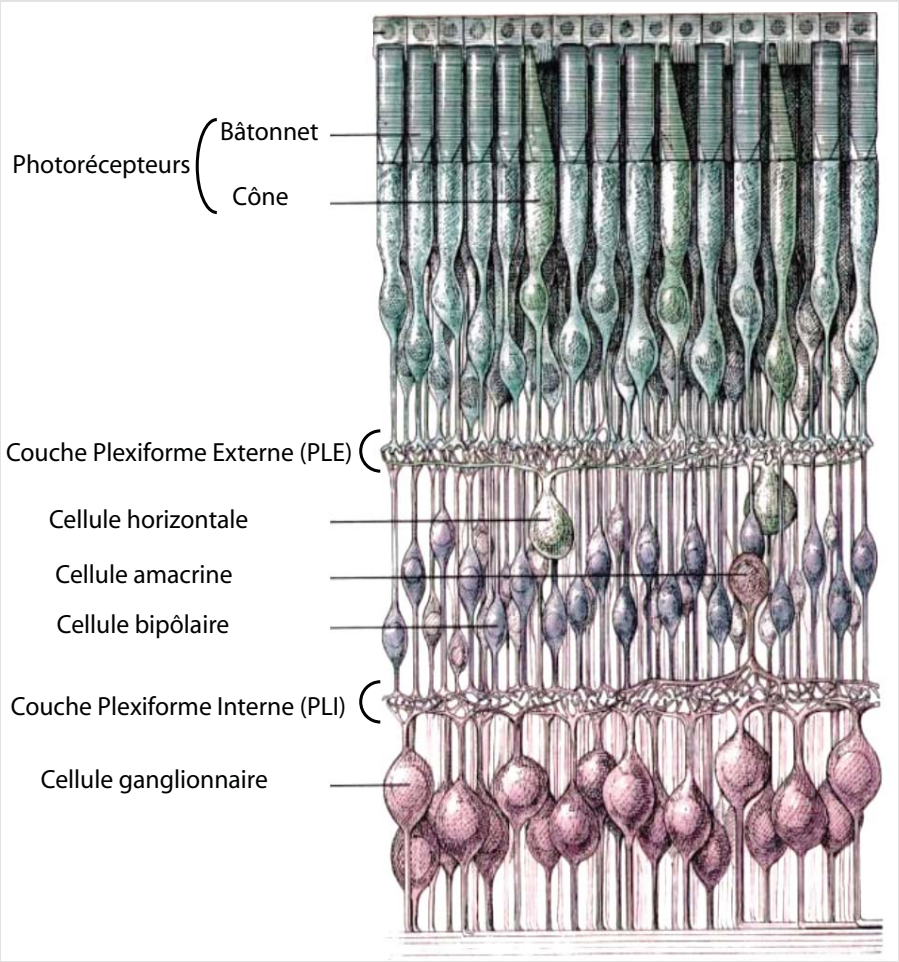


Figure I.3: organisation des couches cellulaires de la rétine [GsbmeSite]

I.3.1. Les Photorécepteurs

I.3.1.1. Sensibilité à la lumière

La lumière traverse les couches neuronales de la rétine et arrive au niveau des photorécepteurs (cônes et bâtonnets). Ces cellules ont pour but d'effectuer la transduction du stimulus lumineux en un potentiel de membrane. Mais, les photorécepteurs ne peuvent donner une réponse fiable que sur une dynamique d'une à deux décades seulement alors que la dynamique à coder par ces capteurs est de l'ordre de 10 unités logarithmiques [Beaudot94]. Les photorécepteurs sont ainsi capables d'adapter (de translater) leur dynamique vers la luminance moyenne locale qu'ils reçoivent, cette capacité est appelée ***compression adaptative (logarithmique)***. La figure I.4 représente les différentes sensibilités que peut avoir un photorécepteur selon la luminance ambiante. La sensibilité moyenne se translate et se centre sur la luminance ambiante, la dynamique reste la même et le photorécepteur travaille sur une gamme de luminance réduite. Si l'on considère toute la gamme de luminance sur laquelle travaille le photorécepteur, le comportement général suit une loi logarithmique (en pointillé sur la figure I.4).

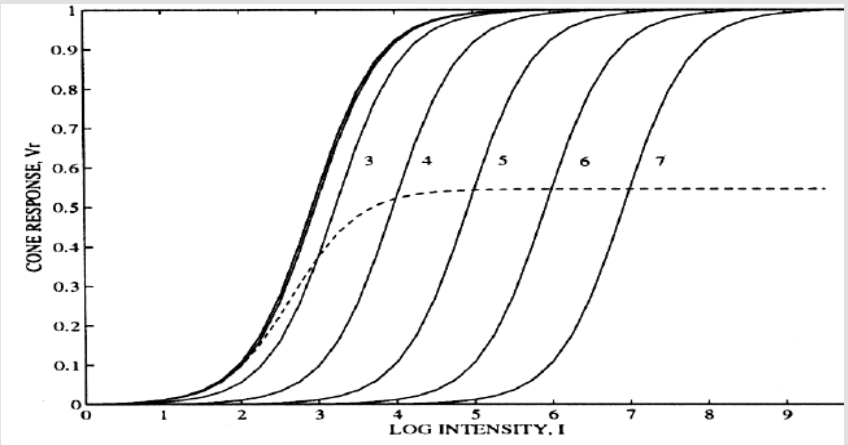


Figure I.4: réponse d'un photorécepteur en fonction de la lumière reçue dans son voisinage [Kolb96]. La sensibilité se translate et se centre sur la luminance locale, le comportement général se rapproche d'une loi logarithmique (en pointillés).

I.3.1.2. Sensibilité à la couleur

Les photorécepteurs présentent des sensibilités spectrales (couleur) et des sensibilités en amplitude (quantité de lumière) différentes. On trouve deux types de photorécepteurs:

→Les bâtonnets :

- ► responsables de la vision scotopique : faible intensité lumineuse (vision de nuit);
- ► très sensibles à la lumière;
- ► de réponse lente aux variations d'illumination.

→Les cônes :

- ► responsables de la vision photopique : vision dans les hautes intensités lumineuses (vision de jour);
- ► responsables de la vision de précision (haute résolution);
- ► déterminent la luminosité des scènes visualisées;
- ► permettent l'extraction des couleurs grâce à 3 types de cônes (cf. fig. I.5):
  - ● Type "L" ("Long", rouge) sensible aux grandes longueurs d'onde visibles;
  - ● Type "M" ("Medium", vert) sensible aux longueurs d'onde visibles moyennes;
  - ● Type "S" ("Short", bleu) sensible aux longueurs d'onde visibles courtes.

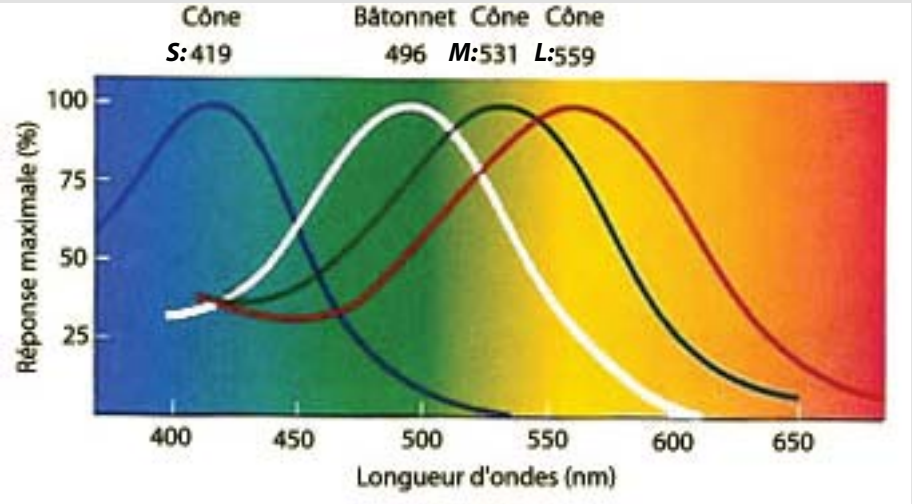


Figure I.5: sensibilité des différents types de photorécepteurs à la longueur d'onde de la lumière [McGillSite]

I.3.1.3. Répartition des photorécepteurs sur la rétine

La population des bâtonnets est nettement plus importante que celle des cônes [Bullier98]. Les bâtonnets seraient au nombre de 120 millions contre 5 millions pour les cônes. La répartition des cônes et des bâtonnets n'est pas uniforme sur la rétine, les bâtonnets étant beaucoup plus nombreux à la périphérie et les cônes plus nombreux près de l'axe optique (à faible excentricité). Cette variation de la concentration des photorécepteurs est représentée sur la figure I.6. La zone centrale (faible excentricité) ne contient que des cônes, elle constitue la fovéa. On y trouve un pas d'échantillonnage des photorécepteurs très faible d'où une vision de haute précision de jour et une vision de très faible sensibilité dans l'obscurité du fait de l'absence de bâtonnets. La zone de moyenne excentricité (parafovéa) comporte une faible proportion de cônes et une population de bâtonnets croissante. Dans cette zone la vision de jour est moins précise, mais la vision en obscurité est performante. Enfin, pour les fortes excentricités, les cônes sont presque absents et laissent la place aux bâtonnets. On remarque aussi le disque optique. C'est une zone de la rétine dépourvue de capteur, car c'est à cet endroit que les axones des neurones qui constituent le nerf otique se rejoignent pour partir en direction du cerveau. Cette zone est aussi appelée "tâche aveugle".

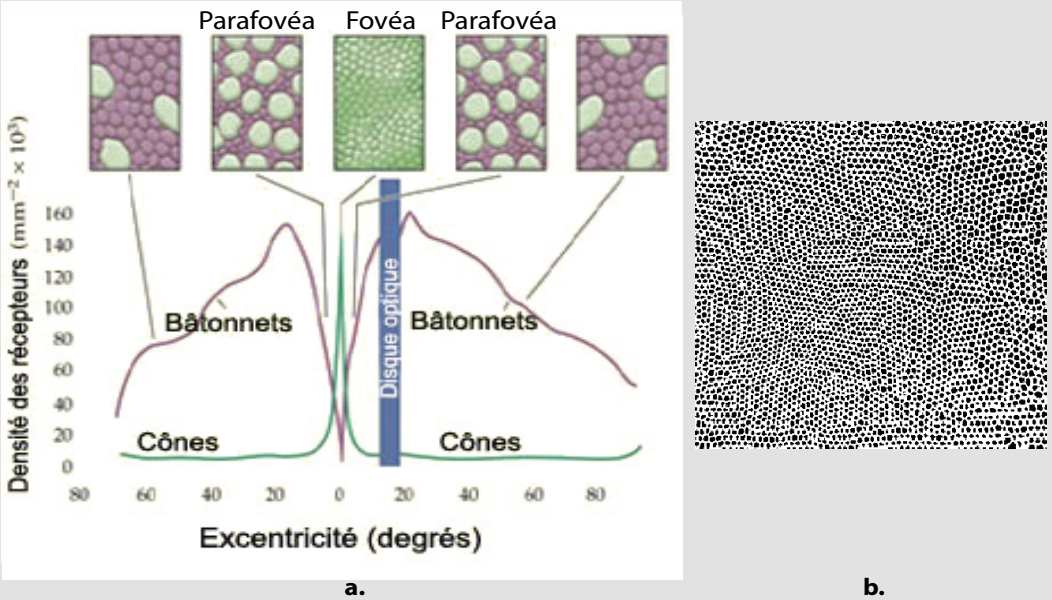


Figure I.6: a, concentration des photorécepteurs en fonction de l'excentricité [McGillSite]. b, aperçu de l'échantillonnage des photorécepteurs sur la rétine [HeraultTeach]

La répartition non régulière et aléatoire (cf. fig. I.6.b) des photorécepteurs est intéressante du point de vue spectral, car l'échantillonnage aléatoire permet une minimisation des problèmes de repliement spectral [Yellott82]. Ces propriétés sont utilisées dans le cadre de l'amélioration des capteurs mono CCD couleur afin d'éviter les problèmes d'apparition de fausses couleurs dans une image [Alleysson05]. Nous ne nous intéresserons pas à cette problématique, car elle débouche d'une part sur la synthèse de capteurs physiques différents ou de filtres spécifiques et d'autre part nous ne travaillerons qu'avec des images en niveau de gris échantillonnées de façon régulière. Nous allons en revanche modéliser les propriétés physiques des photorécepteurs à commencer par leur sensibilité à la lumière afférente. Nous appliquerons ces modélisations aux signaux issus de capteurs CCD couramment utilisés en vision par ordinateur ce qui signifie que nos modélisations sont utilisées sur des capteurs plans plutôt qu'un capteur sphérique comme l'oeil (cf. fig. I.7).

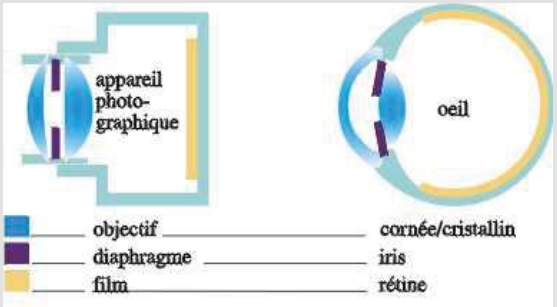


Figure I.7: comparaison entre le capteur naturel , l'oeil et un capteur numérique standard.

I.3.1.4. Compression logarithmique: adaptation locale à la lumière

La compression réalisée par les photorécepteurs peut être modélisée par l'équation de Michaelis-Menten [Beaudot94]:

$$r(p) = \frac{R(p)}{R(p) + R_0(p)} \quad avec \quad R_0(p) = f\left(L(p)\right)$$

(Eq. I.1)

Pour un photorécepteur en position spatiale *p* dans l'image, la luminance corrigée *r(p)* dépend de la luminance *R(p)* reçue et modulée par le paramètre $R_0(p)$ lié à la luminance locale *L(p)* autour de ce capteur. Par action de $R_0(p)$ la courbe de sensibilité de chaque photorécepteur est ajustée aux caractéristiques d'éclairement locales de la scène visualisée : si la luminance locale *L(p)* est faible, $R_0(p)$ est faible et il en résulte une augmentation de la sensibilité des faibles luminances. A l'opposé, une luminance locale *L(p)* forte entraînera un $R_0(p)$ fort, ce qui donne une sensibilité à tendance linéaire. $R_0(p)$ module ainsi le taux de compression. Comme nous travaillons avec des images en niveau de gris codées sur 8 bits, nous ajustons l'équation I.1 de manière à étaler la loi de compression sur tout l'intervalle de travail, ce qui donne la courbe d'équation I.2 présentée sur la figure I.8.

$$r_c(p) = C(p) = \frac{R(p)}{R(p) + R_0(p)} \cdot \left(255 + R_0(p)\right)$$

(Eq. I.2)

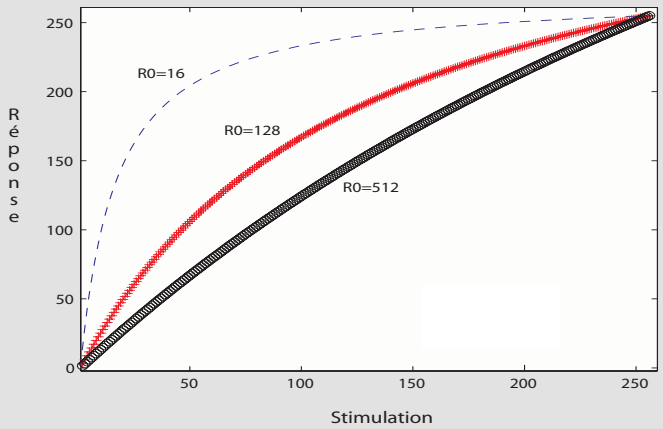


Figure I.8: ajustement de la sensibilité des photorécepteurs par R_0 lié à la luminance locale.

Les zones sombres voient leur dynamique augmenter et les zones claires sont peu modifiées, cette propriété est particulièrement intéressante pour les contre-jours. En effet, dans ce type de situation, notre système visuel est capable de distinguer les détails des régions à faible ou forte luminance contrairement à un appareil photographique standard qui sature les zones sur exposées et donne des zones sombres dépourvues de détails. Cette adaptation est liée à $R_0(p)$, sa valeur est définie pour chaque pixel en fonction de la luminance locale

moyenne *L (p)*, selon la loi linéaire suivante [Durette05]:

$$R_0(p) = \frac{V_0}{256}.L(p) + \left(255 - V_0\right)$$

(Eq. I.3)

Le paramètre V_0 permet d'ajuster la force de la compression. V_0 est ajusté de façon expérimentale à 230, une valeur moindre donne une compression moindre. [Durette05] montre que V_0 peut être choisi dans une gamme de valeurs comprises entre 160 et 250. La valeur de la luminance locale *L(p)* semble être donnée par les couches de cellules en aval des photorécepteurs (le réseau de cellules horizontales). Ceci constitue la première rétroaction du système visuel.

La figure I.9 montre deux exemples d'adaptation à la luminance par les photorécepteurs selon la modélisation présentée précédemment. Sur la figure I.9.a, on observe que la luminance acquise par un appareil photographique est très faible dans les zones de contre-jour au point que l'on n'y perçoit qu'un noir uniforme dans les zones sous exposées. Néanmoins, en appliquant la compression logarithmique, on corrige le défaut de luminance et on se rapproche de ce que le photographe perçoit de ses yeux, mais n'a pu retrouver sur sa photographie. Cette méthode permet de corriger de manière efficace certains défauts des appareils d'acquisition d'images tant que l'information existe (c.-à-d. tant que les zones sombres contiennent un peu d'information sans être saturées).

La figure I.9.b montre l'effet de la compression logarithmique de la luminance sur une image de synthèse sur laquelle des flèches de luminance nulle sont disposées en cercle sur un fond variant progressivement de *255* à *0*. On observe que les motifs sont conservés, la luminance du fond de l'image étant modifiée. Dans la zone sombre persistante, bien que les informations aient à l'origine un niveau de luminance très bas et une dynamique faible, le contraste est renforcé.

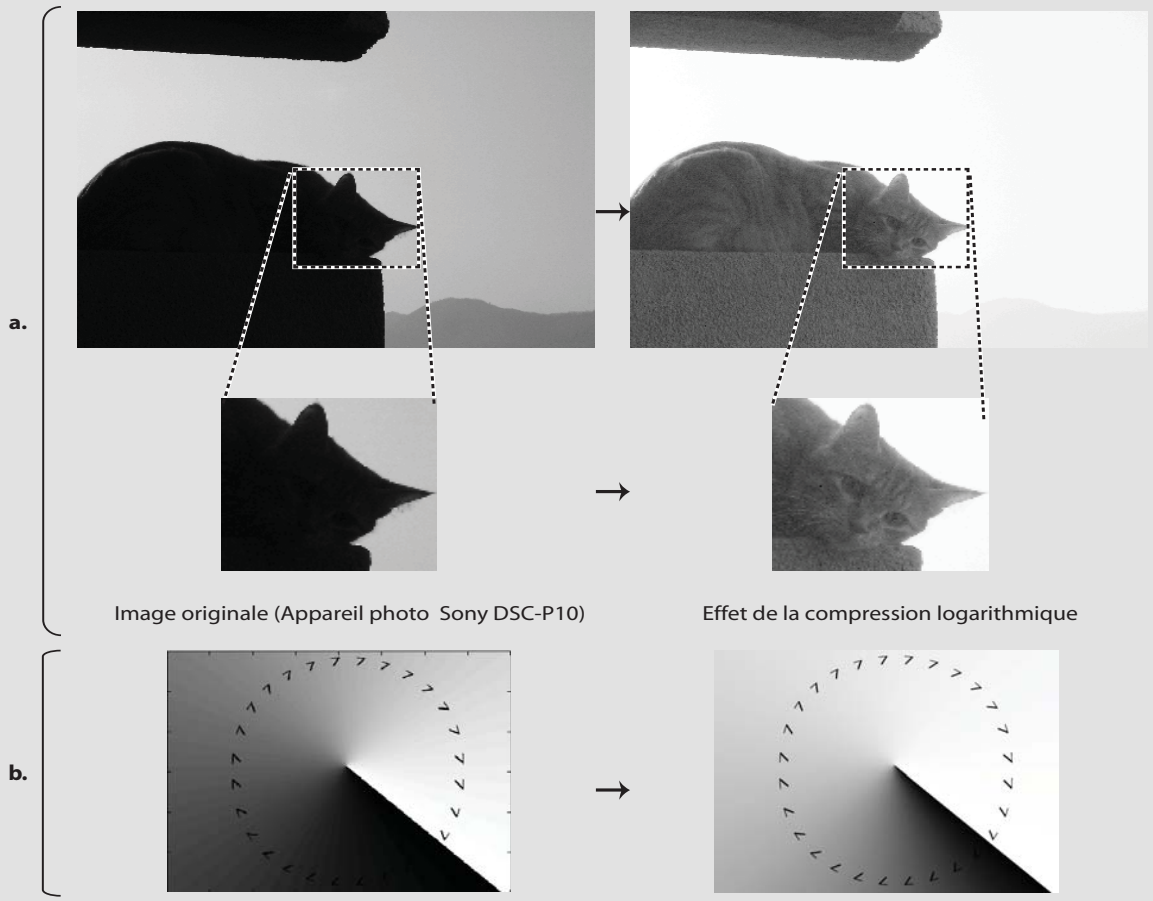


Figure I.9: exemples d'effets de la compression logarithmique: a, comparaison entre une image au format JPG acquise par un appareil photographique "grand public" (Sony DSC-P10) et sa correction par l'algorithme de compression logarithmique. b, effet de la compression logarithmique sur une image de synthèse montrant des motifs noirs sur un fond variant linéairement de blanc à noir.

Notons tout de même que, sur la figure I.9.a, les niveaux de forte luminance (ici, le fond) sont légèrement écrasés par la compression. Ceci est dû au fait que les capteurs CCD et la compression JPG introduisent aussi une compression logarithmique (non adaptative) qui limite l'efficacité de la compression proposée. On remarque également que le résultat après compression est bruité dans les zones qui à l'origine étaient sombres. Ceci est dû au fait que le capteur CCD donne un rapport signal sur bruit faible dans ces zones, car peu d'information de luminance s'y trouve. Après compression logarithmique, la luminance est corrigée, mais le bruit est conservé. Ce bruit très haute fréquence sera éliminé par la suite.

Symbole associé au module de compression logarithmique

La compression logarithmique représente un des premiers traitements effectués par la rétine sur l'information visuelle. Elle considère en entrée l'image en niveaux de gris et une information locale de luminance pour la moduler. La sortie donne une image dans laquelle la luminance de chaque pixel a été réajustée en

fonction de l'illumination de son voisinage proche. Nous associons à ce correcteur réalisant l'ajustement de luminance *C(p)* (cf. Eq.I.2), le symbole défini sur la figure I.10.

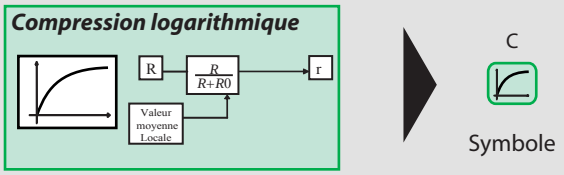


Figure I.10: symbole associé au module de compression logarithmique

Après cette étape de traitement par les photorécepteurs, l'information visuelle résultante se propage dans les couches cellulaires suivantes.

I.3.2. La Couche Plexiforme Externe (PLE)

La couche plexiforme externe correspond à la zone de jonctions entre les photorécepteurs, les cellules bipolaires et les cellules horizontales (cf. figure I.3). Chaque jonction appelée triade synaptique (cf. fig. I.11) permet le traitement de l'information issue des photorécepteurs via des interactions chimiques et électriques entre les différents éléments.

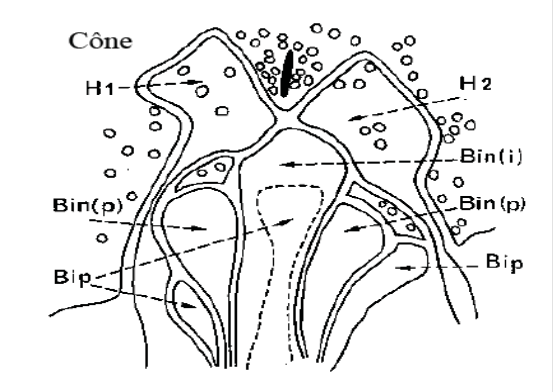


Figure I.11: triade synaptique au sein d'un cône [Buser87]: Bin(i), Bin(p) et Bip sont les dendrites de cellules bipolaires de différents types, H1 et H2 font référence à 2 dendrites de cellules horizontales distinctes.

Chaque triade synaptique permet des interactions entre les signaux de luminance reçus par les photorécepteurs et les signaux porteurs de l'information de luminance locale délivrés par les cellules horizontales. Les cellules bipolaires participent à ces interactions et transmettent ensuite le résultat vers les couches de cellules suivantes.

I.3.2.1. Les cellules horizontales

Ces cellules, proches des cônes et des bâtonnets (cf. figure I.3), sont bien moins nombreuses que les photorécepteurs. Néanmoins, leurs dendrites (prolongations de leur axone permettant les connexions avec les

cellules voisines) s'étendent sur un large périmètre et permettent un nombre très important d'interconnexions entre cellules voisines. Cette stratégie de connexion vise à réduire la quantité d'information à transmettre, suivant en cela le principe d'économie souvent rencontré dans les systèmes biologiques. Ainsi, les dendrites de ces cellules permettent de capter l'information électrique d'un grand nombre de photorécepteurs via des jonctions électriques résistives («Gap») [Mead88]. Chaque cellule horizontale effectue un lissage spatio-temporel de l'information lumineuse des photorécepteurs auxquels elle est rattachée. Les cellules horizontales possèdent donc une information sur l'intensité lumineuse locale de leur voisinage, cette information permettant la compression logarithmique de la luminance au niveau des photorécepteurs (cf. la valeur *L(p)* de l'équation I.3). Nous verrons dans la suite que les cellules horizontales ont également une action dans la transmission en direction des cellules bipolaires de l'information provenant des photorécepteurs.

I.3.2.2. Les cellules bipolaires

Toujours en se référant à la figure I.3, on observe que les cellules bipolaires relient l'ensemble photorécepteurs/cellules horizontales aux cellules ganglionnaires. Elles constituent un relai de l'information. Il en existe deux familles: les cellules bipolaires reliant les bâtonnets aux cellules ganglionnaires et celles reliant les cônes aux cellules ganglionnaires. D'autre part, les cellules bipolaires reçoivent à la fois une information provenant des photorécepteurs et une information provenant des cellules horizontales. Afin de traiter ces deux informations, il existe deux types de cellules bipolaires, les ON réagissant à une excitation des photorécepteurs et à une inhibition des cellules horizontales, et les OFF réagissant à une excitation des cellules horizontales et à une inhibition des photorécepteurs. Ceci nous amène à la notion de champ récepteur.

I.3.2.3. Notion de champ récepteur

Une cellule bipolaire reçoit des connexions synaptiques directes d'un certain nombre de photorécepteurs situés plus ou moins proches d'elle: de un au centre de la fovéa, jusqu'à des milliers dans la périphérie. En plus de ces connexions directes avec les photorécepteurs, les cellules bipolaires reçoivent des afférences de cellules horizontales, elles-mêmes reliées à un grand nombre de photorécepteurs. L'arrangement de ces trois types de cellules est présenté sur la figure I.12. Une cellule bipolaire est reliée directement à des photorécepteurs formant un groupe central et à des cellules horizontales connectées à des photorécepteurs entourant ceux du premier groupe. Par conséquent, le champ récepteur des cellules bipolaires comprend deux parties :

→ Un champ récepteur central recevant une information qui passe directement des photorécepteurs aux cellules bipolaires.

→ Un champ récepteur périphérique qui a au préalable transité par les cellules horizontales. Ceci a été mis en évidence par les travaux de Hartline [Hartline38], ainsi que ceux de Kuffler [Kuffler53] chez le chat et de Barlow [Barlow53] chez la grenouille. Leurs recherches ont permis de caractériser les propriétés des champs récepteurs de la rétine.

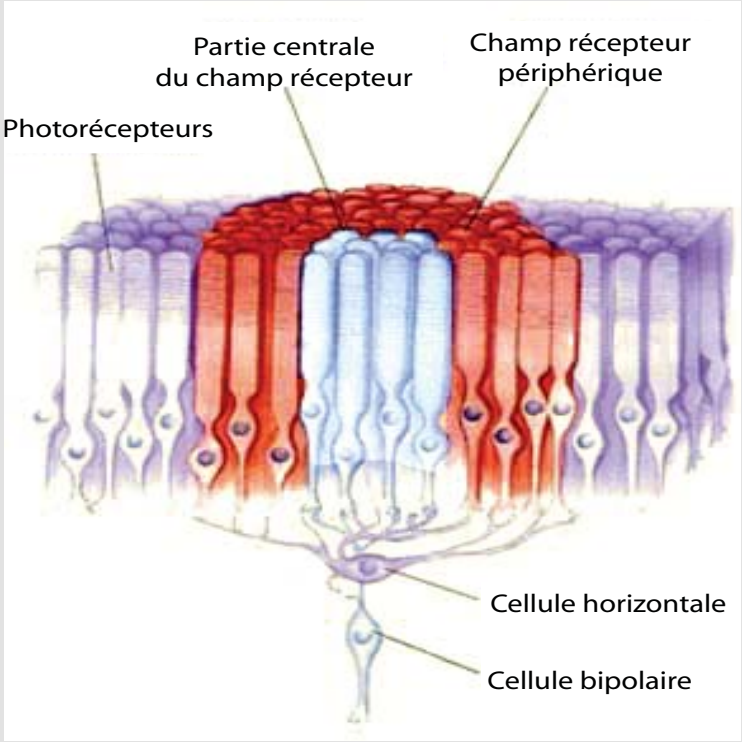


Figure I.12: notion de champ récepteur pour une cellule bipolaire connectée à une cellule horizontale et un ensemble de photorécepteurs [McGillSite]

Au sein de la PLE, les cellules bipolaires de type ON reçoivent une stimulation de la zone centrale et une inhibition de la zone périphérique et inversement pour les cellules bipolaires de type OFF. Cette suppression sélective de certains signaux nerveux s'appelle ***inhibition latérale*** et son rôle est d'augmenter l'acuité d'un signal sensoriel. Dans le cas de la vision, quand une source lumineuse atteint la rétine, elle peut illuminer fortement certains photorécepteurs et d'autres beaucoup moins. En supprimant le signal des photorécepteurs les moins illuminés, les cellules horizontales assurent que seul le signal des photorécepteurs les plus illuminés est transmis aux cellules ganglionnaires, améliorant ainsi le contraste et la définition du stimulus visuel. Les cellules bipolaires ON et OFF sont présentes en quantité égale et offrent deux voies parallèles pour le traitement de l'information visuelle. Par ailleurs, les cellules neuronales ne pouvant coder que les signaux positifs, l'existence des voies parallèles ON et OFF permettent de coder chacune un signal de signe opposé, elles sont donc complémentaires et assurent dans tous les cas une transmission du signal visuel.

I.3.2.4. Modélisation de couches neuronales homogènes

Les travaux de Carver Mead [Mead88] ont constitué une des premières étapes de la mise en correspondance entre les connexions intercellulaires et des modèles électriques équivalents. Le but était de modéliser le filtrage ayant lieu au niveau de la PLE. Dans cette modélisation, les éléments électriques modélisent les connexions entre photorécepteurs, cellules horizontales et cellules bipolaires et les caractéristiques membranaires des cellules horizontales. Ce modèle ne considère qu'un seul réseau résistif capacitif au niveau de la PLE, celui des cellules horizontales.

Un filtre passe-bas spatio-temporel pour modéliser une partie de la PLE

Depuis les travaux pionniers de Mead, Hérault et Beaudot [Beaudot94] ont montré que les connexions au niveau de la PLE sont plus complexes. Ils aboutissent à ***un réseau diffusant plus complet*** qui permet de modéliser un réseau de cellules homogènes au sein de la PLE selon le schéma présenté sur la figure I.13 (pour des raisons de simplicité, nous nous limitons à une représentation 1D). ***Ce réseau n'est pas la modélisation finale de la PLE, il n'en constitue qu'une partie: une couche de cellules homogènes (de même type)***. Chaque neurone d'un réseau de cellules homogènes (qu'elles soient de type photorécepteurs ou cellules horizontales) est représenté par un circuit électrique équivalent. Chaque cellule en position spatiale k reçoit le signal issu d'une source électrique pure à travers la résistance interne r. La jonction avec ses cellules voisines est représentée par une résistance R (jonction électrique résistive "Gap") et les caractéristiques membranaires de ces cellules sont modélisées par la résistance r_m et la capacité mise en parallèle C_m. Les sorties des cellules sont les $s(k)$.

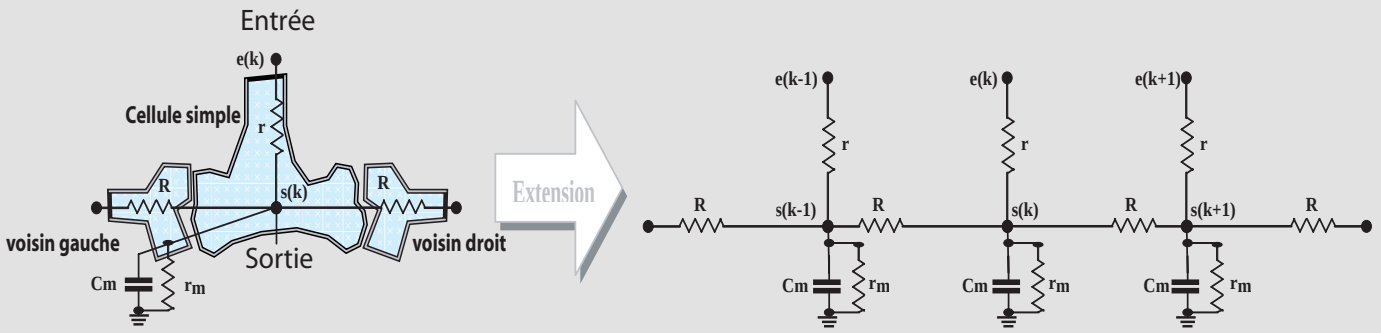


Figure I.13: modélisation des connexions entres cellules de même type au sein de la PLE (photorécepteurs ou cellules horizontales) [HeraultTeach] : a. modèle biologique, b. modèle électrique équivalent

Du fait de la capacité qui introduit une réponse temporelle variable, nous considérons des signaux d'entrée dépendants de la position spatiale k et du temps t notés $e(k, t)$. La sortie est elle aussi dépendante de l'espace et du temps et est notée $s(k, t)$. Ce type d'écriture considère donc une variable spatiale discrète k et un temps continu t. Par ailleurs, nous allons considérer que chaque neurone est espacé de façon régulière vis-à-vis de ses voisins (le pas Δ d'échantillonnage spatial est constant et est fixé à 1). Ceci nous permet de nous rapprocher des capteurs CCD actuels dont les pixels suivent un agencement régulier. On s'éloigne du modèle biologique sur ce point, mais ceci est préférable puisque nous ne travaillons qu'avec des capteurs CCD.

L'équation I.4 résulte de l'écriture de la loi des noeuds pour une cellule du réseau :

$$\frac{e(k,t)-s(k,t)}{r}+\frac{s(k-1,t)-s(k,t)}{R}+\frac{s(k+1,t)-s(k,t)}{R}-\frac{s(k,t)}{r_m}-C_m\frac{ds(k,t)}{dt}=0 \qquad \textit{(Eq. I.4)}$$

Afin de faciliter l'analyse d'une telle équation, on se place dans le domaine fréquentiel. Soient $E(fs, ft)$ et $S(fs, ft)$ les transformées de Fourier du signal d'entrée et de sortie de chaque cellule. Les nouvelles variables sont fs, fréquence spatiale (unité : cycle par unité de longueur Δ) et ft, fréquence temporelle exprimée en cycle

par unité de temps.

En posant $\alpha=r/R$, $\beta=r/r_m$ et $\tau=rC_m$, on aboutit à la fonction de transfert suivante [Beaudot94, Herault01]:

$$G(fs,ft)=\frac{S(fs,ft)}{E(fs,ft)}=\frac{1}{1+\beta+2\alpha\left(1-\cos\left(2\pi fs\right)\right)+j2\pi.\tau.ft} \qquad \textit{(Eq. I.5)}$$

Dans cette équation, les 3 paramètres α, β et τ sont ajustés différemment selon les cellules modélisées. Le paramètre α, constante d'espace, contrôle les caractéristiques de diffusion spatiale, β contrôle le gain à fréquence nulle et τ, la constante de temps, contrôle le comportement temporel.

La figure I.14 donne une représentation graphique de cette fonction de transfert. Comme les cellules occupent des positions spatiales discrètes à pas d'échantillonnage constant, le spectre de la fonction de transfert est périodique selon l'axe *fs*. Le tracé de la fonction de transfert est donc limité à un intervalle de fréquences spatiales réduites de -0.5 à 0.5.

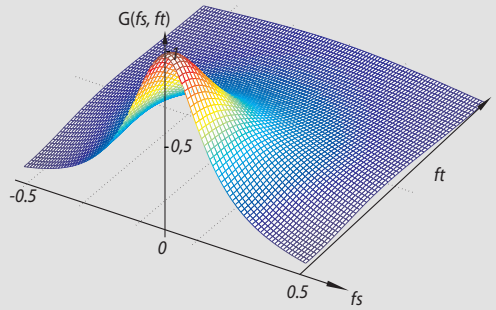


Figure I.14: fonction de transfert du filtre passe bas spatio-temporel

On observe un effet de filtre passe-bas spatio-temporel. Le gain est maximal et unitaire pour les fréquences spatio-temporelles nulles ($fs=ft=0$). Le gain s'atténue quand les fréquences spatiales et/ou temporelles augmentent. Cela signifie par exemple que si l'entrée de ce filtre est un stimulus statique, les contours nets qui constituent les hautes fréquences spatiales sont atténués ce qui donne un rendu plus flou. Par ailleurs, si l'on applique à l'entrée de ce filtre un stimulus en mouvement, les informations qui varient dans le temps (mouvement et bruit temporel) et qui constituent les hautes fréquences temporelles sont atténuées par ce filtrage. Ce filtre privilégie les zones statiques (faibles fréquences temporelles) et homogènes (faibles fréquences spatiales) d'une scène.

Il est important de remarquer que le filtre spatio-temporel proposé n'est pas séparable d'où deux conséquences:

→ Le filtrage spatial évolue dans le temps: après application du stimulus d'entrée, on note un régime transitoire dont la durée est liée au paramètre temporel τ.

→ Lorsque ce filtre est stimulé par un signal variant spatiotemporellement, par exemple par un objet en mouvement translationnel tel que $x(k,t)=x(k-vt, t)$, la sortie suivra la variable $k-vt$, le terme ft étant remplacé

par *vfs* dans l'équation I.5. Cela signifie que l'information de vitesse est préservée.

La figure I.15 montre des exemples d'un tel filtrage. Dans le cas simple du carré et du rond noirs statiques (cf. fig.I.15.a) sur fond blanc, les bords sont lissés, illustrant le comportement typique d'un filtre passe-bas spatial. La figure I.15.b montre l'effet du filtre sur un objet en mouvement translationnel, les hautes fréquences spatiales et temporelles sont atténuées, l'objet apparaît très flou et l'on observe un effet de traînée selon la direction du déplacement. La figure I.15.c montre que le bruit haute fréquence qui nuit à la visualisation des détails de l'image est fortement atténué.

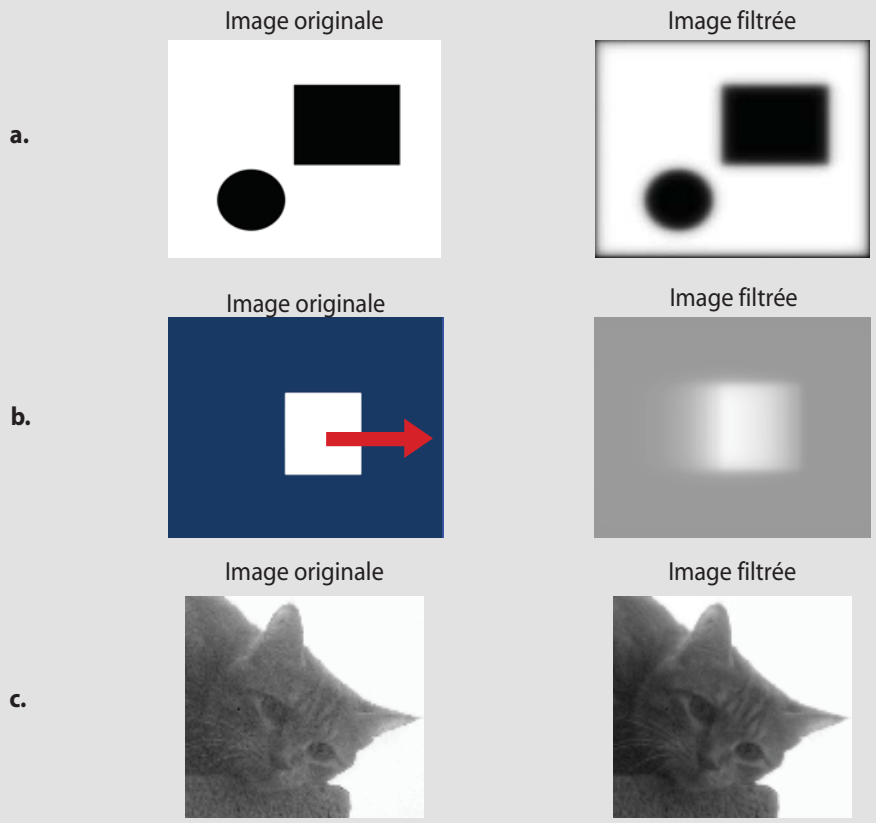


Figure I.15: illustration de l'effet du filtre passe-bas sur une image de type binaire (noir et blanc) (a), sur un objet en translation (b) et sur une image bruitée (c): le bruit haute fréquence est éliminé, l'information visuelle est renforcée.

Mise en oeuvre du filtre

Pour la mise en oeuvre de ce filtre passe-bas spatio-temporel nous utilisons l'algorithme proposé dans [Beaudot94] et repris par [Durette05]. On montre tout d'abord que l'équation de récurrence du filtre peut se mettre sous la forme (I.6).

$$s(k_1,k_2,t)=\frac{1}{1+\beta+4\alpha+\tau}.u(k_1,k_2,t)+\frac{\alpha}{1+\beta+4\alpha+\tau}.\left[s(k_1-1,k_2,t)+s(k_1+1,k_2,t)+s(k_1,k_2-1,t)+s(k_1,k_2+1,t)\right]$$

avec

$$u(k_1,k_2,t)=e(k_1,k_2,t)+\tau.s(k_1,k_2,t-1) \qquad (Eq.\ I.6)$$

Ceci permet de mettre en évidence un algorithme composé d'un filtrage purement spatial rebouclé temporellement comme le montre la figure I.16.

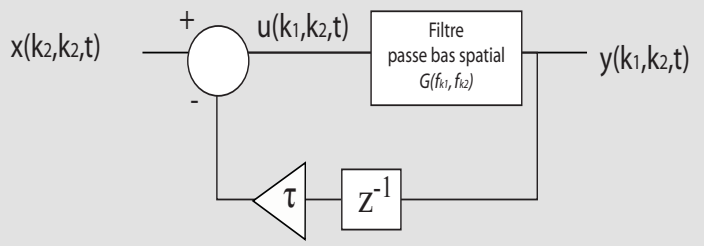


Figure I.16: diagramme pour la mise en oeuvre du filtre passe-bas spatio-temporel associé aux couches neuronales homogènes

Beaudot [Beaudot94] montre que l'équation du filtre passe-bas purement spatial en 2 dimensions est de la forme (I.7), dans laquelle f_{k1} et f_{k2} sont respectivement les fréquences horizontale et verticale.

$$G(f_{k_1},f_{k_2})=\frac{1}{1+\beta+4\alpha-2\alpha\cos\left(2\pi f_{k_1}\right)-2\alpha\cos\left(2\pi f_{k_2}\right)} \qquad (Eq.\ I.7)$$

Ceci est approché par le produit de deux fonctions de transfert unidimensionnelles de direction perpendiculaire : l'une verticale, l'autre horizontale. On introduit dans chaque fonction de transfert le paramètre μ fixé à 0.8 permettant d'améliorer la finesse de l'approximation:

$$G(f_{k_1},f_{k_2})=G_{k_1}(f_{k_1})\times G_{k_2}(f_{k_2}) \qquad (Eq.\ I.8)$$

$$avec\quad G_i(f_i)=\frac{1}{1+\beta+2\alpha\mu\left(1-\cos\left(2\pi f_i\right)\right)}$$

On aboutit à la fonction de transfert (cf. Eq. I.9) réalisée à partir du produit des deux filtrages monodimensionnels discrets et de direction opposée. Le filtre réalisé consiste en quatre filtrages successifs, verticaux et horizontaux, causaux et anticausaux. Plus précisément, on filtre successivement l'image d'entrée par un filtre horizontal causal puis un filtre horizontal anticausal suivi d'un filtre vertical causal et d'un filtre vertical anticausal, le résultat final étant multiplié par le gain constant du filtre.

$$G(z_{k_1}, z_{k_2}) = \left[\frac{(1-a)^2}{\sqrt{1+\beta}} \cdot \frac{1}{1+az_{k_1}^{-1}} \cdot \frac{1}{1+az_{k_1}} \right] \cdot \left[\frac{(1-a)^2}{\sqrt{1+\beta}} \cdot \frac{1}{1+az_{k_2}^{-1}} \cdot \frac{1}{1+az_{k_2}} \right]$$

soit

$$G(z_{k_1}, z_{k_2}) = \frac{(1-a)^4}{1+\beta} \cdot \frac{1}{1+az_{k_1}^{-1}} \cdot \frac{1}{1+az_{k_1}} \cdot \frac{1}{1+az_{k_2}^{-1}} \cdot \frac{1}{1+az_{k_2}} \qquad \textit{(Eq. I.9)}$$

L'implantation de ce filtre est alors simple puisqu'elle consiste à décomposer le filtre en quatre filtrages récursifs successifs (cf. [Beaudot94]) d'où un coût de calcul réduit de 5 produits par pixels.

A titre de comparaison, en prenant un filtre gaussien passe-bas *H* de taille n*n, on aurait un coût de calcul par pixel lié à la taille du filtre, à savoir un coût de *n*n* opérations.

Symbole associé au filtre passe-bas spatio-temporel générique

Dans toute la suite de ce manuscrit, nous appellerons ce filtre "filtre passe-bas spatio-temporel ". Nous lui assignons le symbole présenté sur la figure I.17.

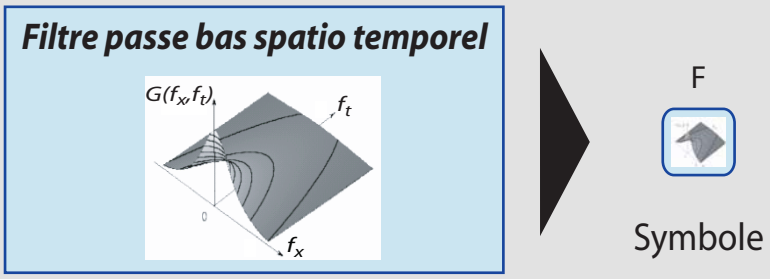


Figure I.17: symbole associé au filtre passe-bas spatio-temporel

I.3.2.5. Modélisation de l'ensemble de la PLE

La PLE étant la zone de jonctions entre le réseau des photorécepteurs, le réseau des cellules horizontales et les cellules bipolaires, sa modélisation tient compte de l'interconnexion entre ces réseaux cellulaires (cf. fig. I.18) [Beaudot94, Hérault01]. On aboutit à un schéma dans lequel on trouve deux filtrages passe-bas spatio-temporels. Un premier au niveau des photorécepteurs et un second au niveau des cellules horizontales. Le premier filtre a pour but de limiter le bruit d'acquisition (deux capteurs voisins n'ont pas exactement les mêmes caractéristiques et ce filtrage permet d'apporter une certaine homogénéité). Le second permet d'extraire la luminance locale. Enfin, au sein d'une même triade synaptique, les cellules bipolaires de type ON font la différence entre la réponse des photorécepteurs et celle des cellules horizontales et inversement pour les cellules bipolaires de type OFF. Les voies ON et OFF portent chacune une information exclusivement positive, car ***les neurones ne peuvent coder que des valeurs positives***. On fusionne les voies ON et OFF en une seule voie par l'opération différence sur les schémas présentés sur les figures I.18-b-c. La partie positive du signal de sortie représente le signal des cellules bipolaires ON, la partie négative représente le signal des cellules bipolaires OFF.

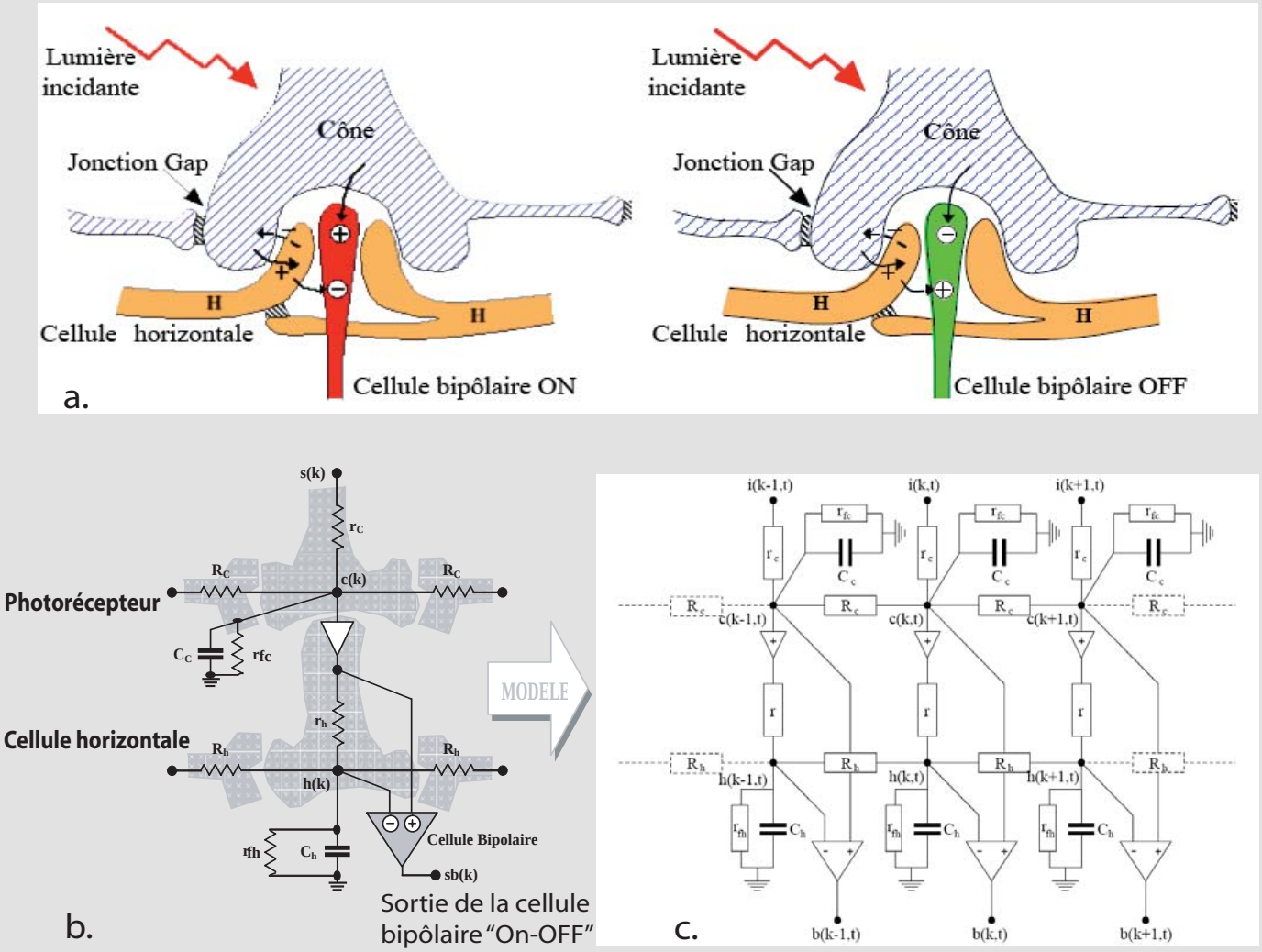


Figure I.18 :modèle de PLE [Beaudot94]: a, interactions entre les différentes cellules au niveau de la PLE de la rétine. b, modélisation au niveau d'un capteur (pixel) des liens électriques entre un photorécepteur, une cellule horizontale de même position spatiale et les cellules voisines auxquelles ils sont connectés (qui correspondent aux pixels voisins). c, vue plus large du modèle: application du modèle sur un réseau de trois photorécepteurs et trois cellules horizontales.

Le schéma de la figure I.19 propose une modélisation des deux réseaux interconnectés. Sur cette figure, un exemple de signal entrant de type échelon illustre l'action de chacun des filtrages.

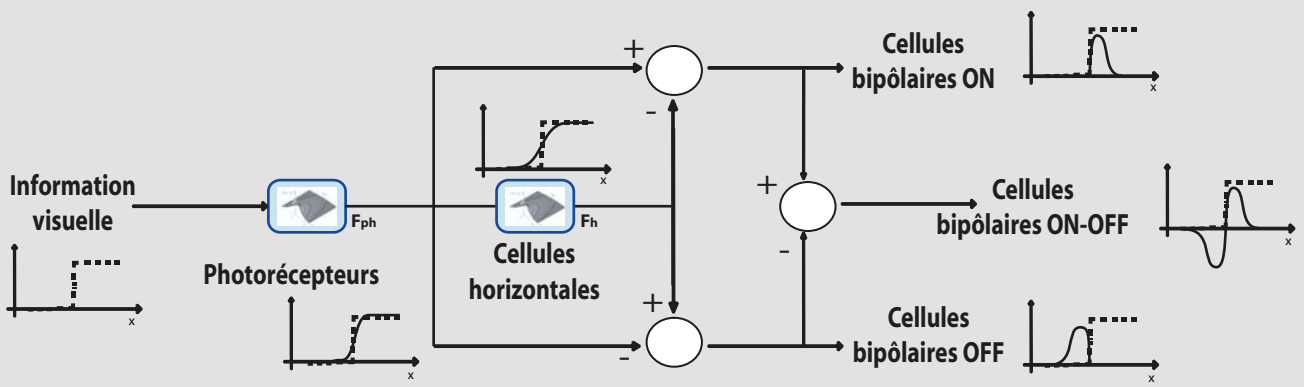


Figure I.19: Modélisation et illustration avec un signal spatial 1D de la PLE à l'aide des modules élémentaires

Analysons maintenant la fonction de transfert de ce système, pour cela, nous considérerons la sortie bipolaire ON-OFF pour plus de commodité. Les photorécepteurs sont modélisés par un premier filtre passe-bas spatio-temporel dont la fonction de transfert et les paramètres porteront l'indice *"ph"* et les cellules horizontales sont modélisées par un second filtre de même type dont l'indice des paramètres sera *"h"*.

Pour chaque réseau, on a:

$$F_{ph}(fs,ft)=\frac{1}{1+\beta_{ph}+2\alpha_{ph}\left(1-\cos\left(2\pi fs\right)\right)+j2\pi.\tau_{ph}.ft}$$

$$F_{h}(fs,ft)=\frac{1}{1+\beta_{h}+2\alpha_{h}\left(1-\cos\left(2\pi fs\right)\right)+j2\pi.\tau_{h}.ft}$$

(Eq. I.10)

Comme chaque sortie des cellules bipolaires ne code que la partie positive du signal, l'ensemble du système décrit par le schéma de la figure I.19, donne la fonction de transfert suivante.

$$G_{PLE}(fs,ft)=G_{BipolaireON}(fs,ft)-G_{BipolaireOFF}(fs,ft)$$

avec

$$\begin{cases} G_{BipolaireON}(fs,ft)=F_{ph}(fs,ft).\left[1-F_{h}(fs,ft)\right] \text{ si } F_{ph}(fs,ft).\left[1-F_{h}(fs,ft)\right]>0 \\ G_{BipolaireON}(fs,ft)=0 \text{ sinon} \end{cases}$$

et

$$\begin{cases} G_{BipolaireOFF}(fs,ft)=-F_{ph}(fs,ft).\left[1-F_{h}(fs,ft)\right] \text{ si } F_{ph}(fs,ft).\left[1-F_{h}(fs,ft)\right]<0 \\ G_{BipolaireOFF}(fs,ft)=0 \text{ sinon} \end{cases}$$

soit

$$G_{PLE}(fs,ft)=F_{ph}(fs,ft).\left[1-F_{h}(fs,ft)\right]$$

(Eq. I.11)

Nous pouvons interpréter cette équation comme la différence entre deux filtres passe-bas spatio-temporels de paramètres différents. La fonction de transfert résultante est représentée sur la figure I.20.

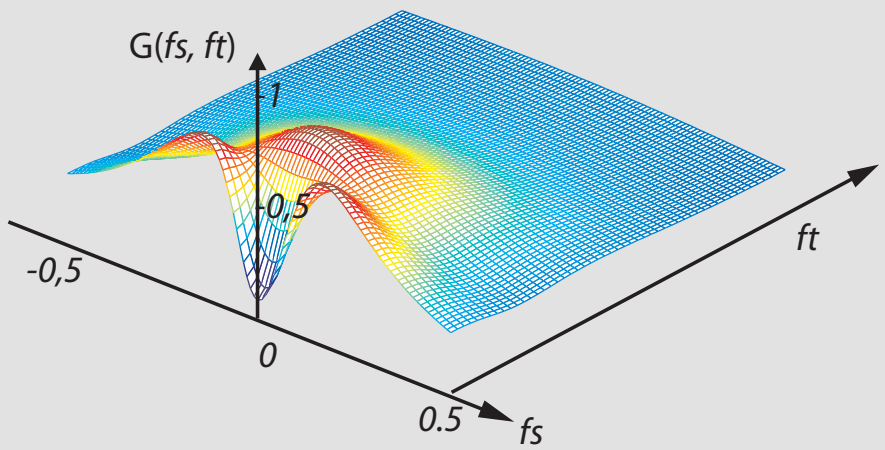


Figure I.20: fonction de transfert de la couche PLE de la rétine.

Ce filtre agit à faibles fréquences temporelles ou spatiales comme un filtre passe-bande. Le filtre montre également une tendance passe-bas spatiale pour les hautes fréquences temporelles ainsi qu'une tendance passe-bas temporelle pour les hautes fréquences spatiales. Nous retiendrons que la sensibilité à fréquence spatiale nulle dépend du paramètre β_h, que (α_{ph}, α_h), les constantes d'espace, et (τ_{ph}, τ_h) les constantes de temps contrôlent les fréquences de coupure des filtres passe-bande spatiales et temporelles. Le paramétrage des fréquences de coupure des filtres est explicité dans la thèse de Beaudot et est résumé ci-dessous :

Le premier filtre ayant pour but de minimiser le bruit spatio-temporel, ses fréquences de coupure spatiale et temporelle sont hautes. La constante d'espace α_{ph} et la constante de temps τ_{ph} sont donc basses. A l'opposé, le second filtre modélisant l'action des cellules horizontales extrait la luminance locale, sa fréquence de coupure spatiale est donc inférieure à celle des photorécepteurs, d'où une constante d'espace α_h supérieure à α_{ph}. Si l'on veut une adaptation rapide à la luminance locale, la constante de temps τ_h doit être réduite. Enfin, la composante continue c.-à-d. la valeur de luminance sur la scène visualisée est éliminée si la valeur de β_h est nulle. Nous reprendrons précisément ce paramétrage dans le chapitre suivant pour une utilisation en vision par ordinateur.

Propriétés de la PLE: comportement statique et renforcement des contours

Le réseau des photorécepteurs réalise un premier filtre passe-bas spatial uniquement puisque l'image considérée est statique. Ceci aboutit à une image dans laquelle le bruit spatial haute fréquence est fortement atténué et l'information visuelle basse fréquence préservée (cf. fig. I.21.b). Pour cela, on ajuste la fréquence de coupure spatiale de manière à éliminer le bruit apparaissant entre pixels voisins, de manière très localisée (constante d'espace α_{ph} du filtre fixée à 1 pixel).

L'information est ensuite filtrée par le réseau de cellules horizontales qui effectue un second filtrage passe-bas spatial uniquement (stimulus statique). Ce filtrage a une fréquence de coupure inférieure à celle du filtre des photorécepteurs (constante d'espace α_h fixé à 5 pixels). Autrement dit, à la sortie du filtre, l'image est plus floue et ne donne en fait qu'une information sur la luminance locale moyenne de la scène visualisée

(cf. fig I.21.c).

Enfin, les cellules bipolaires ON effectuent la différence entre la réponse des photorécepteurs et celle des cellules horizontales et les cellules bipolaires OFF effectuent l'opération inverse (cf. fig I.21.d et I.21.e), chacune ne codant que l'information résultante positive. En rassemblant ces deux dernières réponses (différence ON-OFF), on observe une image dont les contours sont renforcés, l'information de luminance moyenne et le bruit ayant été supprimés (cf. fig. I.21.f).

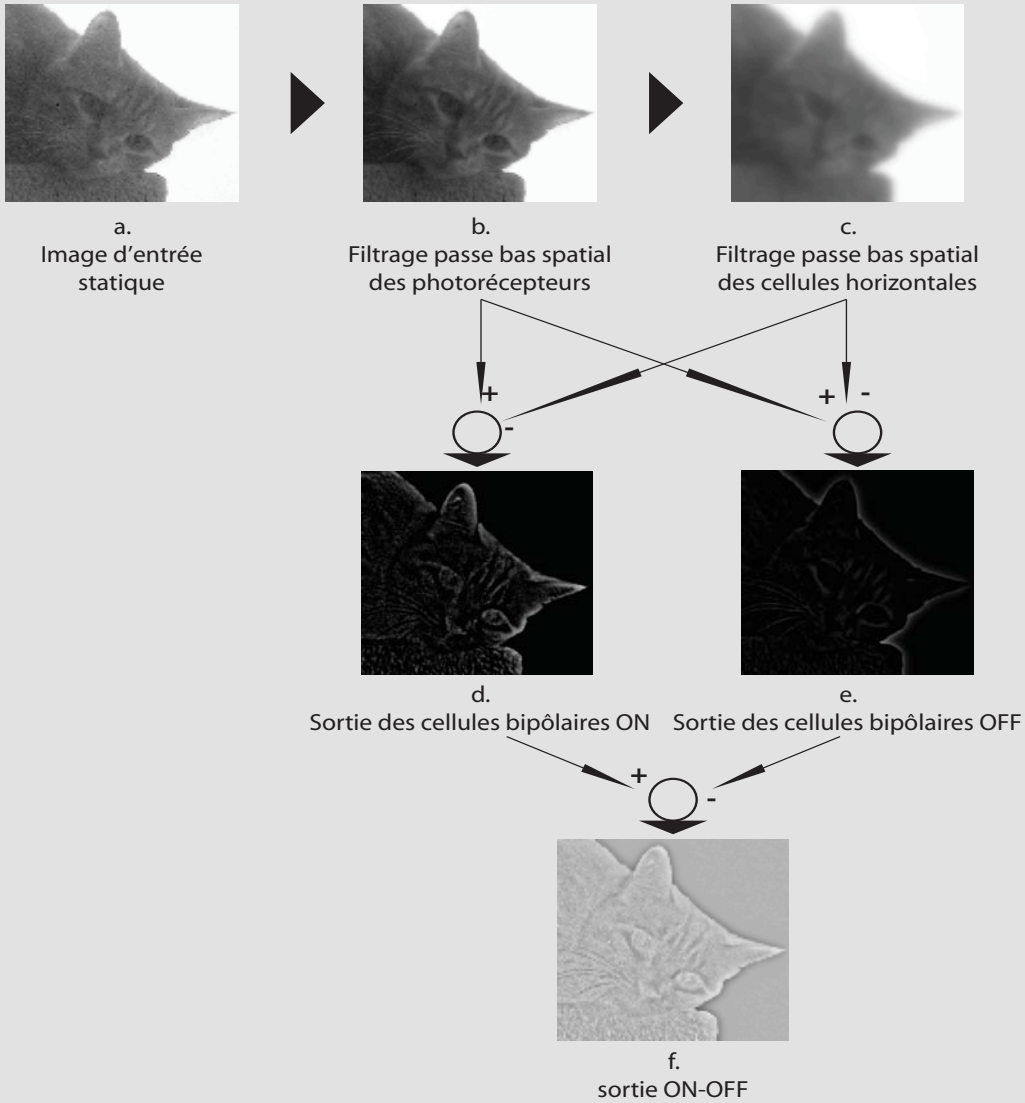


Figure I.21: effet de l'enchaînement des traitements entre les différentes cellules au sein de la PLE dans le cas d'une image statique.

Si on fait précéder le module PLE par le module de compression logarithmique de luminance des photorécepteurs (cf. fig. I.22.a), l'ensemble des traitements réalisés est illustré sur la figure I.22.b pour laquelle

une image d'entrée sous-exposée a été traitée pour en faire ressortir tous les contours.

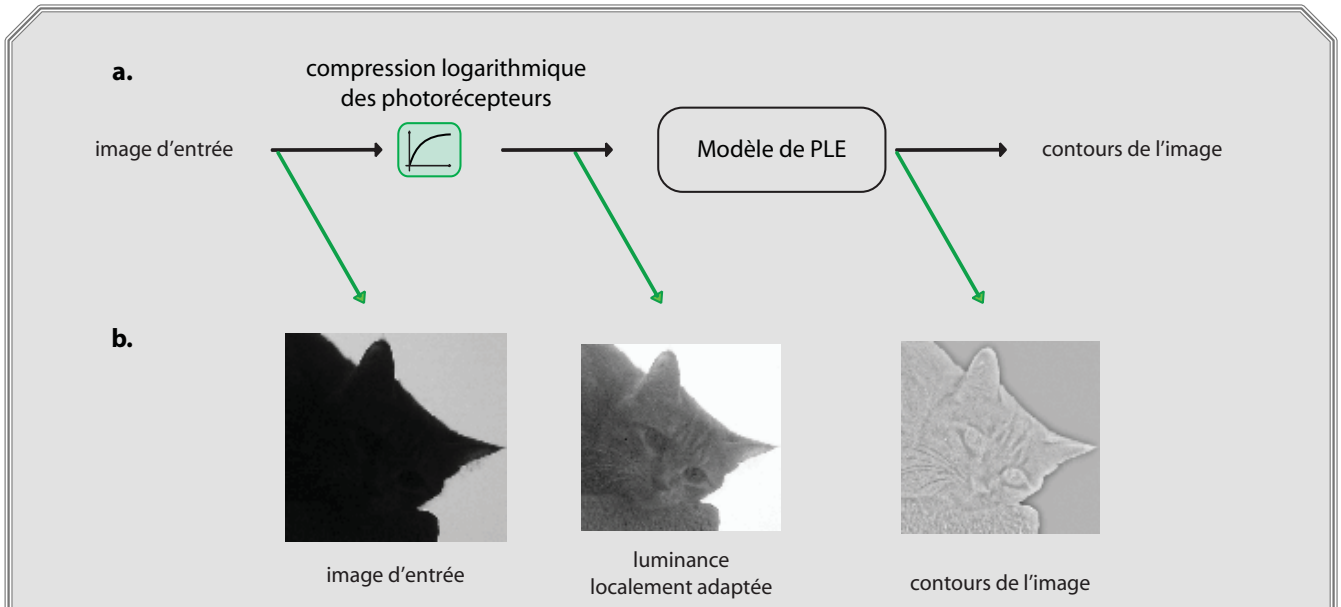


Figure I.22: extraction des contours dans une image par assemblage du module de compression logarithmique des photorécepteurs et du modèle de la PLE.

Propriété de la PLE: blanchiment spectral

Comme les images naturelles ont un spectre qui décroît selon une loi de type 1/|f| [Atick92] et comme la PLE rehausse les hautes fréquences, il en résulte un effet de blanchiment spectral. Par conséquent, les hautes fréquences associées aux détails de l'image sont rehaussées. La figure I.23.a montre le principe du blanchiment spectral sur un spectre de base de forme 1/f et le spectre résultant après multiplication par la fonction de transfert de la rétine. Il est également montré sur la figure I.23.b les spectres avant et après filtrage PLE sur les imagettes du chat présentées précédemment. Le module du spectre initial a une information dominante en basse fréquence (centre de l'image du spectre). Après filtrage, l'énergie du spectre à plus haute fréquence est rehaussée donnant ainsi accès à plus de précision sur les détails de l'image.

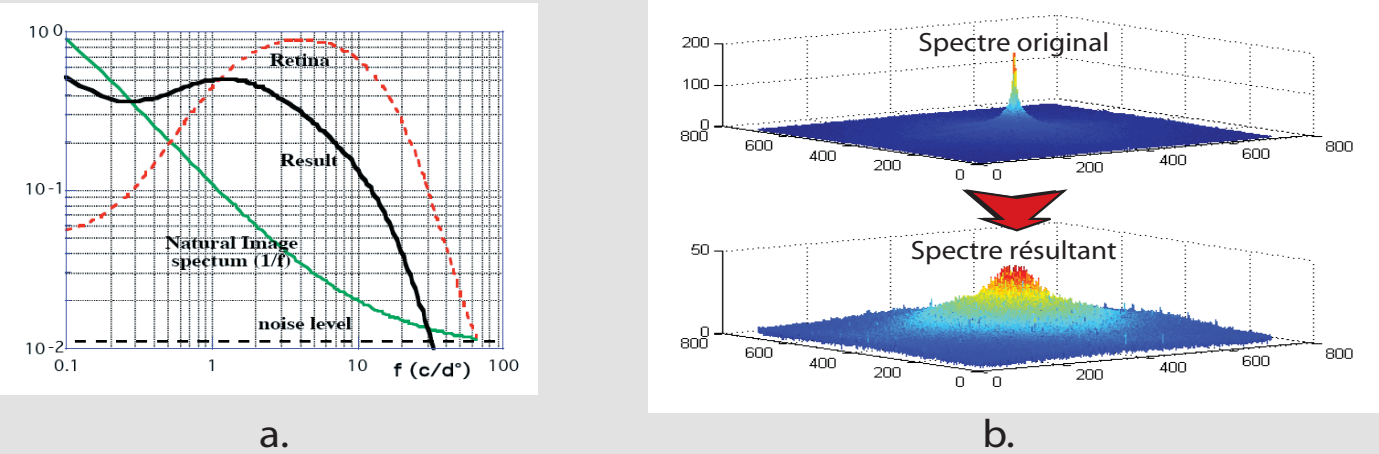


Figure I.23: blanchiment spectral de la rétine : a. application sur un spectre de forme 1/f de la fonction de transfert du filtre PLE [HeraultTeach]. b. spectre initial et traité par le filtre PLE pour l'imagette "chat"

Propriétés de la PLE: comportement dynamique

D'après [Beaudot94], la réponse impulsionnelle du filtre associé à la modélisation de la PLE s'écrit en 1D:

$$g_{PLE}(f_x,t) = TF^{-1}\left|F_{ph}(fs,ft).\left[1-F_h(fs,ft)\right]\right|$$

soit

$$g_{PLE}(f_x,t) = g_1(f_x,t) + g_2(f_x,t)$$

avec

$$g_1(f_x,t) = \frac{\beta_h + 2\alpha_h\left(1-\cos\left(2\pi f_x\right)\right)}{1+\beta_h+2\alpha_h\left(1-\cos\left(2\pi f_x\right)\right)}.\frac{1}{1+\beta_{ph}+2\alpha_{ph}\left(1-\cos\left(2\pi f_x\right)\right)}.\left(1-e^{-t/\tau_{ph}'}\right).e(t$$

$$g_2(f_x,t) = \frac{1}{1+\beta_h+2\alpha_h\left(1-\cos\left(2\pi f_x\right)\right)}.\frac{1}{\tau_{ph}}.\frac{\tau_{ph}'.\tau_h'}{\tau_{ph}'-\tau_h'}.\left(e^{-t/\tau_{ph}'}-e^{-t/\tau_h'}\right).e(t)$$

avec

$$\tau_{ph}' = \frac{\tau_{ph}}{1+\beta_{ph}+2\alpha_{ph}\left(1-\cos\left(2\pi f_x\right)\right)}$$

et

$$\tau_h' = \frac{\tau_h}{1+\beta_h+2\alpha_h\left(1-\cos\left(2\pi f_x\right)\right)}$$

(Eq. I.12)

Cette réponse impulsionnelle est présentée sur la figure I.24. On observe un comportement dit "coarse to fine". Lors de l'application du stimulus, la PLE se comporte comme un filtre passe-bas spatial puis, avec le temps, le filtre devient passe-bande. L'application du stimulus correspond à un régime transitoire à haute fréquence temporelle, gamme de fréquences pour laquelle la PLE a un comportement passe-bas spatial. Le

comportement passe-bande n'est présent qu'un certain temps après l'application du stimulus, ce qui correspond au régime permanent.

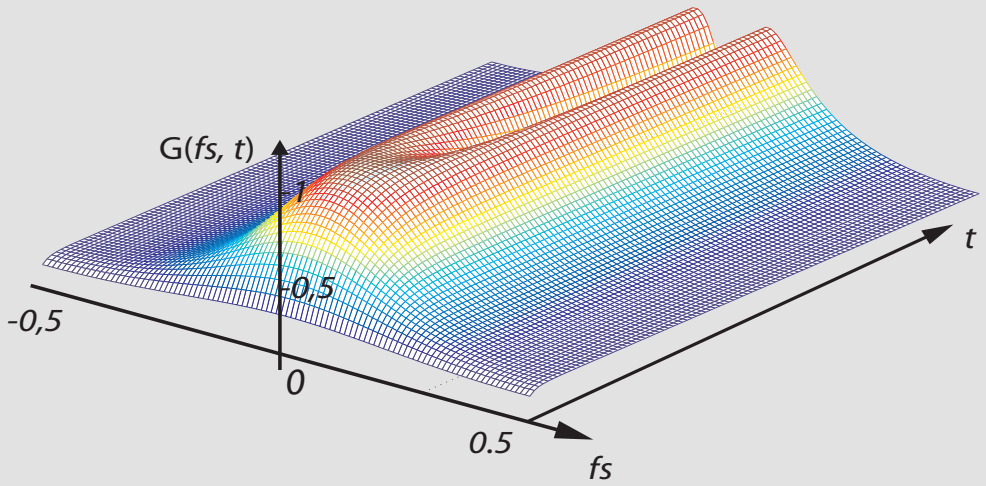
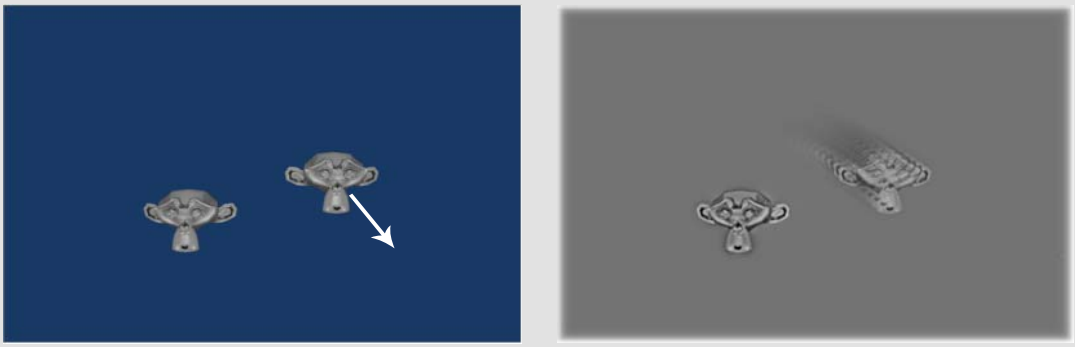


Figure I.24: réponse impulsionnelle du filtre PLE. On observe une tendance passe-bas dans les premiers instants de la réponse (régime transitoire). Le filtre devient ensuite passe-bande en régime permanent.

Ce comportement est illustré sur la figure I.25 sur laquelle on voit une image d'une séquence vidéo sur laquelle deux figurines sont présentes. La première à gauche immobile, donne en sortie de filtre PLE une zone pour laquelle tous les contours apparaissent nettement, on est en régime établi de la fonction de transfert. La seconde figurine mobile (à droite), donne une zone peu détaillée (contours flous), on est dans le régime transitoire de la fonction de transfert.



a. b.

Figure I.25: a. extraits d'une séquence vidéo avec deux figurines, l'une mobile (droite), l'autre immobile (gauche). b, réponse de la PLE, les contours sont bien définis pour la figurine immobile mais plus flous pour l'objet mobile.

En résumé, la PLE extrait les contours statiques et atténue les hautes fréquences des objets mobiles. On perçoit les objets statiques de façon nette et ceux en mouvement sont vu plus flous. Le bruit très haute fréquence spatio-temporel sera donc minimisé, mais l'information de vitesse basse fréquence des objets mobiles est tout de même préservée du fait de la propriété des filtres passe-bas spatio-temporels utilisés (rappel: lorsque ces filtres sont stimulés par un signal variant spatiotemporellement, par exemple par un objet en mouvement translationnel tel que $x(k,t)=x(k-vt, t)$, la sortie suivra la variable $k-vt)$.

I.3.3. La Couche Plexiforme Interne (PLI)

La couche PLI est le dernier étage de traitement au niveau de la rétine avant le nerf optique. L'information entrante résulte des traitements de l'information visuelle par la PLE, elle est transmise par les cellules bipolaires. Au niveau de cette couche, on trouve des interactions entre les cellules bipolaires, les cellules ganglionnaires et les cellules amacrines. Le résultat des interactions au niveau de la PLI est disponible au niveau des cellules ganglionnaires dont les axones forment le nerf optique.

I.3.3.1. Les cellules amacrines

Présentation générale

Chez l'homme, les cellules amacrines sont de 20 types différents dont le rôle de beaucoup reste encore inconnu. Elles sont impliquées dans des processus de modulation de gain de la réponse des cellules bipolaires et des cellules ganglionnaires [Kolb96]. Elles modulent aussi la réponse des champs récepteurs des cellules ganglionnaires en fonction de l'intensité des signaux afférants.

On distingue actuellement deux familles de cellules amacrines: les amacrines dites «phasiques» ne réagissant qu'aux transitions du stimulus afférant et les amacrines dites «toniques» qui se maintiennent polarisées ou dépolarisées pendant la stimulation lumineuse sans développer d'influ nerveux.

Modélisation des cellules amacrines

Le type de cellules amacrines pour lequel il existe une modélisation avancée sont les cellules amacrines de type A-II. Ces cellules amacrines sont phasiques et ont un comportement purement passe-haut temporel [Herault01]. Elles interviennent au niveau de la PLI dans la dyade synaptique [Werblin88]. Leur modélisation consiste à utiliser un filtre passe-haut temporel du premier ordre et de constante de temps τ_A dont la réponse impulsionnelle est la suivante:

$$a(t) = \delta(t) - \left(1/\tau_{\mathrm{A}}\right).\exp(-t/\tau_{\mathrm{A}}) \qquad \textit{(Eq. I.13)}$$

La représentation graphique de cette réponse impulsionnelle est présentée sur la figure I.26.a, nous lui associons le symbole présenté sur la figure I.26.b.

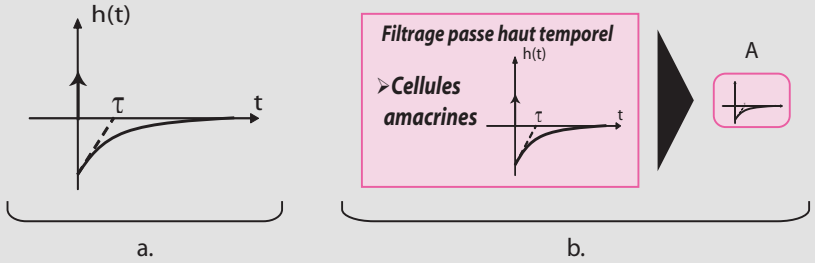


Figure I.26: a, réponse impulsionnelle du filtre passe haut temporel des cellules amacrines b, symbole associé

Dans le domaine des Z, nous prenons un pas temporel unitaire $\Delta t=1$, la constante de temps du système τ_A est ainsi donnée en nombre d'images on aboutit à:

$$A(z) = b.\frac{1-z^{-1}}{1-b.z^{-1}} \text{ avec } b = e^{-\Delta t/\tau_A} \qquad \textit{(Eq. I.14)}$$

Par transformation en Z inverse, nous arrivons à l'équation temporelle discrète suivante :

$$y(t)=b.[y(t-1) + x(t)-x(t-1)] \qquad \textit{(Eq. I.15)}$$

On obtient donc un filtre passe-haut temporel dont la sortie dépend des entrées aux instants *t* et *t-1* et de la sortie à l'instant *t-1* (effet mémoire). On aboutit à un filtre un peu plus évolué qu'une simple différence temporelle *x(t)-x(t-1)* utilisée habituellement en vision par ordinateur.

I.3.3.2. Les cellules ganglionnaires

Les cellules ganglionnaires ne répondent que faiblement à une illumination diffuse du fait de l'organisation centre-périphérie de leur champ récepteur. Elles répondent d'autant plus fortement que les intensités lumineuses du centre et de la périphérie sont très différentes. Ainsi, elles rendent compte principalement des contrastes lumineux plutôt que de l'intensité absolue. On peut se rendre compte de ce comportement sur la figure I.27.

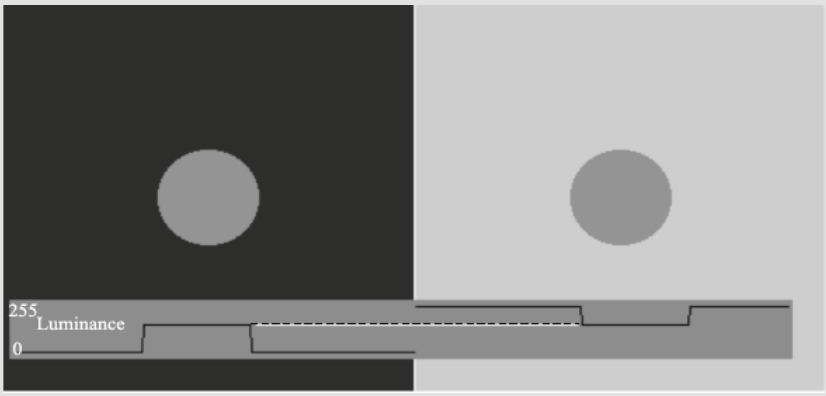


Figure I.27: la perception de la luminance du cercle est différente suivant celle du fond. Pourtant dans les 2 cas, les ronds ont la même luminance [HeraultTeach].

Ce comportement peut s'expliquer par le fait que l'information importante d'une scène visuelle est contenue dans l'arrangement des contrastes. La perception de la brillance des objets repose essentiellement sur l'information de contraste plutôt que sur la quantité absolue de lumière qui peut être influencée par le contraste entre un objet et son environnement.

Différents types de cellules ganglionnaires

D'après les travaux de Kaplan [Kaplan86], on retrouve trois types de cellules ganglionnaires :

→ Les cellules ganglionnaires de type P (Parvocellulaire): elles représentent 80% de la population des cellules ganglionnaires. Ce sont des cellules toniques (réagissant à un signal maintenu) qui ne sont connectées qu'à une seule cellule bipolaire. Ces cellules sont dédiées à l'analyse des contrastes spatiaux (et de la couleur). Elles se retrouvent surtout dans la zone de la fovéa, zone de vision à haute résolution, elles constituent la voie Parvocellulaire.

→ Les cellules ganglionnaires de type M (Magnocellulaire) : ces cellules dont la population est de l'ordre de 10% des cellules ganglionnaires, ne sont connectées qu'à des cellules bipolaires dont le champ récepteur est large (donc situées en zone parafovéale de la rétine). Elles sont couplées aux cellules amacrines et ont un comportement temporel transitoire. La grande majorité des ganglionnaires type M est polarisée de la même manière que les cellules bipolaires auxquelles elles sont connectées (centre ON-périphérie OFF ***ou*** inversement). Elles sont de sous types Magnocellulaire X ON ***ou*** X OFF. D'autres cellules ganglionnaires de type M sont de sous type Y et ont un comportement non linéaire: elles réagissent à des stimulations centre ON-périphérie OFF ***et*** inversement, comme une valeur absolue en quelque sorte. Ces cellules ganglionnaires de type M réagissent essentiellement aux changements temporels hautes fréquences. Elles permettent l'accès à l'information visuelle liée au mouvement. Elles constituent la voie magnocellulaire.

→ Les ganglionnaires type K: ces cellules ont une faible population. Elles forment la voie Koniocellulaire dont le rôle n'est pas encore bien défini [Szmajda05].

En résumé, au niveau de la couche PLI, on distingue trois voies particulières:

→ la voie Parvocellulaire ou Parvo donnant accès à la vision haute résolution de la fovéa très sensible

aux contrastes et à la couleur.

→ la voie Magnocellulaire ou Magno véhiculant l'information contextuelle de la scène visualisée. Du fait de l'interaction avec les cellules amacrines, cette voie est dédiée au mouvement, aux transitions, elle regroupe les informations majoritairement issues de la parafovéa.

→ la voie Koniocellulaire.

I.3.3.3. Modélisation de l'ensemble de la couche d'interactions PLI

La PLI est le siège des interactions entre les cellules bipolaires, amacrines et ganglionnaires. Plus précisément, on y trouve deux réseaux parallèles, l'un est lié aux cellules bipolaires de type ON, l'autre lié aux cellules bipolaires de type OFF. Les signaux issus des cellules bipolaires sont reçus par les cellules ganglionnaires de type M et de type P qui peuvent interagir avec les cellules amacrines. Notons que nous avons vu qu'il existe une grande diversité de cellules ganglionnaires et amacrines et que nous ne pourrons modéliser que les plus connues.

La voie parvocellulaire (réseau de cellules ganglionnaires de type P)

[Herault2001] propose une modélisation de la voie parvocellulaire. Cette modélisation est présentée sur la figure I.28.a. Les cellules ganglionnaires de type P agissent sur l'information issue des cellules bipolaires qui contiennent des informations locales sur les contours. Il a été montré [Smirnakis97] que les cellules ganglionnaires adaptent leur réponse selon l'information locale (les contours locaux) de la même manière que le font les photorécepteurs (adaptation locale de luminance, cf. Eq. I.2) d'où la présence sur le schéma proposé de deux modules de compression logarithmique qui agissent sur les informations des contours, nommés C_{gP}. La figure I.28.b montre l'effet de la voie Parvo sur une image de contours (sortie des cellules bipolaires). Il en résulte un renforcement des contours et donc du contraste. Les voies ON et OFF sont fusionnées en une sortie simple ON-OFF, que nous appellerons sortie parvocellulaire "PARVO ON-OFF".

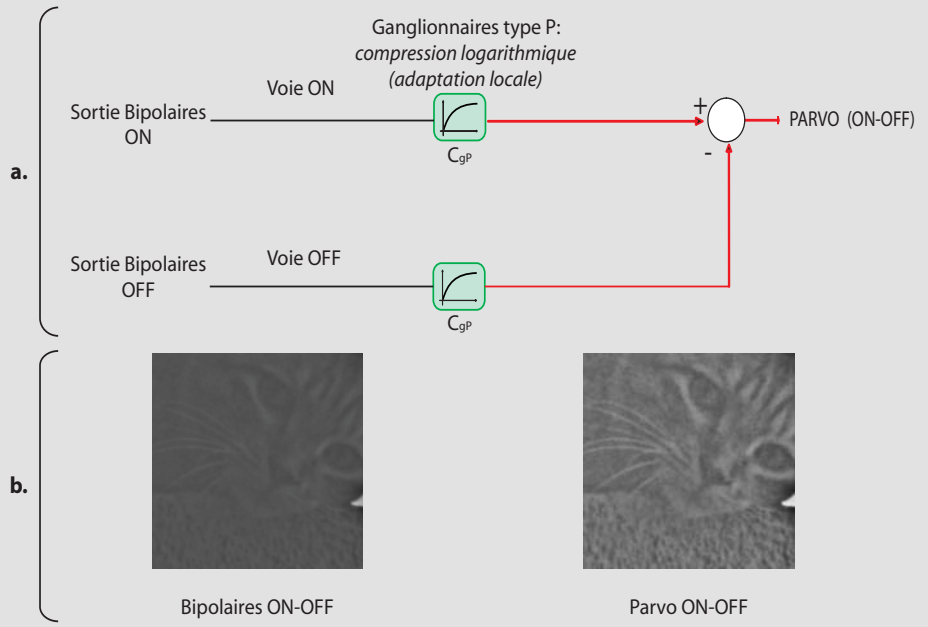


Figure I.28: a, modélisation du traitement de la voie parvocellulaire au niveau de la PLI. b, illustration de l'effet engendré: augmentation du contraste local

La voie magnocellulaire (réseau des cellules ganglionnaires de type M)

La voie magnocellulaire est une voie dédiée à l'analyse du mouvement. Elle est modélisée par le système présenté sur la figure I.29 [Herault01]. La sortie des cellules bipolaires de la PLE est principalement traitée par des cellules ganglionnaires de type M qui sont plus concentrées dans la zone parafovéale de la rétine. Elles ont un champ récepteur étendu et reçoivent donc les afférences d'un grand nombre de cellules bipolaires. L'information visuelle est ainsi lissée spatialement d'où les modules de filtrage passe-bas spatio-temporel (F_{gM}) dont on ajuste la constante de temps à zéro de façon à n'obtenir qu'un filtrage spatial. Ces cellules ganglionnaires sont par ailleurs connectées aux cellules amacrines d'où le traitement de l'information temporelle et la présence du filtre passe-haut temporel (A) sur le schéma. Enfin, les cellules ganglionnaires effectuent une adaptation locale aux contrastes locaux qui est modélisée par le module de compression logarithmique (C_{gM}).

Tout comme pour la voie parvocellulaire, les voies ON et OFF sont traitées séparément: elles conduisent aux sorties Magno X ON et Magno X OFF. Les cellules ganglionnaires de type Y ont quant à elles, un comportement non linéaire: elles sont sensibles à la fois aux signaux issus des voies ON et OFF. On modélise ce comportement comme la somme de ces deux canaux, ce qui conduit à la sortie Magno Y.

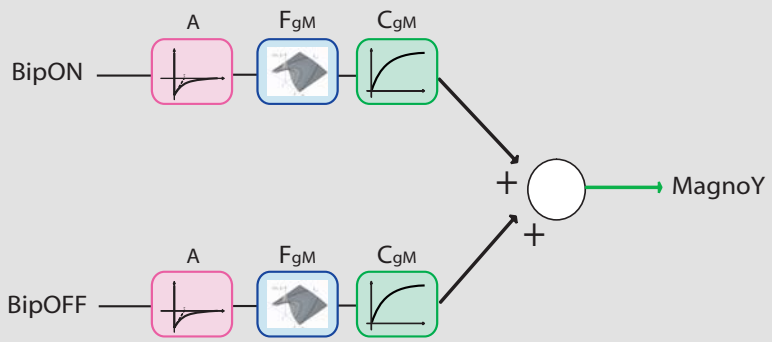


Figure I.29: schéma du traitement de la voie Magnocellulaire de la PLI

La figure I.30 présente chacune des trois sorties de la voie magnocellulaire vis-à-vis d'une scène vidéo de rue. Les voies MagnoX ON et OFF traitent respectivement l'information positive et négative issue de la PLE. Ainsi, chacune maximise le traitement de l'information de façon séparée: les contours en mouvement sur la voie ON sont extraits et renforcés tout comme sur la voie OFF. La voie MagnoY rassemble ces deux informations de façon à maximiser la sensibilité aux changements temporels. On obtient alors la réponse de tous les contours en mouvement.

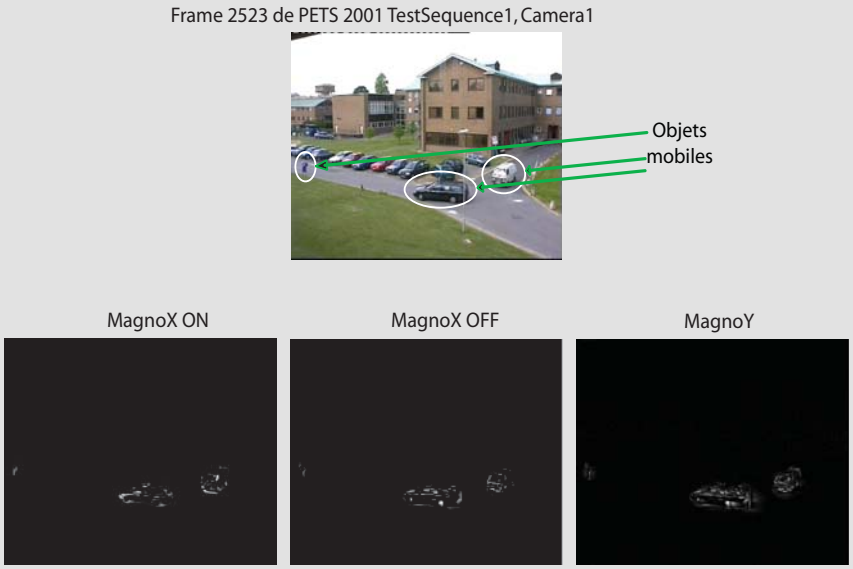


Figure I.30: signaux de sortie de la voie Magnocellulaire

I.3.4. Synthèse de la modélisation de la rétine

La figure I.31.a montre la modélisation complète de la rétine sur laquelle nous allons nous appuyer dans la suite de ce travail. Cette modélisation fait apparaître la compression logarithmique des photorécepteurs, le modèle des interactions de la couche PLE et le modèle des interactions de la couche PLI. En sortie du modèle de rétine, les voies disponibles sont les suivantes:

→ Parvo ”ON-OFF” porteuse des informations de détails et de contrastes locaux qui modélise la voie parvocellulaire de la rétine.

→ MagnoX ON et MagnoX OFF, porteuses de l’information transitoire sur leurs canaux respectifs ON et OFF.

→ MagnoY réagissant aux moindres changements spatio-temporel ou mouvements.

La figure I.31.b montre l’effet de l’ensemble des traitements se produisant sur une séquence lors de son passage par la rétine. Le flux visuel est constitué d’une figurine mobile et d’une figurine statique. Il apparaît clairement que la rétine n’est pas seulement en capteur d’énergie lumineuse. C’est aussi le lieu d’un ensemble de traitements de l’information visuelle.

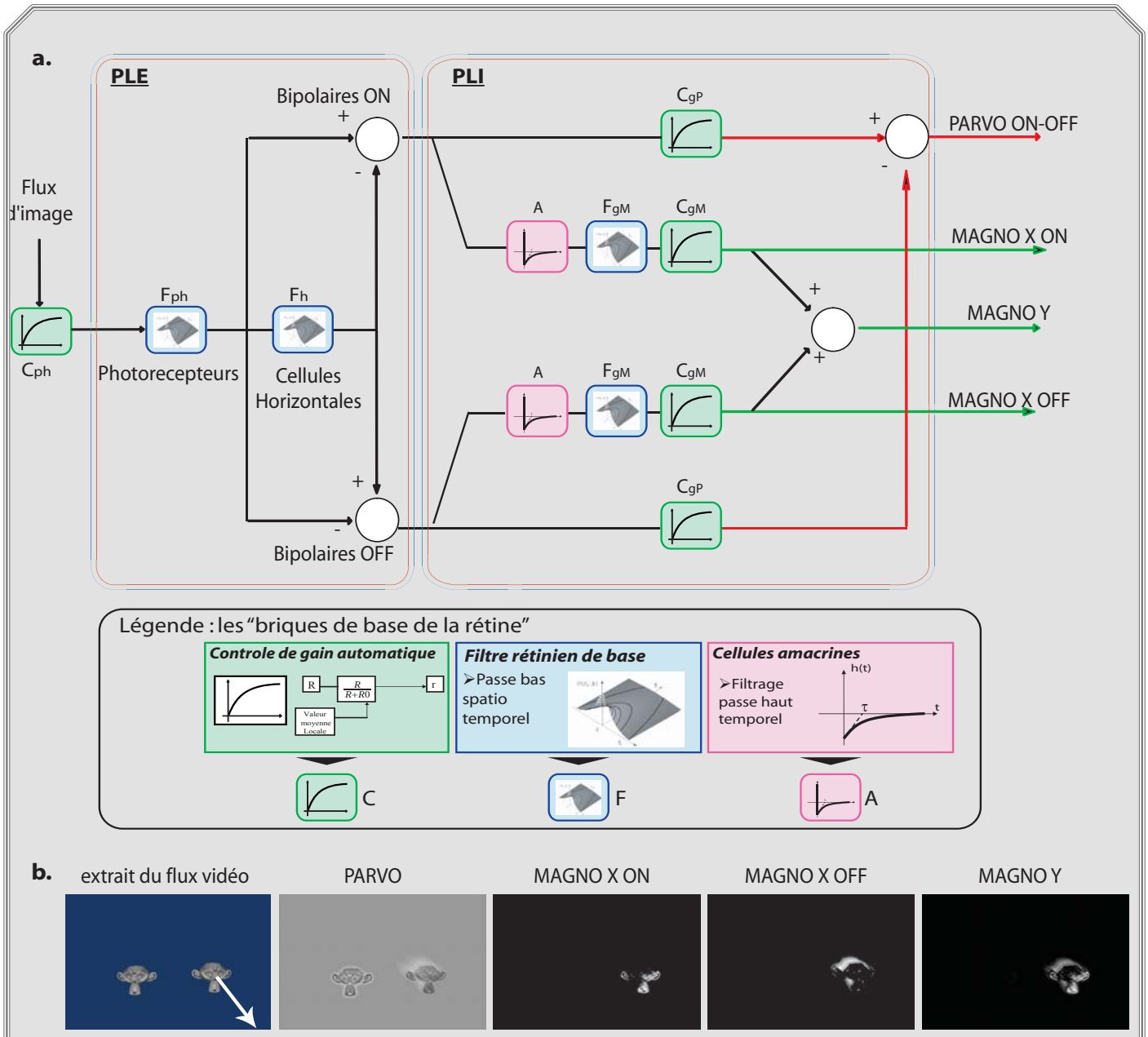


Figure I.31: a, modélisation globale du traitement de l'information visuelle par la rétine. b, illustration des différentes sorties de la rétine

I.4. En direction du cerveau

I.4.1. Transmission de l'information visuelle vers le cerveau

Les voies visuelles que nous venons de décrire partent en direction du cerveau via les nerfs optiques des yeux gauche et droit. Ces derniers contournent le mésencéphale, cheminent sur la face médiane du lobe temporal et se terminent, pour 80% d'entre eux, dans les corps genouillés latéraux (CGL). Les autres fibres nerveuses cheminent en direction du colliculus supérieur [Chauvin03], mais ne feront pas partie de notre étude. Avant d'atteindre les CGL, les nerfs optiques se croisent au niveau du chiasma optique, on y trouve une séparation des champs visuels gauche et droit : la rétine nasale gauche (partie de la rétine à gauche du nerf optique) et la rétine temporale droite (partie de la rétine à droite du nerf optique) sont aiguillées vers le CGL droit alors que la rétine temporale gauche et la rétine nasale droite sont envoyées vers le CGL gauche. Les CGL, situés dans la partie dorsale du thalamus, constituent donc la cible majeure de chaque tractus optique (terminaisons des nerfs optiques).

La distribution des neurones du CGL en différentes couches [Bullier98] indique que des aspects distincts de l'information visuelle (les voies parvocellulaires et magnocellulaires) en provenance de la rétine pourraient être traités séparément au niveau de ce relais synaptique. Cette disposition en couche sépare notamment les voies magnocellulaire et parvocellulaire. Cette description rapide est schématisée sur la figure I.32. Le rôle principal des CGL est une fonction de relais de l'information, mais d'autres fonctionnalités sont supposées comme un rôle de relais de l'information portant sur les commandes motrices par exemple [Sherman02].

Pour notre étude, nous ne considérerons les CGL que comme de simples relais.

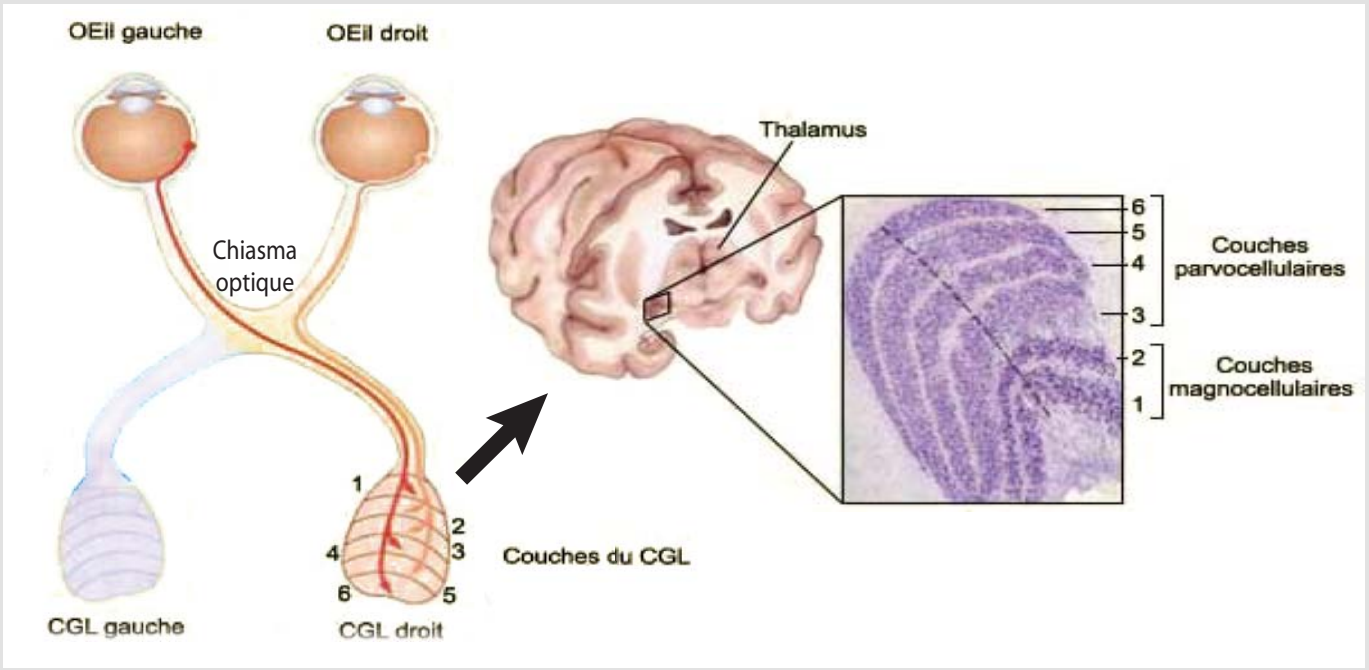


Figure I.32: cheminement de l'information de la rétine au CGL [McGillSite]

I.4.2. Le cortex visuel

Après être arrivée au CGL, l'information visuelle traitée par la rétine part en direction du cortex occipital dans l'aire V1. C'est dans cette région que commence l'analyse haut niveau de la scène visualisée. Néanmoins, l'aire V1 n'est que la première étape du traitement de l'information visuelle par le cerveau. On a découvert jusqu'à ce jour plus d'une trentaine d'aires corticales différentes contribuant à la perception visuelle. Les aires primaire (V1) et secondaire (V2) sont entourées de nombreuses autres aires visuelles tertiaires ou associatives : V3, V4, V5 (ou MT), LO, etc. La figure I.32 montre la répartition de ces aires.

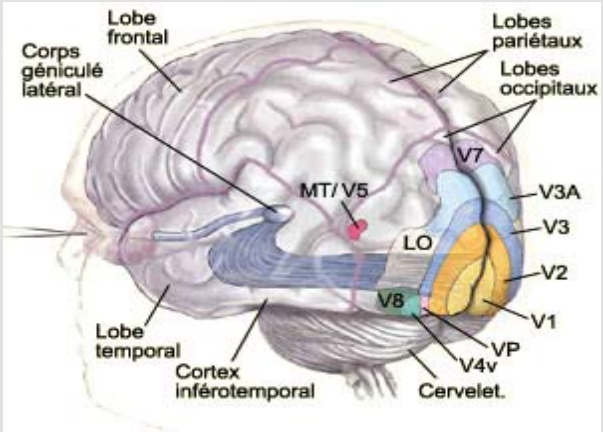


Figure I.33: les aires corticales du système visuel humain [McGillSite].

Après traitement de l'information par l'aire V1, l'information visuelle est ensuite diffusée vers les aires voisines telle l'aire V2 qui apparaît comme un lieu de séparation de l'information visuelle en deux voies distinctes: la voie ventrale et la voie dorsale. La voie dorsale rassemble les aires liées au cortex pariétal: MT, MST et FST et la voie ventrale rassemble les aires liées au cortex inférotemporal comme l'aire V4.

→ La voie ventrale: après avoir transité dans les aires V1, V2, une partie de l'information visuelle chemine ventralement vers le cortex temporal en passant tout d'abord par l'aire V4 (cf. fig. I.34). Cette information visuelle est plus de type "parvocellulaire": large bande riche en informations spatiales. Cette voie ventrale est appelée voie "quoi" car elle semble plus dédiée à l'identification des cibles de la scène visuelle.

→ La voie dorsale: cette voie semble dédiée à l'analyse du contexte. Elle reçoit en effet les informations de type magnocellulaire, porteuses de l'information de mouvement, de transitions temporelles, etc. L'information traitée par cette voie est donc liée à l'information "où" et "quand" [Bullier01]. Elle intervient aussi dans la coordination visio-motrice.

Nous ne présenterons pas plus d'informations sur les aires visuelles supérieures à V1 car il n'en existe pas à l'heure actuelle de modélisation suffisamment avancée pour espérer pouvoir construire des algorithmes de traitement de l'information visuelle qui s'inspirent de leur fonctionnement.

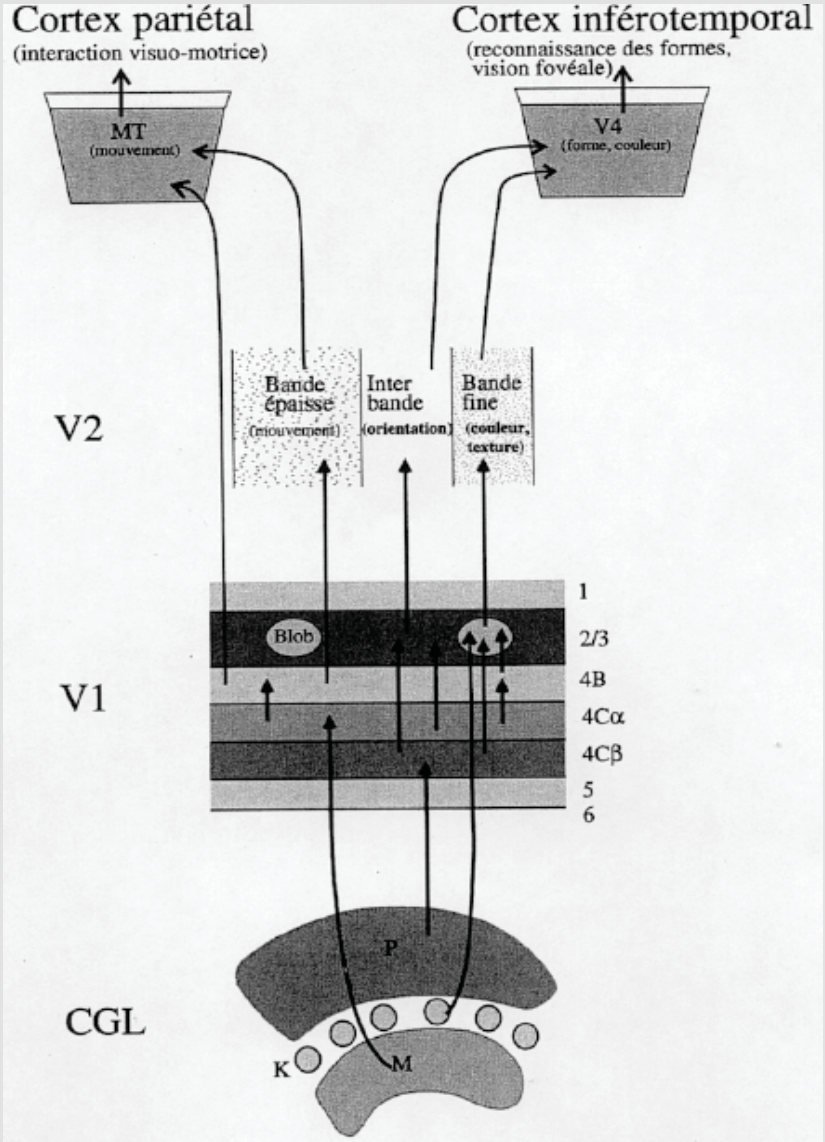


Figure I.34: communication entre les 2 voies de la rétine avec le CGL et les différentes couches du cortex visuel [Guyader04].

Dans ce manuscrit, seule l'aire V1 fera l'objet de notre étude. Cette aire étant la plus connue sa modélisation est plus avancée. L'objet de ce travail sera de montrer les capacités en terme de traitement d'image que l'on pourra mettre en évidence grâce à ce premier niveau d'analyse corticale.

I.5. Fonctionnement et modélisation de l'aire V1

I.5.1. Le cortex V1

Aussi appelée cortex strié, cette aire reçoit les afférences de la rétine via le CGL selon une répartition rétinotopique: les aires spatiales voisines sur la rétine le sont aussi dans leur projection sur V1.

On observe tout d'abord que la moitié de la surface corticale correspond à la zone fovéale de la rétine. Ceci est principalement dû à l'échantillonnage non régulier des photorécepteurs sur la rétine (leur répartition est particulièrement dense dans la fovéa) [Bullier01].

Le cortex V1 est organisé en 6 couches et chaque neurone échange préférentiellement de l'information verticalement i.e. avec les couches voisines (cf. fig.I.34). Chaque axone du CGL se termine dans la couche 4 et les neurones de cette couche se connectent préférentiellement vers les neurones des couches 3 et 5. La classification en couches de V1 est directement liée aux types d'afférents et aux 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. En plus de cette organisation en couches, les travaux de Hubel & Wiesel [Hubel74] montrent que les neurones sont organisés en colonnes verticales, possédant des propriétés proches et notamment une sensibilité à la même orientation vis-à-vis du stimulus d'entrée.

On trouve dans la littérature le terme d'hypercolonne qui représente un ensemble de colonnes juxtaposées regroupant toutes les orientations pour une même dominance oculaire. Les travaux de DeValois [DeValois88] et plus récemment les travaux de Blasdel [Blasdel92a ; Blasdel92b] ont permit de montrer que ce type d'organisation présente l'avantage de rassembler des informations variées sur une même surface corticale en deux dimensions. On y trouve: la représentation spatiale du champ visuel, la dominance oculaire, l'orientation, la fréquence spatiale, la couleur, le mouvement, et les afférences. Cette stratégie permet de maximiser la richesse de l'information et de l'organiser de manière plus efficace pour les traitements plus haut niveau ultérieurs.

Cette organisation montre plus précisément que l'aire V1 procède à une analyse par bandes de fréquences et bandes d'orientations de la scène visualisée. Cette sensibilité aux orientations a été mise en évidence par les travaux de Hubel & Wiesel [Hubel62] et par la suite confirmée par Blakemore & Campbell [Blakemore69], puis De Valois et al., [DeValois82] et Harvey et Doan [Harvey90]. Ces études ont également montré que les orientations verticales et horizontales sont traitées différemment des orientations obliques, le cortex privilégiant en quelque sorte les directions horizontale et verticale.

Enfin, la corrélation entre la sensibilité aux orientations et aux bandes de fréquence est mise en évidence par Webster et De Valois [DeValois85]. Ils ont montré que les cellules sensibles à une orientation sont également sensibles à une certaine bande de fréquences spatiales.

Ces études montrent qu'il existe au sein de l'aire V1 des cellules sensibles aux orientations et aux fréquences qui conduisent à la décomposition de la scène visuelle en bandes d'orientations et en bandes de fréquences. C'est cette propriété qui va être développée dans la suite de ce paragraphe, on peut en effet sup-

poser que cette analyse de l'information visuelle permet ensuite les tâches de catégorisation et d'identification réalisées dans les aires corticales en aval.

I.5.2. Modélisation

I.5.2.1 Différentes modélisations possibles

On peut modéliser la sensibilité aux bandes d'orientations et aux bandes de fréquences de certaines cellules de l'aire V1 à l'aide de filtres de type DOOG (différence de gaussiennes décalées) [Hawken87]. Une modélisation plus précise est l'utilisation de filtres de Gabor monodimensionnels (1D) [Marcelja80] et bidimensionnels (2D) [Daugman88]. On trouve également des modélisations faisant appel aux ondelettes [Mallat99]. Néanmoins, bien que la transformation en ondelettes soit devenue un outil classique en analyse d'images, son utilisation concerne plutôt les tâches de compression et de restauration ou encore d'interprétation-segmentation [Tao03].

Le but est ici d'utiliser une modélisation la plus proche possible du système biologique et qui soit capable de donner une mesure des caractéristiques spectrales de la scène visualisée. Les modélisations les plus précises parmi celles citées ci-dessus sont les ondelettes et les filtres de Gabor 2D. En effet, la modélisation par les filtres de Gabor a l'avantage de permettre de sonder les caractéristiques spectrales de la scène visualisée par orientations et bandes de fréquences. Ils sont couramment utilisés pour la segmentation de textures dans les images [Palagi92] et l'extraction de la forme selon les informations de texture [Massot06], ou encore dans les tâches de catégorisation [LeBorgne04]. [Guyader04] propose une modélisation plus précise de l'aire corticale V1 par l'utilisation de filtres de GloP (Gabor log polaires). Ces filtres ont l'avantage d'être symétriques sur un axe logarithmique contrairement aux filtres de Gabor ce qui se rapproche plus du système visuel et permet en outre une meilleure analyse des effets de zoom sur les images.

I.5.2.2. Analyse d'image par synthèse d'une rosace de filtres GloP

Le spectre d'amplitude donne une information sur la distribution statistique des fréquences et des orientations dans l'image. Ainsi, si l'on place une rosace de filtres GloP sur ce spectre d'amplitude (cf. fig I.35.a), il se retrouve échantillonné d'un point de vue log-polaire. On peut alors analyser la réponse de chaque filtre selon leur orientation et leur bande de fréquence, chaque ondelette GloP jouant le rôle de "sonde locale" dans l'espace fréquentiel. On travaille alors sur un spectre réduit dans le domaine log-polaire (cf. fig. I.35.b).

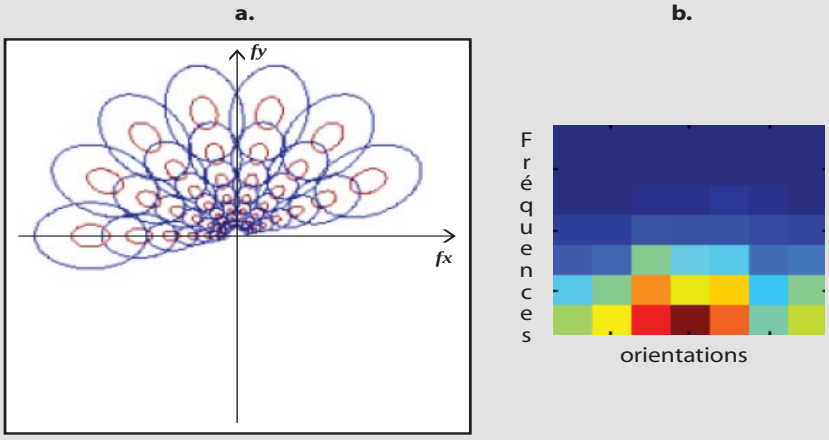


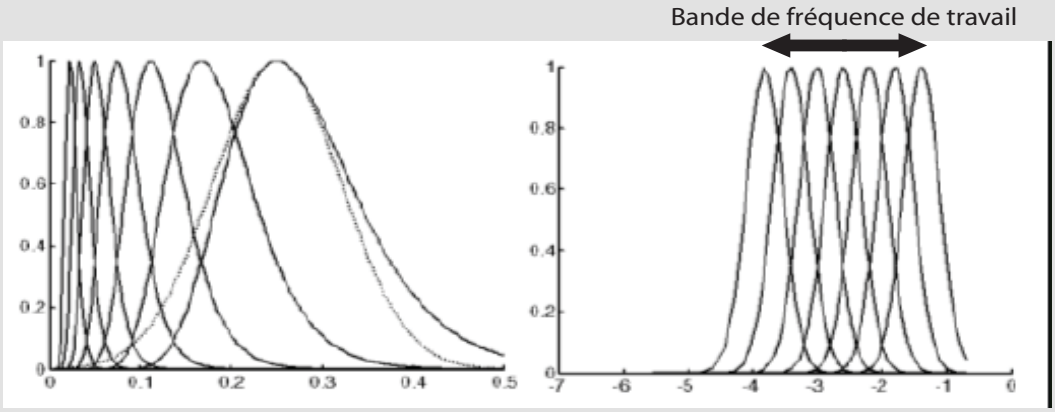
Figure I.35: a, échantillonnage du spectre d'amplitude à l'aide d'une rosace de filtres GloP de 7 orientations et 7 bandes de fréquences. b, spectre échantillonné résultant : 7*7 pixels portant chacun l'énergie selon une bande d'orientation et de fréquence spécifique.

Un filtre GloP est décrit par l'équation (I.16) dans laquelle le filtre centré sur la fréquence f_k, d'orientation θ_i et de facteur d'échelle σ apparaît comme un filtre à variables séparables :

$$G_{ik}(f,\theta) = \frac{1}{\sigma\sqrt{2\pi}}\left(\frac{f_k}{f}\right)^2 \exp\left(-\frac{\ln\left(\frac{f}{f_k}\right)^2}{2\sigma^2}\right).\cos\left(\frac{1+\cos(\theta-\theta_i)}{2}\right)^{50}$$

(Eq. I.16)

Il est recommandé de se reporter à [Guyader04] pour une description plus détaillée de ces filtres. Ces filtres suivent une loi log-normale et permettent de retrouver les propriétés de sonde dans l'espace des fréquences des filtres de Gabor. Ils ont en plus la propriété d'être symétriques en échelle log comme le montre la figure I.36 ce qui n'est pas le cas des filtres de Gabor classiques. Nous utiliserons ces filtres pour toutes nos analyses spectrales dans le domaine log polaire.



a. b.

Figure I.36: extrait de [Guyader04]: a, banc de filtres GloP 1D sur un axe des fréquences en échelle linéaire, leur enveloppe est asymétrique. b, en échelle logarithmique, ils deviennent symétriques, les basses fréquences (à gauche) ne sont plus surévaluées comme c'est le cas avec les filtres de Gabor.

Afin de donner une meilleure caractérisation de l'image ainsi qu'une modélisation fine du système visuel, la conception de la rosace de filtres s'appuie sur des études en neurosciences [Blakemore69; DeValois82] qui montrent que les cellules simples du cortex visuel ont une largeur de bande à mi-hauteur comprise entre 1,2 et 2,5 octaves. Nous choisissons une largeur de bande de 1.2 octaves ce qui permet d'être sélectif et précis en bande de fréquence. On ajuste alors les bandes radiales et transverses des filtres de la manière suivante:

→ Bandes radiales : 1.2 octaves. Pour une même orientation, chaque filtre de la rosace est ajusté sur l'axe des fréquences selon une loi de type $f_i = f_{max}/1.2^i$, en définissant f_{max} le filtre de fréquence maximum et i l'indice du filtre courant. Par exemple, si l'on décompose l'axe des fréquences en 5 bandes avec un filtre de fréquence normalisée maximum $f_{max} = 0.40$, les fréquences centrales normalisées de chaque filtre sont, dans l'ordre décroissant: 0.40, 0.33, 0.28, 0.23 et 0.19.

→ Bandes transverses: on ajuste la largeur de bande transversale de manière à couvrir les 180° d'analyse avec le nombre d'orientations N voulues tout en maintenant un recouvrement entre filtres à la moitié de l'amplitude maximale [Guyader04].

Cette méthodologie aboutit à la création d'une image du spectre échantillonné de façon log polaire grâce à la rosace de filtres de GloP. Ceci permet de «sonder» l'espace des fréquences et de recueillir une énergie totale englobée par chaque filtre. Il est alors possible d'obtenir une image du spectre échantillonné telle celle présentée sur la figure I.37.b, chacun des filtres donnant un "pixel" représentant son énergie. Sur la figure I.37.b, le spectre log polaire résultant montre une énergie plus importante pour les orientations horizontales (autour de 180°). Ceci s'explique par les contours fortement marqués des flans de la carrosserie de la voiture. Les contours verticaux (autour de 90°) qui correspondent principalement à l'avant et à l'arrière de la voiture sont moins énergétiques (car moins nombreux). Du point de vue des bandes de fréquences, l'énergie est plutôt basse fréquence. En effet, les détails ne sont pas très marqués du fait de la faible résolution de l'image (128*128).

L'avantage est que l'on dispose d'une représentation des caractéristiques structurelles de la scène de l'objet analysé. De plus, cette image est de dimensions très faibles comparées à la taille initiale de l'image.

Par exemple, une analyse en 15 orientations et 15 bandes de fréquences représente une “image du spectre log polaire” de 225 pixels.

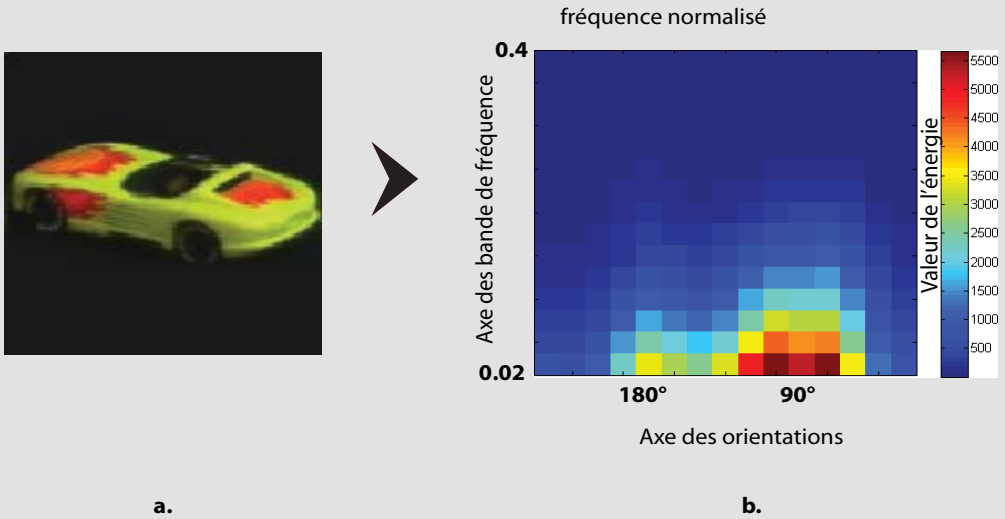


Figure I.37: a. image analysée. b. spectre échantillonné par une rosace de filtres GloP, chaque pixel de cette image a pour valeur l'énergie de la réponse du filtre correspondant de la rosace.

Pour nos études, nous travaillerons avec un spectre log polaire constitué de 15 orientations couvrant les 180° et de 15 bandes de fréquences ce qui représente une image spectrale de 225 pixels. Ce nombre de bandes d'orientations et bandes de fréquences est relativement proche du système visuel [Bullier01] et permet de décrire suffisamment précisément les caractéristiques des images. Le paramétrage est réalisé comme suit:

La bande de fréquence «intéressante» est considérée comme la portion de fréquences du spectre contenant les informations propres à la scène visuelle (les objets, les formes ... c.-à-d. le sujet d'étude) par opposition aux informations liés à la capture (bruit etc.). Dans notre étude, nous nous intéressons aux contours des objets dans la scène. Le bruit haute fréquence et la composante continue (luminance moyenne dans l'image) ne nous intéressent pas. Pour cela, nous fixons la fréquence centrale des filtres les plus haute fréquence à la valeur normalisée $f_{max}=0.35$. Les plus hautes fréquences généralement liées au bruit (entre 0.4 et 0.5) ne sont alors que faiblement ou pas sélectionnées. En fixant la bande radiale à 1.2 octave, la fréquence centrale minimum normalisée est de l'ordre de 0.023 ce qui permet d'obtenir des informations suffisamment basse fréquence sans pour autant sélectionner la composante continue. Ce paramétrage est redonné en annexe.

I.5.2.3. Mise en oeuvre de l'analyse et propriétés du modèle

L'analyse à l'aide des filtres GloP est réalisée dans le domaine fréquentiel. L'analyse commence par le passage du domaine spatial de l'image au domaine spectral. Ceci est réalisé à l'aide de la transformée de Fourier. Pour une image $i(p)$ avec p, vecteur des coordonnées bidimensionnelles de chaque pixel $(p=(x, y))$ sur le support S de l'image, sa transformée de Fourier $I(f)$ est donnée par l'équation I.17, avec f, vecteur des fréquences fx et fy caractéristiques de l'image. Ce spectre résultant comporte une partie réelle et une partie imaginaire. On en extrait alors le spectre d'amplitude $S(f)$.

$$I(f_x, f_y) = \int_S i(x,y).e^{-j2\pi(x.f_x, y.f_y)}.dx.dy$$

dont on déduit:

$$S(f_x, f_y) = \sqrt{\mathrm{Re}^2\left(I(f_x, f_y)\right) + \mathrm{Im}^2\left(I(f_x, f_y)\right)} \qquad \textit{(Eq. I.17)}$$

La transformée de Fourier considère que le signal à analyser est à support non borné c.-à-d. qu'il est de taille infinie. Or, nous travaillons avec des images de taille finie. Par conséquent, afin de limiter les effets entraînés par la taille finie de l'image, nous imposons un fenêtrage de Hanning sur l'image (cf. fig I.38) avant d'effectuer la transformée de Fourier.

Figure I.38: fenêtrage de Hanning de l'image entrante afin de limiter les effets de bord lors du calcul de la transformée de Fourier.

On montre [Guyader04] que le fait d'imposer ce fenêtrage ne perturbe que faiblement le spectre, notamment pour les basses fréquences spatiales. Néanmoins, si cette analyse spectrale est effectuée après le filtrage rétinien qui élimine ces basses fréquences, alors, l'expression de la densité spectrale d'énergie de notre signal est préservée ainsi que la qualité des mesures.

Après le fenêtrage de Hanning de l'image spatiale puis le calcul du spectre d'amplitude, on multiplie ce dernier par la réponse de chaque filtre GloP. Le spectre log polaire résultant présente alors des propriétés intéressantes pour certaines transformations de l'image.

<u>Sensibilité à la rotation</u>

La matrice de rotation *R* appliquée à une image *i(x)* entraîne une rotation identique sur son spectre *S(f)*, comme le décrit la relation suivante:

$$TF\left[i(Rx)\right] = S(Rf)$$

avec

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \qquad \textit{(Eq. I.18)}$$

Une rotation spatiale plane entraîne donc sur le spectre log polaire une translation des énergies caractéristiques de l'image le long de l'axe des orientations (cf. fig. I.39.a).

<u>Sensibilité au zoom</u>

D'une manière générale, le zoom spatial d'un facteur *a* d'un signal *i(x)* dont le spectre est *S(f)* donne un signal résultant *i(ax)* dont le spectre est:

$$TF\left[i(ax)\right]=\frac{1}{a^2}\cdot S\left(\frac{f}{a}\right) \qquad \textit{(Ex. I.19)}$$

En échelle logarithmique, cet effet de zoom apparaît sous la forme:

$$\frac{1}{a^2}\cdot S\left(\ln(f)-\ln(a)\right) \qquad \textit{(Eq. I.20)}$$

On en déduit qu'un zoom se traduit par une translation du spectre log polaire sur l'axe des fréquences (cf. fig. I.39.b). C'est là que l'utilisation de filtres GloP symétriques en échelle logarithmique est pertinente, car ils permettent une meilleure caractérisation des effets de zoom contrairement aux filtres de Gabor qui, du fait de leur asymétrie dans ce type d'échelle surévaluent les basses fréquences.

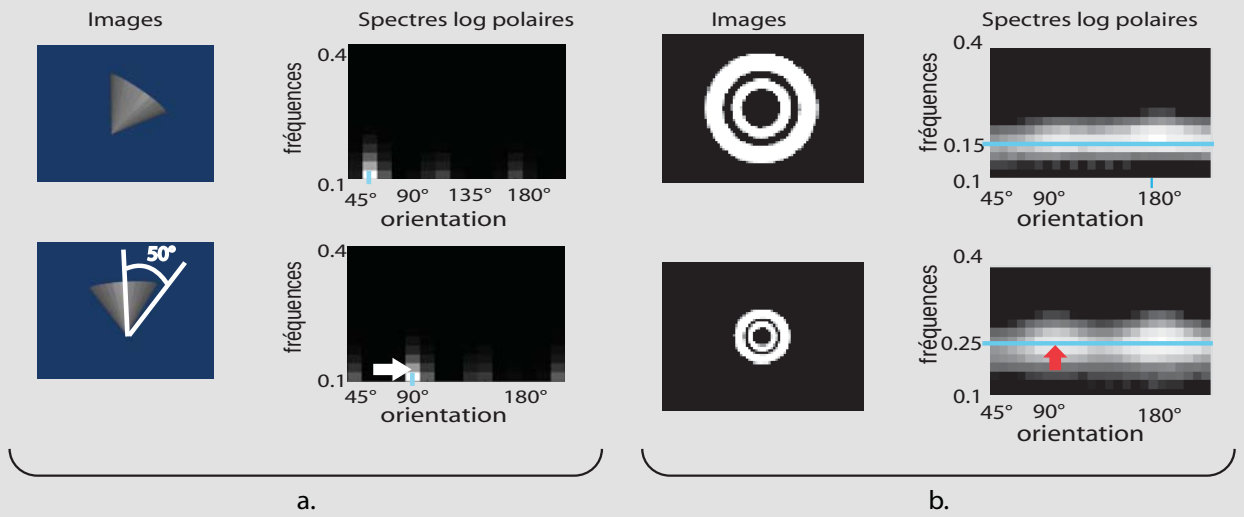


Figure I.39: comportement du spectre log polaire lors d'une rotation dans le plan de la caméra (a) ou lors d'un zoom (b)

Symbole associé à l'analyseur fréquentiel de l'aire V1

Nous associons à l'analyseur de spectre utilisant la batterie de filtres GloP le symbole présenté sur la figure I.40. Ce module prend en entrée une image en niveau de gris et donne en sortie ce que nous allons dorénavant appeler par abus de langage "spectre log polaire" de cette image. Il s'agit en fait du spectre échantillonné de façon log polaire par une rosace de filtres log polaires (GloP). Chaque pixel de cette image représente l'énergie du spectre pour une bande d'orientation et une bande de fréquences données. Cette image du spectre log polaire peut être vue comme caractéristique de la scène visualisée et représentant une image de l'énergie des différents contours orientés et de leurs fréquences spécifiques.

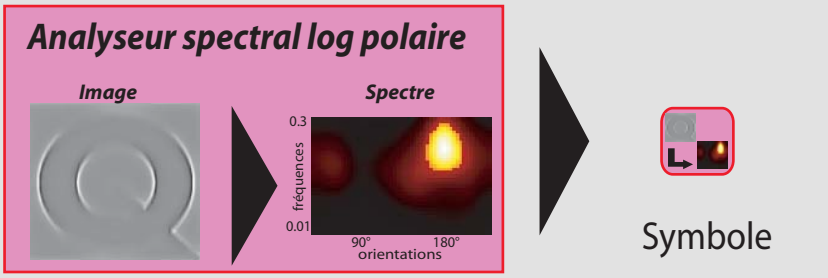


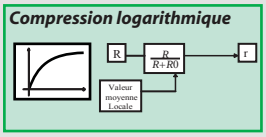
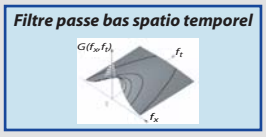
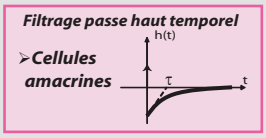
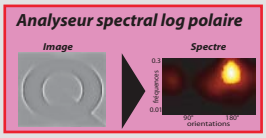
Figure I.40: symbole associé à l'analyseur de spectre "log polaire"

I.6. Conclusion

Dans ce chapitre, nous avons décrit les traitements se produisant sur l'information visuelle au niveau de la rétine et au niveau de l'aire V1 du cortex visuel. Ces deux éléments du système visuel humain sont les deux composantes dont on connaît une modélisation précise du fonctionnement.

Au niveau de la rétine, on trouve deux voies importantes : une voie dédiée à l'analyse spatiale : la voie parvocellulaire et l'autre voie dédiée à l'analyse du mouvement et des transitions temporelles : la voie magnocellulaire. Ces deux voies sont transmises à l'aire V1 qui réalise une analyse spectrale en bande d'orientations et de fréquences. Nous avons montré comment modéliser ces comportements à l'aide d'une combinaison de traitements bas niveau dont le tableau I.I fait la synthèse en ajoutant également les fréquences de traitement associés à chaque module. Ces fréquences évaluées à partir d'un ordinateur standard et une implantation en langage C/C++/Matlab non optimisée sont élevées et montrent que ces algorithmes peuvent être utilisés pour des calculs temps réel.

Table I.1: synthèse des modules associés au système visuel

Nom	***Fonction***	***Symbole***	***Cadence de traitement (pour des images de taille 320*240 pixels et un processeur Pentium IV, 3GHz)***
Compression logarithmique	*Adaptation locale*		*250 images par seconde*
Filtre passe-bas spatio-temporel	*Lissage de l'information dans le temps et l'espace*		*190 images par seconde*
Filtre passe-haut temporel	*Extraction des changements temporels*		*70 images par seconde*
Analyseur de spectre log polaire	*Calcul du spectre d'une image et décomposition en bandes d'orientations et de fréquence*		*50 images par seconde*

La figure I.41 présente l'enchaînement de ces traitements sous forme d'un schéma bloc. Bien que la modélisation du système visuel humain considérée ici soit incomplète (puisque bon nombre de cellules de la rétine et d'aires visuelles ne sont pas prises en compte), nous allons voir dans la suite de ce mémoire quels sont les traitements qu'il est possible de réaliser sur les images avec ce modèle "limité" afin d'obtenir des interprétations de plus haut niveau.

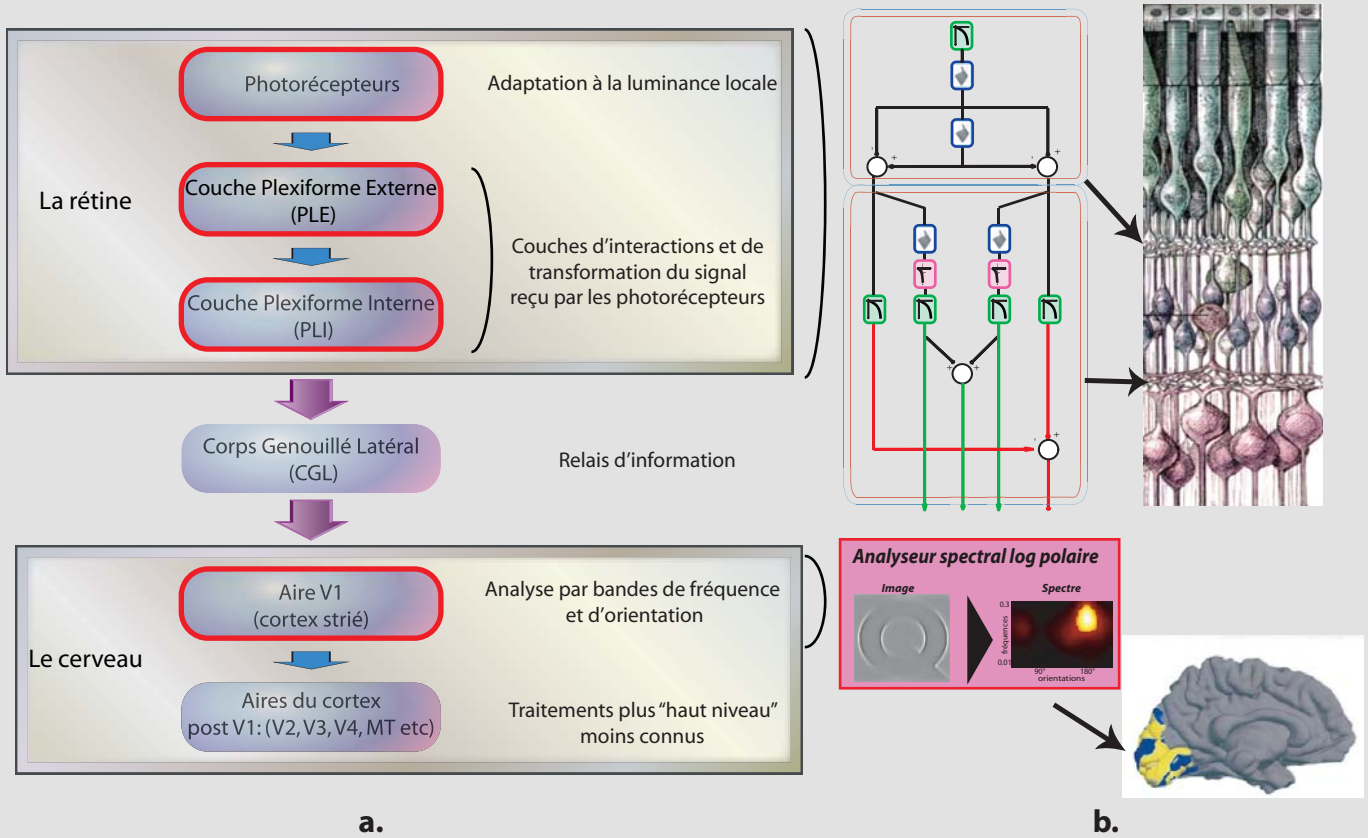


Figure I.41: résumé des principaux "éléments" du système visuel humain avec, (a), en rouge ceux qui ont été modélisés et (b) les symboles associés aux modélisations face au modèle réel.

Chapitre II: Système visuel humain et traitement d'images

II.1. Introduction

Dans ce chapitre, nous allons montrer comment utiliser les modèles de la rétine et du cortex V1 afin de réaliser des traitements bas niveau sur l'information visuelle. Nous proposons des méthodes pour renforcer les contours et les contours en mouvement, pour analyser les orientations présentes dans les scènes, pour détecter les événements de mouvement et pour segmenter les zones en mouvement et estimer leur vitesse. Chacun de ces algorithmes n'a pas été défini à l'origine dans le but d'être meilleur que les algorithmes de l'état de l'art sur le thème considéré, chaque algorithme est une conséquence des propriétés des modèles présentés au chapitre I. Par conséquent, nous ne présenterons pas un état de l'art exhaustif des algorithmes qui existent en vision par ordinateur pour chacun des domaines considérés (extraction de contours, analyse du mouvement, etc.). Notre objectif est de mettre en évidence les traitements que l'on peut faire sur les images en utilisant les modélisations actuelles (bien qu'incomplètes) de la rétine et de l'aire corticale V1.

II.2. Extraction de contours

En reprenant les modélisations des couches PLE et PLI du chapitre I, nous dégageons deux outils d'extractions de contours.

II.2.1. Extraction de tous les contours: filtre Parvo contours

II.2.1.1. Principe

A partir de la figure I.31, nous définissons un module d'extraction de contours que nous appellerons dans toute la suite «filtre Parvo contours» (cf. fig II.1). Ce module regroupe l'adaptation locale de luminance des photorécepteurs, la PLE et la voie parvocellulaire de la PLI. L'adaptation locale de luminance permet de renforcer les contrastes dans les zones sombres, la PLE extrait alors efficacement les contours et élimine la composante continue tout en blanchissant le spectre et enfin l'adaptation aux contrastes locaux qui prend lieu au niveau de la PLI renforce la réponse des contours extraits en amont. On applique en entrée de ce module une image ou une séquence vidéo en niveau de gris et on récupère en sortie le signal Parvo ON-OFF qui est l'image pour laquelle tous les contours ont été renforcés. On extrait également des sorties secondaires qui seront utiles pour d'autres modules: les sorties des cellules bipolaires ON et OFF et la sortie des photorécepteurs (adaptation locale de luminance et filtrage du bruit haute fréquence).

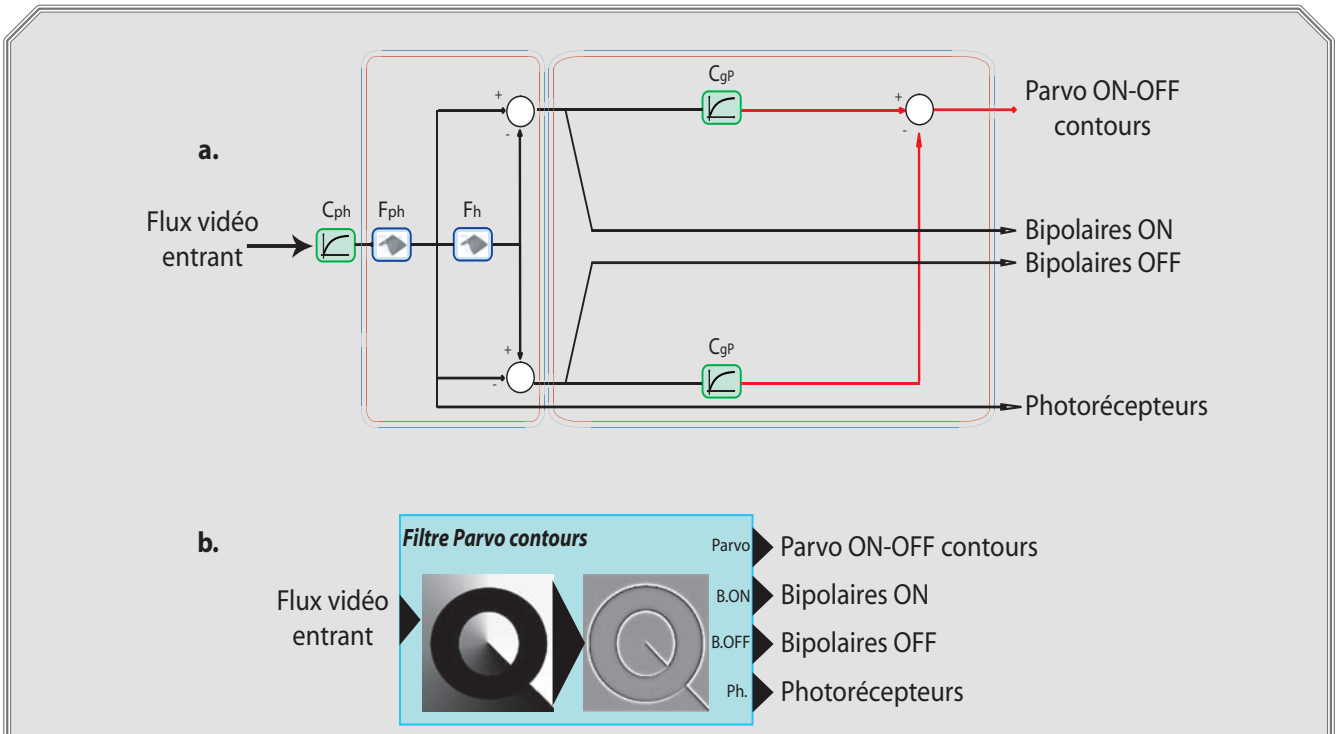


Figure II.1: architecture filtre Parvo contours (a) et symbole associé (b)

II.2.1.2. Propriétés

Comportement vis-à-vis des contours statiques

Comme le montre la fonction de transfert de la figure I.20, la PLE dont la modélisation est incluse dans le filtre Parvo contours a une tendance passe-bas temporelle pour les hautes fréquences spatiales et un gain élevé pour une bande de fréquences spatiales définie autour de la fréquence temporelle nulle. Il en résulte un renforcement des contours statiques par élimination du bruit spatio-temporel. Prenons l'exemple de la figure II.2 qui représente une image dans laquelle les motifs sont des flèches de luminance 0 disposées en cercle et dont le fond est un gradient de luminance variant linéairement de 0 à 255. Sur cette image, on ajoute également un bruit spatio-temporel de variance σ =0,01. Le filtre Parvo contours réalise de manière simultanée l'adaptation locale de luminance des photorécepteurs, le filtrage passe-bande pour l'extraction des détails statiques de l'image et le filtrage passe-bas temporel pour l'élimination du bruit. La figure II.2 montre à titre de comparaison pour cette même image la sortie obtenue suite au calcul du gradient de l'image par un filtre de Sobel. Les contours obtenus sont beaucoup moins nets et sont noyés dans le bruit.

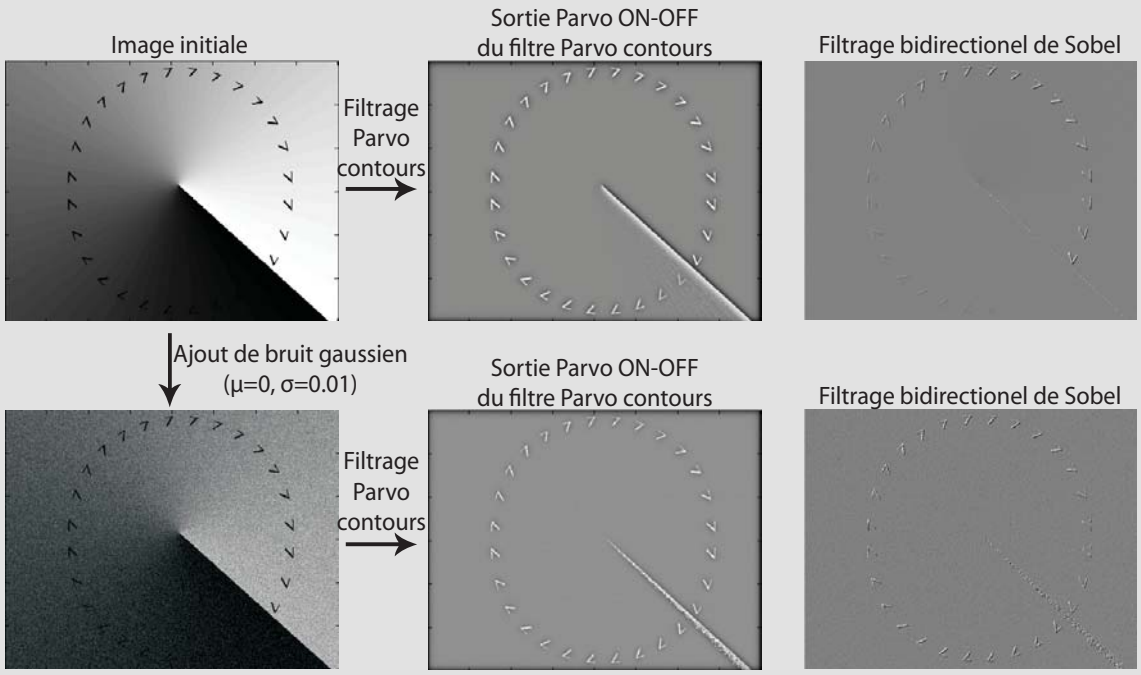


Figure II.2: effet du filtrage Parvo contours (sortie Parvo ON-OFF) pour l'extraction de détails et comparaison avec un filtrage de Sobel bidirectionnel

Une analyse de rapport signal sur bruit permet de quantifier plus précisément les performances du filtre. La table II.1 présente une comparaison de rapports signal sur bruit (SNR) et d'erreurs quadratiques moyennes (MSE). On compare d'une part l'image originale avec sa version bruitée. D'autre part, on compare les sorties du filtrage Parvo contours appliquées à l'image originale et à l'image bruitée. On effectue les mêmes analyses avec le filtrage de Sobel. On présente les SNR et MSE entre:

→ l'image originale et sa version bruitée.

→ la sortie Parvo ON-OFF sur l'image originale et la sortie sur l'image bruitée.

→ la sortie du filtre de Sobel sur l'image originale et la sortie sur l'image bruitée.

Table II.1: comparaison de la réponse au bruit avec ou sans filtrage Parvo contours ou Sobel

Comparaison entre images	*MSE*	*SNR (dB)*
originale v.s. originale bruitée	*2e+004*	*1.3*
sorties filtre Sobel	*1.5e+3*	*1.6*
sorties Parvo contours ON-OFF	*344*	*4.4*

Nous constatons que le rapport signal sur bruit est le plus élevé à la sortie du filtre Parvo contours (sor-

tie ON-OFF). De même, l'erreur quadratique moyenne est nettement abaissée en sortie de ce filtre, la sortie Parvo ON-OFF donne donc une réponse similaire (stable) pour une même image avec ou sans bruit. Notons qu'il ne serait pas cohérent d'effectuer des calculs de MSE et SNR entre les images entrée/sortie de chaque filtre, car l'entrée est une image de luminance et la sortie est une image de contours ce qui constitue 2 espaces différents. En conclusion, le filtre Parvo contours permet un renforcement des contours même en cas de bruit ou de fortes variations de luminance.

Comportement vis-à-vis du mouvement

Comme le montre la réponse impulsionnelle présentée sur la figure I.24, le modèle associé à la PLE montre une tendance passe-bas spatiale pour une information variant temporellement. Autrement dit, on a tendance à voir un objet en mouvement de manière plus floue qu'un objet immobile dont on distingue plus nettement les détails. Nous observons ce phénomène sur la figure II.3: le mouvement constitue un événement spatio-temporel et donne une réponse filtrée passe-bas en sortie PARVO ON-OFF. Ceci correspond à la phase transitoire de la réponse impulsionnelle.

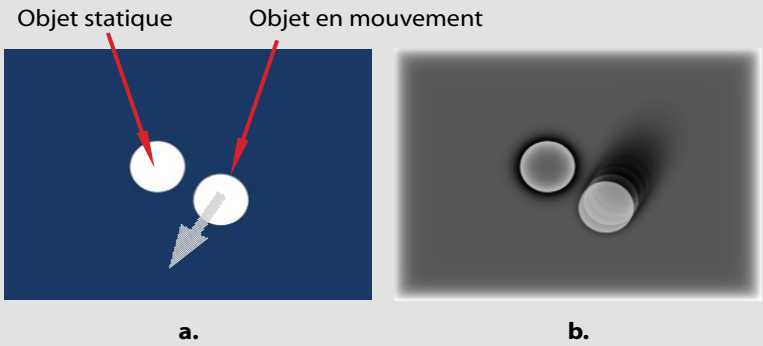


Figure II.3: a. extrait d'une séquence vidéo avec 2 objets l'un mobile, l'autre immobile. b, sortie Parvo ON-OFF du filtre Parvo contours, les contours sont bien définis pour l'objet immobile mais plus flous pour l'objet mobile.

En conclusion, ce module permet une extraction de contours efficace et localement adaptée. Les contours statiques sont renforcés et les basses fréquences des contours mobiles sont extraites. Ce module permet de limiter les effets du bruit spatio-temporel et du manque de contraste dans les images. Les images résultantes présentent des contours renforcés propres aux caractéristiques des objets dans la scène.

II.2.1.3. Paramétrage et coût de calcul

Les étages d'adaptation locale de gain C_{ph} et C_{gP} ont un paramètre de compression V_0 fixé à *230* ce qui permet une compression importante et un renforcement efficace de la sensibilité dans les zones sombres et faiblement contrastées. La constante d'espace α_{ph} du filtre passe-bas spatio-temporel F_{ph} associé aux photo-récepteurs est fixée à $\alpha_{ph}=1$ pixel de façon à limiter le bruit spatial haute fréquence. Celle du filtre passe-bas spatio-temporel F_h associé aux cellules horizontales est fixée à $\alpha_h=5$ pixels de façon à extraire la luminance locale dans un voisinage proche. Les constantes de temps τ_{ph} et τ_p de ces deux filtres sont fixées à 1/25 seconde. Ceci permet de limiter le bruit temporel haute fréquence et conserver une réactivité (temps de réponse faible/régime transitoire court) correcte pour les flux vidéo à 25 images par seconde. Dans la suite, les valeurs

des paramètres proposés ici seront conservées. Il est néanmoins possible de modifier ces valeurs selon les principes suivants:

→ Le paramètre V_0 ajuste le taux de compression des étages C_{ph} et C_{gP}. Sa valeur est fixée par défaut à 230, une valeur moindre donnerait une compression moindre. Cette valeur donne un comportement général satisfaisant. A titre informatif, V_0 donne des résultats intéressants du point de vue visuel dans un intervalle de valeurs entre 160 et 250 [Durette05].

→ Le paramètre α_{ph} permet le filtrage préliminaire des hautes fréquences souvent liées au bruit. Une valeur plus faible de α_{ph} augmente la fréquence de coupure spatiale ce qui peut limiter l'efficacité du filtrage du bruit. Une valeur plus forte de α_{ph} renforce le filtrage du bruit, mais risque d'atténuer la réponse des contours intéressants dans l'image.

→ Le filtrage temporel passe-bas des photorécepteurs est paramétré avec une constante de temps de $\tau_{ph}=1/25$ seconde ce qui permet un filtrage efficace des hautes fréquences temporelles tout en laissant le système suffisamment réactif. Une valeur plus faible augmente la réactivité, mais favorise le passage du bruit haute fréquence. Une valeur plus forte force un moyennage temporel de l'information, ceci minimise le bruit, mais atténue la réponse des contours en mouvement rapide.

→ Le paramètre α_h permet de fixer l'étendue sur laquelle le calcul de la luminance locale moyenne est réalisé pour un pixel donné. Une constante d'espace forte permet d'estimer la luminance locale sur un voisinage large, une valeur faible rend l'estimation très localisée.

→ La constante de temps τ_{ph} est fixée à 1/25 seconde ce qui laisse le système temporellement assez réactif. Une valeur plus forte donnera une estimation de luminance locale moyennée temporellement ce qui entraînera un effet retard entre l'image courante et le calcul de la luminance moyenne. Une valeur plus faible permet de se rapprocher d'une estimation image par image.

Le filtre Parvo contours a un coût de calcul faible. En effet, il est construit à partir de deux filtres élémentaires passe-bas spatio-temporels et de trois modules de compression logarithmique. Il en résulte un coût de l'ordre de 20 produits par pixel quelles que soient les fréquences de coupure caractéristiques des filtres. Il est possible d'atteindre une vitesse de traitement de 60 images par seconde pour des images de taille 320*240 sur un ordinateur équipé d'un processeur de type Intel Pentium 4, 3.0Ghz avec un code C++ non optimisé. Rappelons à titre de comparaison, que pour une extraction des gradients spatiaux par filtre convolutif tel le filtre de Sobel, le coût serait bien plus élevé et variable selon les fréquences de coupure. Par exemple, pour un filtre convolutif approchant uniquement le comportement passe-bande du filtre Parvo contours à fréquence temporelle nulle, sa taille minimum serait de 5*5 pixels ce qui présenterait un coût de l'ordre de 25 produits par pixel.

II.2.2. Extraction de contours en mouvement: filtre MagnoY contours mobiles

II.2.2.1 Principe

Si on fait suivre le filtre Parvo contours des éléments de la voie magnocellulaire selon le schéma de la figure II.4.a, on obtient un système d'extraction des contours en mouvement. En effet, les contours statiques sont totalement éliminés grâce à l'effet de dérivation temporelle créé par le filtre passe-haut temporel modélisant les cellules amacrines. On crée alors le module «MagnoY contours mobiles» (cf. fig II.4.b) qui prend en entrée les sorties des cellules bipolaires ON et OFF du module Parvo contours et donne en sortie la voie MagnoY contours mobiles.

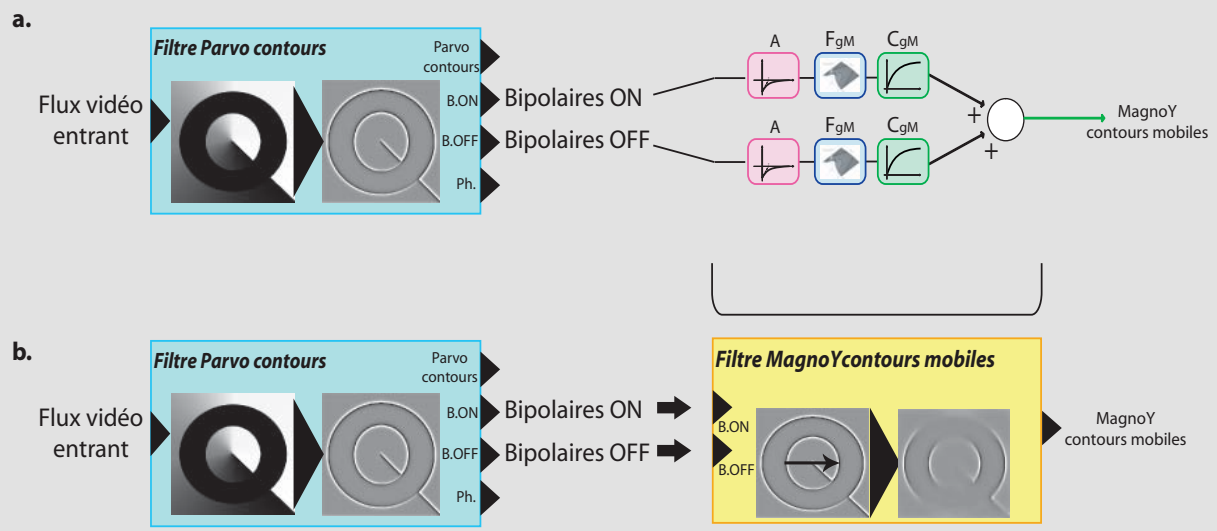


Figure II.4: filtre MagnoY contours mobiles. a, filtre Parvo contours suivi des modules élémentaires de la voie magnocellulaire Y de la couche PLI de la rétine. b, filtres et symboles associés

II.2.2.2. Résultats

La figure II.5 montre le résultat du filtrage par le filtre MagnoY contours mobiles d'une séquence dans laquelle un objet est statique et un autre est en mouvement. On ajoute sur cette séquence un bruit spatio-temporel haute fréquence de moyenne nulle et de variance 0.01. Le filtre MagnoY contours mobiles extrait l'objet en mouvement seulement et lorsque la vidéo est bruitée, peu de bruit est transmis. A titre de comparaison, la figure II.5 montre le résultat obtenu par le calcul du gradient temporel par différence d'image. On observe que la différentielle temporelle est plus bruitée. Au contraire, le filtre MagnoY contours mobiles profite des filtrages spatio-temporels passe-bas F_h.

D'autre part, les filtrages temporels des filtres Parvo contours et MagnoY contours mobiles introduisent un effet de traînée plus prononcé qui permet de lisser l'énergie sur toute la surface des objets lors de leur mouvement ce qui est intéressant en particulier pour les objets à luminance uniforme. La non-séparabilité des filtrages spatio-temporels permet à la fois la réduction du bruit spatio-temporel et l'extraction des contours en mouvement.

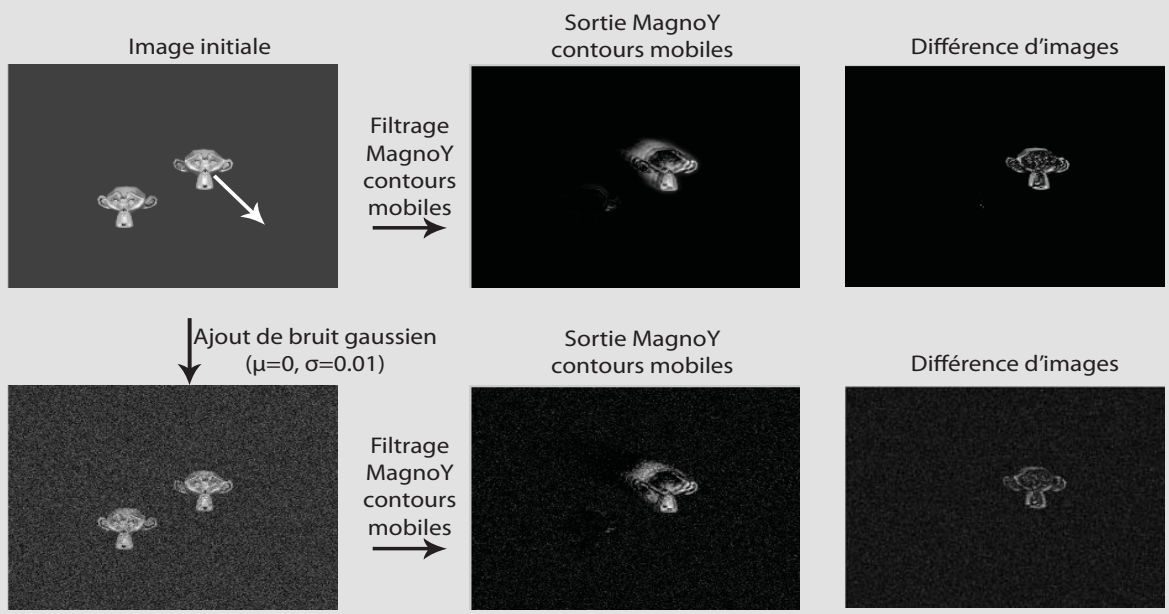


Figure II.5: effet du filtrage MagnoY contours mobiles pour l'extraction des contours en mouvement et comparaison avec une différence temporelle d'images (en valeur absolue)

Afin de quantifier les performances du filtre MagnoY contours mobiles, nous avons effectué une analyse de rapports signal sur bruit (SNR) et d'erreurs quadratiques moyennes (MSE) (cf. table II.2). On présente les SNR et MSE entre:

→ l'image originale et sa version bruitée.

→ la sortie MagnoY contours mobiles sur l'image originale et sur l'image bruitée.

→ la différentielle temporelle sur l'image originale et sur l'image bruitée.

Table II.2: comparaison de la réponse du filtrage MagnoY contours mobiles et de la différentielle temporelle

Comparaison entre images	*MSE*	*SNR (dB)*
originale v.s. originale bruitée	*264.8*	*10.9*
sorties différentielle temporelle	*521.8*	*0.4*
sorties MagnoY contours mobiles	*5.7*	*3.2*

Le rapport signal sur bruit (SNR) est plus élevé à la sortie du filtre MagnoY contours mobiles qu'à la sortie de la différence d'images. De même, l'erreur quadratique moyenne est nettement abaissée en sortie du MagnoY contours mobiles. Ce filtre donne comme le filtre Parvo contours une réponse similaire (stable) pour une même image avec ou sans bruit. En conclusion, le filtre MagnoY contours mobiles permet une extraction optimisée des contours en mouvement même en cas de bruit. De plus, il profite de la correction de luminance

apportée par le filtre Parvo contours en amont ce qui autorise une extraction des contours mobiles même dans les zones sombres.

II.2.2.3. Paramétrage et coût de calcul

La constante de temps τ_A du filtre passe-haut temporel sur chacune des voies ON et OFF est fixée à 1/12 seconde ce qui permet d'extraire les changements temporels supérieurs à 12Hz. Nous fixons la constante d'espace α_{gM} des filtres passe-bas spatio-temporels à 5 pixels et leur constante de temps τ_{gM} à 0 de façon à estimer une moyenne locale de mouvement dans un voisinage faible sans introduire d'effet temporel supplémentaire. Dans la suite, les valeurs des paramètres proposés ici seront conservées. Il est néanmoins possible de modifier ces valeurs selon les principes suivants:

→ La constante de temps τ_A fixe la fréquence minimum autorisée par le filtre passe haut temporel. Une valeur de constante de temps plus haute permet d'extraire des changements plus lents (de fréquence plus faible) (c.-à-d. on analyse le mouvement de façon plus globale temporellement). Au contraire, une constante plus faible permet d'extraire les changements très rapides seulement, plus particulièrement, le bruit résiduel haute fréquence.

→ La constante d'espace α_{gM} permet de fixer l'étendue spatiale de l'estimation de la moyenne locale de mouvement. Une valeur plus forte augmente l'étendue du calcul de moyenne sur la surface de l'image (c.-à-d. on analyse le mouvement de façon plus globale). Une valeur faible permet de calculer la moyenne très localement.

→ Le paramètre V_0 de l'étage de compression logarithmique des étages C_{gM} est fixé à 230. Une valeur moindre donnerait une compression moindre. Cette valeur donne un comportement général satisfaisant. A titre informatif, V_0 donne des résultats intéressants du point de vue visuel dans un intervalle de valeurs entre 160 et 250 [Durette05].

Si l'on considère le filtre MagnoY contours mobiles seul, son coût de calcul est de l'ordre de 16 produits par pixels. Il nécessite néanmoins l'utilisation en amont du filtre Parvo contours. L'ensemble fonctionne à 40 images par seconde pour des images de taille 320*240 sur un ordinateur équipé d'un processeur de type Intel Pentium 4, 3.0Ghz avec un code C++ non optimisé.

II.3. Analyse fréquentielle

Le fonctionnement de l'aire V1 est modélisé par un analyseur spectral (cf. I.5.2) qui agit comme une sonde dans l'espace des fréquences et des orientations. L'image spectrale que l'on extrait rapporte de l'énergie sur les orientations et bandes de fréquences correspondant aux caractéristiques de l'image analysée.

II.3.1. Analyse des orientations dominantes de l'image à partir du spectre log polaire

II.3.1.1. Principe

Nous proposons un algorithme qui s'insère immédiatement après l'analyseur spectral et qui a pour but d'interpréter la répartition de l'énergie selon les orientations (cf. fig. II.6.a). Il consiste à sommer toutes les énergies de l'image spectrale par orientation. Ceci conduit pour chaque image à l'obtention d'une "courbe d'énergie cumulée par orientation". L'analyse de cette courbe donne une idée de la répartition des orientations contenues dans l'image et permet d'en extraire l'orientation dominante par calcul de l'abscisse du maximum

d'amplitude. Les figures II.6.b et II.6.c illustrent le résultat de cette analyse pour deux images particulières. Soulignons que sur l'image spectrale, les "pixels" les plus clairs sont associés aux énergies les plus élevées. Dans le cas de la figure II.6.b qui représente l'analyse spectrale de la sortie du filtre MagnoY sur un oeil dont la paupière se referme, on détecte un maximum autour de l'orientation 90° sur la courbe d'énergie cumulée par orientation. Cela traduit la présence dans l'image analysée de l'orientation dominante 90° (à savoir le contour horizontal de la paupière en mouvement). A l'opposé, la figure II.6.c représente l'analyse de la sortie du filtre MagnoY sur un oeil dont la pupille se translate sur la gauche. L'analyse de la courbe d'énergie cumulée par orientation correspondante traduit la présence dominante de contours verticaux (orientation 180°) liés à la pupille en mouvement.

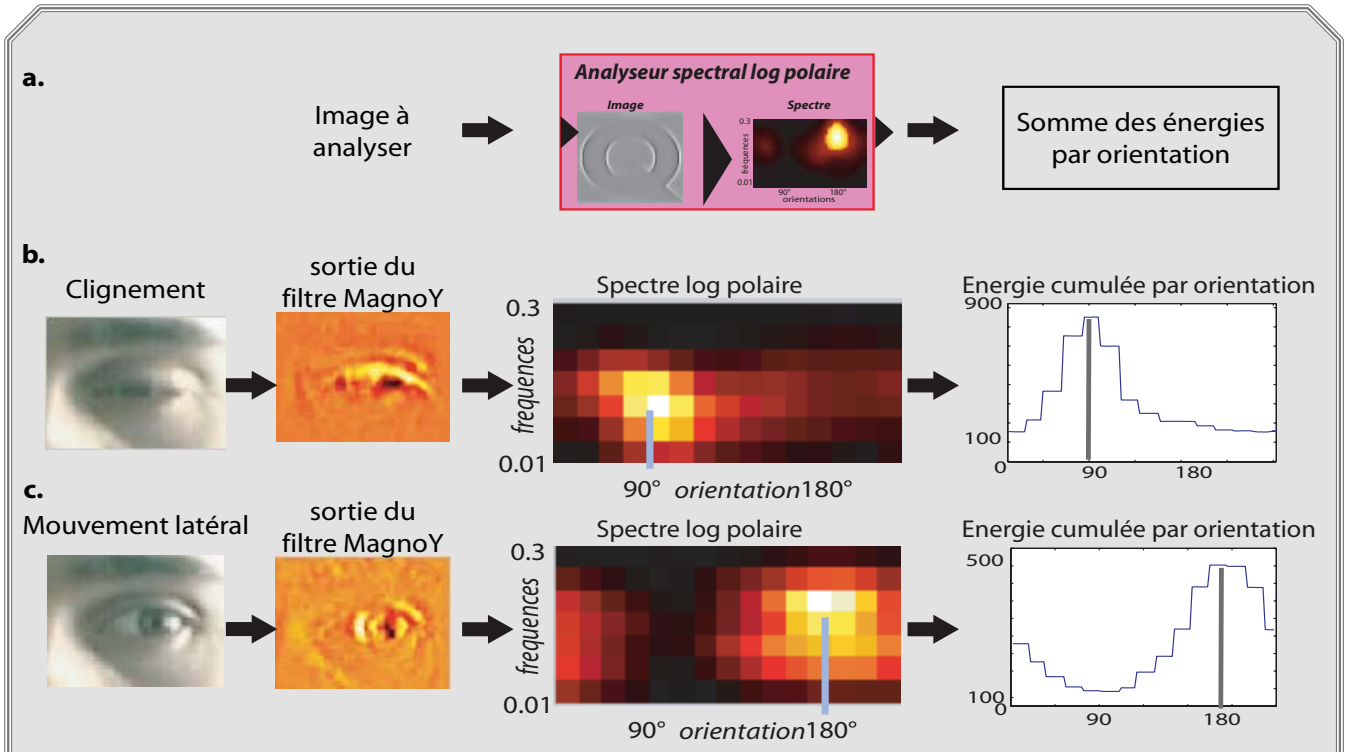


Figure II.6: exemple de calcul de courbe d'énergie cumulée par orientation à partir de différentes images spectrales.

Par ailleurs, cette courbe d'énergie est a priori indépendante des effets de zoom. En effet, son allure ne change pas si l'énergie de l'image spectrale est translatée selon l'axe des fréquences (conséquence d'un effet zoom). La figure II.7 présente une illustration de cette propriété. Si on compare les spectres log polaires de l'image de plage et sa version réduite d'un facteur 3, on constate qu'il y a eu une translation des caractéristiques du spectre selon l'axe des fréquences. En revanche, l'allure générale de la courbe d'énergie cumulée par orientation reste la même même si son amplitude a diminué du fait de la perte de précision au niveau des contours sur l'image réduite. Néanmoins, cette invariance est limitée à des zooms faibles, un effet de zoom trop important apportant un changement de l'information relative aux contours (des détails disparaissant ou apparaissant en grande proportion).

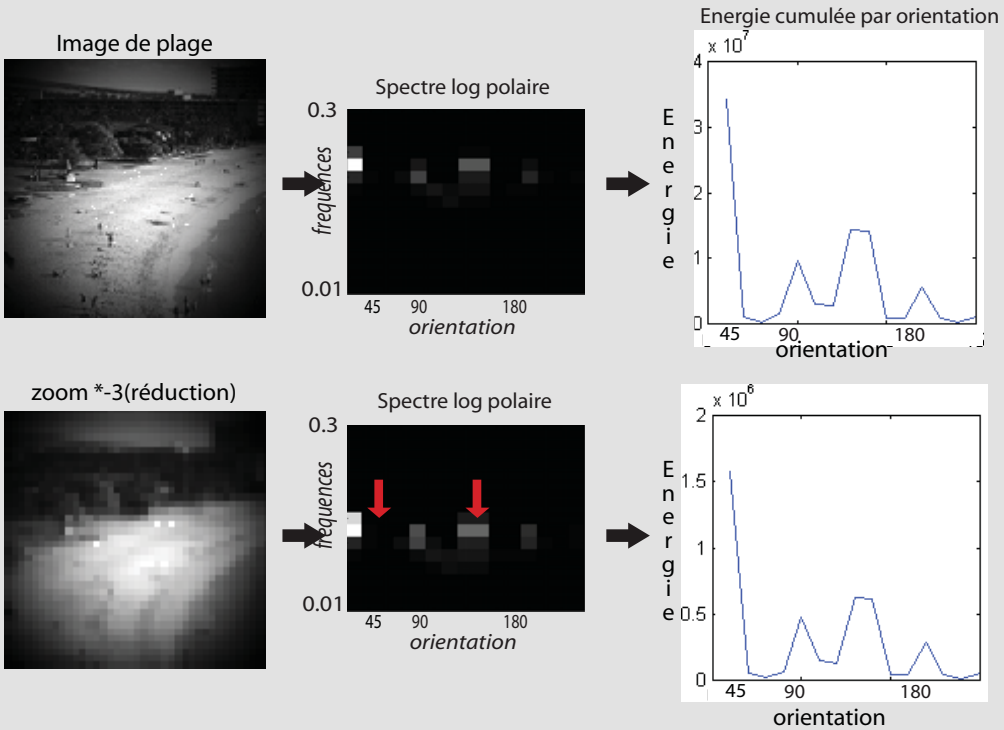


Figure II.7: courbe d'énergie par orientation et effet zoom: l'allure de la courbe ne change pas, seule son amplitude varie du fait de la baisse de précision des contours.

II.3.1.2. Paramétrage et coût de calcul

Aucun paramétrage spécifique n'est nécessaire pour cet algorithme d'analyse des orientations. La résolution de la courbe d'énergie cumulée par orientation dépend uniquement de la résolution de l'analyseur spectral. Une image spectrale décomposée en *n* bandes d'orientations donne une résolution angulaire de *180/n* degrés. La précision augmente avec le nombre d'orientations de l'analyseur spectral. Comme dans ce travail nous décomposons le spectre en 15 orientations, la résolution est de 12°.

Cet algorithme a un coût de calcul négligeable du fait de la taille réduite de l'image spectre log polaire (15 orientations *15 fréquences = 225 "pixels").

Symbole associé à l'analyseur d'orientations

On associe à ce module d'analyse des orientations dans une image le symbole présenté sur la figure II.8. Il prend en entrée un spectre échantillonné log polaire. Sa sortie est la courbe d'énergie cumulée par orientation.

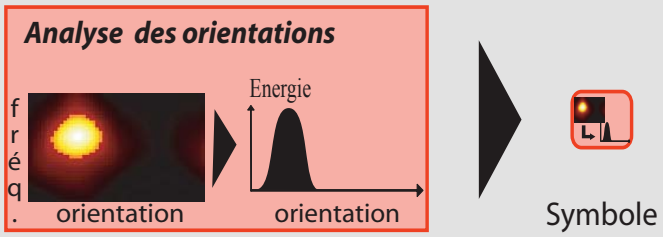


Figure II.8: symbole associé au détecteur d'orientations

II.3.2. Détection des changements temporels: détecteur d'événements

II.3.2.1. Principe

La figure II.9 montre l'architecture de l'algorithme proposé pour détecter les changements spatio-temporels présents dans une séquence vidéo sur la base de l'analyse de l'évolution de l'énergie de son spectre log polaire. Pour chaque image de la séquence, après extraction des contours en mouvement, on calcule le spectre log polaire de l'image ainsi que l'énergie globale associée à ce spectre. Le principe est d'observer l'évolution temporelle de cette énergie pour en déduire les événements liés au mouvement. En effet, on rappelle que la sortie du filtre MagnoY contours mobiles ne présente de l'énergie qu'au niveau des pixels appartenant à des contours mobiles. Donc, en cas d'absence de mouvement, l'énergie du spectre log polaire est très faible, voire nulle et elle augmente en présence de mouvement.

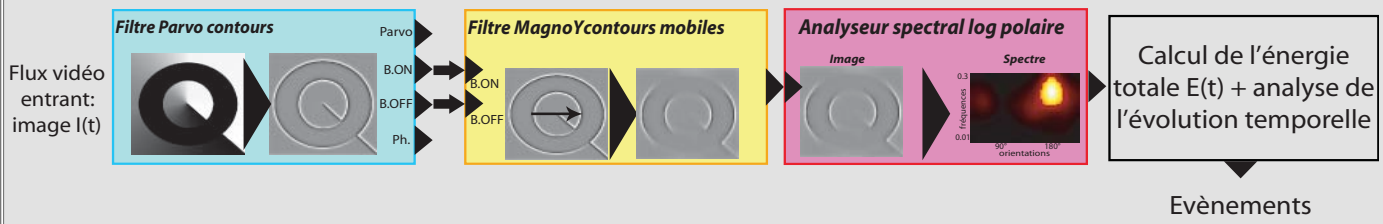


Figure II.9: schéma du détecteur d'événements

La figure II.10.a montre quelques images d'une scène vidéo de rue dans laquelle un cycliste fait irruption entre les images 86 et 123 dans une zone précédemment exempte de mouvement puis ensuite, un groupe de personnes traverse la scène des images 199 à 411. La figure II.10.b montre l'évolution temporelle de l'énergie totale de la sortie de l'analyseur spectral log polaire pour cette scène.

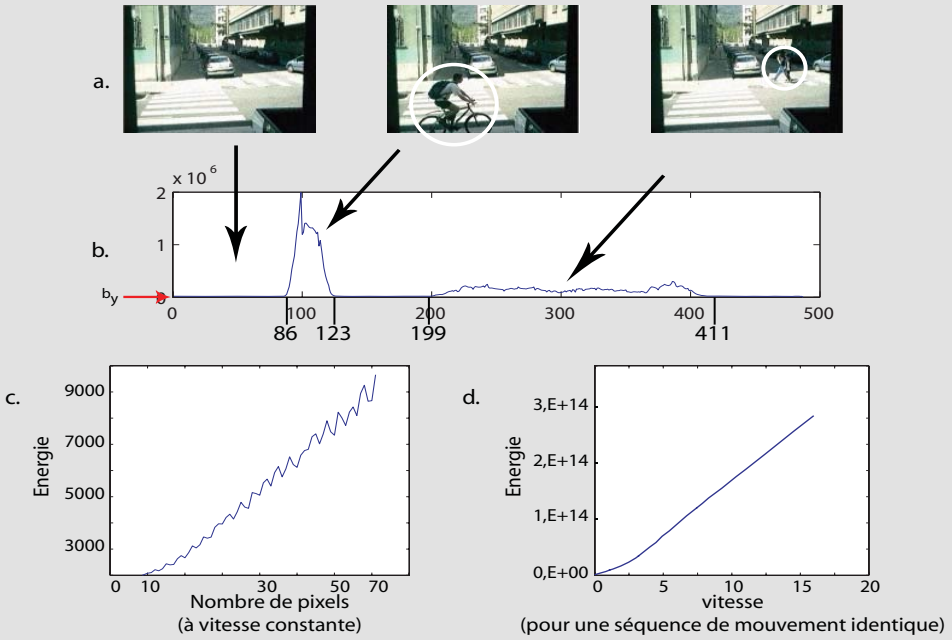


Figure II.10: analyse de l'évolution de l'énergie totale du spectre log polaire de l'image des contours mobiles. a, extraits de la séquence test (scène de rue bruitée). b, énergie totale du spectre. c, évolution de l'énergie en sortie du filtre MagnoY contours mobiles en fonction du nombre de pixels en mouvement à vitesse constante. d, évolution de cette énergie en fonction de la vitesse à nombre de pixels en mouvement constant.

→ Dans les périodes exemptes de mouvement, on observe un niveau d'énergie très faible. Ce niveau d'énergie noté b_y est lié au niveau de bruit résiduel créé par l'acquisition.

→ Lors des périodes de mouvement, l'énergie augmente du fait de la détection des contours mobiles dans la scène. On constate que cette énergie dépasse alors le niveau dû au bruit seul.

Cette augmentation de l'énergie avec le mouvement est liée au filtrage passe-haut temporel du filtre MagnoY contours mobiles. Plus précisément, la figure II.10.c montre qu'en sortie de ce filtre, l'énergie augmente linéairement avec le nombre de pixels impliqués dans le mouvement pour une vitesse constante. De même, l'énergie augmente linéairement en fonction de la vitesse d'un même mouvement (cf. fig. II.10.d). Ceci explique pourquoi le mouvement du cycliste donne une énergie plus importante que le groupe de piétons: il met en jeux plus de pixels sur l'image et a une vitesse de déplacement plus importante.

II.3.2.2. Détecteur de changements temporels

Dans cette partie, nous proposons un système qui permet de détecter les changements temporels liés à des événements de mouvement en les décrivant non pas de manière indépendante, mais plutôt du point de vue de leur enchaînement. Le but est de créer un indicateur normalisé à partir duquel on pourra comparer le mouvement courant aux mouvements précédents temporellement proches. Pour cela, différentes approches sont possibles:

→ Seuillage simple : après avoir estimé le bruit de fond, un seuillage permet de détecter un mouvement. Dans cette méthode se pose le problème du choix du seuil. Par ailleurs, la décision prise à l'instant courant ne tient pas compte des mouvements passés.

→ Calcul de variation de l'énergie sur un intervalle de temps fixe : on calcule la variation des *n* dernières valeurs d'énergie mesurées et quand celle-ci dépasse un seuil, on considère qu'il y a mouvement. La difficulté est une fois de plus le choix du seuil de déclenchement de l'alerte, mais cette fois au niveau de la déviation maximum tolérée. Par suite, comment décrire les fluctuations si le mouvement ralenti, s'accélère?

Dans ce travail, nous proposons une méthode qui permet un choix de seuil simplifié ainsi qu'une description de l'enchaînement des mouvements. Elle est basée sur une intégration temporelle de l'énergie selon un principe similaire à ce que l'on trouve en électronique dans les montages diode/condensateur pour le redressement d'une tension alternative. On utilise un modèle de redresseur simple alternance tel que celui illustré sur la figure II.11. Le principe de fonctionnement est le suivant: pour chaque image $I(t)$ d'une séquence vidéo, on applique l'énergie totale $E(t)$ du spectre log polaire à l'entrée du circuit redresseur. On obtient en sortie le signal $E_1(t)$. Le signal $E_1(t)$ prend des valeurs décroissantes tendant vers zéro lorsque $E(t)$ est faible ou bien suit l'accroissement de $E(t)$ lorsque celui-ci prend de grandes valeurs. Plus précisément, au niveau de la diode, quand au temps t la tension à la cathode ($E_1(t)$) plus la tension de diode caractéristique Vd est inférieure à celle de l'anode, alors la diode est passante et l'on arrive à l'égalité $E_1(t)=E(t)-Vd$. Ceci correspond aux périodes pour lesquelles un fort mouvement est présent ($E(t)$ *fort*). Au contraire quand l'entrée $E(t)$ est inférieure à $E_1(t)+Vd$ (c.-à-d. mouvement faible par rapport aux mouvements passés ou nul), alors la diode est bloquée et la charge accumulée dans le condensateur C est évacuée dans la résistance R. Ceci entraîne une évolution temporelle décroissante de la tension $E_1(t)$ de la forme $E_1(t)=E_0.e(-(t-t_0)/\tau_{E1})$ avec $\tau_{E1}=RC$, la constante de temps et E_0 la tension $E_1(t_0)$ juste avant le blocage de la diode au temps t_0.

L'idée est ensuite de définir l'indicateur d'événements de mouvement α(t) par la relation:

$$\alpha(t)= E_1(t)/(E(t)-Vd) \qquad \textit{(Eq. II.1)}$$

Cet indicateur prend ses valeurs dans l'intervalle [0;1].

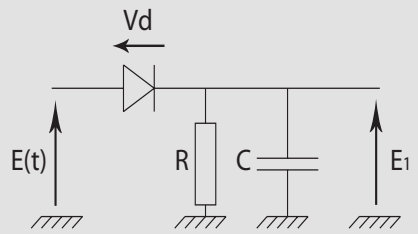


Figure II.11: schéma électrique équivalent permettant de calculer l'énergie $E_1(t)$

Fonctionnement et paramétrage

→ La tension seuil de diode Vd représente le seuil de déclenchement du système. Son but est d'éliminer les fausses alertes de mouvement dues au niveau de bruit résiduel sur $E(t)$. Si l'on évalue le bruit moyen μ_E de $E(t)$ et son écart type σ_E lors d'une absence de mouvement (en début d'analyse par exemple) on peut fixer le seuil Vd de façon adaptée vis-à-vis de la scène analysée. Nos différents tests ont montré que le seuil Vd peut

être fixé à $Vd=\mu_E+3\sigma_E$. Nous appellerons E_{noise} l'énergie seuil jouant le rôle de Vd.

→ Le paramètre τ_{E1} est la constante de temps du système. Plus τ_{E1} est fort, plus le système prendra en compte les événements passés éloignés. Au contraire, plus τ_{E1} est faible, moins les événements passés à long terme seront intégrés. $E_1(t)$ atteint en effet 64% de sa valeur finale en régime permanent τ_{E1} secondes après application du signal E(t). Nous fixons τ_{E1} à 0.5 pour obtenir une constante de temps de 0.5 seconde (c.-à-d. que l'énergie du spectre log polaire est intégrée sur 12.5 images pour une vidéo à 25 images par seconde).

Propriétés

La figure II.12.b montre l'évolution temporelle des énergies E(t) et E_1(t) calculées sur une boite englobante centrée sur un oeil (cf. fig II.12.a). Les événements de mouvement recherchés sont les clignements.

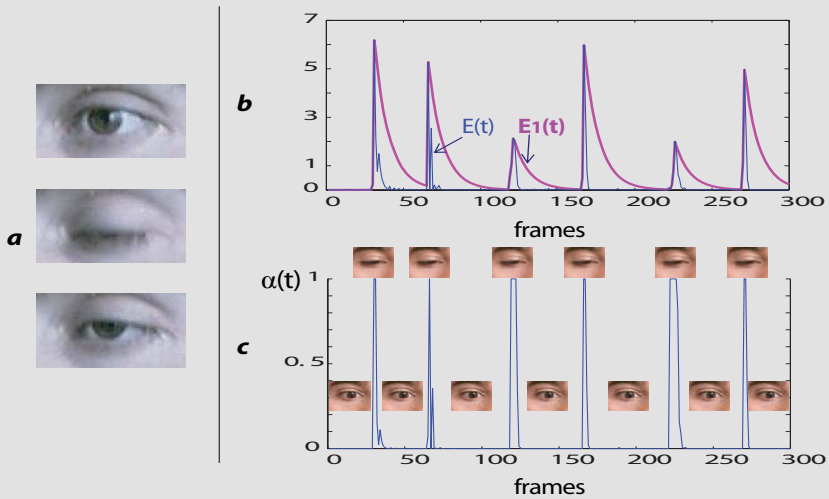


Figure II.12: détection d'événements de mouvement. a, extraits de la séquence de clignements. b, énergies E(t) et E_1(t). c. indicateur de mouvement α(t)

Dans cet exemple, on observe que chaque clignement crée un pic d'énergie sur *E(t)*. Ces pics se retrouvent également sur $E_1(t)$, mais leur amplitude décroît plus lentement du fait de l'effet temporel du système proposé. On observe sur la figure II.12.c que l'indicateur *α(t)* atteint son maximum lors de chaque clignement d'oeil, mais reste à zéro le reste du temps. Cet indicateur constitue un détecteur d'événements temporels très sélectif.

Remarque: dans cette scène vidéo, lors d'un clignement d'oeil, l'amplitude des pics d'énergie sur *E(t)* est toujours différente alors que le mouvement est sensiblement le même. Ceci est dû à la vitesse de capture de la caméra. Elle n'est que de 30 images par seconde ce qui est insuffisant pour analyser de façon précise les clignements. On ne respecte pas les conditions de Shannon, la fréquence du clignement est supérieure à la moitié de la période d'échantillonnage (15 images/s). Pour le confirmer, la figure II.13 décompose image par image un clignement d'oeil, on constate que la fermeture de l'oeil peut se faire en 2 images. Avec la méthode

proposée, nous arrivons tout de même à déclencher une alerte de mouvement: l'indicateur $\alpha(t)$ prend sa valeur maximale. Il serait néanmoins intéressant d'utiliser une caméra plus rapide pour d'autres analyses plus fines.

Figure II.13: décomposition image par image d'un clignement d'oeil filmé à 30 images par seconde

L'exemple de la figure II.14 montre l'évolution des énergies $E(t)$, $E_1(t)$ et de l'indicateur de mouvement $\alpha(t)$ pour une scène plus globale dans laquelle une personne effectue une séquence de gestes avec sa main.

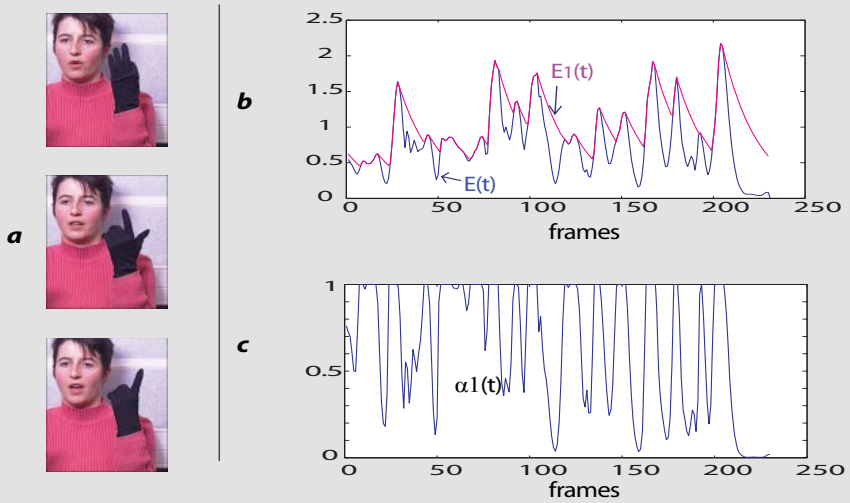


Figure II.14: a, extraits d'une séquence de gestes de main. b, énergies E(t) et E_1(t). c, indicateur de mouvement α(t)

Cet exemple est plus complexe, car plusieurs mouvements sont mélangés: mouvements des doigts, mouvements de l'ensemble (bras+main) et mouvements du visage. Le mouvement de l'ensemble (bras+main) est dominant, car de plus forte amplitude et mettant en jeux plus de pixels. On observe que l'énergie totale $E(t)$ est sujette à une forte croissance lors de chaque accélération, la vitesse maximale d'un mouvement correspond à une énergie maximale et chaque arrêt ou transition correspond à une énergie minimum. L'indicateur $E_1(t)$ suit les accélérations quand elles ne précèdent pas un mouvement immédiat de plus forte amplitude. La décroissance temporelle de $E_1(t)$ permet en quelque sorte de filtrer certaines accélérations trop faibles par rapport aux mouvements précédents de plus fortes amplitudes. La conséquence est un comportement particulier de l'indicateur $\alpha(t)$:

→ Lorsque le mouvement précédent est faible ou temporellement trop éloigné du mouvement courant,

alors $\alpha(t)$ prend une valeur élevée.

→ Si le mouvement précédent est du même ordre de grandeur que le mouvement courant, alors $\alpha(t)$ prendra une valeur équivalente à celle du mouvement précédent.

→ Enfin, si le mouvement courant est faible devant les mouvements précédents, alors $\alpha(t)$ aura une valeur faible, voire nulle.

Ceci permet une comparaison temporellement adaptative. On a un descripteur de mouvement courant avec un effet d'oubli progressif des mouvements précédents. Lorsqu'un nouveau mouvement est détecté, on sait de plus si il est du même ordre de grandeur, plus rapide ou plus lent que les précédents.

II.3.2.4. Analyse de performances

Afin de détecter automatiquement les événements de mouvement, on seuille l'indicateur $\alpha(t)$. Celui-ci étant dans l'intervalle [0; 1], le seuillage est facilité car la normalisation du signal a déjà été effectuée en respectant le niveau de bruit et l'amplitude des signaux. Nous choisissons par défaut un seuil de détection α_{min} bas fixé à 0,2. Cette valeur permet de sélectionner les mouvements dont l'amplitude est supérieure de 20% à celle des mouvements précédents (en prenant en compte l'effet d'oubli temporel introduit par $E_1(t)$). Nous obtenons un signal temporel binaire que l'on note $Mv(t)$ qui suit la relation II.2. Ce signal est très sensible aux mouvements, y compris aux mouvements donnant une faible valeur de $\alpha(t)$ (i.e. les mouvements de plus faible amplitude que le mouvement précédent temporellement proche).

$$\begin{cases} Mv(t) = 1 \text{ si } \alpha(t) > \alpha_{\min} \\ Mv(t) = 0 \text{ sinon} \end{cases} \qquad \textit{(Eq. II.2)}$$

Nous avons mené une campagne de mesures visant à évaluer le taux de fiabilité de l'indicateur $Mv(t)$. Pour cela, nous avons considéré une base de vidéos pour lesquelles les événements de mouvement sont repérés manuellement. La base contient 410 minutes de vidéo constituée de scènes acquises dans des conditions différentes: éclairage standard (de bureau par exemple), éclairage faible et conditions bruitées. Le contenu des scènes filmées est également varié: scènes de rue (passage de voitures, etc.), scènes dans lesquelles un visage est par moment en mouvement (mouvement global de la tête, clignements d'yeux, mouvement de bouche) et scènes de bureau (passage de personne, etc.). 2037 alertes de mouvement ont été repérées. La figure II.15 montre quelques exemples de ces vidéos tests.

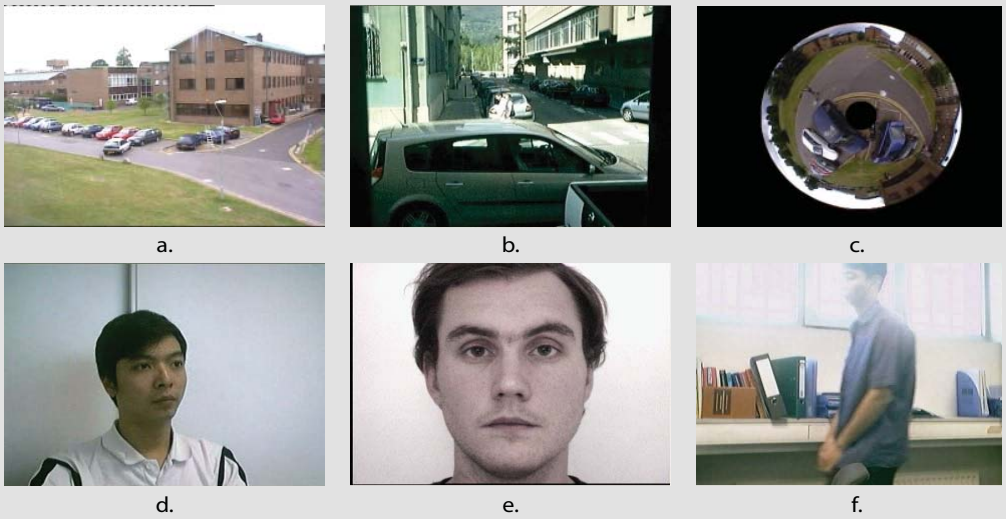


Figure II.15: extrait de la base de test pour l'évaluation du détecteur d'événements de mouvement. a, b et c, exemples de séquences de rue ((c) avec objectif 360°). Mouvement de tête (d) et d'yeux (e). f, scène de bureau

Nous avons évalué le taux de succès de ce système, les taux d'échec et d'oubli. Nous avons également étudié la capacité de cet algorithme à estimer la durée d'un mouvement. Pour estimer la précision de la mesure de durée de mouvement, nous comparons la durée détectée par le système par rapport à la vérité terrain (étiquetée manuellement). De même, nous évaluons la synchronisation entre le début des alertes effectives et le début extrait par l'algorithme. Cette donnée figure sous le terme "décalage", elle représente en valeur absolue le nombre d'images de décalage entre la vérité terrain et la réponse de l'algorithme. Le tableau II.3 synthétise les performances globales de ce système. La méthode de calcul est la suivante, soit *GD(i)* l'image *i* dans une séquence vidéo pour laquelle on détecte manuellement une alerte de mouvement (vérité terrain). On note *S(i)*, les images pour lesquelles les alertes détectées correspondent à la vérité terrain, *F(i)*, les images pour lesquelles l'algorithme a effectué une fausse détection et *O(i)*, les images pour lesquelles l'algorithme a oublié de détecter une alerte.

On extrait alors les taux de performance T_S, T_F et T_O respectivement taux de succès, de fausse alarme et d'oubli selon la relation :

$$\begin{cases} T_S = \dfrac{\sum S(i)}{\sum GD(i)} \\ T_F = \dfrac{\sum F(i)}{\sum GD(i)} \\ T_O = \dfrac{\sum O(i)}{\sum GD(i)} \end{cases} \qquad \textit{(Eq. II.3)}$$

Table II.3: performances du détecteur d'événements							
	Nombre d'alertes	***T_S***	***T_E***	***T_Q***	***Moyenne des décalages***	***Ecart type des décalages***	***Précision de l'estimation de la durée***
Conditions standard	*1202*	*95.38%*	*1.28%*	*2.80%*	*1.28*	*2.7*	*79.8%*
Eclairage faible	*835*	*94.97%*	*1.81%*	*2.9%*	*1.60*	*2.1*	*81.61%*

On observe que le taux de réussite T_S est élevé que ce soit en conditions standard ou en conditions d'éclairage faible. Près de 95% des mouvements présents dans les vidéos ont été détectés. Les échecs (oublis ou fausses détections) se produisent surtout lorsque le rapport signal sur bruit est très faible (de l'ordre de 5dB), la réponse des contours en mouvements restant noyée dans le bruit. Parallèlement, la détection des événements est synchronisée avec la vidéo, on note un retard inférieur à 2 images. L'estimation de la durée des événements de mouvement est correcte.

II.3.2.5. Comparaison avec une différence temporelle

La détection de mouvement est un sujet qui a donné lieu à nombre de recherche. Ce type de détection est généralement basé sur une analyse pixel par pixel des changements de luminance. Cette évolution temporelle de la luminance de chaque pixel amène à identifier les zones dans lesquelles un changement se produit. Historiquement, la détection des changements temporels s'est faite par analyse de la différence inter-image.

Sur la figure II.16 est présentée une comparaison entre la méthode proposée et la différence d'images pour la même séquence vidéo que celle présentée sur la figure II.10. La figure II.16.b montre l'évolution temporelle de l'énergie totale *E(t)* du spectre log polaire ainsi que la détection binaire de mouvement *Mv(t)*. La détection de mouvement par différence d'images est présentée sur la figure II.16.c qui montre l'énergie de la différence entre deux images successives ainsi que la détection de mouvement par seuillage au dessus du niveau de bruit mesuré b_d.

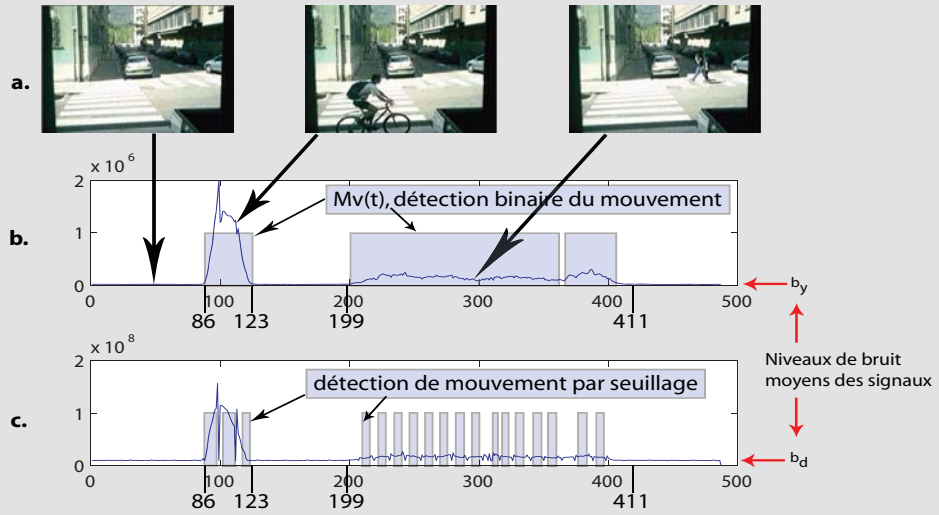


Figure II.16: comparaison de méthodes de détection de mouvement. a, extraits de la séquence test (scène de rue bruitée). b, énergie mesurée en sortie de l'analyseur spectral et représentation de Mv(t). c, énergie de la différence d'images et seuillage binaire.

Au niveau de l'évolution temporelle de l'énergie, les deux méthodes ont un comportement similaire: de fortes valeurs lors de la présence d'un mouvement et des valeurs faibles en l'absence de mouvement. En revanche, la détection binaire de mouvement est plus efficace avec notre méthode, car elle maintient l'alerte de mouvement tout le long de celui-ci en introduisant moins de fragmentation. Ceci s'explique par la méthode employée qui privilégie un prétraitement efficace du signal à analyser. Les filtrages Parvo contours et MagnoY contours mobiles renforcent les contours en mouvement tout en minimisant le bruit spatio-temporel. L'effet de ces filtrages s'observe notamment dans les périodes sans mouvement. On mesure sur la courbe d'énergie II.16.b, un bruit de fond b_y d'amplitude très faible et constante donnant un rapport signal sur bruit de l'ordre de 19dB. A l'opposé, par utilisation de la différentielle temporelle, le niveau de bruit b_d donne un rapport signal sur bruit de l'ordre de 6dB. Donc il est plus aisé de faire la distinction entre l'énergie due au bruit et celle due à la présence effective d'un objet en mouvement

Les principales différences entre la méthode proposée et la différence d'image classiquement utilisée concernent deux points importants:

→ La méthode proposée minimise le bruit, réduisant ainsi le risque de fausse alarme.

→ La différence d'images suppose que l'illumination de la scène analysée est constante de manière à attribuer à un mouvement, toute variation temporelle de la fonction de luminance. La méthode que nous proposons s'affranchit de cette hypothèse grâce à la non prise en compte de la valeur de luminance moyenne de l'image au niveau de filtre Parvo contours ainsi qu'au niveau de l'analyseur spectral. Par voie de conséquence, toute variation temporelle de la fonction de luminance peut être imputée à la présence d'un mouvement.

II.3.2.6. Résumé

Nous venons de proposer un système capable de détecter les événements de mouvement dans les vidéos. Cet algorithme est capable de donner en plus une information adaptative sur les mouvements détectés: il compare le mouvement courant par rapport aux événements précédents avec un effet d'oubli progressif. Nous savons si le mouvement courant est plus rapide, plus lent ou équivalent aux mouvements précédents.

Cet algorithme bas niveau est plus approprié pour la détection des événements de mouvement que pour l'estimation de leur durée. C'est avant tout un détecteur de transitions temporelles. Il donne néanmoins un bon ordre de grandeur de la durée des mouvements présents et peut les décomposer suivant les phases d'accélérations et de décélérations rencontrées.

Son paramétrage consiste à ajuster le paramètre τ_{E1} régissant le comportement d'atténuation temporelle du signal $E_1(t)$: plus τ_{E1} est grand, plus le système tient compte des événements temporellement éloignés (effet d'oubli atténué). De même, l'ajustement du seuil de déclenchement E_{noise} permet de maintenir un niveau de sensibilité minimisant les problèmes de bruit. Le choix du seuil de binarisation de $\alpha(t)$ est flexible, une valeur proche de 1.0 permet la sélection des mouvements de grande amplitude seulement ou des mouvements temporellement isolés, des valeurs plus faibles permettront d'extraire tous les mouvements.

Du point de vue du coût de calcul, cette méthode est très peu coûteuse une fois les filtrages Parvo contours, MagnoY contours mobiles et analyse spectrale effectués. La première étape est le calcul de l'énergie totale à partir du spectre log polaire ce qui représente 225 additions par image pour des spectres de 15 orientations par 15 bandes de fréquence. Le calcul de $E_1(t)$ et $\alpha(t)$ est ensuite réalisé en 3 opérations par image. Au final, le coût de calcul associé au détecteur d'événements proprement dit est négligeable au regard des autres calculs réalisés en amont (filtres Parvo contours, MagnoY contours mobiles et analyseur spectral). Si l'on considère toute la chaîne de traitement, cet algorithme fonctionne à 30 images par seconde sur un ordinateur

équipé d'un processeur de type Intel Pentium 4, 3.0Ghz avec un code C++ non optimisé.

Nous associons à ce module le symbole présenté sur la figure II.17. Ce module prend en entrée l'énergie *E(t)* du spectre log polaire de la sortie du filtre MagnoY contours mobiles (cf. fig. II.9). On dispose en sortie de l'indicateur *α(t)* et de l'indicateur binaire *Mv(t)*.

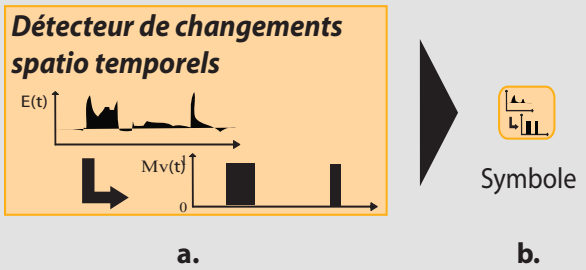


Figure II.17: symbole du détecteur d'alertes de mouvements (changements spatio-temporels)

II.4. Segmentation de mouvement

Nous allons maintenant voir comment effectuer une segmentation des zones en mouvement présentes dans une séquence vidéo en tenant compte des informations bas niveau fournies par la rétine. Nous nous plaçons dans le cas où la caméra utilisée pour l'acquisition des images est fixe.

II.4.1. Principe

Le but est d'isoler des zones de mouvement. La sortie du filtre MagnoY contours mobiles donne l'énergie des contours en mouvement ce qui permet la localisation spatiale des zones mobiles. Nous allons utiliser cette information pour mener à bien la tâche de segmentation. Le principe est le suivant: si un objet est en mouvement, alors son énergie en sortie du filtre MagnoY contours mobiles est localement plus forte que celle des pixels statiques dans la zone alentour (en dehors de la zone de l'objet mobile). Nous allons donc évaluer localement l'énergie en sortie MagnoY contours mobiles et la comparer à l'énergie moyenne dans un voisinage plus étendu. Si l'énergie locale est supérieure à l'énergie moyenne dans un voisinage plus étendu, alors il s'agit d'une zone de mouvement.

Sur ce principe, nous construisons le système décrit sur la figure II.18.a. Nous effectuons deux filtrages passe-bas spatio-temporels successifs F_1 et F_2 sur **l'énergie** E_y de la sortie du filtre MagnoY contours mobiles. Le premier filtrage de fréquence de coupure spatiale élevée donne l'énergie locale *Seg_p*. Le second filtrage de fréquence de coupure spatiale plus faible donne l'énergie moyenne *Seg_h* dans un voisinage plus large. Selon le principe énoncé précédemment, un pixel est en mouvement si son énergie est localement supérieure à l'énergie moyenne avoisinante. La segmentation du mouvement consiste à sélectionner les pixels *(x, y)* tels que *Seg_p(x, y) > Seg_h (x, y)*. **L'avantage de cette méthode est que la segmentation est localement adaptée à la quantité de mouvement**. La figure II.18.b schématise ce processus dans le cas d'un signal carré en translation.

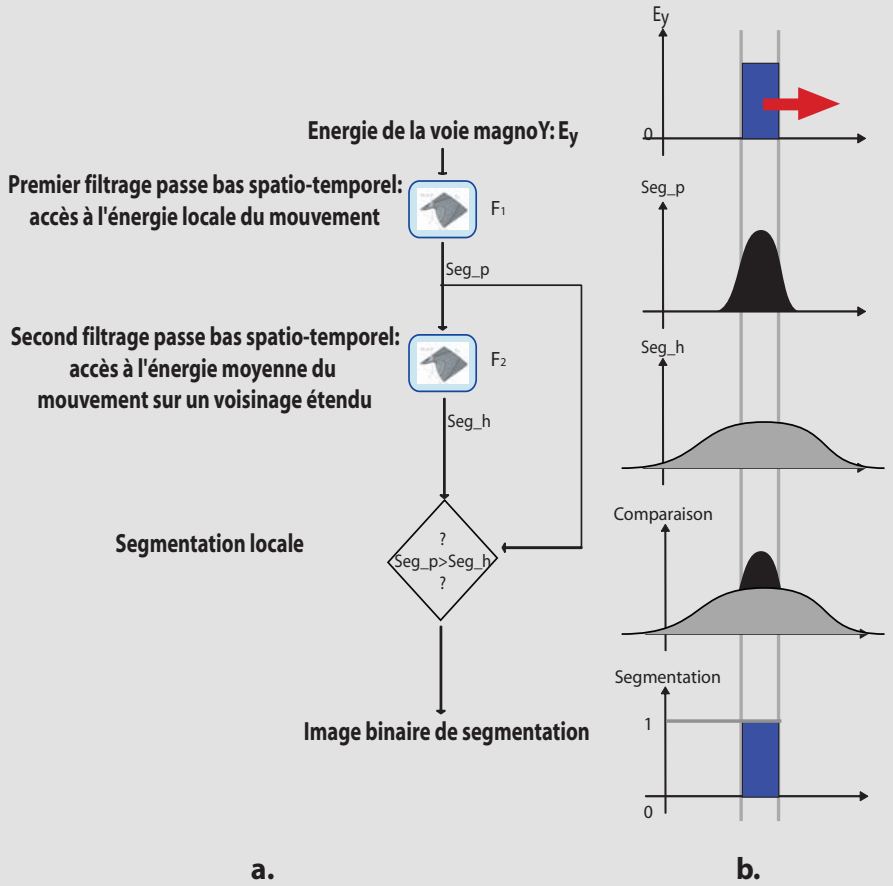


Figure II.18: segmentation de zones en mouvement. a, système proposé. b, réponses des différentes étapes du modèle pour un carré en translation

Remarque: cette méthode est, de par son architecture, inspirée de la couche PLE de la rétine. Le premier filtrage donne en sortie l'image *Seg_p* permettant de lisser les maximums et minimums locaux de l'énergie de la sortie MagnoY contours mobiles. Ceci donne une information sur la quantité locale de mouvement, et permet en quelque sorte de lisser et de minimiser le "bruit" comme le font les photorécepteurs avec l'information de luminance. Le second filtrage donne l'image *Seg_h* dont la fonction est de donner une information sur le mouvement moyen sur un voisinage étendu, par analogie avec les cellules horizontales de la PLE qui donnent la luminance moyenne locale. Enfin, la segmentation par comparaison des 2 filtrages peut être assimilée à l'action des cellules bipolaires ON de la rétine.

II.4.2. Limitations initiales de la méthode

La figure II.19 montre sur un exemple 1D le fonctionnement de l'algorithme proposé. Sur la figure II.19.a, nous avons la sortie de l'énergie du filtre MagnoY sur une ligne d'une scène vidéo. On y remarque un bruit résiduel non négligeable et deux objets en mouvement. Le premier objet (pixels 60 à 80) donne des valeurs d'énergie légèrement supérieures au niveau de l'énergie du bruit et présente un décrochement en son milieu (dû à une zone uniforme). Le second objet (pixels 140 à 180) présente un mouvement plus important donc une énergie en sortie MagnoY contours mobiles également plus forte que le premier objet. Ce second objet est texturé et son énergie est quasi constante sur toute sa surface.

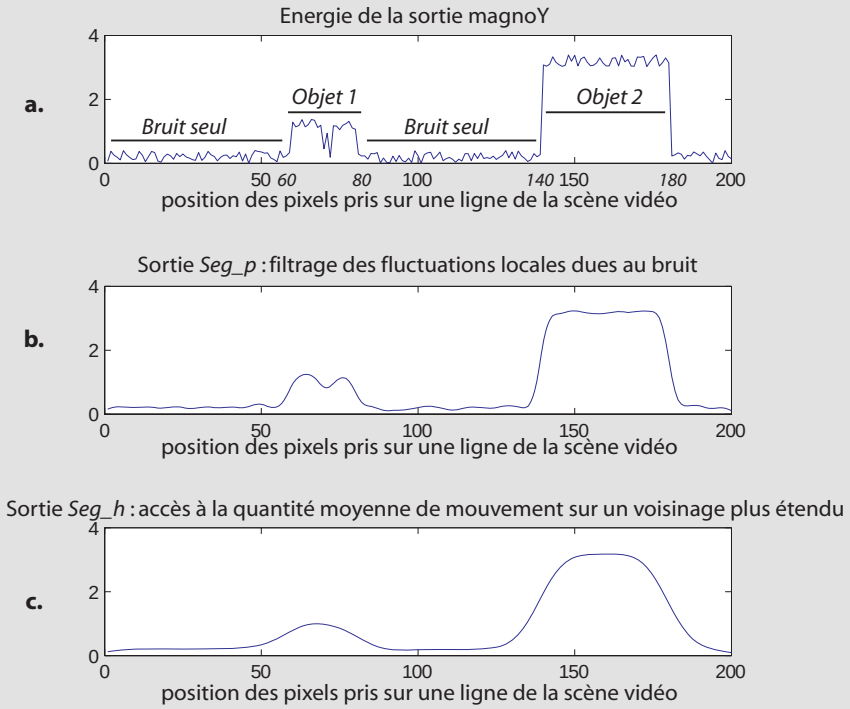


Figure II.19: réponse des filtres F_1 et F_2 à l'énergie de la voie MagnoY contours mobiles. a, énergie E_y de la voie magnoY. b, énergie Seg_p: bruit et fluctuations locales minimisés. c, énergie Seg_h: moyenne du mouvement sur un voisinage plus étendu.

Observons maintenant les réponses des différents étages du système:

→ Le filtre F_1(cf. fig. II.19.b) atténue le bruit tout en conservant la forme globale du signal.

→ Le filtre F_2(cf. fig. II.19.c) montre des valeurs plus fortes dans les zones de mouvement, mais cette énergie est spatialement étalée (diffusée).

Si l'on effectue une segmentation par comparaison des deux énergies *Seg_p(x, y)* et *Seg_h(x, y)* (cf. fig. II.20), nous segmentons effectivement les deux objets en mouvement. Mais quelques problèmes apparaissent:

→ Le décrochement au milieu du premier objet dû à une zone uniforme entraîne une absence locale de segmentation. Ceci est dû à l'amplitude de sortie du second filtrage *Seg_h* qui est trop importante dans la zone en mouvement.

→ Le second objet est segmenté en deux parties distinctes. Ceci est également dû à une amplitude trop forte du second filtrage *Seg_h* dans la zone en mouvement. Les amplitudes de *Seg_p* et *Seg_h* donnent un résultat similaire au milieu d'un objet large, empêchant alors par endroit la segmentation.

→ Les zones dans lesquelles aucun mouvement n'apparaît donnent également des zones de segmentation. Ceci est dû au bruit résiduel qui force *Seg_p* à passer au dessus de *Seg_h* car les fluctuations locales sont parfois supérieures à la moyenne locale calculée sur un voisinage plus étendu. Ainsi, il faudrait augmenter la valeur de *Seg_h* dans les zones dépourvues de mouvement de manière à éviter ce problème de sur-segmentation dû au bruit.

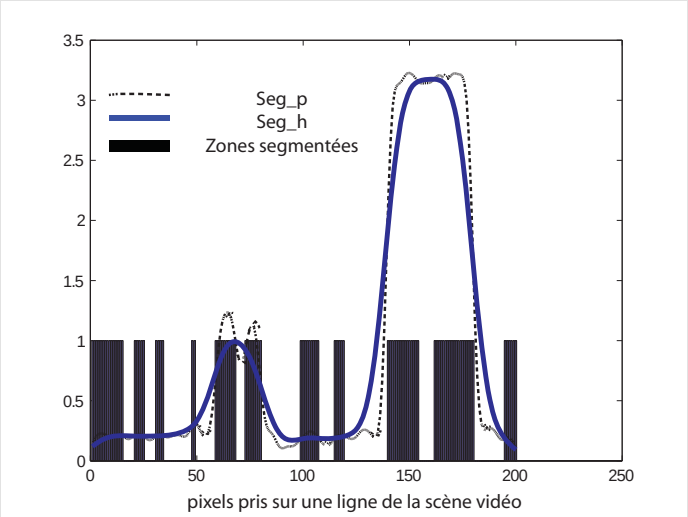


Figure II.20: résultat de la segmentation par comparaison de Seg_p et Seg_h

Deux problèmes apparaissent donc au niveau de la sortie du second filtrage *Seg_h*. Cette sortie doit être supérieure à *Seg_p* dans les zones dépourvues de mouvement de manière à limiter l'influence du bruit. En revanche, cette sortie doit être atténuée par rapport à *Seg_p* dans les zones en mouvement de manière à segmenter toute la surface d'un objet mobile et gérer le cas des zones uniformes.

II.4.3. Correction du système

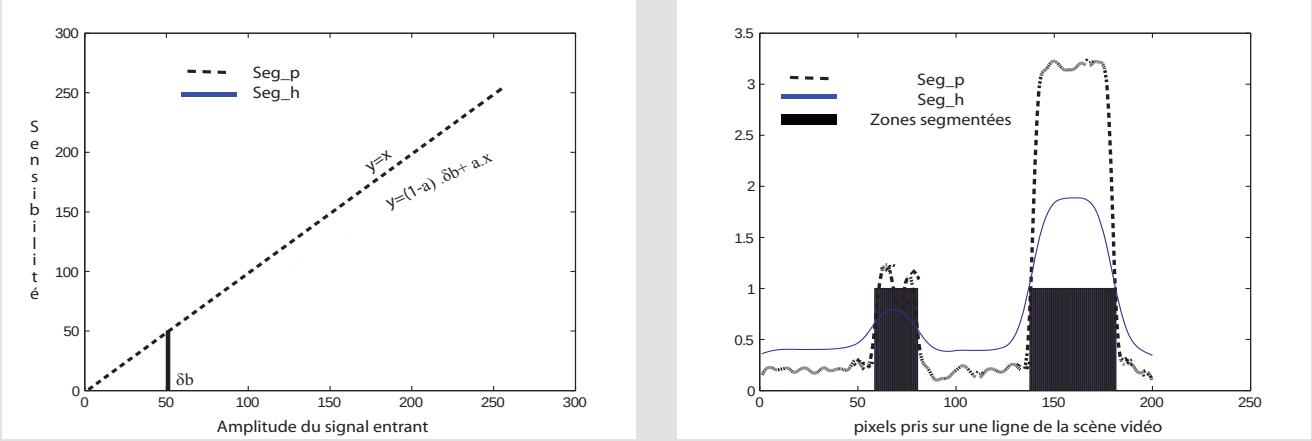
II.4.3.1. Principe

La solution proposée consiste à introduire une pondération sur la sortie *Seg_h* par une relation linéaire. La figure II.21.a montre la pondération des sorties *Seg_h* et *Seg_p*. *Seg_p* est inchangé et suit une loi linéaire du type $y=x$. En revanche, *Seg_h* est pondéré par une relation du type $y=ax+b$ dont nous allons décrire le paramétrage. L'intersection des 2 courbes de gain se situe en l'abscisse $x=\delta_b$, ce paramètre correspond au niveau de bruit maximum admissible. Par déduction, le paramètre b de la pondération de *Seg_h* est égale à δ_b $(1-a)$. L'effet de cette modification entraîne les évolutions suivantes:

→ Pour les valeurs d'énergie d'entrée inférieures à δ_b, la sortie *Seg_h* a un gain supérieur à la réponse de *Seg_p*. Comme nous l'avons vu sur la figure II.19.a, les valeurs d'entrée faibles correspondent aux zones sans mouvement dans lesquelles seul subsiste le bruit. Ceci va permettre de ne pas segmenter le bruit lors de la comparaison *Seg_p* v.s. *Seg_h*.

→ Pour les valeurs d'énergie d'entrée supérieures à δ_b, le gain de *Seg_h* devient inférieur à celui de *Seg_p*. L'effet escompté est de pouvoir segmenter toute la surface de l'objet en mouvement, même si sa surface est importante ou faiblement texturée.

La figure II.21.b montre le résultat de la segmentation avec cette pondération. On observe effectivement une segmentation des objets en mouvement seuls, sans problème de sous-segmentation de leur surface. Par ailleurs, on ne constate pas de sur-segmentation du bruit.



a. **b.**

Figure II.21: segmentation en différentiant les pondération des sorties Seg_p et Seg_h. a, courbes de sensibilité attribuées aux sorties Seg_p et Seg_h en fonction du niveau d'entrée. b. résultat de la segmentation.

II.4.3.2. Paramétrage

Paramétrage du filtre passe-bas spatio-temporel F_1: énergie locale Seg_p

Le filtre passe-bas F_1 a pour but d'homogénéiser très localement l'énergie de la sortie MagnoY contours mobiles. Ainsi, suivant la taille des objets à segmenter, la fréquence de coupure spatiale du filtre devra être ajustée. Une valeur élevée de cette fréquence de coupure ne minimise que faiblement les fluctuations locales de l'énergie. Cela signifie que la sortie de ce filtre (*Seg_p*) conserve des fluctuations d'énergie sur les zones en mouvement, que ce soit pour des objets réellement en mouvement ou pour le bruit résiduel. On risque alors de segmenter le bruit et de décomposer un objet en plusieurs régions. Au contraire, une fréquence de coupure trop faible risque d'étendre les zones segmentées, trop largement autour de l'objet en mouvement, mais le bruit lui sera beaucoup moins facilement segmenté. Par défaut, nous utilisons la même fréquence de coupure spatiale que celle utilisée pour les cellules horizontales de la rétine (constante d'espace de 5 pixels). Ce paramètre de lissage est donc fort, néanmoins, rappelons-nous que tout objet en mouvement est visuellement perçu de manière floue, sa zone de segmentation n'est donc pas précisément définie.

On donne également une réponse temporelle à ce filtre ce qui permet de lisser temporellement l'énergie et surtout d'étendre la réponse des contours en mouvement sur les surfaces homogènes de l'objet en mouvement. La constante de temps du filtre F_1 est fixée à 1/25 seconde pour des séquences acquises à 25 images par seconde. Une valeur faible de cette constante de temps permet au système de répondre rapidement, mais segmente surtout les contours de l'objet sans ses surfaces homogènes. Au contraire, une constante de temps de forte valeur rend le système moins réactif, mais permet d'étendre la zone de segmentation à toute la surface de l'objet.

Paramétrage du filtre passe-bas spatio-temporel F_2 :énergie moyenne Seg_h

Le filtre F_2 a pour fonction d'estimer le mouvement moyen dans un voisinage plus étendu. Sa fréquence de coupure spatiale doit donc être inférieure à celle du premier filtrage. Une fréquence de coupure trop

élevée conduit à un risque plus élevé de segmentation du bruit et à une décomposition des objets en plusieurs parties suivant la concentration locale d'énergie sur leur surface. Une fréquence tendant vers 0 conduit à se rapprocher d'une estimation de l'énergie moyenne sur toute l'image. Nous avons fixé la fréquence de coupure spatiale à une valeur basse, avec une constante d'espace de 25 pixels. Les valeurs de ces paramètres spatiaux seront réduites dans le but de segmenter des objets plus petits et augmentées pour segmenter des objets de taille plus importante.

Comme pour le filtre F_1, on ajuste la constante de temps de ce filtre à 1/25s dans le but d'étendre la zone de segmentation sur la surface de l'objet.

Paramétrage de la pondération de l'algorithme de segmentation

Le paramètre δ_b est directement lié au bruit maximum admissible en sortie magnoY contours mobiles. Il représente la limite entre le signal utile de mouvement et le bruit résiduel au niveau de la sortie *Seg_p*. Une étude préalable du niveau de bruit en sortie MagnoY est requise pour une segmentation optimale. Si l'on considère le niveau de bruit comme constant, le calcul de sa moyenne μ_b et de son écart type σ_b sur une série d'images statiques de la scène à analyser permettra d'ajuster δ_b. Dans ce travail, on fixe $\delta_b = \mu_b + 3\sigma_b$ ce qui revient à considérer le bruit comme gaussien et on se place alors à trois écarts types de la valeur moyenne de façon à limiter la segmentation du bruit.

Le gain *a* de la pondération sur la sortie du filtre *Seg_h* est fixée par défaut à *0.5*. Une valeur plus faible a pour effet une baisse de l'efficacité de l'adaptation locale. Cela tend à effectuer un seuillage à valeur fixe sur toute l'image d'où un risque de segmentation large autour des objets en mouvement fort et un risque de non-segmentation des objets à mouvement lent. A l'inverse, une valeur proche de *1.0* tend à produire un comportement tel que celui décrit sur la figure II.20 avec une sur-segmentation du bruit et sous-segmentation des objets texturés de grande taille (cf. objet 2 sur l'exemple de la figure II.19).

II.4.4. Performances

II.4.4.1. Qualité de la segmentation

La figure II.22 montre deux résultats de segmentation avec la méthode proposée pour un même paramétrage.

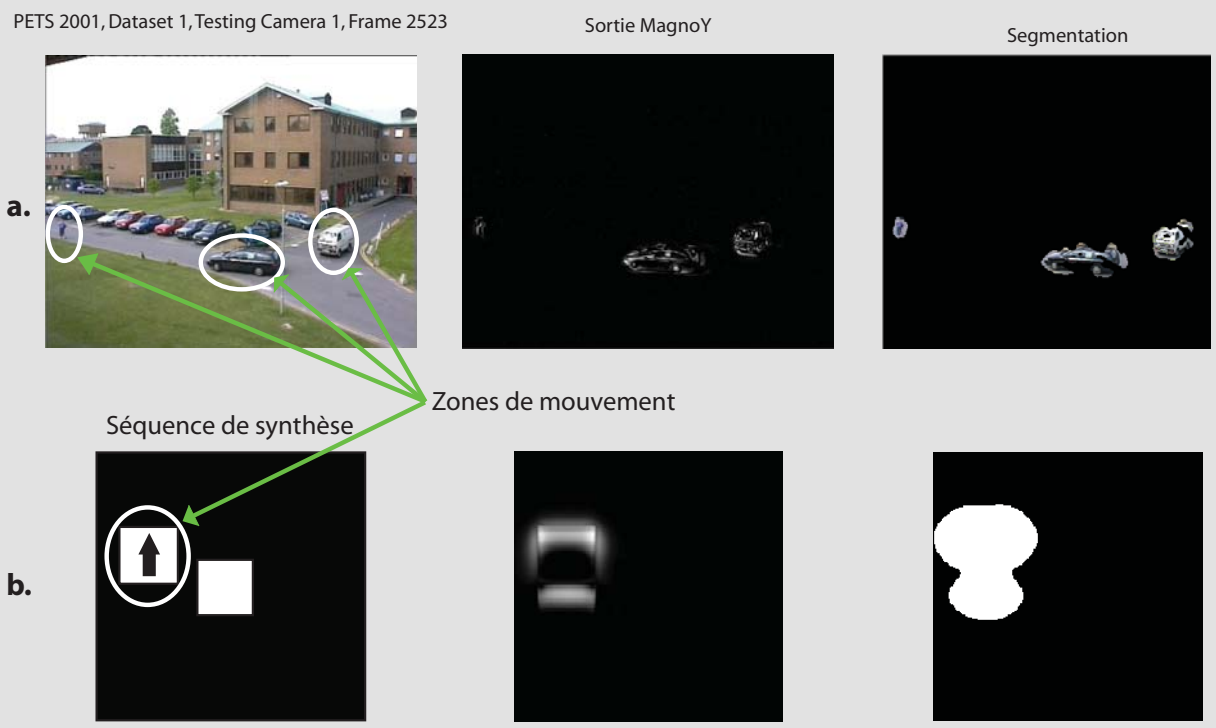


Figure II.22: exemple de segmentation avec la méthode proposée: a, segmentation sur une séquence vidéo d'extérieur PETS [PETSsite]. b, segmentation d'un objet à texture uniforme en mouvement.

Le premier exemple (cf. fig. II.22.a) montre une segmentation de zones de différente taille sur une séquence vidéo de jour (extrait de [PETSsite]). Ce premier exemple montre que la segmentation est efficace dans le cas d'objets texturés. Pour les objets à surfaces homogènes séparées par des arêtes (cf. fig. II.22.b), la segmentation de l'objet dans son ensemble est efficace, mais moins précise. Néanmoins, l'effet passe-bas temporel du filtre Parvo contours génère un effet mémoire qui permet de conserver un niveau d'énergie suffisant à l'intérieur de la surface des objets. Ceci est complété par l'effet passe-bas temporel des filtres *Seg_p* et *Seg_h*. Il est alors possible de segmenter une partie importante des surfaces homogènes en mouvement qui se trouvent en arrière des contours qui les entourent par rapport au mouvement. On constate néanmoins que la zone de segmentation est diffuse, ceci s'explique par la méthode employée. Comme nous l'avons vu dans la présentation des propriétés de la rétine, lorsque nous observons un objet en mouvement, nous le voyons "flou" du fait de la tendance passe-bas spatiale de la couche PLE de la rétine vis-à-vis du mouvement, la méthode de segmentation proposée ici donne le même type de comportement.

II.4.4.2. Comparaison

On trouve dans la littérature de nombreuses solutions aux problèmes de segmentation [Tsaig02, Zhang01, Freixenet02]. Nous proposons sur la figure II.23 une comparaison de notre méthode avec des méthodes classiques (FD: différence d'image et MFD: différence avec image temporelle moyenne) et d'autres plus récentes. Il est possible de formuler quelques remarques qualitatives. Comparé aux méthodes de segmentation présentées (FD et MFD), l'algorithme proposé ne segmente que les zones en mouvement, mais pas le bruit. Des opérations morphologiques de dilatation et d'érosion ne sont donc pas requises à postériori pour limiter les problèmes de sur-segmentation. D'autre part, du fait de sa segmentation diffuse, il donne des

zones de segmentation maximisant la surface des objets en mouvement, les objets sont donc segmentés en un faible nombre de zones de mouvement. Si l'on compare cette méthode à d'autres algorithmes tels qu'une méthode de segmentation hybride basée sur la combinaison d'opérations morphologiques couplées et d'opérations linéaires récursives [Richefeu04], on constate que les résultats sont relativement proches. Néanmoins, notre algorithme donne des zones de segmentation plus larges mais moins dispersées. Enfin, comparé à des méthodes basées sur des tenseurs 3D [Kühne01] ou des méthodes basées sur des contours actifs géodésiques [Collins04], notre algorithme montre des performances similaires.

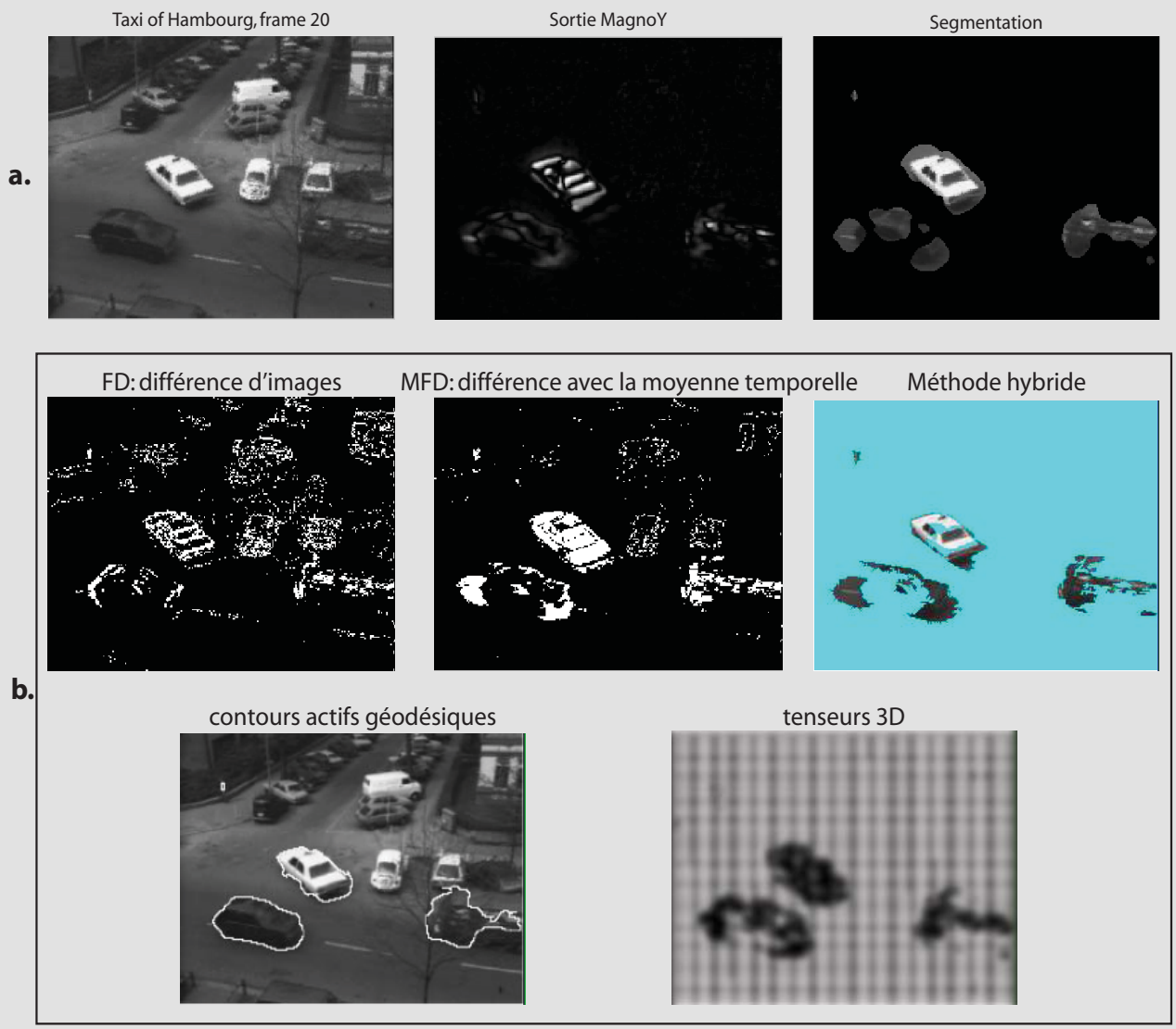


Figure II.23: comparaison de différentes méthodes de segmentation pour la scène de référence «Taxi of Hambourg». a, résultat de segmentation obtenu avec la méthode proposée. b, résultats des méthodes FD, MFD, de la méthode différentielle hybride de détection de mouvement [Richefeu04], de la méthode de segmentation par contours actifs géodésiques [Collins04] et de la méthode basée sur l'emploi de tenseurs 3D[Kühne01]

II.4.4.3. Coût de calcul

Du point de vue du coût de calcul, cet algorithme utilise deux filtrages passe-bas spatio-temporels et une méthode de segmentation basée sur la comparaison des sorties des deux filtrages moyennant une pondération de la seconde. Il en ressort un coût de calcul de l'ordre de 12 multiplications par pixel ce qui est faible. L'algorithme de segmentation dans son ensemble fonctionne à 30 images par seconde sur un ordinateur équipé d'un processeur de type Intel Pentium 4, 3.0Ghz avec un code C++ non optimisé.

II.4.4.4. Avantages et limites du système

Avantages

Cette méthode présente les atouts suivants:

→ Une segmentation générique: sans a priori sur l'objet à détecter, elle est susceptible de détecter tout objet en mouvement.

→ Une segmentation localement adaptée qui sélectionne les mouvements les plus importants dans une région donnée.

→ Une méthode qui utilise les propriétés de la rétine d'où un nombre d'opérations supplémentaires limitées et une correction de la scène visuelle qui rendent la détection optimale quelques soient les conditions (variation de luminance, filtrage du bruit, etc.).

→ Un algorithme peu coûteux en temps de calcul.

→ Cette méthode a aussi l'avantage d'être robuste au bruit et d'être autosuffisante. En effet, la segmentation sélectionne peu de zones de bruit, il n'y a donc pas besoin d'avoir recours à des post-traitements pour limiter les problèmes de sur-segmentations de petites zones introduites par le bruit. Aussi, cette méthode est temporellement stable, car elle profite des effets temporels des filtres de la rétine ce qui permet de retrouver des zones segmentées de forme très proches d'une image à la suivante dans les scènes vidéo.

Limites

Les zones de segmentation de mouvement différent peuvent se chevaucher, car on ne fait pas à ce niveau de différence entre des mouvements spatialement proches (on ne caractérise pas l'orientation de leur déplacement). Enfin, les zones de segmentation calculées sont diffusées sur les bords extérieurs de l'objet en mouvement, du fait de la réponse de la rétine aux contours en mouvement.

Cette méthode ne travaille que sur une analyse des niveaux de gris de la scène visuelle. Néanmoins, la couleur est également considérée comme une base d'information pour la segmentation et ouvre également la voie à d'autres méthodes [Dorea05]. Il sera donc intéressant d'enrichir la méthode proposée en lui ajoutant une analyse de la couleur.

Symbole associé

Nous associons à cet algorithme de segmentation de zones en mouvement le symbole présenté sur la figure II.24. Ce module prend en entrée la sortie du filtre MagnoY contours mobiles et donne en sortie les régions en mouvement sous forme d'une carte binaire.

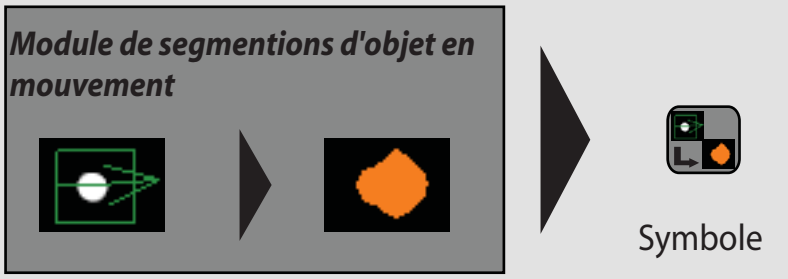


Figure II.24: symbole associé l'algorithme de segmentation de mouvement

II.5. Estimation de vitesse

II.5.1. Introduction

II.5.1.1. Présentation générale

Dans cette partie nous nous intéressons au problème de l'estimation du mouvement dans les scènes visuelles. L'analyse de vitesse est un domaine massivement exploré [Barron94]. Le mouvement en analyse vidéo est associé au "flot optique" qui représente la projection sur le plan image du champ de vecteurs vitesse 3D.

Pour cette problématique, on utilise un algorithme faisant également partie de la classe des algorithmes "bio-inspirés" qui font le lien entre la biologie du système visuel et la vision par ordinateur. Cette partie ne fait que décrire une méthode existante que nous allons utiliser pour développer des applications pour lesquelles l'information de vitesse est nécessaire. Cette partie ne présente donc pas de nouvelle contribution dans l'estimation de vitesse. Elle décrit l'outil d'estimation basé sur les circuits neuromorphiques développé dans [Torralba99]. Cet algorithme est basé sur une analyse du mouvement dans le domaine des fréquences.

II.5.1.2. Analyse de vitesse par approche fréquentielle

L'estimation de vitesse par analyse fréquentielle suppose classiquement que le mouvement étudié est constant et translationnel d'une image à la suivante. Partant de ce postulat, on observe que le spectre spatio-temporel d'un objet en mouvement est disposé sur un plan dont l'orientation dépend de la vitesse et de la direction du mouvement [Jepson90].

Dans le cas d'un objet en mouvement dans les scènes vidéo (dans le plan *(x, y)*), son spectre reste dans le plan (f_x, f_y) s'il est immobile ou se place dans le plan d'équation $f_t+v_x.f_x+v_y.f_y=0$ s'il est en translation à la vitesse (v_x, v_y). Ceci est illustré sur la figure II.25.

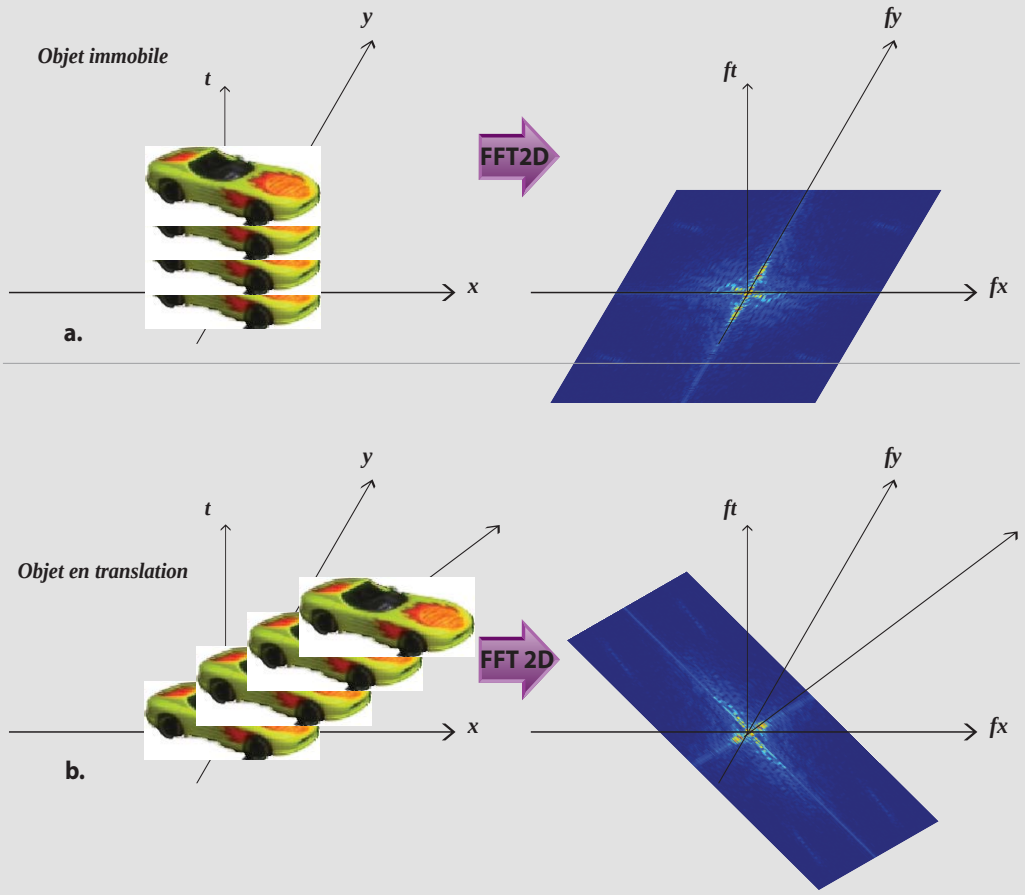


Figure II.25: particularité du spectre d'un objet variant temporellement a, objet immobile et son spectre. b, l'objet est en translation et son spectre s'oriente selon le plan $f_t+v_x.f_x+v_y.f_y=0$

Estimer le mouvement correspond alors à estimer l'orientation du plan du spectre dans le domaine spatio-temporel (f_x,f_y,f_t). Les méthodes couramment employées utilisent des bancs de filtres de Gabor [Adelson85, Heeger87, Watson85, Spinei98] (cf. fig. II.26.a). Chaque filtre est sensible à une gamme de fréquences spatiales et temporelles précise. L'espace spatio-temporel est pavé par toute une batterie de filtres et le plan de l'objet en mouvement est extrait par analyse de la réponse de chaque filtre. Ces méthodes sont connues pour leur grande robustesse face au bruit, elles ont néanmoins un coût de calcul très élevé, car elles ont recours à un nombre de filtres important. Une méthode fonctionnant sur le même principe, mais plus économique en terme de coût de calcul a été proposée par Torralba. Comme le montre la figure II.26.b, elle utilise des filtres large bande permettant de conserver les avantages de l'approche énergétique, mais présentant un coût de calcul bien plus intéressant, car utilisant beaucoup moins de filtres.

C'est cette approche que nous allons utiliser pour accéder à la mesure de vitesse. Nous décrirons donc rapidement cette méthode et illustrerons son comportement sur quelques exemples.

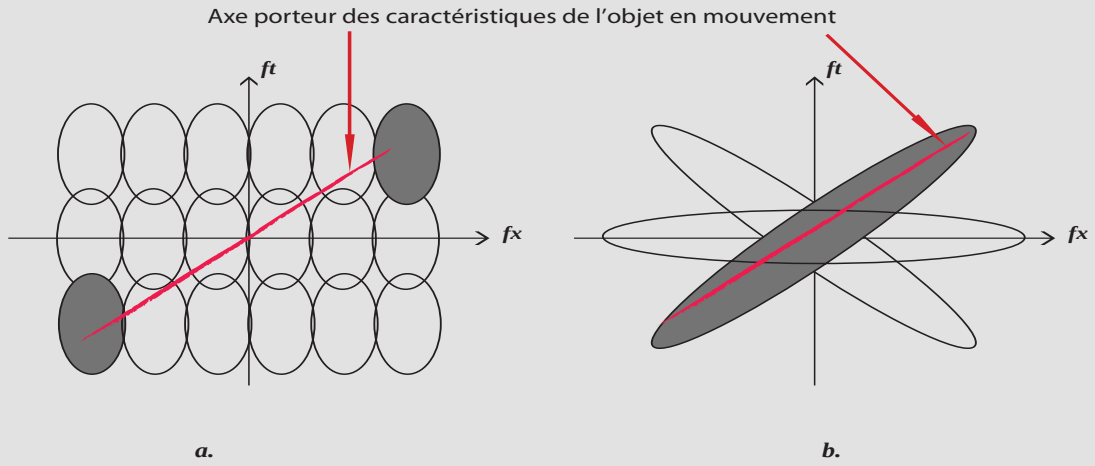


Figure II.26: estimation de la vitesse par utilisation de filtres placés dans le domaine spatio-temporel. a, utilisation de bancs de filtres de type Gabor. b, utilisation de filtres larges bande proposée par Torralba

II.5.2. Filtres de vitesse large bande monodimensionnels

[Torralba99] présente une méthode de synthèse de filtres basée sur une architecture «neuromorphique» (c.-à-d. inspirée de l'architecture de circuits neuronaux [Douglas95]) et dérivée de l'architecture des filtres passe-bas présentés au chapitre I. Cette méthode utilise une modélisation de réseaux neuronaux présentant des inhibitions et des excitations latérales permettant une certaine sélectivité au mouvement. Cette modélisation aboutit à la synthèse de filtres sensibles à la vitesse dont la fonction de transfert 1D est la suivante :

$$G(f_x, f_t) = \frac{1}{1 + 2.(1 - \cos(2.\pi.f_s)) + j.2.\pi.\tau.\left(f_t + \frac{(b-1)}{\pi.\tau}.\sin(2.\pi.f_s)\right)}$$

$$avec \left| \begin{array}{l} a + b = 2 \\ v_0 = (a-b)/\tau \end{array} \right.$$

(Eq. II.4)

Dans cette fonction de transfert, le paramètre v_0 représente la vitesse centrale à laquelle est sensible le filtre, et est ajusté par τ, la constante de temps du système et les paramètres a et b dont la somme doit être égale à 2.

La représentation d'une telle fonction de transfert mono dimensionnelle est donnée sur la figure II.27.

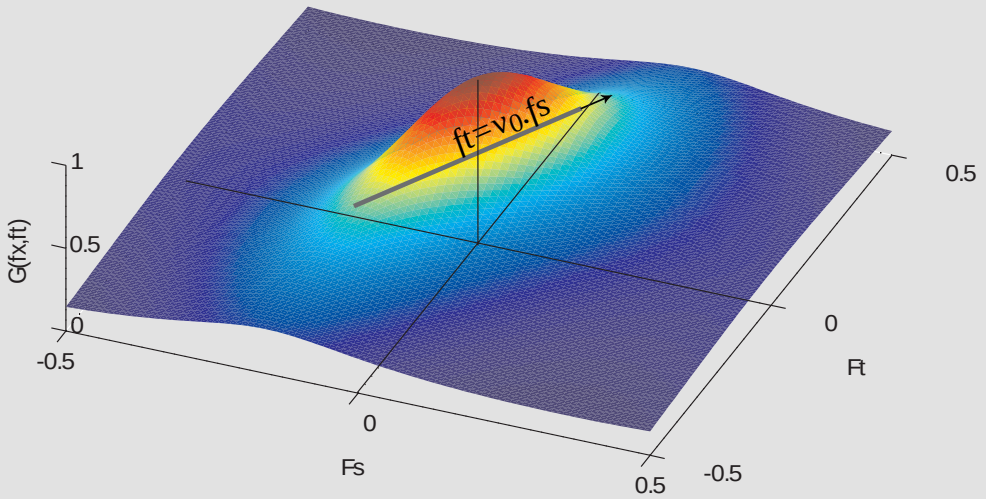


Figure II.27: fonction de transfert du filtre de vitesse unidimensionnel

Lorsqu'un motif dans l'image se déplace à la vitesse v_0, son énergie est située sur la droite d'équation $f_t+v_0.f_x=0$. Le filtre ajusté à la même vitesse donne une réponse maximale, qui est aussi dépendante de l'énergie du signal d'entrée. Notons que ce type de filtre suppose un spectre blanchi pour une maximisation de son efficacité.

II.5.3. Estimation de vitesse 2D

II.5.3.1. Modèle utilisé

L'idée de Torralba était de combiner ces filtres monodimensionels de façon à réaliser une analyse de vitesse en deux dimensions efficace. Nous utilisons son système d'analyse présenté sur la figure II.28.b. Ce système prend en entrée un flux d'images de contours. Il permet de calculer le mouvement en analysant indépendamment ses composantes verticale et horizontale. Pour chaque composante, le principe est d'utiliser 2 filtres sensibles à la même vitesse, mais de signe (direction) opposé. On utilise donc un premier filtre sensible à la vitesse v_0 notée v_{0+} et le second sensible à la vitesse $-v_0$ notée v_{0-}. Chacune des sorties des filtres de vitesse est mise au carré (calcul de l'énergie) et est lissée spatialement à l'aide d'un filtre passe-bas spatio-temporel tels ceux présentés au chapitre I de façon à homogénéiser la vitesse sur la zone de calcul et limiter les problèmes d'ouverture. Le calcul de la vitesse est ensuite effectué en comparant l'énergie de chaque filtre à l'aide d'un opérateur de type diviseur de tension ce qui correspond à un phénomène d'inhibition [Nabet92]. On obtient finalement les deux composantes v_x et v_y du vecteur vitesse pour chaque pixel de l'image traitée.

La contrainte est que le spectre des contours appliqué en entrée du système soit blanchi et que leur énergie soit maximale dans la direction du mouvement. Afin de respecter cette contrainte, nous appliquons en entrée de ce système la sortie du filtre MagnoY contours mobiles précédé du filtre Parvo contours (cf. fig. II.28.a).

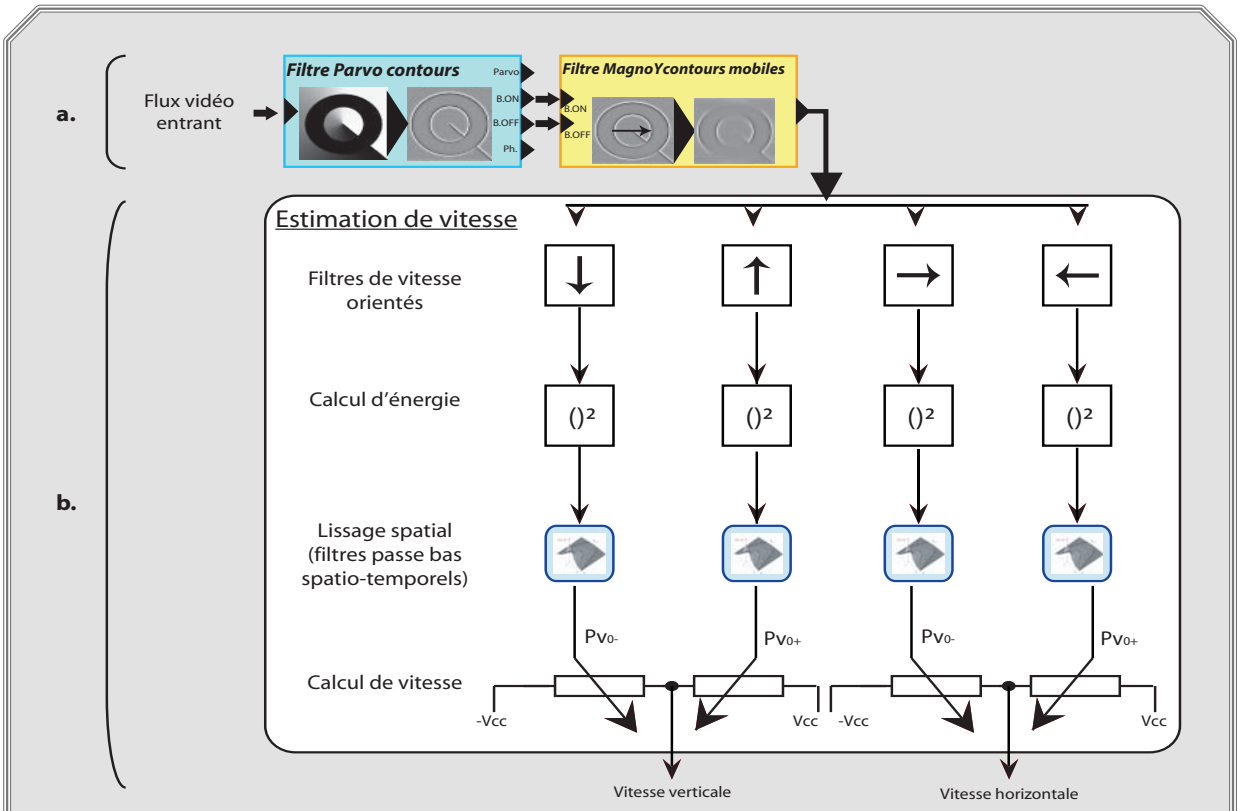


Figure II.28: chaîne complète du calcul de flot optique

II.5.3.2. Performances

Précision

La figure II.29 montre des résultats de calcul de flot optique sur une séquence de synthèse. On y trouve 4 balles se déplaçant horizontalement vers la droite, chacune 30% plus vite que son homologue placée au dessus. Chacun des objets a été segmenté manuellement de façon à en sélectionner précisément la surface. On observe (cf. fig. II.29.b) l'effet du filtrage passe-bas spatial qui étale la réponse de chaque pixel sur toute la surface d'intégration. Le vecteur vitesse moyen de chaque objet (cf. fig. II.29.a) montre leur orientation et vitesse. Cette vitesse estimée montre une évolution linéaire de la vitesse entre les objets (cf. fig. II.29.c), la précision est de 95%.

Coût de calcul

Les tests menés par Torralba montrent que les performances sont similaires aux autres méthodes énergétiques utilisant des ondelettes de Gabor. Néanmoins, le coût de calcul est nettement à l'avantage de cette méthode, car elle n'utilise que 4 filtres large bande (deux filtres verticaux opposés et deux autres filtres horizontaux opposés) contre un nombre beaucoup plus important pour les autres méthodes (jusqu'à 36 filtres dans [Heeger87]). Une description plus détaillée est disponible dans [Torralba97, Torralba99].

Cet algorithme exige 40 opérations par pixel auquel on ajoute le coût de calcul des filtres Parvo contours et MagnoY contours mobiles. Il fonctionne alors à 17 images par seconde pour des images de taille 320*240 pixels sur un ordinateur équipé d'un processeur de type Intel Pentium 4, 3.0Ghz avec un code C++/Matlab non optimisé.

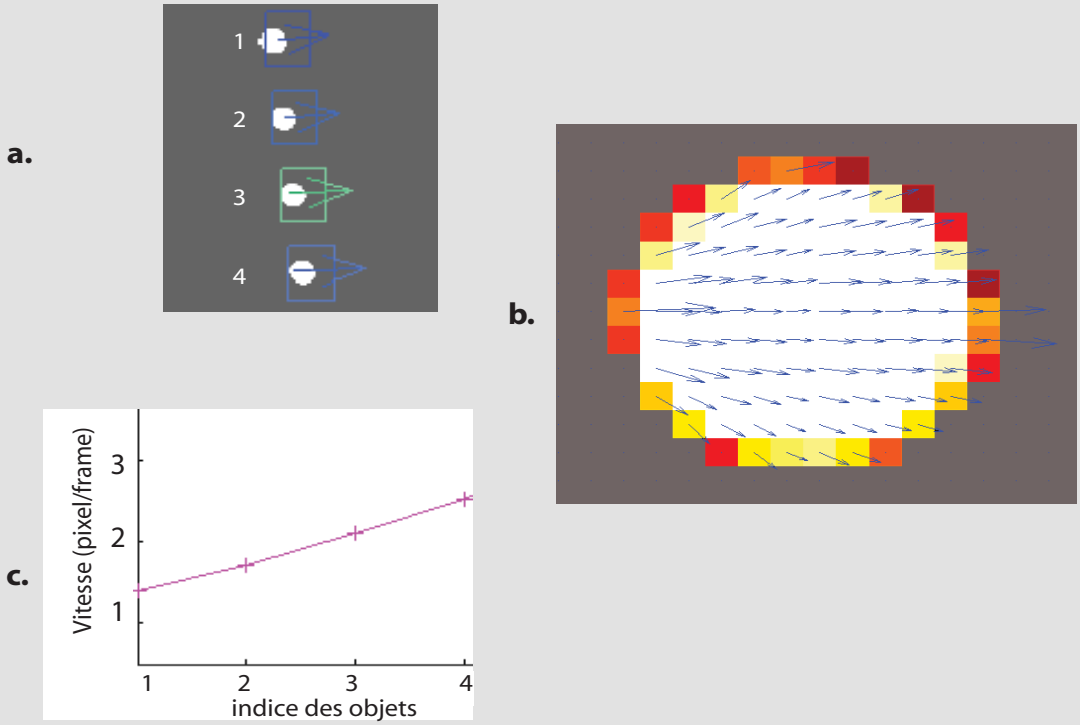


Figure II.29: exemple de calcul de flot optique avec la méthode proposée dans [Torralba99]. Dans cet exemple, 4 balles se déplacent vers la droite (a) avec une vitesse croissante (augmentation de +30% d'un objet par rapport à celui placé au dessus). Le flot optique calculé pour un objet (b) est calculé pour chaque pixel. Le calcul de la vitesse moyenne de chaque objet (c) montre une croissance linéaire de la vitesse.

Limites

Les limites de cet estimateur sont liées au problème d'ouverture. Ce problème bien connu illustre le fait qu'un estimateur de vitesse ne calcule pas correctement la direction du mouvement lorsqu'un contour n'est pas perpendiculaire à celui-ci. Ce problème est en général minimisé dans le cas d'objets texturés notamment grâce au filtrage passe-bas précédant l'estimation finale de la vitesse. Néanmoins, dans le cas de droite en translation, la contrainte de spectre blanchi n'est plus satisfaite et l'estimation est faussée (cf. fig II.30).

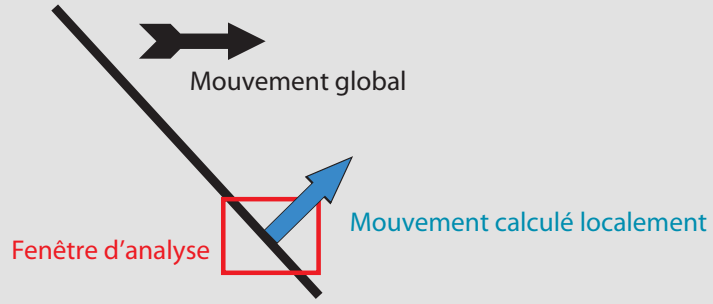


Figure II.30: illustration du problème d'ouverture: le mouvement local ne correspond pas au mouvement global de l'objet.

<u>Symbole associé à l'estimateur de flot optique</u>

Nous associons à cet estimateur de flot optique le symbole présenté sur la figure II.31. Il prend en entrée un flux d'images dont le spectre est blanchi (en l'occurrence dans notre cas la sortie du filtre MagnoY contours mobiles) ainsi que la zone de segmentation de chaque objet en mouvement, il en ressort la vitesse moyenne de chaque objet.

Nous décrivons ici rapidement les paramètres associés aux filtres de vitesse. Se référer à la thèse de Torralba pour une description plus fine des paramètres.

→ La constante d'espace de ces filtres (k^2) est fixée à 1 de façon à respecter $a+b=2$ [Torralba99].

→ La constante de temps est fixée à 1/25 cycle par seconde et le paramètre a (le même que celui énoncé précédemment) est fixé à 0.99. Ce qui permet au système d'être sensible aux grandes vitesses.

→ Les filtres passe-bas spatio-temporels placés en sortie des filtres de vitesse doivent intégrer les valeurs de vitesse de chaque pixel au sein d'une zone en mouvement. Ils permettent d'homogénéiser la vitesse sur tout l'objet. Si le but est d'estimer le vecteur vitesse moyen, alors, il est nécessaire d'effectuer un lissage très fort de l'information. Pour cela, nous fixons la constante d'espace à environ trois fois la taille de l'objet. Les constantes de temps sont en revanche nulles de façon à ne pas introduire d'effet temporel dans l'intégration.

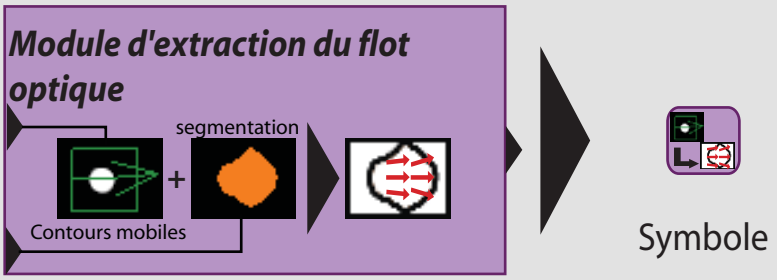
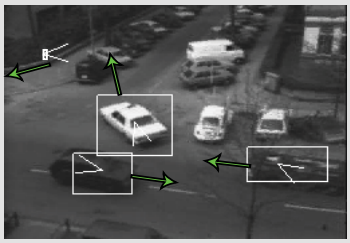
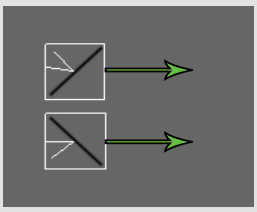


Figure II.31: symbole associé l'estimateur de flot optique

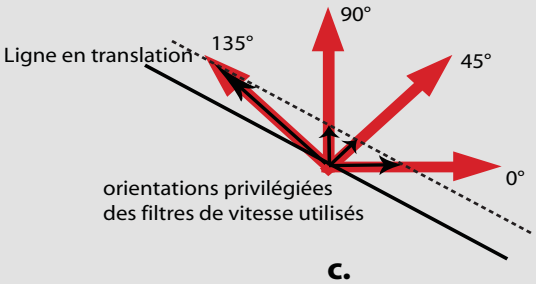
La figure II.32 montre quelques exemples de résultats d'analyse de vitesse avec ce module. On observe que le flot optique moyen obtenu sur chaque objet segmenté correspond au déplacement réel de l'objet dans la mesure ou l'objet analysé respecte les contraintes imposées par l'algorithme (cf. fig. II.32.a-b). Néanmoins, ce système reste sensible aux problèmes d'ouverture comme le montre la figure II.32.c. Une extension de cet algorithme est présentée dans [Torralba99], elle fait appel à un plus grand nombre de filtres orientés et permet de limiter les problèmes d'ouverture. Elle est basée sur l'analyse des orientations 45° et 135° en plus des orientations 90° et 180° déjà utilisées par le système proposé (cf. fig. II.32.d). Son coût de calcul est néanmoins plus élevé (deux fois plus important). L'utilisation de cette extension pourra être envisagée dans le cas où les problèmes d'ouverture deviennent trop importants ou bien lorsqu'une plus grande précision de l'estimation de vitesse est requise.



a.



b.



c.

Légende:

orientation réelle du mouvement :

Figure II.32: exemples de calcul de flot optique moyen sur des objets en mouvement : les boites englobantes des zones de segmentation des objets mobiles sont représentées en blanc et une flèche blanche indique l'orientation du flot optique estimé, la flèche verte indique la direction réelle.

a, calcul correct des vecteurs vitesse moyens sur la séquence "Taxi of Hamburg".

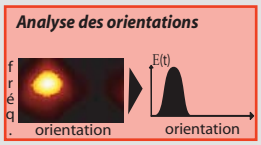
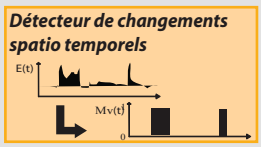
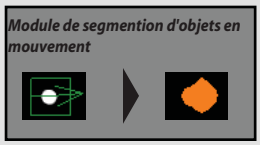
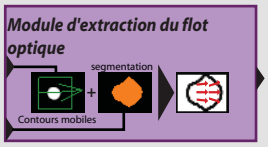
b, erreur d'estimation due à l'analyse d'objet dont le spectre n'est pas blanchi (problème d'ouverture).

c, extension du modèle étudié permettant de limiter les problèmes d'ouverture par utilisation de filtres supplémentaires sensibles aux orientations 45° et 135°. Les 4 composantes de vitesse estimées permettent de déterminer la direction exacte du mouvement [Torralba99].

III Conclusion

Nous venons de voir comment synthétiser un ensemble d'outils d'analyse de scènes vidéo à partir de la modélisation de la rétine et de l'aire corticale V1 du système visuel humain. Le tableau II.4 résume les informations relatives à ces outils.

Table II.4: synthèse des modules d'analyse d'image bio-inspirés

Nom	*Fonction*	*Symbole*	*Cadence de traitement (pour des images de taille 320*240 pixels)*	*Performances/Limites*
Filtre ParvoContours	*Extraction de tous les contours de la scène*		*60 images par seconde*	*générique et robuste au bruit*
Filtre MagnoY contours mobiles	*Extraction des contours mobiles*		*40 images par seconde avec le filtre Parvo contours en amont*	*générique et robuste au bruit*
Analyseur spectral des orientations	*Extraction de l'orientation dominante*		*50 images par seconde=coût de calcul du filtre amont: analyseur spectral*	*générique. Précision 12°*
Détecteur d'événements de mouvement	*Détection des instants pour lesquels une transition de mouvement est observée.*		*30 images par seconde=coût de calcul du filtre Parvo contours+ filtre MagnoY contours mobiles+analyseur spectral*	*générique et robuste*
Module de segmentation	*Définit les zones dans l'image dans lesquelles un mouvement est présent.*		*30 images par seconde avec la chaîne de filtres en amont (Parvo contours+MagnoY contours mobiles)*	*générique. Ne peut segmenter précisément les contours d'une zone en mouvement (segmentation diffuse)*
Estimateur de flot optique	*Estime la vitesse et la direction du mouvement d'une zone d'analyse.*		*17 images par seconde avec la chaîne de filtres en amont (Parvo contours+MagnoY contours mobiles)*	*Analyseur générique. Sensible aux problèmes d'ouverture.*

Tout cet ensemble est générique et fonctionne en temps réel sur les ordinateurs actuels (de type Intel Pentium IV, 3.0GHz ou équivalent) et ouvre de nombreuses perspectives d'applications. Dans les chapitres suivants, nous allons voir comment assembler ces différents éléments pour en dégager des outils d'analyse de plus haut niveau tel qu'un système d'analyse des mouvements du visage et un système de suivi d'objets en mouvement.

Partie II: Applications bio-inspirées, les modélisations du système visuel humain au service de la vision par ordinateur

Dans cette seconde partie, nous allons voir comment utiliser et particulariser les modules de traitement proposés dans la partie I dans le but de faire des traitements de plus haut niveau tels qu'une analyse des mouvements du visage, un suivi d'objets en mouvement et de la classification ou de l'identification d'objets.

Les applications présentées profitent toutes des avantages apportés par le filtre Parvo contours et, le cas échéant par le filtre MagnoY contours mobiles issus de la modélisation de la rétine. Les méthodes appliquées exploitent également l'analyse spectrale modélisant les traitements effectués au sein du cortex visuel primaire (aire V1). Notons néanmoins que pour chaque application considérée, la stratégie algorithmique proposée bien que proche du schéma biologique ne correspond pas à une modélisation de celui-ci. L'objectif de cette partie est de montrer qu'en s'appuyant sur les modèles issus de la rétine et de l'aire V1, il est possible de développer des algorithmes capables d'extraire un large panel d'informations et des interprétations haut niveau (le mouvement de forme (objets, traits du visage etc.), une description de leur état (oeil ouvert/fermé), etc.).

Chapitre III: Analyse du visage

III.1. Introduction

Dans ce chapitre, nous proposons des techniques pour analyser et interpréter les mouvements typiques de la tête et du visage. Ces mouvements peuvent être classés en deux catégories: d'une part les mouvements globaux ou rigides (mouvements de la tête dans son ensemble) et d'autre part les mouvements locaux ou non rigides du visage (mouvements des yeux et de la bouche).

Nous proposons tout d'abord une méthode de localisation des yeux (cf. III.2). Ensuite une méthode d'identification du type de mouvement global v.s. local sera décrite (cf. III.3). Nous exposerons par la suite des méthodes permettant de traiter les mouvements des yeux et de la bouche (cf. III.4-5) puis les mouvements globaux de la tête (cf. III.6). Enfin, les sections III.7 et III.8 présentent deux applications: un système de détection de signes d'hypovigilance chez un conducteur et un outil d'apprentissage de la langue des signes pour les malentendants.

Notre étude est basée exclusivement sur l'analyse des contours présents dans la scène analysée. Pour les extraire, nous appliquerons systématiquement l'ensemble des filtrages présentés sur la figure III.1.a. Le filtrage Parvo contours extrait les contours en limitant les problèmes de variation d'éclairage et de bruit. Le filtrage MagnoY contours mobiles rend possible l'extraction robuste des contours en mouvement. Nous nous focalisons alors sur la sortie des deux filtres Parvo contours ON-OFF et MagnoY contours mobiles associés à la zone du visage (cf. fig. III.1.b). Nous disposons donc de deux sources d'information:

→ L'information Parvo contours qui va servir à faire une analyse des structures du visage (localisation des yeux, état ouvert/fermé des yeux et de la bouche, etc.).

→ L'information MagnoY contours mobiles dont l'analyse va permettre la description des mouvements globaux (mouvements de tête) ou locaux (bâillements, parole, clignements, etc.).

La localisation de visage est réalisée à l'aide de la Toolbox MPT [MpisearchSite] ou son équivalent OpenCV [OpenCVSite] basés sur les travaux de Viola&Jones [Viola02]. Ces détecteurs utilisent une cascade de filtres chargés d'extraire des indices caractéristiques du visage (cf. fig. III.1.c). Cette cascade de filtres est optimisée par une phase d'apprentissage par méthode dite "boostée" [Bartlett03]. Notons que d'autres algorithmes présentant des performances très intéressantes peuvent également être utilisés pour la détection de visage. [Garcia05] propose une méthode hiérarchique comprenant trois étapes: une détection rapide du visage, une seconde étape de détection des caractéristiques plus fines du visage (les yeux, la bouche, le nez) et enfin, une analyse fine de ces traits (contours des yeux, de la bouche, etc.). Cette méthode est très précise, mais est plus coûteuse en temps de calcul que celle énoncée précédemment. Le choix du détecteur de visage s'est orienté vers un algorithme de type Viola&Jones pour sa rapidité d'exécution. Cette méthode est efficace lorsque le visage fait face à la caméra, les performances baissent de façon significative lorsque le visage s'incline de plus de 20° dans le plan de la caméra ou effectue une rotation 3D de plus de 25°. De plus, la localisation du visage reste grossière puisque d'une image à la suivante, la boite englobante donnée par le détecteur peut subir

un changement de taille de l'ordre de 20% et un décalage spatial de l'ordre de 10% de la taille du visage.

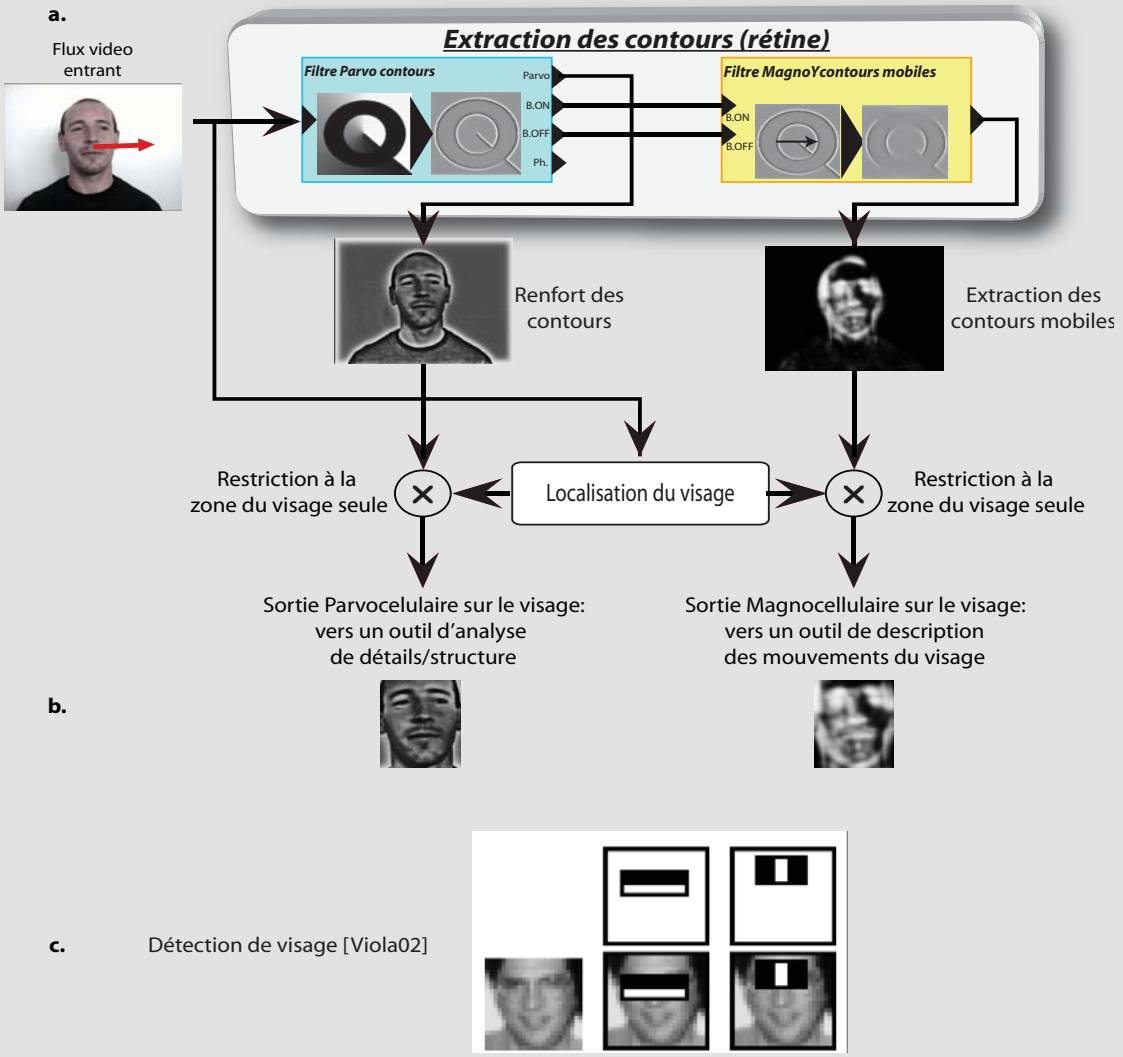


Figure III.1: Informations de base pour l'analyse d'un visage dans une scène vidéo (a): après localisation du visage, deux voies d'information sont extraites sur la zone du visage (b). (c) : méthode de détection du visage à l'aide de détecteurs boostés [Viola02]

III.2. Localisation des yeux

Ayant à disposition une boite englobante autour du visage, l'objectif est de détecter automatiquement une boite englobante autour de chaque oeil. Ces boites délimiteront des zones d'analyse pour déterminer l'état ouvert ou fermé ainsi que les mouvements de chaque oeil.

On fait l'hypothèse que chaque oeil se trouve dans un quart supérieur du visage (droit ou gauche) et on recherche chacun d'eux de façon indépendante (cf. fig. III.2). La conséquence immédiate de cette hypothèse est que le visage doit être vertical dans le plan de la caméra. Le système peut tolérer des rotations jusqu'à 45° maximum dans le plan de la caméra.

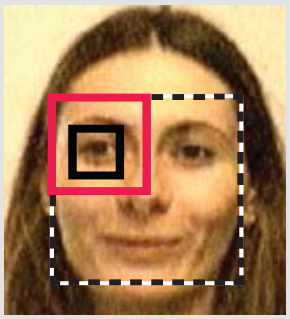


Figure III.2: boite englobante du visage (en pointillés), zone de recherche de chaque oeil (en rouge, un des quarts supérieurs du visage) et boite d'analyse de l'oeil (en noir) de taille 1/8 de la taille de la boite englobant le visage

III.2.2. Principe

Dans la zone de recherche prédéfinie pour chaque oeil, l'objectif est de centrer une boite englobante carrée de taille 1/16 de la taille du visage (cf. fig. III.2). Conformément à ce qui a été présenté dans [Hammal06], on suppose que la zone englobant chaque oeil est 16 fois plus petite que celle englobant le visage.

En supposant que l'oeil est ouvert, la méthode de localisation de l'oeil consiste à faire coïncider le carré englobant l'oeil avec la zone de la fenêtre de recherche qui contient le plus d'orientations horizontales et verticales. En effet, dans le cas d'un visage vertical, les caractéristiques des contours présents dans chaque quart supérieur du visage sont les suivantes:

→ Les sourcils présentent des contours principalement horizontaux.

→ La frontière peau/cheveux proche des oreilles est principalement verticale et la frontière front/cheveux est horizontale.

→ Les yeux ouverts présentent des contours à la fois horizontaux (contours des paupières, etc.) et verticaux (frontière iris/blanc de l'oeil latérale).

Les yeux se différencient donc du reste dans la zone de recherche par leur richesse en orientation de contours. Notre objectif est alors de trouver le point de la zone de recherche pour lequel les contours verticaux et horizontaux donnent l'énergie la plus forte. Pour ce faire, on utilise l'information Parvo ON-OFF du filtre Parvo contours. De cette information de contours, on calcule indépendamment sur chaque zone de recherche les moyennes spatiales horizontales et verticales de façon à obtenir la répartition de l'énergie des contours horizontaux et verticaux. Ce calcul des moyennes spatiales mono-dimensionnelles est réalisé à l'aide de deux filtres monodimentionels dont la fonction de transfert est la suivante (chaque filtre travaillant selon une direction *i* donnée) :

$$G(z_{ki}) = \frac{(1-a)^2}{\sqrt{1+\beta}} \cdot \frac{1}{1+az_{k_1}^{-1}} \cdot \frac{1}{1+az_{k_1}} \qquad \textit{(Eq. III.1)}$$

Il s'agit du même filtre que celui présenté au chapitre I lors de la modélisation de la PLE mais en une seule dimension (cf. I.3.2.4). On obtient deux images de la répartition de l'énergie des contours, l'une pour

les contours verticaux (cf. fig III.3.b), l'autre pour les contours horizontaux (cf. fig III.3.c). Comme on cherche le point donnant l'énergie maximale pour l'ensemble des deux directions, on multiplie les deux énergies moyennes (cf. fig III.3.d). Le pixel d'amplitude maximale est alors le centre recherché, ce centre coïncidant avec le centre de l'oeil (cf. fig III.3.e).

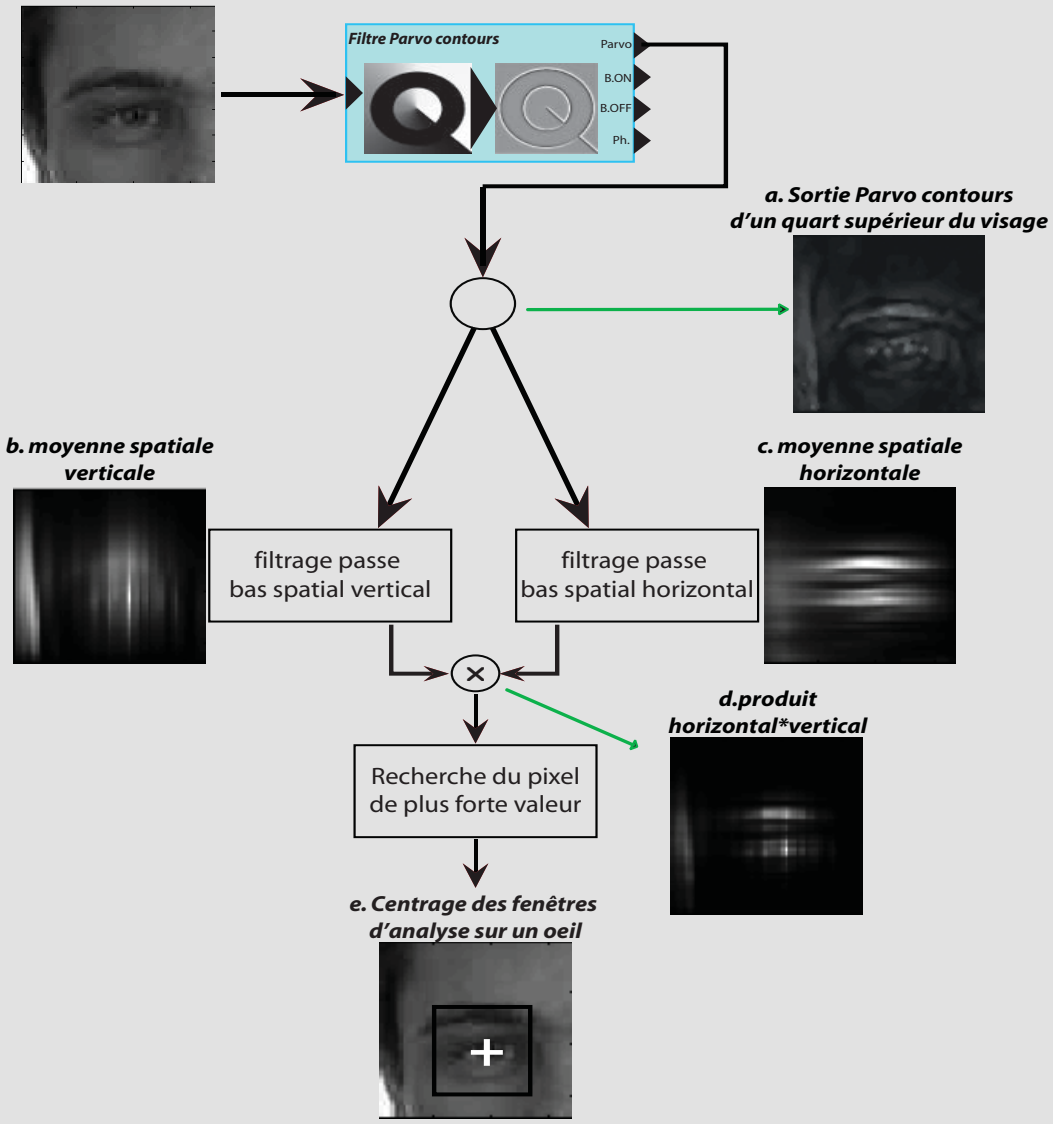


***Figure III.3:** schéma de détection de la position d'un oeil. a, la donnée initiale est la sortie du filtre Parvo contours sur le quart supérieur du visage. b et c, résultats des filtrages horizontal et vertical. d, produit des filtrages mono directionnels, le pixel donnant la réponse maximum représente le centre de la zone dans laquelle se trouve l'oeil. e, centrage des fenêtres d'analyse sur chacun des yeux.*

III.2.3. Résultats

Cet algorithme a été testé sur trois bases de visage dont la position du centre de chaque oeil est disponible suite à un étiquetage manuel. Nous avons utilisé la base Feret [FeretSite] (2370 visages) et la base BioID [BioIdSite] (1071 visages). Ces deux bases proposent une série de visages en vue de face pris sous diverses

conditions d'éclairage. Nous ajoutons également une troisième base de visages acquise au laboratoire. Elle contient 220 images de visage sans lunettes de plus basse résolution (160*120) de façon à tester l'efficacité de l'algorithme pour des tailles d'iris faibles. Ces tests ont tous été effectués en prenant comme constantes d'espace pour les filtres mono-dimensionnels, une valeur égale à la taille de la boite englobante de l'oeil (hauteur et largeur fixées à 1/4 de la hauteur et de la largeur de la boite englobante du visage) de façon à effectuer un lissage de l'information adapté à la zone de recherche. La figure III.4 présente différents résultats.

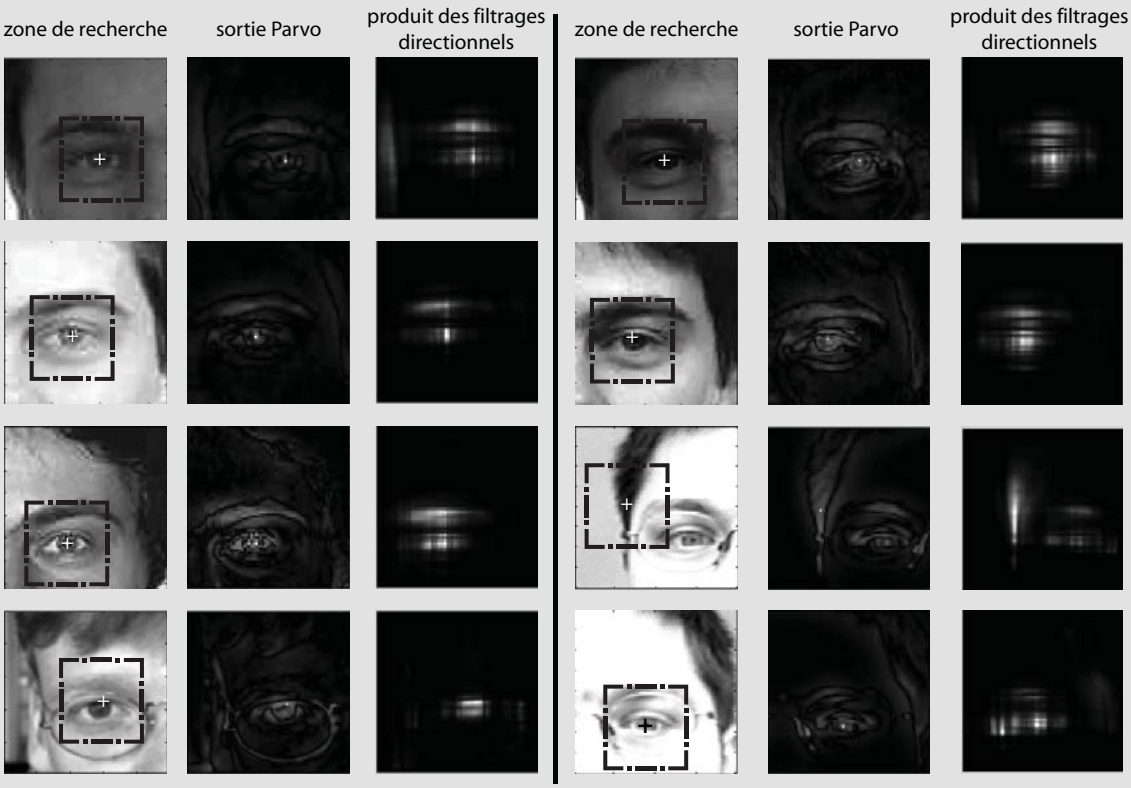


Figure III.4: détection de la position de l'oeil. Sont représentés les fenêtres de recherche, la sortie du filtre Parvo contours correspondante ainsi que le résultat du produit des deux filtrages directionnels à partir duquel la position de l'oeil est estimée. On représente sur l'image originale la croix indiquant la position détectée ainsi que la boite englobante associée.

On remarque que la position du maximum d'énergie sur le produit des deux filtres directionnels correspond bien à la position du centre de l'oeil. Par ailleurs, la zone donnant l'énergie des contours horizontaux et verticaux la plus forte est bien celle de l'oeil. La détection peut être faussée dans certains cas de prise de vue surexposée, particulièrement lorsque la personne porte des lunettes. La fausse détection est générée par une forte augmentation de la réponse des contours correspondant à la frontière peau/ cheveux ou sur certaines montures de lunettes. Ceci est dû au fait que cet algorithme ne recherche que le gradient maximum selon deux directions.

La table III.1 présente une synthèse quantitative des résultats en séparant le cas de personnes portant ou non des lunettes de vue. On définit le taux de succès comme le pourcentage d'images pour lesquelles la position trouvée est comprise dans la zone réelle de l'oeil (c.-à-d. dans l'iris ou le blanc de l'oeil). On relève également la moyenne et l'écart type de la distance en pixels qui sépare la position de l'oeil établie en vérité

terrain et celle trouvée par l'algorithme.

Table III.1: performance de l'algorithme de détection des yeux

Base de test	*Taille moyenne de l'iris*	*Nombre d'images*	*Taux de succès*	*Ecart moyen (pixels)*	*Ecart type (pixels)*
Feret A et B sans lunettes	10 pixels	2370	95%	3.95	4.02
Feret + lunette		161	83%	4.49	4.89
BioID sans lunettes	8 pixels	1071	94%	3.42	3.71
BioID+lunettes		450	81%	3.56	3.88
Base de test propre	4 pixels	220	90%	3.8	3.1

Grâce aux propriétés du filtre Parvo contours, cet algorithme n'est pas influencé par les problèmes de variation d'éclairement et réalise la détection de l'oeil efficacement avec un jeu de paramètres constant. Les performances sont satisfaisantes dans le cas de personnes sans lunettes sur différentes bases avec des tailles de visage différentes. Néanmoins, la précision baisse en cas de port de lunettes. Ceci est principalement dû à l'atténuation de la réponse des contours de l'oeil du fait du verre.

Comparée à la méthode décrite dans [Hammal06], notre algorithme donne une précision légèrement inférieure pour la localisation de l'iris (précision de l'ordre de 4 pixels comparés à 2 pixels pour la méthode [Hammal06]) elle est néanmoins plus rapide et plus tolérante vis-à-vis de la qualité des images analysées (fonctionne à résolution plus faible et est moins sensible au bruit d'acquisition). Le diamètre minimum de l'iris pour maintenir un niveau de performances satisfaisant est de l'ordre de 4 pixels contrairement aux 7 pixels annoncés pour la méthode [Hammal06]. En dessous de cette valeur, les contours horizontaux et verticaux créés par les contours de l'iris en particulier donnent une énergie trop faible ce qui entraîne une baisse significative des performances.

Le coût de calcul supplémentaire de la méthode proposée est très faible : elle requiert seulement 8 opérations par pixel sur un quart de visage. Si l'on considère maintenant le système complet sans la détection de visage, la détection d'un oeil peut être effectuée à environ 400 images par seconde pour des images de taille 100*100 pixels sur un ordinateur équipé d'un processeur de type Intel Pentium 4 fonctionnant à 3.0Ghz avec un code C/Matlab non optimisé.

III.3 Mouvements de tête globaux versus mouvements locaux

III.3.1. Principe

Nous nous intéressons à deux types de mouvements faciaux. Les mouvements globaux de la tête (c.-à-d. les mouvements du crâne réalisés au niveau des cervicales) d'une part, et les mouvements locaux internes au visage qui correspondent aux mouvements des yeux et de la bouche d'autre part.

L'objectif est de détecter dans un premier temps la présence d'un mouvement au niveau du visage et

ensuite de déterminer si ce mouvement correspond à un mouvement de tête (global) ou à un mouvement facial (local). Pour cette analyse, l'information prise en compte est l'énergie à la sortie du filtre MagnoY contours mobiles. Cette information est utilisée pour détecter les événements de mouvement à l'aide du module de détection d'événements spatio-temporels. Ensuite, la répartition spatiale sur le visage de cette énergie est utilisée pour classer le type de mouvement. Le principe de la méthode proposée repose sur le fait qu'il existe un lien entre la répartition spatiale de l'énergie en sortie du filtre MagnoY contours mobiles et le type de mouvement recherché. En effet :

→ Si l'énergie est faible sur la zone du visage, alors il n'y a pas de mouvement.

→ Si l'énergie est localisée principalement sur la zone des yeux, alors il y a un mouvement des yeux.

→ Si l'énergie est localisée principalement sur la zone de la bouche, alors il y a un mouvement de la bouche.

→ Si l'énergie est localisée sur la zone des yeux ***et*** la zone de la bouche, alors il y a un mouvement des yeux et de la bouche.

→ Si l'énergie est répartie sur toute la surface du visage, alors il y a un mouvement de tête.

La figure III.5.a présente l'architecture complète de l'algorithme proposé et la figure III.5.b montre un exemple de cette répartition d'énergie pour une séquence vidéo donnée.

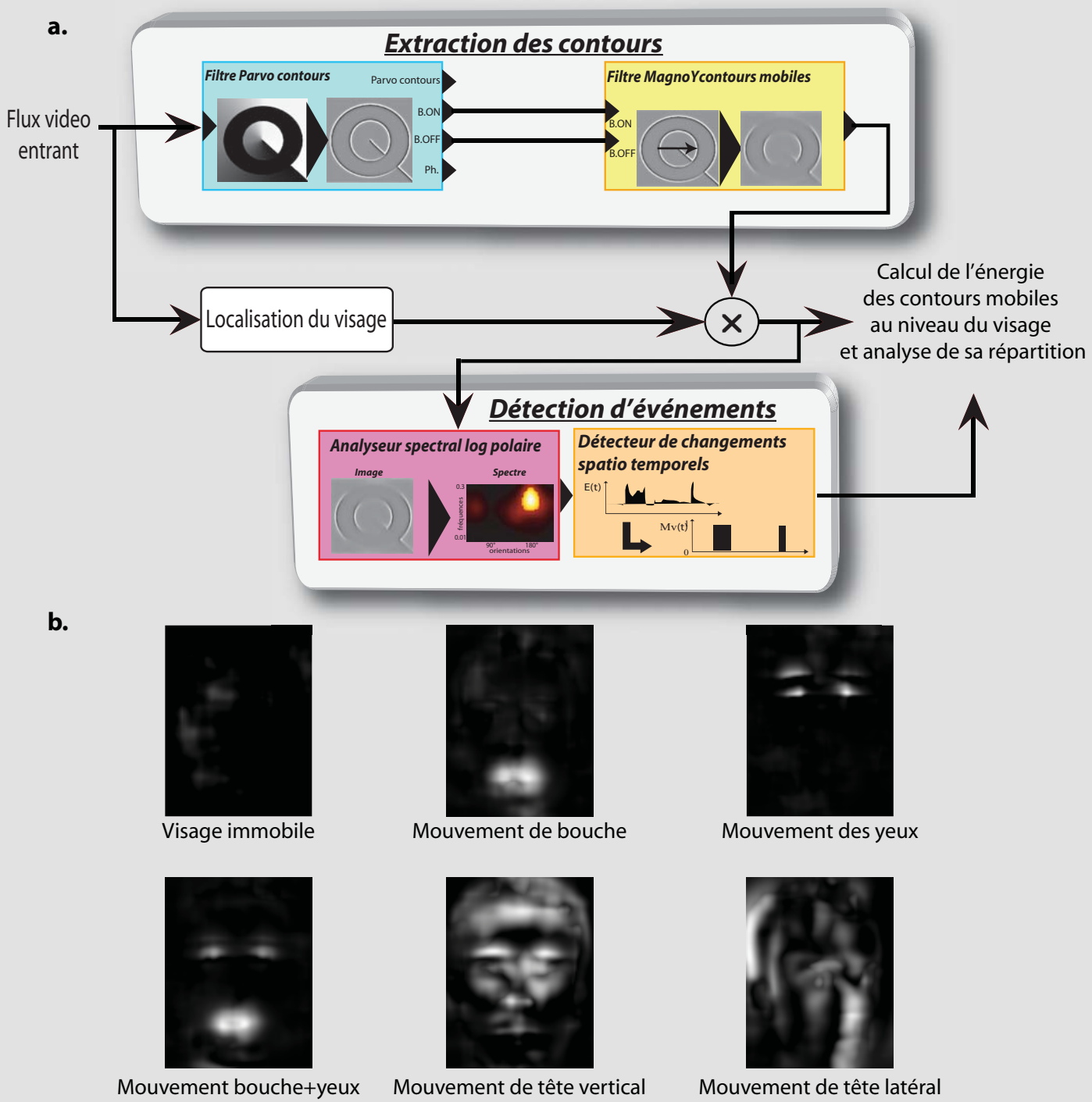


Figure III.5: classification des mouvements de tête (a) et (b) illustration des différentes répartitions de l'énergie en sortie du filtre MagnoY dans la zone du visage selon le type de mouvement. Sur chaque image, les pixels blancs correspondent à une zone de forte énergie, les pixels noirs à une énergie nulle.

On constate qu'en absence de mouvement, l'énergie est bien minimale. Au contraire, un mouvement d'oeil se traduit par une augmentation de l'énergie dans la zone des yeux et des sourcils, de même pour un mouvement de la bouche. A l'opposé, des mouvements de tête entraînent une augmentation de l'énergie sur toute la surface du visage, quel que soit le type de mouvement.

III.3.2. Classification du type de mouvement

Pour identifier le type de mouvement présent, on définit tout d'abord les zones d'analyse suivantes par rapport au cadre englobant le visage (cf. fig. III.6):

→ A_{haut_vis} et A_{bas_vis}, les deux moitiés hautes et basses du visage dont les énergies totales respectives sont notées E_{Ahaut_vis} et E_{Abas_vis}.

→ A_{yeux}, une zone autour des yeux: il s'agit du rectangle regroupant les boites englobant chaque oeil (cf. III.2) auquel on ajoute 1/8 de la hauteur du visage par rapport au haut de la boite englobant l'oeil le plus haut de façon à bien englober également les sourcils. Cette zone rassemble donc toute l'information de contours en mouvement liée aux yeux et aux sourcils, son énergie totale est notée E_{Ayeux}.

→ A_{bouche}, une région autour de la bouche définie comme la moitié centrée de la moitié inférieure du visage, son énergie est notée $E_{Abouche}$.

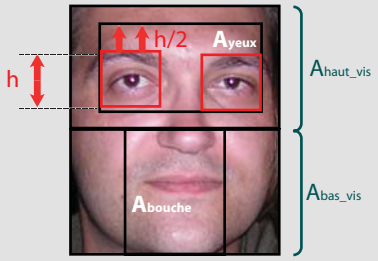


Figure III.6: définition des zones d'analyse permettant la classification des mouvements détectés sur le visage: la zone autour des yeux (Ayeux) en noir est définie à l'aide des zones d'analyse autour des yeux en pointillés

A partir de ces différentes zones et de l'énergie qu'elles contiennent, nous élaborons l'algorithme suivant:

```
Si un mouvement est détecté
alors{
        Condition_1= (E_Ayeux/E_Ahaut_vis)>x; // le mouvement dominant est dans la zone des yeux
        Condition_2= (E_Abouche/ E_Abas_vis)>x; // le mouvement dominant est dans la zone de la bouche

        Si Condition_1 ET Condition_2 FAUSSES // vérifie si l'énergie est diffuse sur tout le visage
        alors{
                mouvement de tête
        }sinon{
                Si Condition_1 VRAI
                        Alors mouvement d'oeil

                Si Condition_2 VRAI
                        Alors mouvement de bouche
                }

}Sinon
        Pas de mouvement
```

Cet algorithme vérifie tout d'abord qu'un mouvement est présent sur la zone du visage. Pour ce faire, on utilise le module de détection d'événements (cf. II.3). Ensuite, on vérifie si dans chaque moitié de visage le mouvement est localisé ou réparti. S'il est réparti dans les deux sous-parties du visage, alors on interprète le mouvement comme un mouvement global de la tête, sinon, il s'agit d'un mouvement de bouche et/ou des yeux. *x* représente un seuil fixé expérimentalement à 30%.

III.3.3. Performances

Les performances de ce système ont été évaluées sur une base de 52 minutes de vidéos réalisées avec 10 personnes différentes. Il leur était demandé d'exécuter de manière libre des mouvements de tête et/ou des mouvements faciaux. Les résultats ont été comparés à une vérité terrain établie manuellement, ils sont présentés sur la table III.2.

Table III.2: performances de l'identification des mouvements du visage

	Taux de succès	*Taux de fausse alarme*	*Taux d'oubli*
mouvement global	***93%***	***6%***	***1%***
mouvement de bouche seul	***92%***	***3%***	***5%***
mouvement des yeux seuls	***97%***	***2%***	***1%***
mouvement des yeux et de la bouche seuls	***93%***	***4%***	***3%***

Le taux de succès est globalement satisfaisant, on note un taux d'oubli faible. La présence de fausses alarmes s'explique par des cas de confusion entre différents mouvements tels que bâillement et mouvements de tête. En effet, lors d'un bâillement, le mouvement de bouche peut être réalisé simultanément avec un mouvement de tête qui devient dominant. La question de la prise de décision entre un mouvement de tête ou un mouvement de bouche est alors plus difficile. Il serait en fait intéressant de pouvoir traiter les mouvements non rigides (bouche et yeux) en même temps que les mouvements de tête ce qui n'est pas possible avec l'algorithme proposé.

III.4 Détection de l'état ouvert ou fermé de la bouche et des yeux

Nous proposons une méthode qui permet de traiter indifféremment le cas des yeux ou de la bouche. Cette approche repose sur une analyse de l'évolution temporelle de l'énergie des contours dans leur ensemble et des contours mobiles dans la boite englobante centrée sur la bouche ou sur un oeil.

III.4.1. Principe

Nous partons du fait que la quantité de contours dans la boite englobante (d'un oeil ou de la bouche) est moins importante lorsque l'oeil (ou la bouche) est fermé(e) que lorsqu'il (elle) est ouvert(e). Par conséquent, l'énergie totale des contours E_t est plus forte dans le cas ouvert que dans le cas fermé. Afin de différencier un oeil (ou une bouche) ouvert(e) d'un oeil (ou d'une bouche) fermé(e), on s'appuie sur l'étude de l'information d'énergie du filtre Parvo contours d'où l'architecture de la figure III.7.a. La figure III.7.b montre l'évolution de l'énergie des contours en fonction de l'ouverture.

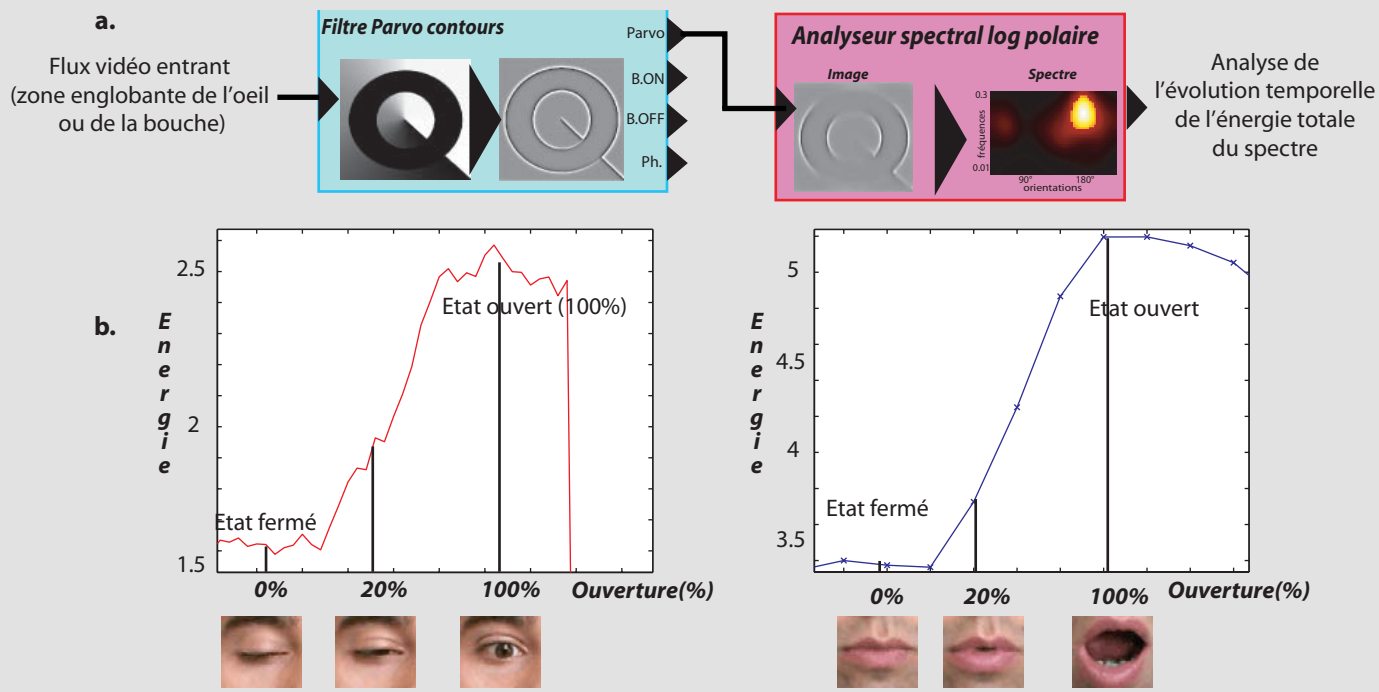


Figure III.7: a, système d'analyse de l'état ouvert ou fermé. b, évolution de l'énergie totale du spectre en sortie de filtre Parvo contours pour une fenêtre d'analyse centrée sur un oeil ou sur la bouche

On constate que l'énergie augmente avec le degré d'ouverture et que les niveaux d'énergie lorsque la bouche (ou l'oeil) est fermé(e) ou ouvert(e) sont très différents. Le tableau III.3 précise la différence d'énergie entre l'état ouvert ou fermé d'un oeil ou de la bouche dans une situation de mesure donnée. L'écart énergétique entre les deux états est suffisamment important pour envisager une reconnaissance de l'état sur la base de cette information.

Table III.3: relation entre niveau d'énergie dans les zones d'analyse et état d'ouverture

	Zone d'analyse	Energie totale en sortie de PLE	Etat binaire
Oeil ouvert		4.7	Energie Haute
Oeil fermé		2.8	Energie Basse
Bouche ouverte		1.9	Energie Haute
Bouche fermée		1.1	Energie Basse

Dans le cas de l'oeil, le bord de l'oeil, le blanc de l'oeil (la sclérotique) et l'iris apparaissent lorsqu'il est ouvert. Ils forment des contours associés à une transition blanc/sombre très marquée et donc très riches en énergie sur une large gamme de fréquences. Plus l'oeil est ouvert, plus on distingue de contours (les transitions blanches de l'oeil/iris/contour de l'oeil) et donc, plus l'énergie est élevée. Par ailleurs, la position de l'iris (liée à la direction du regard) pour un niveau d'ouverture donné ne modifie pas de façon significative le niveau d'énergie, les variations étant inférieures à 20%.

Dans le cas de la bouche, lorsque celle-ci est fermée, la jonction entre les deux lèvres forme un contour proche d'une droite horizontale donnant une énergie propre à l'état «Fermé». Lors de l'ouverture de la bouche, ce contour évolue en se dédoublant et prend une forme différente, plus ovale, ronde et la longueur de ce contour est bien plus importante que lorsque la bouche est fermée. L'énergie est donc plus importante. Néanmoins, deux cas de figure peuvent se présenter suivant les conditions d'acquisition:

→ Si l'éclairage est suffisant et permet d'obtenir des détails sur l'intérieur de la bouche, les dents et autres détails peuvent apparaître. La conséquence directe est une nette augmentation du nombre de contours et donc de l'énergie en sortie du filtre Parvo contours.

→ Si l'éclairage ne permet de voir qu'une zone sombre à l'intérieur de la bouche, cette zone sombre entre les deux lèvres forme un contour de longueur importante (le périmètre du contour délimité par les contours intérieurs des deux lèvres). Ce contour est formé d'une transition de luminance claire des lèvres à un noir marqué à l'intérieur de la bouche. Il donne alors également une énergie nettement supérieure à celle obtenue dans le cas d'une bouche fermée.

Pour la suite, on associe à l'énergie de l'état ouvert la notation E_{HO} et à l'énergie de l'état fermé la notation E_{BF}.

Afin de déterminer à chaque instant l'état de la bouche ou des yeux, il est nécessaire de pouvoir créer un système capable de détecter ces niveaux et de les adapter temporellement si nécessaire. Comme le montre la figure III.7.b, l'état "ouvert" est valable pour une gamme de degrés d'ouverture large (de 20% à 100%) c.-à-d. de faiblement ouvert à grandement ouvert. Au contraire, l'état fermé reste dans un intervalle beaucoup plus restreint, les fluctuations de l'énergie étant dues au bruit résiduel. Enfin, selon les conditions d'éclairage, selon la résolution ou encore selon les personnes, les niveaux absolus d'énergie E_{HO} et E_{BF} sont d'amplitude différente, mais conservent le même comportement relatif.

III.4.2. Méthode de détection

Nous venons de mettre en évidence que les états ouverts et fermés se distinguent par une énergie de la sortie Parvo contours très différente (haute pour l'état ouvert et basse pour l'état fermé). Néanmoins, la valeur absolue des deux énergies E_{HO} et E_{BF} dépend des individus, du degré d'ouverture et des conditions d'acquisition. D'où la nécessité d'initialiser ces deux valeurs et de les mettre à jour au cours du temps.

III.4.2.1. Initialisation des niveaux de référence

Afin d'estimer les niveaux E_{HO} et E_{BF}, nous utilisons l'information de mouvement de la bouche ou de l'oeil. En effet, si l'on détecte que la bouche s'ouvre, alors on sait qu'avant elle était fermée et qu'elle est maintenant ouverte. Le niveau E_{BF} correspond alors à la valeur d'énergie en sortie du filtre Parvo contours avant le début du mouvement et le niveau E_{HO} correspond à l'énergie en sortie de ce même filtre tant que le

sens du mouvement reste le même.

Initialiser les niveaux de référence revient donc à détecter les mouvements d'ouverture et de fermeture d'où la présence du module de détection des événements de mouvement sur la figure III.8.a. Ce module fournit les alertes de mouvement. Il est ensuite couplé au module d'analyse des orientations afin d'estimer la direction du mouvement dominant. En effet, ces ouvertures/fermetures de la bouche et des yeux sont associées à un mouvement vertical. Ainsi, il est possible de différencier les mouvements d'ouverture/fermeture verticaux des mouvements autres, tels que le changement de direction du regard (cf. fig. III.8.b-c). Lorsqu'un mouvement vertical est détecté, les niveaux E_{HO} et E_{BF} sont initialisés.

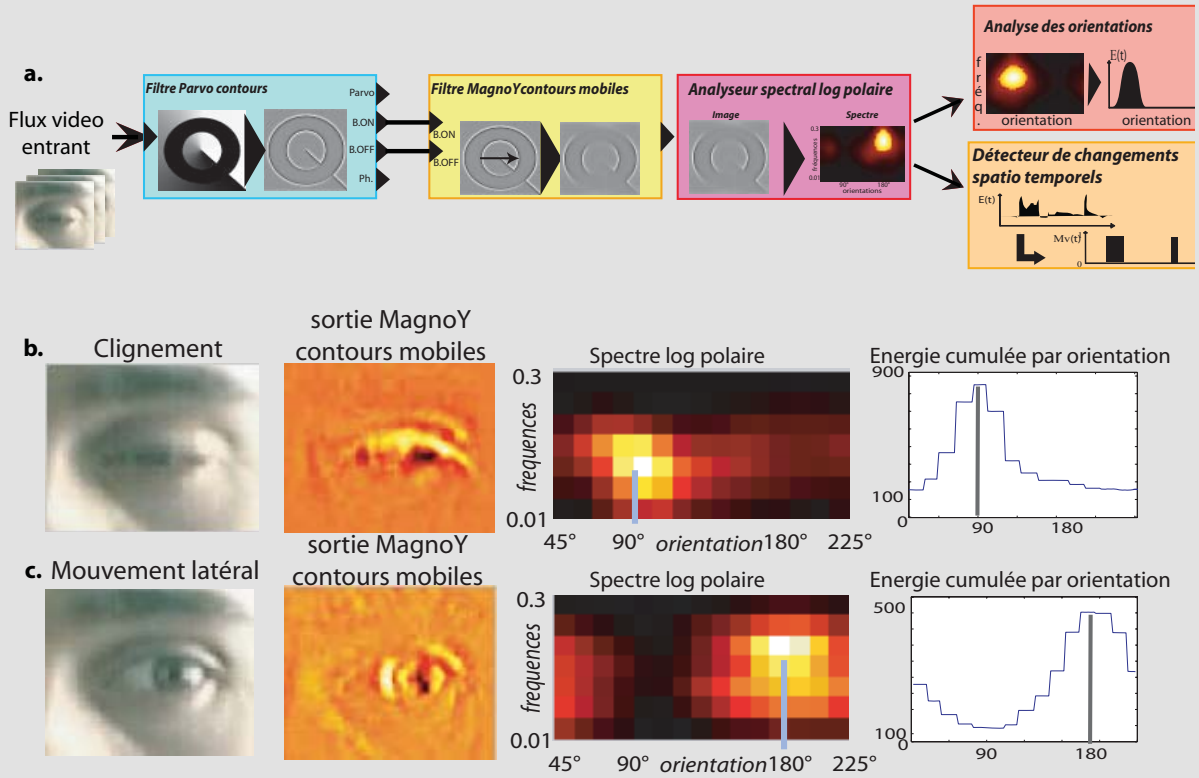


Figure III.8: a. système de détection des ouvertures et fermetures d'un oeil ou de la bouche. b,c, spectre log polaire et orientations de la sortie du filtre MagnoY contours mobiles pour différents mouvements d'oeil: clignement et changement de la direction du regard

III.4.2.2. Architecture du système complet et mise à jour de E_{HO} et E_{BF}

Pour l'identification de l'état ouvert/fermé de la bouche ou des yeux, nous proposons le système d'analyse décrit sur la figure III.9. Les deux voies d'information (Parvo contours et MagnoY contours mobiles) sont indépendamment prises en charge par un analyseur spectral log polaire. Celui associé à la voie Parvo contours donne l'énergie des contours qui va permettre la détection de l'état ouvert ou fermé tandis que l'analyseur spectral log polaire associé à la voie MagnoY contours mobiles sert à détecter les mouvements et estimer leur orientation.

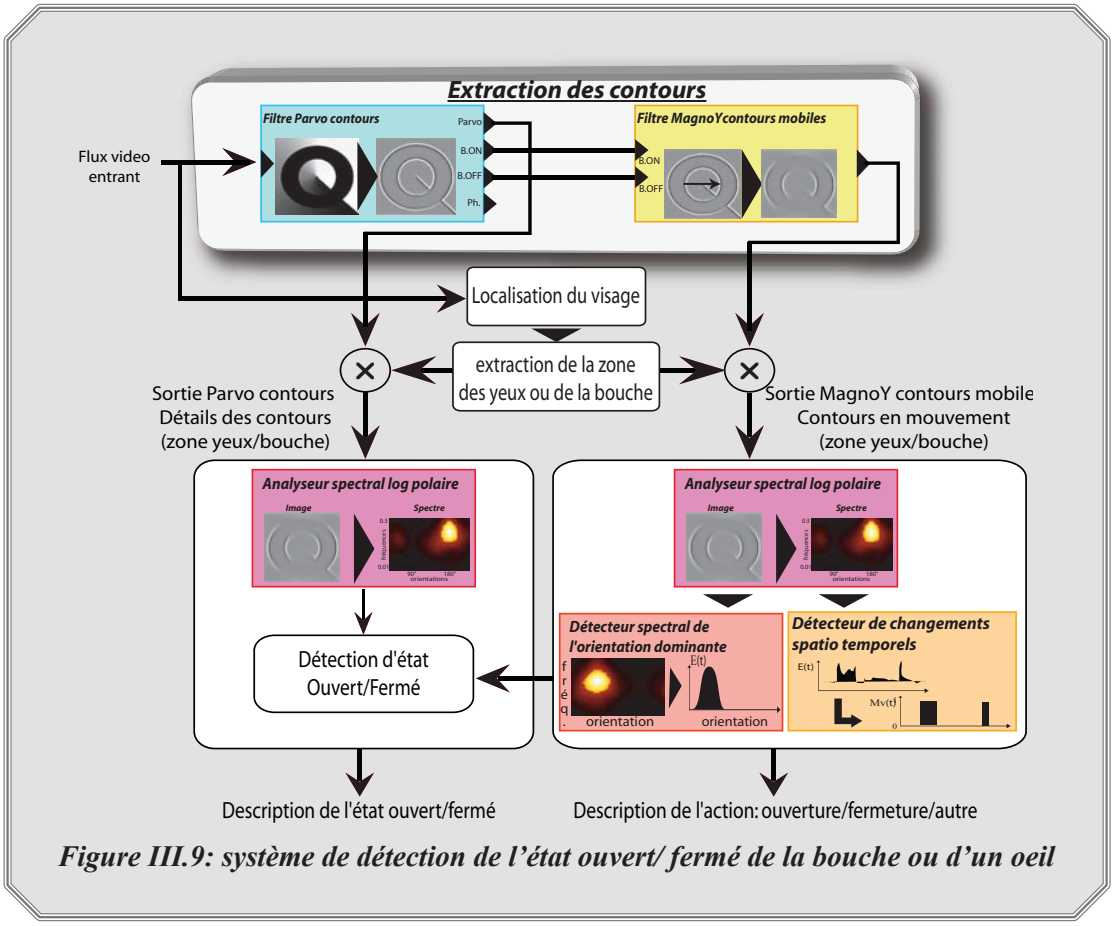


Figure III.9: système de détection de l'état ouvert/ fermé de la bouche ou d'un oeil

Nous avons maintenant accès aux informations essentielles nécessaires à un fonctionnement autonome du détecteur d'état ouvert/fermé d'un oeil ou de la bouche, à savoir:

→ L'énergie totale de tous les contours dans la boite englobante considérée.

→ Les alertes de mouvement.

→ L'orientation des mouvements.

L'algorithme décrit sur la figure III.10 est utilisé pour initialiser et remettre à jour les niveaux d'énergie E_{HO} et E_{BF} et pour en déduire l'état ouvert ou fermé du trait considéré.

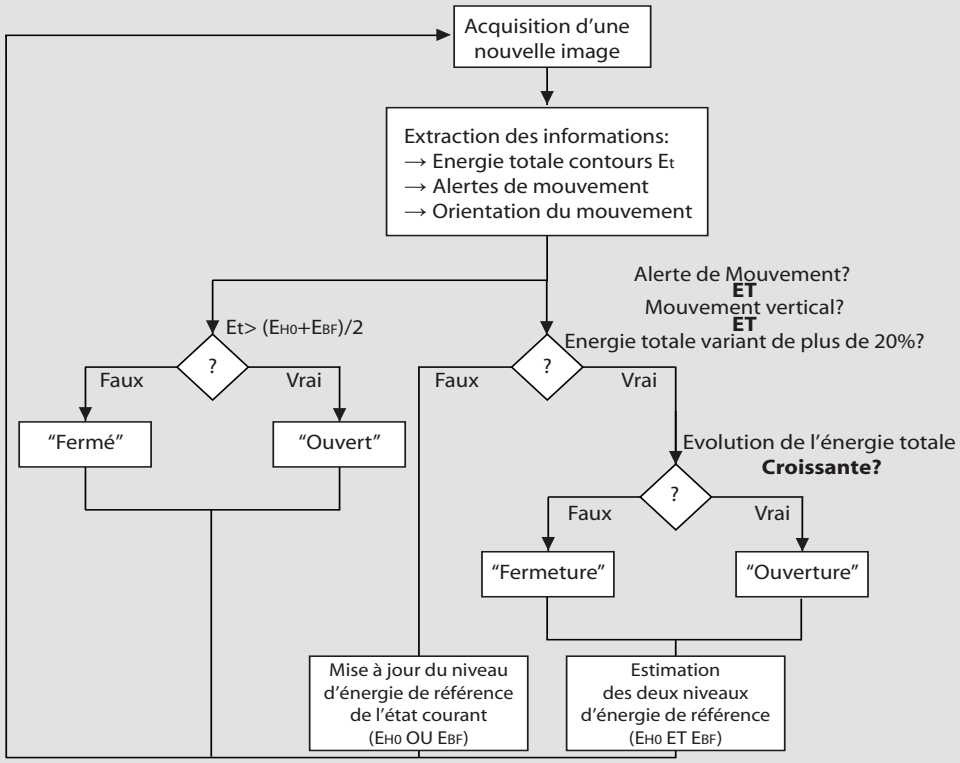


Figure III.10: algorithme de détection de l'état ouvert/fermé lors d'alerte de mouvement

La mise à jour des niveaux de référence E_{HO} et E_{BF} est réalisée en continu ce qui permet une adaptation aux conditions d'acquisition (niveau de bruit, etc.). Elle est néanmoins réalisée de deux façons différentes.

→ Lorsqu'un mouvement d'ouverture ou de fermeture est détecté, les deux niveaux d'énergie de référence sont mis à jour. Lors d'une ouverture, E_{BF} prend la valeur du niveau d'énergie avant ouverture et E_{HO} prend la valeur de l'énergie après mouvement et inversement dans le cas d'une fermeture.

→ Lorsqu'il n'y a pas de mouvement, seul le niveau d'énergie de référence correspondant à l'état courant est mis à jour. Ceci est réalisé à l'aide d'un lissage temporel permettant de donner la moyenne des dernières valeurs de l'énergie. Cette valeur moyenne est calculée grâce à un filtre adaptatif dont l'équation III.2 montre la réponse: pour une énergie $e(t)$ appliquée en entrée on obtient la valeur moyenne adaptative $\mu(t)$, le paramètre α ajustant l'effet mémoire de cette moyenne temporelle (nous fixons $\alpha = 0.6$ pour des séquences vidéo à 25 images par seconde). Ainsi, $\mu(t)$ correspond à $E_{HO}(t)$ dans le cas d'un état ouvert et $\mu(t)$ correspond à $E_{BF}(t)$ dans le cas d'un état courant fermé.

$$\mu_e(t) = \alpha.\mu_e(t-1) + (1-\alpha).e(t) \qquad \textit{(Eq. III.2)}$$

III.4.2.3. Performances du système de détection d'état

La méthode d'analyse utilisée ici se démarque des méthodes que l'on trouve dans la littérature. Celles-ci sont généralement basées sur une analyse du flot optique ou font appel à une phase de segmentation des contours des lèvres et des yeux à l'aide de modèles déformables [Eveno03, Bailly01] à divers degrés de complexité. La plupart de ces méthodes sont généralement basées sur une approche spatiale: l'analyse des caractéristiques des yeux et de la bouche dépend d'une phase de segmentation relativement précise de certains paramètres de la zone à analyser (commissures, iris, etc.). Ces méthodes sont souvent sensibles au bruit spatio-temporel qui limite les performances des algorithmes de suivi des points caractéristiques ou encore entraîne des changements locaux trop importants perturbant les mesures locales. Notre méthode, grâce au filtre Parvo contours limite les problèmes de bruit et de luminance, de plus, l'analyse est globale et aucune recherche de points caractéristiques n'est nécessaire.

Détection de l'état d'un oeil

Nous avons testé les performances de notre algorithme pour la détection de l'état des yeux. La figure III.11 montre les relevés d'énergie en sortie de filtre Parvo contours pour une séquence centrée sur un oeil. On note dans un premier temps la présence de deux niveaux d'énergie distincts, E_{BF} dont le niveau est régulier et E_{HO} qui est quant à lui plus variable ce qui correspond à toute une gamme de degrés d'ouverture de l'oeil. Les niveaux d'énergie de référence haute et basse s'adaptent aux conditions d'analyse dès les premiers clignements d'yeux. Lors de chaque clignement, le système détecte le changement d'état et ajuste les niveaux de référence qui correspondent aux conditions d'analyse courantes. Le seuil de séparation entre l'état ouvert ou fermé est déterminé à chaque instant et correspond à la moyenne des deux niveaux E_{BF} et E_{HO}, d'où la détection d'état par comparaison à ce seuil.

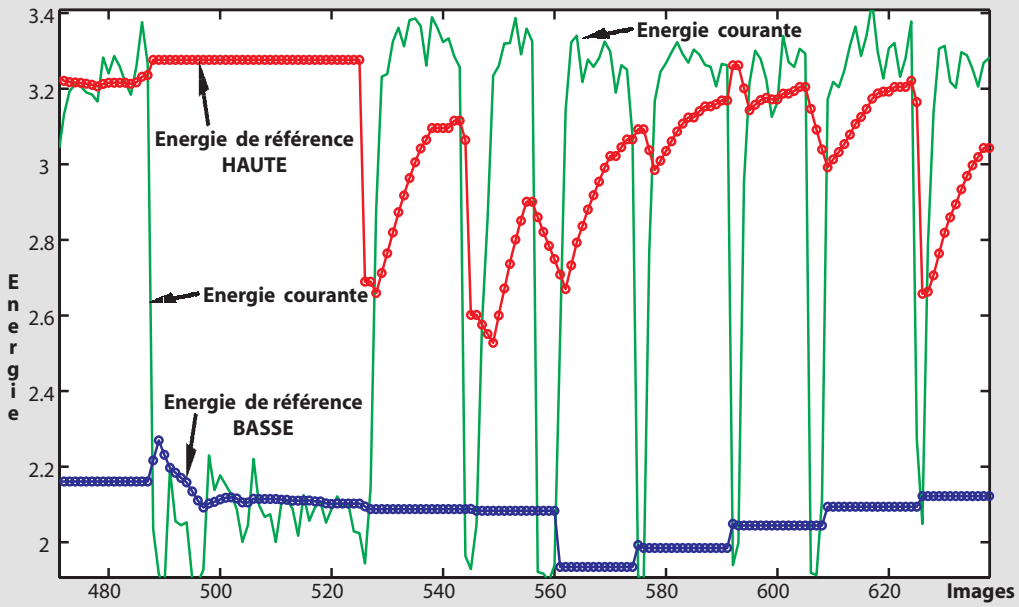


Figure III.11: exemple d'évolution temporelle de l'énergie E(t) en sortie du filtre Parvo contours pour l'analyse de l'état de l'oeil

Les performances de cet algorithme ont été évaluées en terme de rapidité à l'initialisation (c.-à-d. en combien de changements d'état ouvert/fermé, l'algorithme est initialisé) et en performance pour la détection de l'état. Les changements d'état considérés sont simulés et suffisamment lents pour qu'ils soient visibles à une cadence d'acquisition de 30 images par seconde. Les résultats sont présentés dans la table III.4. Cette table représente les tests réalisés sur dix personnes différentes. Une première étude a été menée dans le cas idéal c.-à-d. avec l'oeil centré dans la zone d'analyse, puis ont été ajoutées des contraintes telles que le port de lunettes transparentes ou une analyse sur boite englobante excentrée par rapport à l'oeil de manière à ne couvrir que 50% de sa surface. Ces analyses ont été menées avec différentes conditions d'éclairage allant des conditions standard aux lumières faibles (lumière tamisée). Les séquences ont été acquises avec une caméra de type webcam (basse qualité et bruit fort). La vérité terrain annotée à la main recense 686 changements d'état pour une durée totale de 76 minutes de vidéo.

Table III.4: performances de l'analyseur d'état de l'oeil

	Taux de succès	***Taux de fausse alarme***	***Taux d'oubli***
Oeil entier	96%	2%	2%
Analyse à 50%	94%	4%	2%
Oeil + lunettes	95%	3%	2%
lunette et oeil à 50%	92%	5%	3%

Les performances sont très bonnes pour les conditions idéales quelques soient les conditions d'éclairage (normales ou lumière tamisée) et l'on constate qu'elles diminuent très faiblement même si l'oeil n'est analysé que sur 50% de sa surface. Cette propriété permet de pallier d'éventuelles erreurs lors du centrage de la boite englobante de l'oeil. Il faut néanmoins qu'une partie de l'iris soit présente dans la zone d'analyse.

De même, les résultats avec lunettes sont satisfaisants. La présence de reflets sur le verre peut néanmoins avoir des effets négatifs sur les performances, car ceux-ci cachent les détails de l'oeil et donc atténuent ou annulent la réponse de ses contours.

La résolution de la zone d'analyse de l'oeil a peu d'influence sur les performances de l'algorithme. Néanmoins, il existe une limite basse de l'ordre de 4 pixels pour le diamètre de l'iris. En dessous de cette taille, il devient difficile de distinguer précisément les contours et leur réponse énergétique devient trop faible par rapport au niveau de bruit.

Détection de l'état de la bouche

La figure III.12 montre l'évolution temporelle de l'énergie totale *E(t)* du spectre du filtre Parvo contours ainsi que les niveaux de référence E_{HO} et E_{BF} pour une fenêtre d'analyse centrée sur la bouche. Sur la courbe *E(t)*, on distingue nettement des variations importantes avec des valeurs basses correspondant à l'état fermé et des valeurs hautes correspondant à l'état ouvert. Le niveau bas est, comme pour l'oeil, relativement régulier dans le temps (l'état fermé étant quasiment "unique"). L'état ouvert donne quant à lui un niveau d'énergie haut variant sur toute une gamme de valeurs d'où la nécessité d'adapter au cours du temps E_{HO} et dans une moindre mesure E_{BF}. Lors de séquence de paroles, *E(t)* prend une valeur maximale presque 2 fois supérieure à la valeur de l'état fermé. Lors d'un bâillement, l'énergie *E(t)* croît de façon caractéristique: une longue durée d'accroissement jusqu'à une très forte valeur.

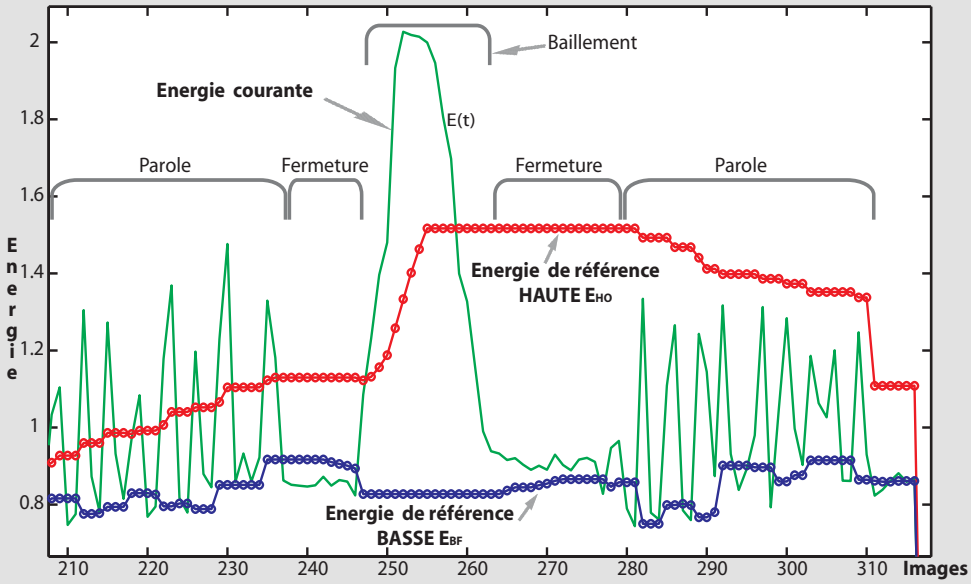


Figure III.12 exemple d'évolution de l'énergie E(t) en sortie du filtre Parvo lors de mouvements de bouche

Les performances de la détection de l'état ouvert ou fermé de la bouche sont présentées sur la table III.5. Les résultats sont issus de tests effectués sur une base vidéo de 1219 mouvements de bouche pour une durée totale de 80 minutes. Comme pour l'oeil, nous estimons les performances dans le cas idéal puis dans deux cas plus contraignants dans lesquels la bouche est excentrée par rapport à la fenêtre d'analyse. Dans ces cas de figure, la bouche n'est visible qu'à respectivement 50% ou 30% dans la zone boite englobante. Ceci permet de simuler une mauvaise estimation de la localisation de la bouche afin d'étudier la perte de performances que cela entraîne sur notre détecteur. Rappelons ici que la bouche est supposée se trouver dans la partie basse au centre de la boite englobant le visage A_{bouche} (cf. fig.III.6).

Table III.5: performances de la détection des états ouvert et fermé de la bouche

	Taux de succès	*Taux de fausse alarme*	*Taux d'oubli*
Anayse complète	97%	1%	2%
50% de la bouche	93%	5%	2%
30% de la bouche	82%	10%	8%

On observe que les performances restent correctes même si la bouche n'est pas entièrement incluse dans la zone d'analyse. Ceci suppose néanmoins que la bouche soit le seul élément mobile dans la zone d'analyse (les régions annexes étant les joues, le menton et le nez, cette hypothèse peut être considérée comme valable). Le point fort de cette méthode est qu'elle ne nécessite pas l'extraction de points caractéristiques de la bouche, mais se base sur une analyse globale ce qui la rend moins sensible à d'éventuelles mauvaises détections.

Les tests présentant la détection d'état dans le cas excentré permettent de montrer que l'on peut se

contenter d'une fenêtre d'analyse centrée de façon peu précise. La figure III.13 le confirme: lorsque la bouche se translate sur la fenêtre d'analyse durant l'expérimentation, le système se réadapte rapidement empêchant ainsi toute fausse alarme.

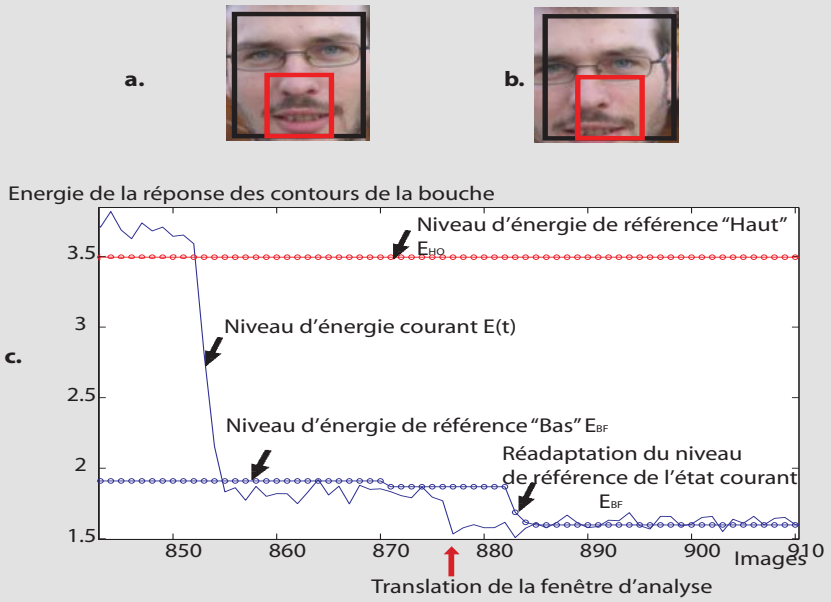


Figure III.13: exemple d'évolution de l'énergie en sortie du filtre Parvo contours (c) lors de la translation de la bouche dans son cadre d'analyse (a) vers (b) (en rouge)

Coût de calcul

Du point de vue du temps de calcul, si l'on ne prend pas en compte le système de détection du visage, cet algorithme fonctionne à plus de 500 images par seconde pour une zone d'analyse de 50*50 pixels sur un Pentium4 à 3.0GHz sous Matlab. Ces performances sont dues au fait que la zone d'analyse est réduite.

III.5. Interprétation des mouvements locaux du visage

Dans ce paragraphe, nous faisons l'hypothèse que les seuls mouvements détectés sur la zone du visage sont des mouvements locaux c.-à-d. des yeux et/ou de la bouche. On se place donc dans le cas où l'algorithme présenté au paragraphe III.3 n'a sélectionné que ce type de mouvement.

III.5.1 Description des clignements d'yeux

L'algorithme de détection des états ouvert ou fermé permet d'extraire:

→ L'ensemble des mouvements verticaux des yeux d'où la possibilité d'estimer le nombre de clignements.

→ La durée de fermeture des yeux.

→ L'intervalle de temps entre deux clignements d'où une estimation de leur fréquence.

Nous avons effectué des tests d'évaluation de performance de notre algorithme pour estimer le nombre

de clignements sur une base de tests rassemblant 982 clignements pour une durée totale de 86 minutes de vidéo. La vérité terrain est définie manuellement et l'acquisition est réalisée à l'aide d'une webcam fonctionnant à 30 images par seconde. On distingue dans ces tests des clignements simulés (70%) et des clignements naturels (30%). Les résultats sont présentés sur la table III.6.

Table III.6: performances de la détection des clignements

	Taux de succès	***Taux de fausse alarme***	***Taux d'oubli***
simulés	93%	2%	5%
naturels	68%	3%	29%

L'analyse de ces résultats montre qu'il y a une différence significative entre le taux de succès pour les clignements simulés et celui obtenu pour des clignements naturels. Pour les clignements simulés, le taux de succès est plus élevé, car ces clignements sont moins rapides. En revanche, le taux assez moyen obtenu pour les clignements naturels est à lier à la vitesse d'acquisition de la caméra qui est alors insuffisante. En effet, des travaux [Caffier03] montrent que la durée moyenne de fermeture des yeux lors de clignements est comprise entre 150ms et 500 ms. L'utilisation de caméras à haute vitesse est préférable. En conclusion, la relative contre-performance en cas de clignements naturels n'est pas due à l'algorithme de détection proposé.

III.5.2. Interprétation des mouvements de bouche

Un mouvement de bouche est associé à de la parole, à un bâillement ou à une mimique (grimace). Ce paragraphe montre des méthodes de caractérisation de tels mouvements ; les mimiques ne sont pas considérées ici.

III.5.2.1 Détection de bâillements

Comme nous l'avons vu sur la figure III.12, en cas de bâillement (de l'image 246 à 263), l'ouverture de bouche est extrême et l'augmentation de l'énergie en sortie du filtre Parvo contours est importante et caractéristique. L'idée est donc de détecter ces grandes variations caractéristiques pour détecter les bâillements. Pour cela, une expertise a montré que l'amplitude des variations d'énergie lors de bâillement est plus de 1.5 fois supérieure à celle correspondant à des mouvements de parole (le cri n'est pas pris en considération) et plus de 2 fois supérieure au niveau d'énergie de l'état fermé. Donc, la détection de bâillement se base sur la détection d'un mouvement vertical d'ouverture ou de fermeture associé à une évolution de l'énergie courante du spectre log polaire $E(t)$ respectivement d'un facteur 2 (lors de l'ouverture) ou d'un facteur 0.5 lors de la fermeture.

La table III.7 montre les tests réalisés sur une dizaine de personnes réalisant des bâillements volontaires ou simulés. La seule contrainte est que la main ne doit pas occulter la bouche. La base de tests de 152 minutes représente 203 bâillements répartis également entre simulation et naturel. Ces vidéos tests sont entrecoupées de périodes d'absence de mouvement (silence), de périodes de parole avec ou sans bâillements.

Table III.7: performances de la détection des bâillements

	Taux de succès	***Taux de fausse alarme***	***Taux d'oubli***
naturels	85%	3%	13%
simulés	87%	2%	11%

Les performances sont correctes et comparables entre les bâillements naturels et simulés. On note un faible taux de fausses alarmes, il y a en effet peu de cas de confusion entre la parole et le bâillement dans le cas normal. En revanche, le taux d'oubli est plus important, il s'explique par une confusion entre bâillements et les mouvements de parole lors de bâillements de faible amplitude.

III.5.2.2. Détection de parole

Plus précisément, nous proposons un algorithme de détection d'absence de parole. En effet, tout mouvement de lèvres ne peut pas être associé de façon systématique à de la production de parole [Rivet06]. En revanche, on considérera qu'une absence de mouvement labial est une marque de silence (on exclut ici le cas des ventriloques).

Le système d'extraction des temps de non-parole consiste alors à détecter les images pour lesquelles aucun mouvement n'est présent. Ceci se fait facilement en utilisant le détecteur d'événements temporels et en analysant l'évolution temporelle de l'indicateur *α(t)*. Il s'agit donc de détecter les périodes pour lesquelles l'indicateur *α(t)* reste à une valeur nulle. Cette méthode est illustrée sur la figure III.14 sur laquelle est représentée l'évolution temporelle de l'indicateur *α(t)* pour une séquence d'analyse des mouvements de bouche lors d'une conversation standard. On distingue les intervalles de temps pendant lesquels la personne parle (notés *P*), durées pour lesquelles notre indicateur est variable, et au contraire, des périodes d'absence de parole (notées *AP*) pour lesquelles l'indicateur reste constamment à 0.

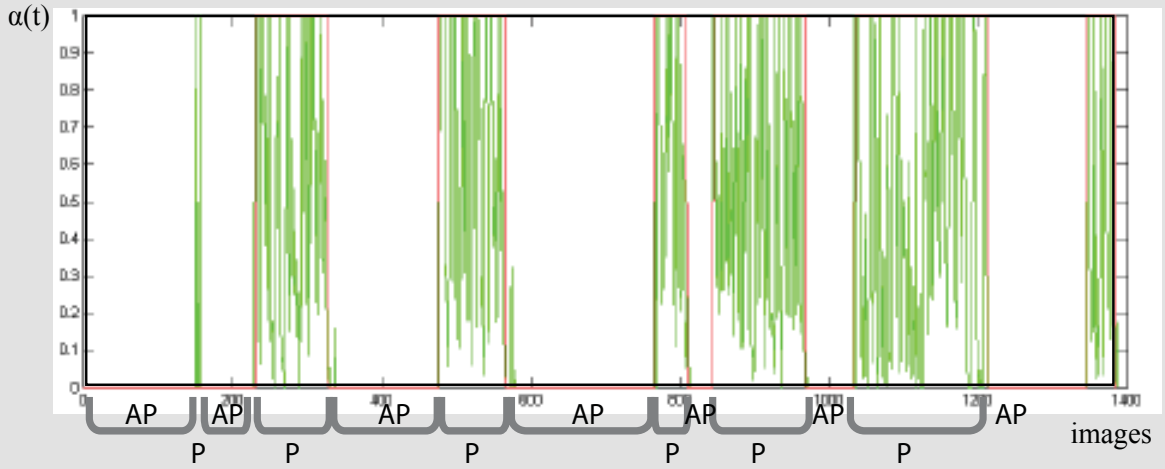


Figure III.14: exemple d'évolution de l'indicateur *α(t) pour une fenêtre d'analyse centrée sur la bouche durant une séquence de conversation.*

La table III.8 présente les résultats quantitatifs de détection de non-parole. Ces tests ont été effectués sur 45 minutes de séquences vidéo pour lesquelles ont été étiquetées manuellement les périodes d'absence de parole. On présente sur ce tableau le taux de reconnaissance des durées de non-parole (taux de succès, d'oubli et de fausses alarmes) ainsi que la moyenne et l'écart type de la synchronisation entre les périodes détectées et la vérité terrain. Enfin, on présente le taux de recouvrement des périodes de non-parole estimées par rapport aux durées réelles.

Table III.8: performances du détecteur de périodes d'absence de parole

Taux de succès	*Taux de fausse alarme*	*Taux d'oubli*	*Synchronisation Moyenne (+/-écart-type)*	*Recouvrement de la durée: moyenne (+/-écart-type)*
94%	2%	6%	2.1 (+/-2.2) images	93%(+/-7%)

La sensibilité de l'indicateur $\alpha(t)$ permet d'extraire le moindre mouvement de bouche. Cette sensibilité se révèle parfois trop importante et entraîne des oublis dans la détection des périodes de non-parole pour certains petits mouvements de bouche. Ceci explique un taux d'oubli plus élevé que le taux de fausses alarmes. Parallèlement, la synchronisation entre la mesure et les données réelles est bonne ; si l'on considère un flux vidéo à 30 images par seconde, la désynchronisation est alors de l'ordre de 60ms. Le recouvrement de la durée de non-parole et la vérité terrain est satisfaisant. Vue la sensibilité de la détection de mouvement, l'algorithme a tendance à sous-estimer les durées de non-parole ce qui autorise un bon niveau de confiance sur les périodes de non-parole détectées.

Une application intéressante de notre système est l'estimation de bruit ambiant durant les périodes de non-parole pour les systèmes de communication en milieu bruité [Rivet06]. Dans la littérature, les recherches sont plutôt orientées sur un aspect détection de parole (à l'opposé de non-parole) dans le but de localiser des locuteurs. Ces études se basent sur une analyse du signal sonore pour aboutir à la localisation 2D/3D de locuteurs dans des espaces restreints (salle de réunion, etc.) [Saric04]. Cette détection n'est néanmoins pas suffisante, elle ne peut être dissociée d'une analyse visuelle [Rivet06].

III.6 Extraction de l'information de mouvement global de la tête

III.6.1. Système proposé

Dans cette partie, nous faisons l'hypothèse que les mouvements étudiés sont des mouvements rigides de la tête dans son ensemble. On s'intéresse à la détection des changements d'orientation et des hochements. Nous ne traitons pas le problème de l'estimation de la pose 3D de la tête.

Afin d'analyser ces mouvements de tête, nous proposons le système de la figure III.15 basé sur une analyse des contours en mouvement. La première étape consiste à extraire ces contours en mouvement par l'usage des filtres Parvo contours et MagnoY contours mobiles. La détection des mouvements se fait ensuite grâce à l'utilisation du module de détection des événements spatio-temporels placé après le module d'analyse spectral. Une fois qu'un mouvement est détecté sur la zone englobant le visage, les caractéristiques d'amplitude, d'orientation et de direction sont extraites grâce aux modules d'analyse des orientations et d'estimation de flot optique. En sortie du système, trois données sont accessibles: les événements de mouvement, leur orientation et leur vecteur vitesse. Un système de fusion a la charge de donner l'interprétation haut niveau du

mouvement considéré. Le détecteur d'événement contrôle les entrées de ce système de fusion: les informations liées à l'orientation du mouvement et de la vitesse ne sont estimées que si un mouvement est détecté.

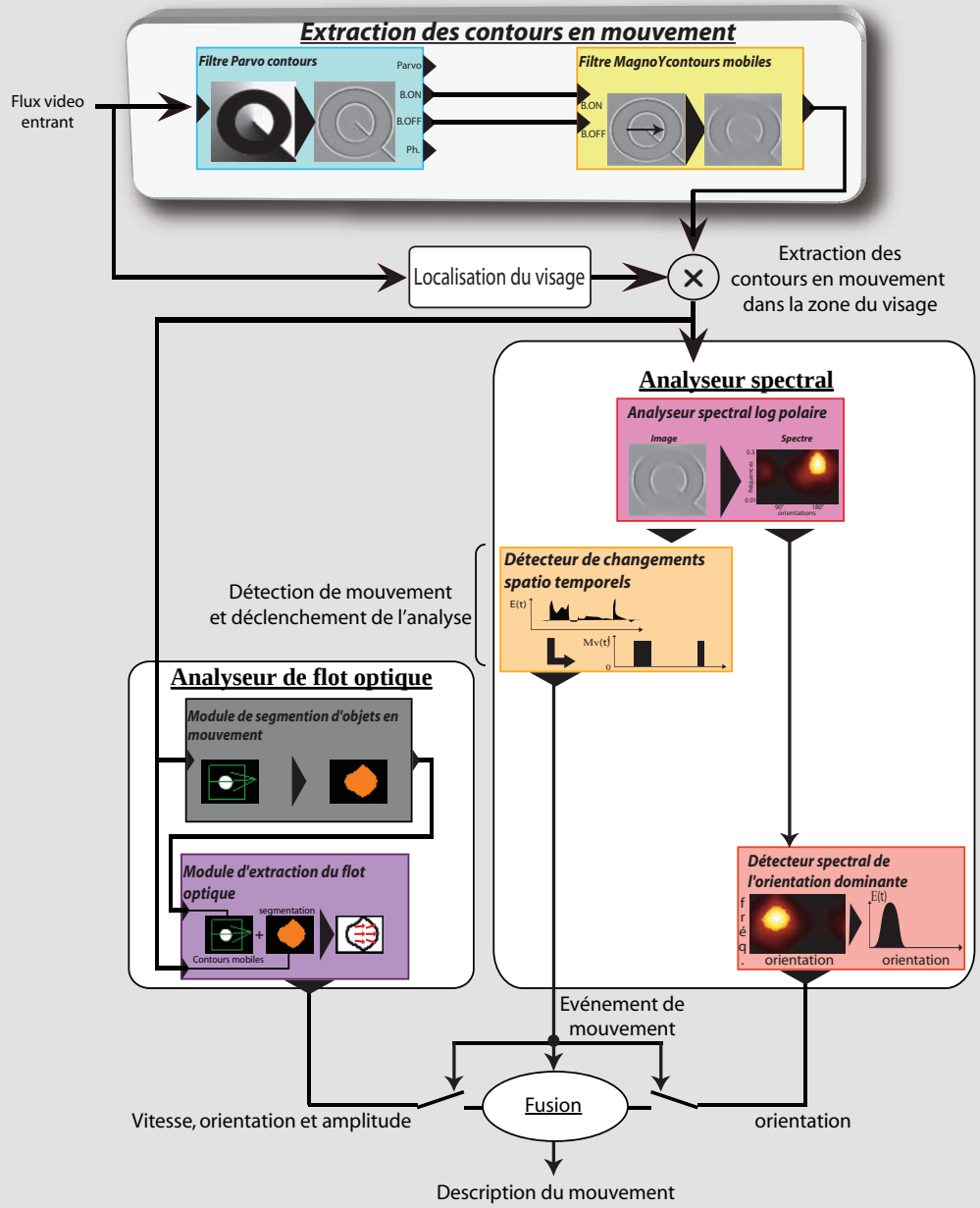


Figure III.15: schéma global du processus d'analyse des mouvements globaux de la tête. Les deux interrupteurs sont connectés uniquement lorsqu'un mouvement est détecté

III.6.2. Principe

III.6.2.1. Extraction d'informations sur l'orientation du mouvement

→ Estimation du flot optique

Le module flot optique donne à chaque instant l'orientation du mouvement. On note θ_f l'orientation du vecteur vitesse moyen du visage calculé par le module flot optique. Cette orientation est la moyenne des orientations des vecteurs vitesse calculés en chaque pixel de la zone du visage.

→ Estimation de l'orientation globale du mouvement par analyse spectrale

La sortie de l'analyseur spectral des orientations fournit l'ensemble des orientations associées aux mouvements présents sur le visage. A titre d'illustration, la figure III.16 montre la courbe d'énergie cumulée par orientation obtenue pour différents mouvements de tête. Les figures III.16.I.a-b-c montrent que lors d'un mouvement horizontal ou vertical, il n'apparaît qu'un seul maximum, ce maximum correspondant à l'orientation dominante perpendiculaire au déplacement. Dans le cas d'un mouvement de biais (cf. fig. III.16.I.d), on constate que la courbe d'énergie possède deux maximums.

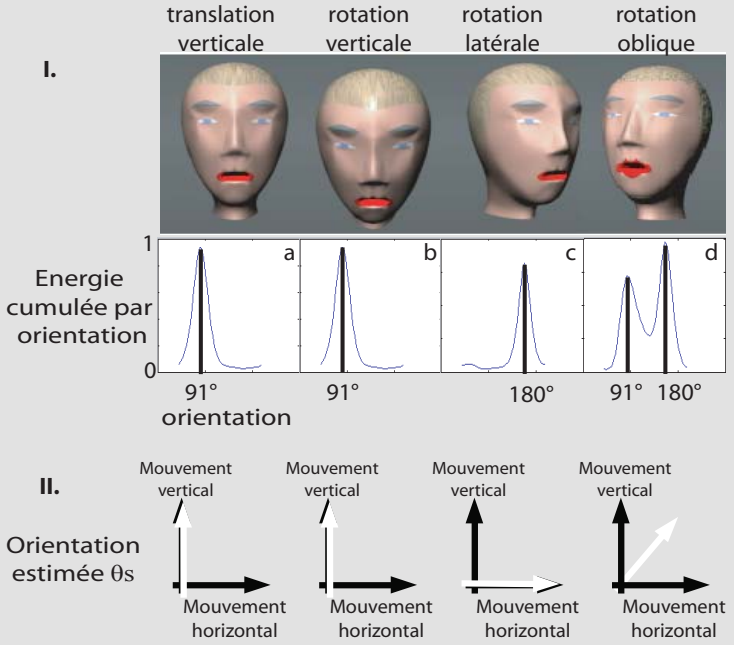
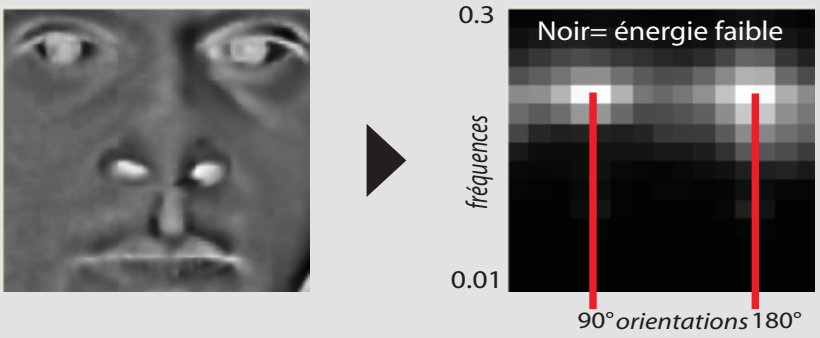


Figure III.16: (I) mouvements globaux simples et composés de la tête dans son ensemble. I.a, mouvement de translation verticale. I.b,. rotation verticale. I.c, rotation latérale et I.d, rotation oblique. II, représentation de l'orientation du mouvement sur une fenêtre [90°;180°]

Pour ce dernier exemple, l'apparition de deux pics s'explique par les caractéristiques des contours du visage. En observant le spectre log polaire de la réponse des contours au niveau du filtre Parvo contours

(cf. fig. III.17) c.-à-d. le spectre du visage, sans information de mouvement, on remarque qu'il existe deux orientations dominantes: les orientations verticale (90°) et horizontale (180°) ce qui n'est pas une surprise. Ces orientations sont créées par les contours du nez, de la bouche et des yeux. Les contours présentant ces orientations sont impliqués dans le mouvement même s'ils ne sont pas exactement perpendiculaires à celui-ci (ce qui correspond au problème d'ouverture récurrent en analyse de mouvement).



a. **b.**

Figure III.17: répartition de l'énergie des contours sur le spectre log polaire d'un visage immobile après filtrage Parvo contours. a, sortie du filtre Parvo contours sur le visage. b, spectre log polaire correspondant

Il est possible d'estimer l'orientation des mouvements sur une fenêtre [90°;180°] en effectuant une moyenne pondérée de l'énergie par orientation. Sachant que le spectre log polaire est échantillonné selon n orientations, en notant E_i l'énergie liée à l'orientation θ_i on extrait l'orientation moyenne selon la relation:

$$\theta_s = \frac{\sum_{i=1}^{n} \theta_i . E_i}{\sum_{i=1}^{n} E_i}$$

(Eq. III.3)

Cet analyseur des orientations des contours en mouvement donne une autre information sur l'orientation du mouvement. Il décrit son orientation dans une fenêtre [90°;180°] (cf. fig. III.16.II) du fait des deux orientations principales du visage. On sait donc pour chaque mouvement s'il est horizontal ou vertical ou oblique.

θ_s et θ_f représentent ainsi la même information, mais obtenue de deux façons différentes.

III.6.2.2. Confiance sur les orientations estimées

Lorsqu'un mouvement est détecté sur le visage, on dispose des deux mesures θ_s et θ_f décrivant l'orientation du mouvement. Ces deux mesures vont être comparées afin d'en déduire un niveau de confiance vis-à-vis de ces estimations. On définit alors la fonction $T_{s,f}$ de confiance selon l'expression III.4 avec $\theta_{s'}$ et $\theta_{f'}$ les orientations estimée θ_s et $\theta_{f'}$ ramenées à un intervalle [0; π/2]:

$$T_{s,f} = \cos\left(\theta_{s'} - \theta_{f'}\right) \qquad \textit{(Eq. III.4)}$$

$T_{s,f}$ évalue la différence entre les deux orientations estimées par les indicateurs θ_s et θ_f. Au final, le cosinus de la différence des deux valeurs permet d'exprimer le niveau de confiance accordé aux deux mesures en donnant une valeur comprise entre 0 et 1. Si $T_{s,f}$ tend vers 0, alors les estimations d'orientation par deux méthodes différentes sont très différentes d'où un niveau de confiance bas. Si $T_{s,f}$ tend vers 1 alors, les deux calculs d'orientation sont similaires (c.-à-d. que les deux méthodes différentes ont donné le même résultat) et le niveau de confiance est considéré comme bon.

On définit un niveau de confiance minimum permettant de valider l'estimation de l'orientation. Si le niveau de confiance est supérieur au seuil ρ, la mesure est validée et le système de fusion donne en sortie l'orientation du mouvement ainsi que sa direction donnée par θ_f (cette mesure θ_f étant plus précise que θ_s qui est basé sur un spectre échantillonné par bande de 12°, cf. chapitre I.5). Dans la pratique, on fixe le seuil ρ à 0.93, ce qui représente une tolérance à l'orientation de 20°. Si le niveau de confiance est inférieur à ρ alors, la mesure n'est pas validée, la description du mouvement consiste alors simplement à signaler qu'un mouvement a été détecté, sans qu'il puisse être décrit précisément.

Performances

Nous avons évalué les performances de ce système sur une base vidéo de 123 minutes contenant 75 minutes de mouvements globaux de tête effectués de façon libre et dont les caractéristiques ont été étiquetées à la main. La table III.9 montre les résultats obtenus. On montre la précision de la mesure d'orientation donnée par θ_f lorsque la mesure est validée c.-à-d. lorsque $T_{s,f}>\rho$. On donne également le niveau de confiance moyen accordé à toutes les mesures effectuées qui est donné par la moyenne de $T_{s,f}$. Pour rappel, on ne donne pas d'estimation de l'orientation lorsque les mesures ne sont pas valides (c.-à-d. $T_{s,f}<\rho$).

Table III.9: performances du système de détection de mouvement rigides du visage

	Précision de l'estimation de la direction lorsque $T_{s,f}>\rho$	***Niveau de confiance moyen pour l'ensemble des mesures***
Taux de succès	9°	$T_{s,f_moyen} = 88\%$

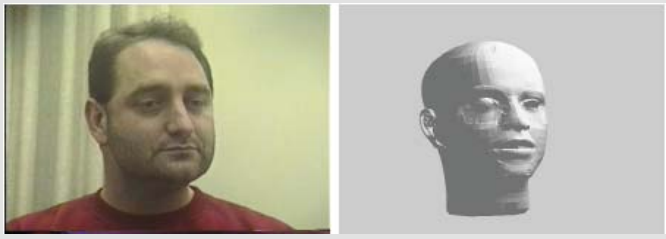
On remarque d'une part que l'estimation de l'orientation par le flot optique est correcte lorsque la confiance est suffisante. D'autre part, le niveau de confiance moyen indique une cohérence généralement correcte entre les deux systèmes d'estimation de l'orientation. Le niveau de confiance montre tout de même que certains mouvements posent problème ce qui conduit à une erreur d'estimation de l'orientation. Ces cas particuliers sont principalement dus aux mouvements obliques, qui, du fait des deux orientations principales du visage créent des problèmes d'ouverture lors de l'estimation du flot optique. Ceci entraîne une différence importante entre les résultats des deux systèmes d'analyse et invalide la mesure. On note donc ici l'avantage d'avoir recours à deux systèmes parallèles qui permettent de détecter les estimations faussées.

Dans la littérature, cette tâche d'analyse est généralement abordée par des méthodes classiques d'analyse directe du flot optique [Baron94] sur le visage. La précision est du même ordre de grandeur. Elle est généralement employée pour une caractérisation fine de la pose du visage par l'ajustement de modèles 3D projetés sur le plan 2D de la caméra. Comme illustrés sur la figure III.18, différents modèles à divers niveaux de complexité sont utilisés. On trouve des modèles simples comme une représentation cylindrique du visage

[Xiao03] (cf. fig. III.18.a), des modèles non déformables (cf. fig. III.18.b) ainsi que de nombreuses variantes utilisant notamment le filtrage particulaire [Bogdan06] (cf.fig III.18.c). Ces méthodes n'ont pas la même finalité, elles ont plus pour but de déterminer la pose 3D du visage, notre méthode étant axée sur la description du mouvement.

a.

b.



c.

Figure III.18: exemples de modèles utilisés pour l'estimation de la pose et des mouvements de la tête. a, utilisation d'un modèle cylindrique [Xiao03]. b, utilisation de modèle géométrique générique [Zivkovic01]. c, utilisation de filtrage particulaire [Braathen01].

III.6.3. Détection des hochements d'approbation et de négation

L'objectif dans ce paragraphe est de proposer un algorithme de détection automatique des hochements d'approbation et de négation. Selon les coutumes européennes:

→ Un hochement d'approbation se caractérise par des oscillations verticales périodiques.

→ Un hochement de négation se caractérise par des oscillations horizontales périodiques.

Deux informations sont requises: l'orientation du mouvement de la tête et une description de l'enchaînement des mouvements. On utilise donc le même système que celui présenté précédemment (cf. fig. III.15), nous allons néanmoins voir que le module d'estimation du flot optique n'est pas nécessaire dans le principe proposé. Les deux informations requises sont:

→ L'analyse de l'évolution temporelle de l'orientation dominante du spectre log polaire en sortie du filtre MagnoY contours mobiles permet d'estimer l'orientation du mouvement dominant et son évolution selon les mouvements effectués. On rappelle que l'orientation dominante correspond à l'abscisse du maximum sur la courbe d'énergie cumulée par orientation (cf. chapitre II).

→ La détection de changements périodiques du sens du mouvement (oscillations) est obtenue à partir de l'étude de l'évolution de l'énergie totale *E(t)* du spectre des contours en mouvement. On extrait une information sur l'évolution du mouvement: accélérations, décélérations, arrêts (cf. chapitre II.3). On détecte également l'indice temporel de chacun des maximums ou bien des minimums d'amplitude à l'aide du détecteur de changements spatio-temporels (les indices temporels pour lesquels la différentielle temporelle de l'énergie *E(t)* du spectre de la sortie MagnoY contours mobiles s'annule).

La figure III.19 montre l'analyse réalisée sur une séquence vidéo de synthèse (cf. fig. II.19.a) dans laquelle une tête artificielle réalise un enchaînement de hochements de têtes de négation (mouvements horizontaux) entre les images 30 et 255 puis d'approbation (mouvements verticaux) entre les images 290 à 490. La figure III.19.b montre l'évolution de l'orientation dominante des contours en mouvement durant la séquence. On distingue deux paliers : le premier correspond à une orientation de l'ordre de 180° donc à un mouvement dominant horizontal (négation) et le second correspond à une orientation de l'ordre de 90° donc à un mouvement dominant vertical (approbation). Par ailleurs, la mesure de l'évolution temporelle de l'énergie du spectre en sortie du filtre MagnoY contours mobiles est périodique durant les hochements (cf. fig. III.19.c). Chaque maximum correspond à un mouvement de grande amplitude tandis que chaque minimum correspond à un ralentissement voire à un arrêt complet du mouvement dans la séquence.

Donc, finalement, l'analyse conjointe des deux courbes traduit la présence dans la séquence d'un mouvement horizontal périodique donc une négation puis d'un mouvement vertical périodique, donc une approbation.

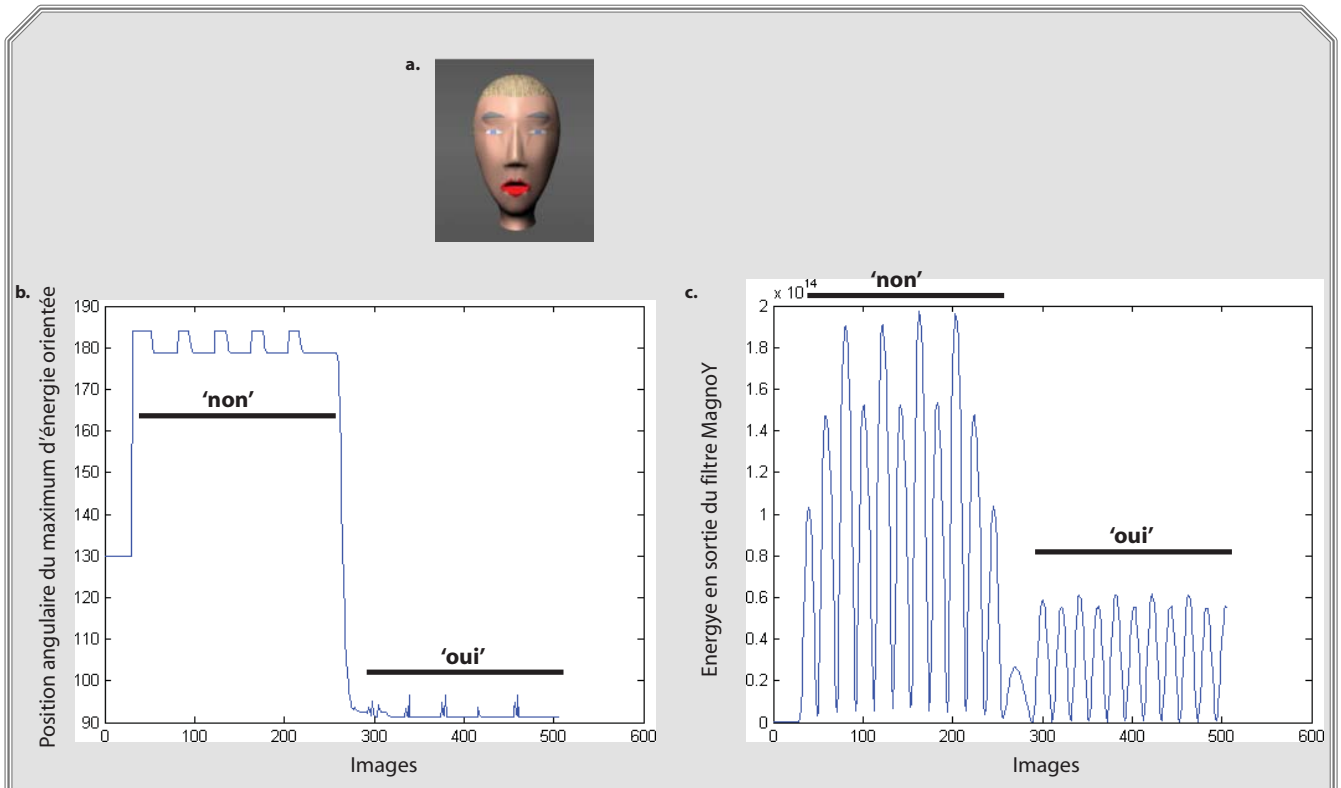


Figure III.19: séquence de hochements de tête d'approbation et de négation. a, extrait de la séquence. b, évolution temporelle de la position du maximum de la courbe d'énergie cumulée par orientation. c, évolution temporelle de l'énergie du spectre en sortie du filtre MagnoY contours mobiles

III.6.4.3. Algorithme de détection d'approbation ou de négation

A partir des deux informations temporelles que nous venons de voir, il est possible d'élaborer un algorithme de détection de «Oui» ou de «Non». La figure III.20 montre l'algorithme proposé.

Dans le but de détecter des approbations ou des négations, sans générer de fausses alarmes par confusion avec d'autres gestuelles simples (par exemple le fait de tourner la tête rapidement pour regarder de côté), nous fixons expérimentalement un nombre minimum de mouvements opposés *n* égal à *4* (soit 2 oscillations) dont la fréquence minimale d'inversion est fixée à f_{min}=1.2Hz (0.83 aller-retour par seconde). La plausibilité que des mouvements plus lents et moins nombreux puissent être liés à un "OUI" ou un "NON" n'est alors pas considérée. Toujours dans le souci de limiter le nombre de fausses alertes dues à des gestes différents, on autorise une tolérance sur la valeur des orientations et de la période de la séquence de hochements. Celles-ci sont fixées respectivement à 15° et 0.3s.

Ainsi, dans le cas où un mouvement global est détecté sur le visage, on enregistre dans une liste la valeur de l'orientation et de l'amplitude du mouvement dominant ainsi que l'indice de l'image. A partir de cette liste, on définit $\theta\mu$ et $\theta\sigma$ respectivement la moyenne et l'écart type sur l'orientation des *n* derniers mouvements dominants rencontrés (couramment, *n=4*). On définit également $T\mu$ et $T\sigma$ la moyenne et l'écart type de la période temporelle entre deux alertes successives de mouvement.

Comme le montre la figure III.20, le système détecte l'expression d'un OUI ou d'un NON si l'enchaînement des alertes de mouvement est à la fois régulière en période et régulière en orientation. Si ces conditions sont respectées, alors il s'agit de l'un des deux gestes recherchés, et à partir de la valeur de l'orientation estimée, l'algorithme détermine s'il s'agit d'une approbation ou d'une négation.

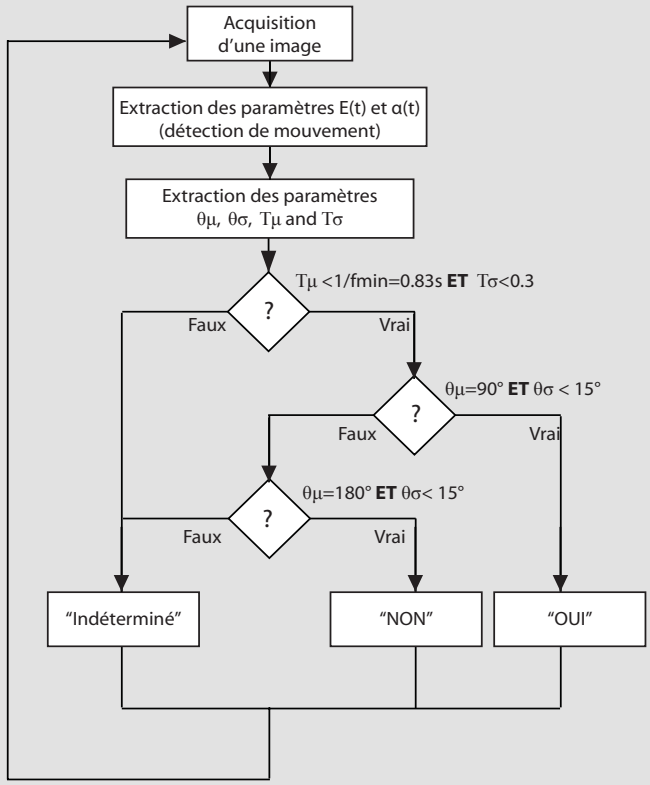


Figure III.20: algorithme de reconnaissance des hochements de tête

III.6.4.2. Performances

La figure III.21 montre les résultats obtenus pour une séquence réelle complexe. La figure III.21.a montre un extrait d'une séquence vidéo dans laquelle le sujet effectue des hochements de tête d'approbation de l'image 370 à 400 et des hochements de négation entre les images 400 et 462. On remarque d'une part que sa main est devant la bouche et d'autre part que la caméra ne fait pas exactement face au visage. L'arrière plan est texturé, mais reste statique (caméra fixe). Cet exemple est un cas difficile (mauvaise orientation et main perturbatrice). La figure III.21.b, montre l'évolution temporelle de l'orientation du mouvement dominant pour cette séquence. On constate que les orientations dominantes sont celles associées successivement aux contours horizontaux (90°, entre les images 371 et 398, mouvement vertical (tilt)) puis aux contours verticaux (180°, entre les images 400 et 463, mouvement horizontal (pan)). On remarque la présence d'une erreur d'estimation plus importante autour de l'image 415, celle-ci étant due au mouvement parasite temporaire de la main qui ne suit pas exactement le mouvement du visage. La figure III.21.c présente l'évolution temporelle de l'énergie totale du spectre log polaire des contours mobiles pour cette même séquence. Elle montre que cette énergie évolue périodiquement avec des maxima et minima à intervalles réguliers. La période, correspondant aux

alternances de mouvement est ici de l'ordre de 6Hz (c'est à dire 3 allers-retours par seconde). L'algorithme interprète les hochements de manière correcte même si les conditions d'analyse ne sont pas favorables. La reconnaissance n'est effective qu'après les *n=4* hochements réguliers détectés.

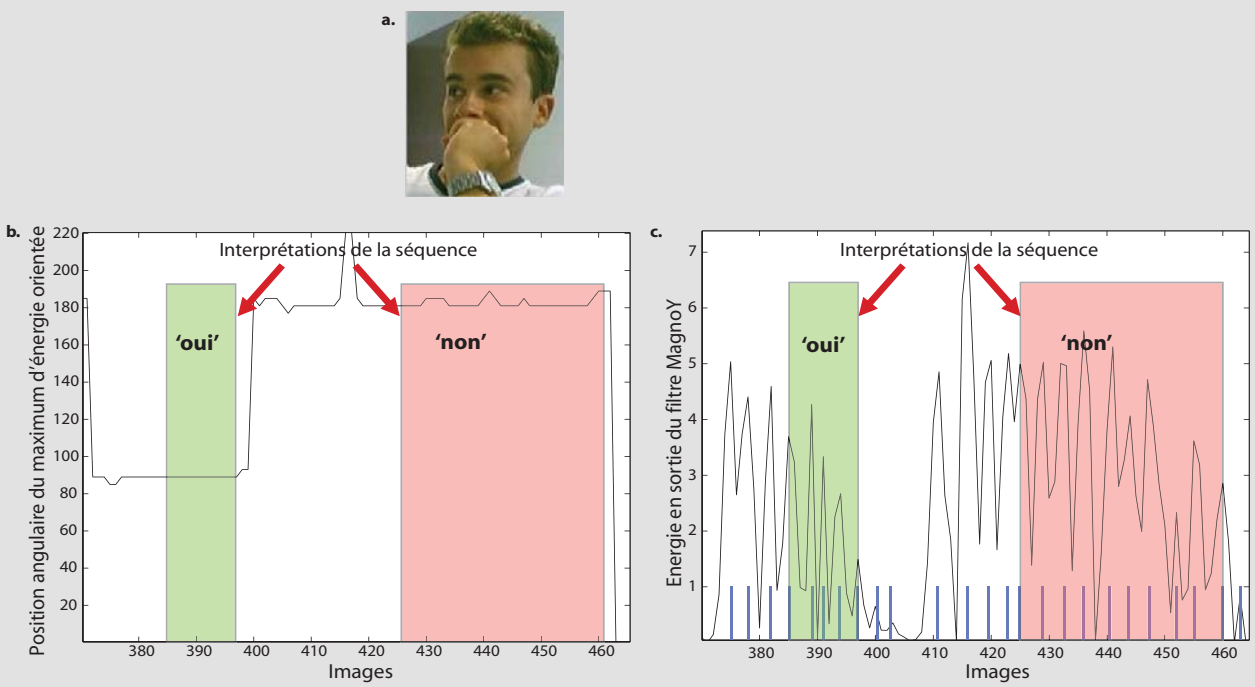


Figure III.21: séquence de hochements de tête d'approbation et de négation. a, extrait de la séquence. b, évolution temporelle de l'orientation du mouvement dominant. c, évolution temporelle de l'énergie du spectre en sortie du filtre MagnoY contours mobiles.

Cet algorithme a été testé sur une base vidéo contenant 415 séquences de hochements de tête (210 d'approbation et 205 de négation). Les sujets devaient exprimer "Oui" ou "Non" de façon libre. La figure III.22 montre différents extraits de la base de test, on y trouve des séquences de synthèse et des séquences vidéo à différentes résolutions (visages de taille 80*80 pixels à 200*200 pixels) dans différentes conditions d'éclairage (standard, tamisée ou bruitée). Ces vidéos ont été acquises dans deux cadres différents: expérimentations contraintes (individu devant la caméra et exécutant des séquences de mouvement) ou dans des phases de conversation libre en scènes d'intérieur. On évalue par ailleurs les performances dans le cas où le visage est occulté à 50% et lorsque l'axe de la caméra forme un angle de 45° latéral par rapport au visage en vue de face.

Figure III.22: extraits de la base de tests pour l'évaluation du détecteur de hochements de tête

Les performances sont présentées dans la table III.10. Les taux de bonnes détections sont satisfaisants dans le cas idéal d'un cadrage correct du visage. Ces performances restent également correctes lorsque le cadrage n'est plus centré sur le visage. Les taux de fausses alarmes et d'oublis sont principalement dus à un manque d'information sur les contours dans la direction du mouvement en cas de conditions de tests (mauvais cadrage: analyse sur 50% du visage et caméra orientée à 45° latéral par rapport à l'axe du visage).

Table III.10 : performance du détecteur de hochement de tête

	Taux de succès	***Taux de fausse alarme***	***Taux d'oubli***
Analyse globale de la tête	95%	2%	3%
Analyse de 50% de la tête	80%	5%	15%
Orientation de la caméra à 45° de l'axe du visage	82%	5%	13%

Ces hochements de tête sont difficiles à interpréter avec des méthodes classiques de flot optique ou des méthodes de suivi de points caractéristiques, car les mouvements peuvent être très rapides ou au contraire de faible vitesse. De plus, des événements perturbateurs peuvent introduire une erreur d'estimation, comme par exemple, la présence d'une main devant la bouche qui peut produire des mouvements parasites. Aussi, les méthodes de calcul par flot optique sont peu adaptées à une tâche de ce type. Ceci explique le fait que l'on n'ait pas utilisé le module flot optique pour cette analyse.

Cet algorithme est mis en oeuvre sous Matlab, sur un ordinateur de bureau standard (Pentium IV, 3.0GHz) équipé d'une webcam. Il fonctionne à la cadence de 30 images par seconde pour des images de taille 320*240 pixels. La figure III.23 présente l'interface graphique développée, on retrouve à gauche le flux vidéo entrant et à droite l'interprétation associée. Aucune contrainte n'est imposée à l'utilisateur. Néanmoins, dans la phase d'initialisation, une absence de mouvement est requise durant la première seconde de façon à évaluer le niveau de bruit moyen présent dans la scène. Ce niveau de bruit moyen étant nécessaire à l'initialisation du module de détection de mouvement.

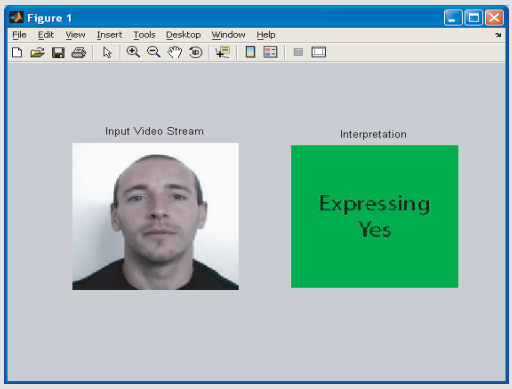


Figure III.23: interface graphique de l'application de détection d'approbation et de négation.

III.7. Intégration dans un système de détection de stress et d'hypo-vigilance chez les conducteurs

III.7.1. Présentation

Les différents modules d'analyse des mouvements du visage ont été utilisés lors des deux premières éditions du Worshop Enterface en 2005 et en 2006 [EnterfaceSite] dans un projet visant à créer un système de détection de l'hypovigilance au volant.

Le workshop eNTERFACE est une émanation du réseau d'excellence européen SIMILAR [SimilarSite]. Il a été créé dans le but d'établir des collaborations concrètes de recherche et de développement entre chercheurs, doctorants et étudiants d'origines variées en les réunissant pendant 4 semaines dans un même lieu afin de travailler autour d'un projet commun. Les participants sont regroupés en équipes rattachées à des projets spécifiques liés aux interfaces multimodales et aboutissant à des applications sous licence libre. Le premier workshop s'est déroulé en 2005, en Belgique (Mons), le second s'est déroulé en 2006 en Croatie (Dubrovnik).

Dans le cadre de ces deux workshops, nous avons travaillé au développement d'un simulateur de conduite augmenté. L'objectif est de pouvoir estimer sur la base de l'analyse d'informations vidéo et d'informations physiologiques des états de stress et d'hypovigilance chez un conducteur. En retour, des alertes appropriées sont déclenchées pour informer le conducteur de son état. Ce projet, en réunissant des chercheurs des communautés Traitement du Signal et Interaction Homme-Machine, vise à tester la faisabilité d'une phase d'estimation multimodale (stress et hypovigilance) qui lie des mesures visuelles et biologiques au travers une plate-forme à interactions utilisateur-machine également multimodales.

Plus précisément, on utilise dans ce projet différents modules de traitement appelés composants. On trouve:

→ un composant d'analyse vidéo basé sur les algorithmes décrits précédemment. Ce composant génère un premier diagnostic sur l'état de fatigue du conducteur. Pour ce faire, il extrait trois indices caractéristiques d'une hypovigilance: présence de bâillements, augmentation de la fréquence des clignements et mouvements de la tête. Les deux premiers indices donnent une information sur l'état de fatigue du conducteur. Le dernier

indice donne plutôt une information sur son état de concentration vis-à-vis de la route.

→ un composant de mesures de paramètres biologiques tels que le rythme cardiaque (ECG) et la sudation (GSR) qui informent sur l'état de stress du conducteur.

→ un composant de prédiction des risques d'accident basé sur l'analyse des clignements d'yeux.

Ces composants sont réalisés sous forme de modules indépendants écrits dans un langage donné (C/C++, Matlab, etc.) qui communiquent entre eux grâce à la plate-forme OpenInterface [OpenInterfaceSite]. Celle-ci est basée sur une description de chaque composant sous forme XML. Cette plate-forme permet alors une transmission d'informations entre différents composants de façon unifiée et transparente et permet de faire communiquer différents programmes écrits dans des langages différents. Cet outil est couplé à la plate-forme ICARE [Bouchet04] qui permet de gérer la fusion des données et les interactions multimodales entre la machine et l'utilisateur. Ce composant conceptuel permet alors un retour adapté du système vers le conducteur en fonction de son état d'hypovigilance et de son stress.

Architecture matérielle du projet

Le système utilise le simulateur de conduite (de type jeux vidéo) "Opensource" (code source libre) TORCS [TorcsSite] qui permet à l'utilisateur de se plonger dans un environnement de conduite en toute sécurité. Ce simulateur a été modifié de façon à pouvoir diffuser des messages textuels sur l'interface tel que le ferait un système de visualisation tête haute (Head Up Display, HUD).

Des hauts parleurs, un volant à vibrations contrôlables ainsi qu'un miniécran secondaire ont été ajoutés de manière à obtenir différentes modalités d'interaction machine/utilisateur. Une caméra couplée à un ordinateur permet de filmer et d'analyser le visage du conducteur afin de détecter des signes caractéristiques d'une hypovigilance. Un système de capteurs biologiques multiples BIOPAC [BiopacSite] permet d'enregistrer des paramètres physiologiques pour estimer l'état de stress.

La figure III.24 montre les différents éléments du système: le simulateur est projeté sur un écran large ou écran d'ordinateur portable tandis que le conducteur fait face à cet écran, les mains sur le volant, le visage face à la caméra.

Figure III.24: illustration du système de détection d'hypovigilance chez le conducteur utilisé lors des workshop eNTERFACE 2005 et 2006.

Architecture logicielle du système

Le logiciel est articulé autour d'une architecture à composants ICARE et OpenEnterface communicant grâce à OpenEnterface. On trouve tout d'abord les composants générant des alarmes sur l'état de fatigue du conducteur et les risques d'hypovigilance. Les autres composants d'entrée sont un microphone et une série de boutons sur le volant qui permet de jongler entre les différentes modalités de sortie de manière à ajuster les alertes de la façon la plus appropriée pour un utilisateur donné.

Les modalités de sortie sont décrites sous forme de composants dont voici les différents types:

→ un composant d'affichage de messages sur écran (Head Up Display).

→ un composant gérant l'affichage d'alertes sur le miniécran secondaire.

→ un composant de génération d'alarmes sonores

→ un composant de génération et de contrôle des vibrations du volant pour réveiller le conducteur assoupi.

Les données issues des composants d'entrée sont prises en charge par l'architecture à composants ICARE capable de gérer en temps réel les différents flux d'informations et de commandes. Plus de détails sont disponibles dans [DriverSimulator05, DriverSimulator06].

II.7.2. Analyse visuelle de l'état de fatigue du conducteur

L'analyse vidéo est réalisée grâce aux algorithmes présentés au paragraphe III.2-6. La figure III.25 montre l'architecture retenue. Son objectif est de générer des alarmes liées à des événements détectés au niveau du visage. Trois événements différents sont considérés: les bâillements, des fermetures longues des yeux et des mouvements de tête de grande amplitude. Chacun de ces critères génère une alarme qui est redirigée

grâce à OpenInterface vers le module de fusion de données plus haut niveau non représenté sur cette figure. Une liste des durées des derniers clignements est également extraite pour une utilisation par le module de prédiction des risques d'accident.

Plus précisément, ce système d'analyse vidéo a été découpé en deux composants distincts. Le premier, le composant "Rétine" est constitué des filtres Parvo contours et MagnoY contours mobiles. Il donne pour chaque image trois sorties distinctes:

→ Une image en niveau de gris de luminance localement corrigée (cf. module de compression logarithmique inclus dans le filtre Parvo contours).

→ Une image de contours (cf. sortie Parvo ON-OFF du filtre Parvo contours).

→ Une image de contours en mouvement (cf. sortie du filtre MagnoY contours mobiles).

Ces différentes sorties sont prises en charge par les modules du second composant "Détection de l'état de fatigue" (cf. fig. III.25). Il est composé d'un module de détection de visage et des méthodes de localisation des yeux et de la bouche (cf. III.2). Quatre analyseurs spectraux indépendants extraient des informations sur l'état et le mouvement de chaque oeil et de la bouche ainsi que les mouvements liés au visage dans son ensemble. En parallèle, le module d'analyse de flot optique décrit précisément l'orientation et la direction des mouvements associés à la tête. Enfin, des alarmes sont générées selon les différentes informations recueillies. Les paragraphes suivants décrivent plus précisément chaque module de ce système.

Détection du visage et des yeux et stabilisation temporelle de leur position

Le premier module du composant "détection de l'état de fatigue" permet de localiser le visage du conducteur. Une fois le visage détecté, les zones des yeux et de la bouche sont extraites pour les traitements ultérieurs. Pour cela, nous utilisons les méthodes présentées en section III.2. Comme nous travaillons sur un flux vidéo et pour limiter les instabilités entraînées par une détection instable de la boite englobante du visage, nous proposons d'effectuer un filtrage temporel de la position spatiale de la boite englobant le visage et des boites englobant les yeux. Pour cela, chaque coordonnée $(x(t), y(t))$ d'une boite englobante à l'instant t est filtrée grâce à l'équation (III.5), et la boite englobante est placée à la coordonnée $(\mu_x(t), \mu_y(t))$ selon un filtrage de moyenne adaptatif. L'adaptation temporelle est paramétrée par α fixé à 0.7.

$$\left| \begin{array}{l} \mu_x(t) = \alpha.\mu_y(t-1) + (1-\alpha).x(t) \\ \mu_y(t) = \alpha.\mu_y(t-1) + (1-\alpha).y(t) \end{array} \right. \qquad \text{(Eq. III.5)}$$

Détection des mouvements du visage et synthèse d'alarmes

Une fois le visage détecté, deux modules travaillent en parallèle: un estimateur de flot optique et un ensemble d'analyseurs spectraux locaux. Ils ont été décrits dans les paragraphes III.3 à III.6. et permettent de générer des alarmes dès lors que:

→ les yeux se ferment plus de 0.5 seconde.

→ un bâillement est détecté.

→ un mouvement de tête dont la durée est supérieure à 1 seconde est détecté.

Plusieurs de ces phénomènes peuvent être détectés en même temps.

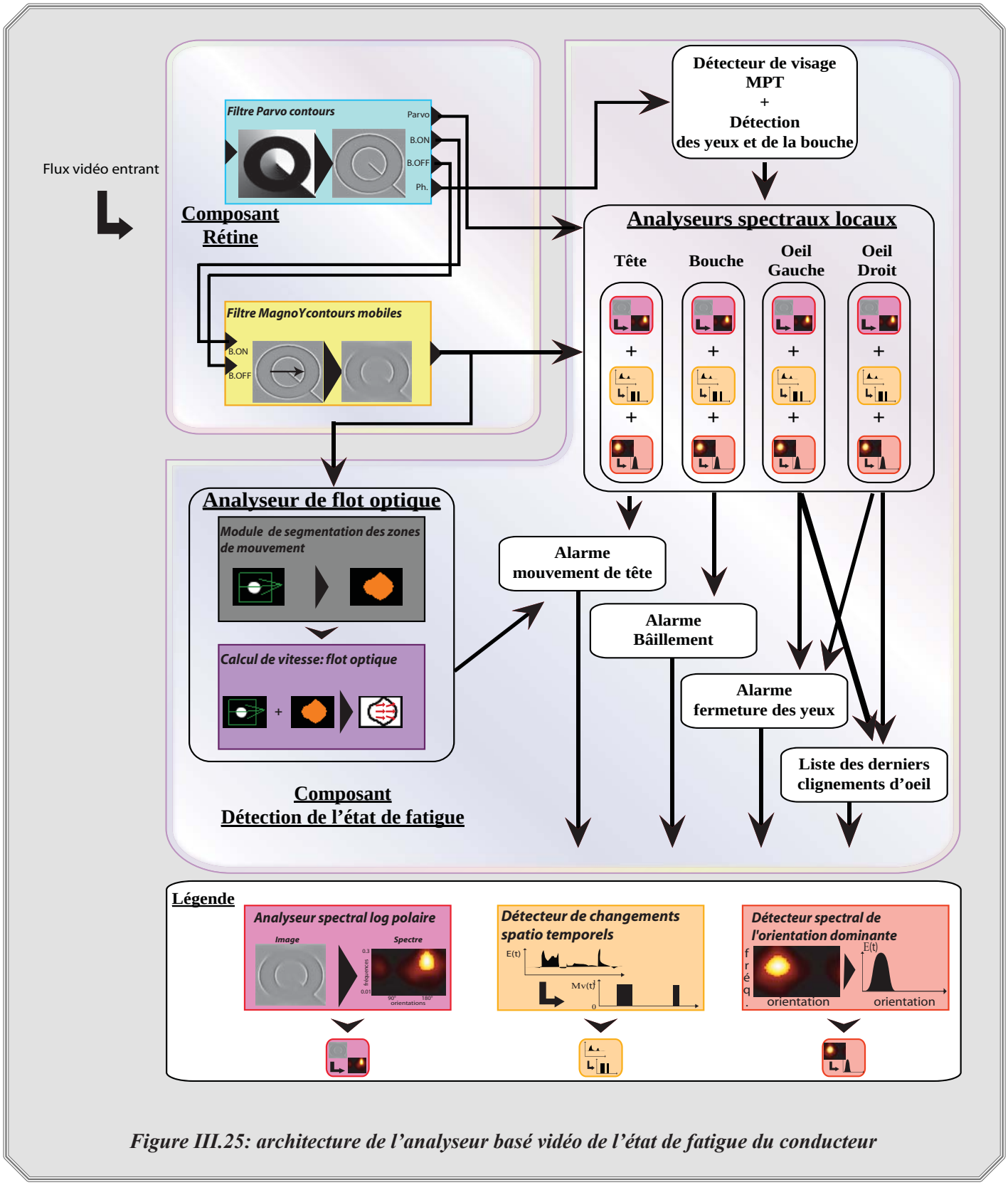


Figure III.25: architecture de l'analyseur basé vidéo de l'état de fatigue du conducteur

Ces alarmes sont envoyées au composant ICARE qui les fusionne en une réponse adaptée via les différentes modalités de sortie (vibration du volant, alarme sonore et visuelle). Par exemple, une alerte de fermeture des yeux génère une alarme sonore ainsi que l'activation des vibrations sur le volant de façon à réveiller le conducteur assoupi. Une alarme liée à un bâillement entraîne une alarme visuelle sur les deux écrans ainsi qu'une alarme sonore.

Ce système montre la faisabilité d'une telle architecture. Les différents composants du système complet (analyse visuelle, OpenInterface, ICARE etc.) sont synchronisés en temps réel et l'utilisateur n'est soumis à aucune contrainte mis à part lors de la première seconde de test pour laquelle il doit rester immobile (initialisation des détecteurs de mouvement).

Actuellement, la vitesse de traitement de 12 images par seconde limite les possibilités d'analyse, au niveau des mouvements des yeux en particulier.

III.7.3. Extensions à envisager

Du point de vue du système d'analyse visuelle, certains points seraient intéressants à approfondir:

→ Il serait intéressant de pouvoir analyser les mouvements faciaux en même temps que les mouvements de tête plutôt que de les analyser indépendamment. Pour cela, un autre système de fusion des informations de mouvement doit être envisagé.

→ L'utilisation d'un système d'acquisition vidéo plus rapide permettrait de renforcer et de mieux décrire les analyses du mouvement des yeux et de la bouche notamment.

III.8. Intégration dans un système d'apprentissage de la langue des signes

Lors du workshop eNTERFACE 2006 [EnterfaceSite], nous avons également intégré nos algorithmes d'analyse des mouvements de tête dans une application d'apprentissage de la langue des signes [SignLanguageTutoringTool06]. Cette application permet à un utilisateur d'apprendre ce langage par visualisation de gestes de référence qu'il doit mémoriser et reproduire de la façon la plus fidèle possible. Le système enregistre et analyse la séquence de mouvements de l'élève. Si la séquence de mouvements n'est pas correctement reproduite, l'application indique à l'utilisateur ce qu'il doit corriger. L'algorithme utilisé est capable de reconnaître des mouvements complexes de la main et de la tête. En effet, la langue des signes a pour particularité le fait que le message est transmis par le mouvement des mains et par différents mouvements de têtes et expressions faciales ce qui en fait un langage multimodal. Pour un même geste manuel, des mouvements de tête différents peuvent aboutir à un message différent. Comme le montre la figure III.26, le mot "Here" ("ici") met en jeu un mouvement circulaire de la main droite. Cependant, différents mouvements de tête aboutissent à une signification différente, à savoir:

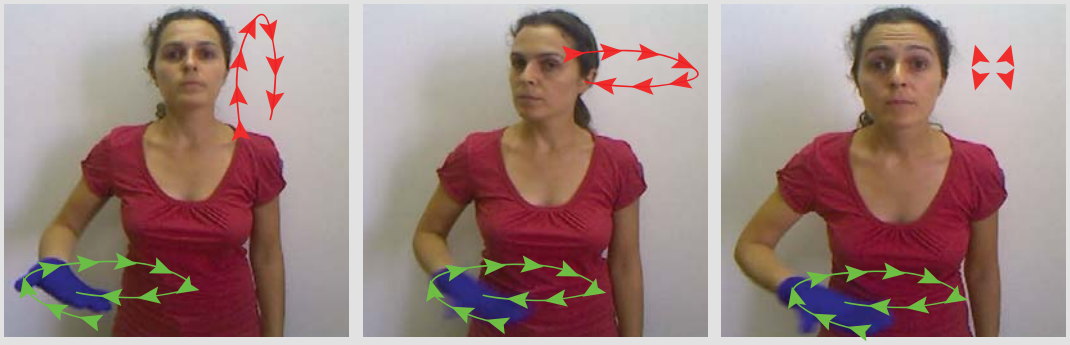
→ Un hochement de tête positif exprime "is here" ("est ici").

→ Un hochement de tête négatif exprime "not here" ("n'est pas ici").

→ Un mouvement de tête vers l'avant avec écarquillement des sourcils exprime la forme interrogative

de “here” (“est-ce/il ici?”).

Dans cette application, la trajectoire des mains est analysée par utilisation de HMM (Hidden Markov Models) après une phase de segmentation de chacune de celles-ci [Aran05]. Leur trajectoire pour un geste donné suit un motif reconnaissable qui justifie l’utilisation de HMM. Les algorithmes de classification utilisés, décrits dans [Aran05] donnent jusqu’à 99% de reconnaissance sur une base de signes dont l’information est entièrement contenue dans la gestuelle des mains. En revanche, les taux de reconnaissance chutent à 67% lorsqu’il s’agit de reconnaître des signes caractérisés conjointement par une gestuelle des mains et des mouvements de la tête ou du visage.



a. b. c.

Figure III.26: exemple de signes. Pour un même geste de la main, 3 mouvements de tête différents entraînent 3 messages différents. a, “ici” affirmatif (est ici), hochement de tête positif. b, “ici” négatif (n’est pas ici), hochement de tête négatif. c, “ici?” (est-ce ici?), mouvement de tête vers l’avant et expression interrogative

III.8.1. Description de l’application développée

Dans ce projet a été mis en place un système de reconnaissance de 19 signes dont certains mettent en jeu à la fois des mouvements des mains et de la tête. L’originalité du système proposé réside dans la prise en compte et la fusion des deux modalités. La base est constituée de 19 signes reproduits par 8 personnes différentes, chacune répétant chaque signe 5 fois. La table III.11 donne une description des différents signes considérés.

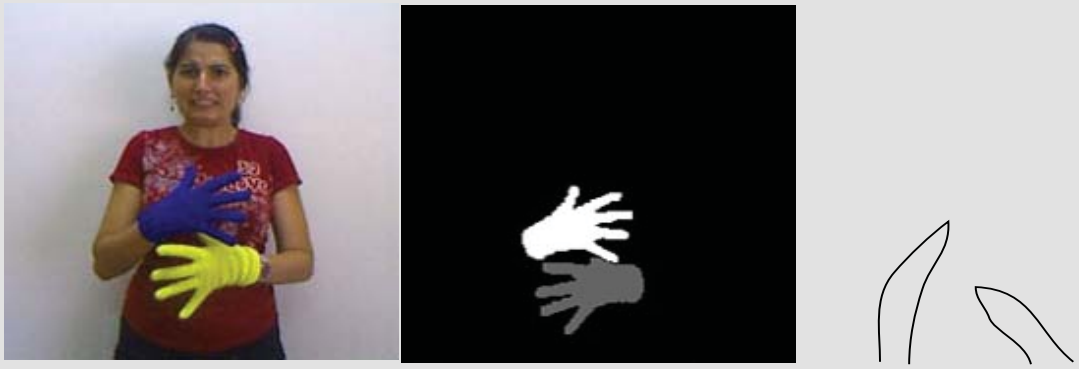
Table III.11: présentation des différents signes de la base de test

Signe	***Mouvement de tête***	***Mouvement de main***
"here"/ est ici	hochement affirmatif	mouvement circulaire de la main droite, parallèle au sol
"is here"/ est-(il/elle/ce) ici ?	mouvement de tête vers l'avant	
"not there"/ n'est pas ici	hochement négatif	
"clean"/ propre	-	main droite face vers le bas effleurant la main gauche tournée face vers le haut
"very clean"/ très propre	lèvres fermées+rotation latérale rapide de la tête	
"afraid"/ effrayé	-	les mains se rejoignent au niveau du buste, main gauche sous la main droite
"very afraid"/ très effrayé	mouvement oscillant de la tête rapide+ expressions faciales (lèvres et yeux ouverts)	
"fast"/ vite	-	les 2 bras tendus face au visage se replient et les mains se rapprochent du buste, les mains sont partiellement fermées
"very fast"/ très vite	mouvement "vibratoire" de la tête rapide+ expressions faciales (lèvres et yeux ouverts)	
"drink"/ boire (verbe)	mouvement vers le haut	mouvement propre à l'action de boire, au niveau de la bouche
"drink"/ boire (nom)	-	mouvements de poignets répétitifs au niveau de la bouche
"open door"/ ouvrir la porte	-	mains au niveau du visage, paumes face à la caméra, la main gauche fait un mouvement d'ouverture
"door open"/ porte ouverte	-	mouvement identique au précédent, mais répété
"study" étudier	-	main gauche à plat face vers le haut, main droite, au dessus, orientée perpendiculairement et immobile + mouvements de doigts
"study continuously"/ étudier continuellement	mouvements de tête circulaires	main gauche à plat face vers le haut, main droite perpendiculaire, ouverte en mouvements circulaires verticaux + mouvements de doigts
"study regulary"/ étudier régulièrement	mouvements oscillants verticaux	main gauche à plat face vers le haut, main droite perpendiculaire en mouvement vertical oscillant
"look at"/ regarde	-	les doigts au niveau des yeux s'éloignent du visage vers l'avant
"look at continuously"/ regarder continuellement	mouvements de tête circulaires	les doigts au niveau des yeux s'éloignent du visage vers l'avant de façon répétée, mais circulaire verticale
"look at regulary"/ regarder régulièrement	mouvements oscillants verticaux	les doigts au niveau des yeux s'éloignent du visage vers l'avant de façon répétée

III.8.1.1. Analyse du mouvement des mains

Chaque main porte un gant de couleur différente et saillante ce qui facilite les problèmes d'occultations et de suivi. Les mains sont extraites grâce à une phase de segmentation couleur suivie d'un filtrage de Kalman [Aran05]. Un HMM analyse les informations de trajectoire et en déduit le signe associé aux gestes de la main par comparaison à la trajectoire typique des gestes appris lors de la phase d'apprentissage.

La figure III.27.a montre une image de séquence de gestes, III.27.b montre la segmentation des mains et la figure III.27.c donne la trajectoire normalisée du centre de gravité analysée par le HMM.



a. b. c.

Figure III.27: exemple de séquence de langage des signes: a, image de la séquence. b, image de segmentation des mains après segmentation couleur. c, image de la trajectoire des mains

III.8.1.2. Analyse des mouvements de tête

Une analyse experte des signes considérés montre qu'ils peuvent être catégorisés selon les critères suivants:

→ présence d'un mouvement détecté ou non.

→ s'il y a un mouvement de tête, son orientation est critique pour différencier deux signes utilisant la même gestuelle manuelle (par exemple "here" et "not there").

→ certaines expressions faciales apparaissent comme éléments permettant de différencier deux signes (par exemple, "fast" et "very fast").

Construction de l'analyseur de visage

En vertu des critères énoncés précédemment, l'analyseur de visage est construit selon l'architecture présentée en figure III.28.

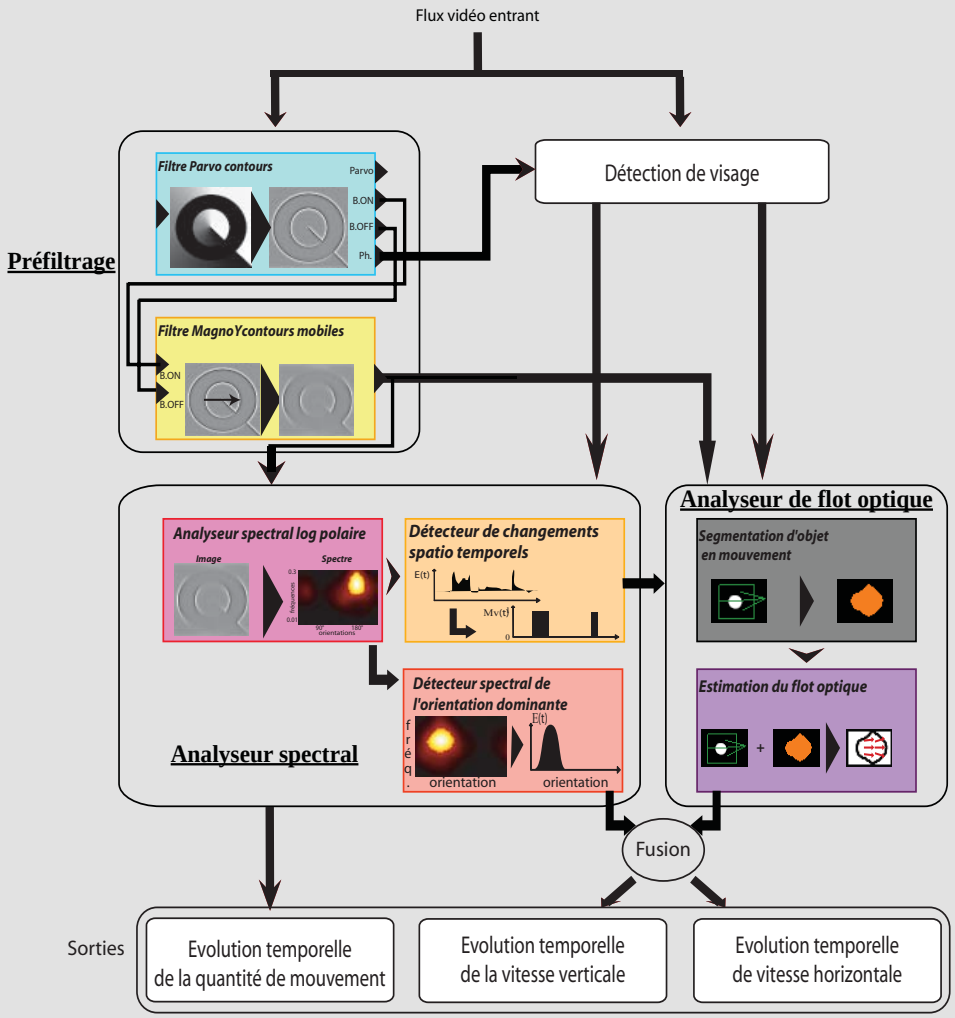


Figure III.28: système d'analyse des mouvements de tête utilisé pour l'analyse du langage des signes: les flèches indiquent les liens entre les blocs

A partir de chaque image du flux vidéo analysé, deux modules fonctionnent en parallèle. Le premier est le pré filtrage rétinien. De celui-ci est extraite l'information de contours en mouvement (sortie du filtre MagnoY contours mobiles). Le module fonctionnant en parallèle est le détecteur de visage. Celui-ci donne les coordonnées dans chaque image de la position du cadre englobant le visage et permet de concentrer l'analyse sur cette zone.

Une fois le visage repéré, la sortie du filtre MagnoY contours mobiles est prise en charge par l'analyseur spectral qui permet à la fois de détecter les alertes de mouvement et d'en analyser l'orientation. Un module de segmentation de zones en mouvement et un module d'analyse de flot optique sont utilisés sur la zone du visage de manière à donner une description supplémentaire de l'orientation du mouvement (redondance de l'information) et de direction.

De cet algorithme nous extrayons trois informations:

→ une information sur la quantité de mouvement: l'énergie totale du spectre en sortie de filtre MagnoY contours mobiles.

→ les composantes horizontale et verticale du vecteur vitesse associé aux mouvements de tête.

La figure III.29 montre l'évolution des différentes données considérées pour deux gestes différents: "ici" affirmatif (Here) et "très propre" (Very clean). Pour le premier signe, l'utilisateur effectue un hochement de tête d'approbation ce qui conduit à une vitesse horizontale nulle et une variation périodique de la valeur de la vitesse verticale. Cet enchaînement de mouvement se retrouve également sur l'évolution temporelle de l'énergie totale en sortie du filtre MagnoY contours mobiles qui donne une image de l'amplitude et de la période du mouvement. Le geste "très propre" implique quant à lui une alternance simple de mouvement latéral opposé (un grand mouvement vers la droite puis un retour de la tête face à la caméra par un mouvement vers la gauche). Plus précisément, le grand mouvement vers la droite est précédé d'un court et rapide mouvement vers la gauche de façon à amplifier ce grand mouvement (cf. une grande amplitude sur la courbe d'énergie pour ces deux gestes, mais une valeur de vitesse opposée). Après l'exécution du grand mouvement vers la droite, le visage revient doucement vers la caméra d'où une vitesse plus faible (et de signe opposé) et donc une énergie également plus faible.

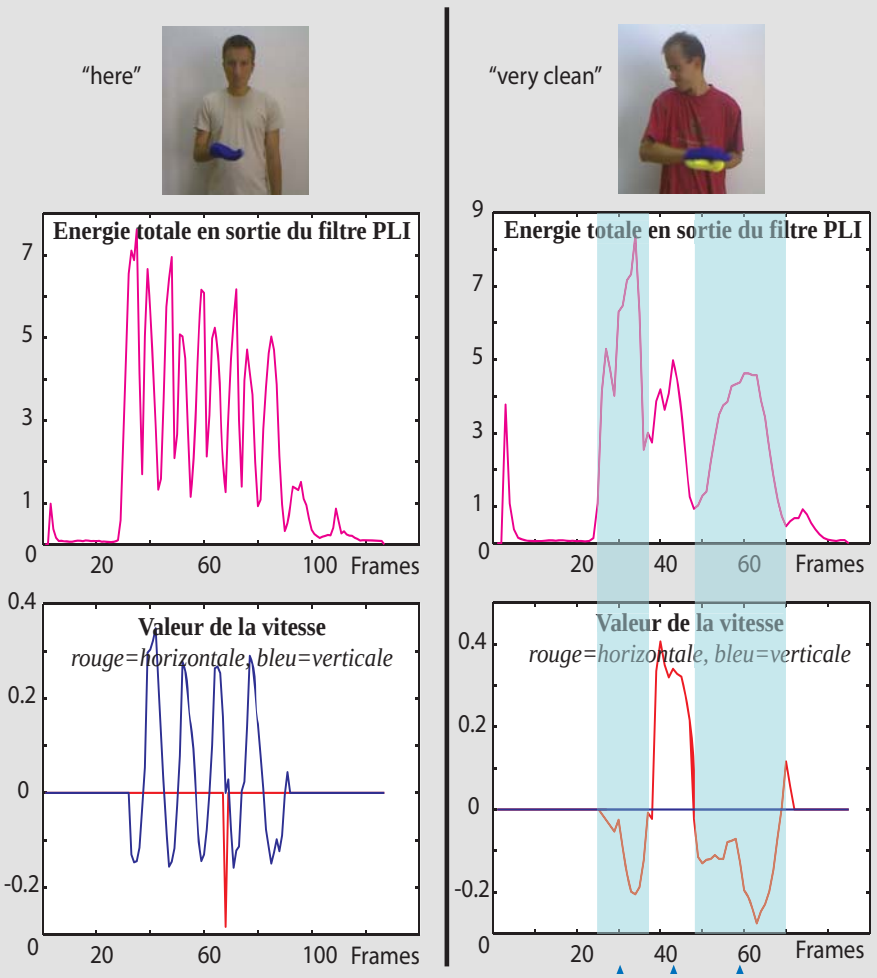


Figure III.29: exemple d'évolution temporelle des indicateurs de mouvement issus de l'analyseur de visage dans le cas de deux gestes de langage des signes: "ici" et "very clean"

Intégration des données issues de l'analyseur des mouvements de tête dans le système complet

La reconnaissance multimodale (gestes manuels+faciaux) se fait par utilisation de HMM [Aran06], le but est de reconnaître un motif caractéristique dans l'évolution temporelle des différents indicateurs appliqués en entrée. Ainsi, les données de mouvement de tête sont intégrées dans le vecteur de caractéristiques associé à chaque geste. La reconnaissance atteint alors un taux de reconnaissance de 85% de réussite pour des gestes multimodaux soit 18% de plus que le système n'intégrant pas les gestes associés à la tête.

L'ajout de l'information de mouvements de tête permet une identification plus précise des gestes tels que "Here", "Clean" et "Fast". On remarque cependant que certains gestes tels que "Study", "Loot at" et leurs variantes posent quelques problèmes. Ceci est principalement dû au fait que pour ces gestes, les mains se placent devant le visage et peuvent fausser l'analyse des mouvements de tête. Aussi, et c'est là le point le plus important, pour ces gestes, seuls les mouvements manuels portent le message. Néanmoins, l'analyse 2D de leur trajectoire n'est pas suffisante, car cette trajectoire est réalisée dans l'axe de la caméra. Aussi, l'ajout de la troisième dimension (la profondeur) permettrait d'améliorer l'analyse [Aran05].

III.8.2. Logiciel proposé

Un logiciel d'apprentissage de la langue des signes a été réalisé. Il permet à un utilisateur isolé de visualiser le geste à effectuer, de le réaliser lui même et après une phase d'évaluation de la qualité de son geste, il lui est proposé de revisualiser son geste via un avatar 3D. La figure III.30 montre l'interface du logiciel. Il peut être téléchargé sur le site www.enterface.net.

Figure III.30: interface du logiciel d'apprentissage de la langue des signes. En haut à droite, fenêtre de visualisation du geste de référence que l'utilisateur doit reproduire. En bas à gauche, fenêtre de visualisation du geste reproduit par l'utilisateur et à droite, geste reproduit par un avatar 3D. En haut à gauche, l'utilisateur est informé de la qualité du geste réalisé.

III.8.3. Extensions à envisager

Ce système est un prototype qui pourra être amélioré sur divers points notamment la richesse des signes ainsi que les méthodes d'analyse:

→ Enrichissement de la base de gestes: lors de la création de cet outil d'apprentissage, seuls 19 signes ont été pris en compte. Une extension de la base de gestes apparaît comme essentielle de façon à rendre cet outil réellement utilisable à des fins d'enseignement de ce langage.

→ Enrichissement des paramètres analysés: actuellement ne sont traités que les mouvements de tête. L'analyse des expressions faciales permettrait de renforcer la reconnaissance de certains mouvements tels que "Very afraid" et "Very fast" qui en plus d'un mouvement de tête sont caractérisés par une expression faciale caractéristique.

VII. Conclusion du chapitre

Nous venons de voir dans ce chapitre qu'à partir d'un assemblage des différents modules issus de la modélisation du système visuel, il est possible de créer des systèmes haut niveau de traitement d'image. Cet assemblage permet l'extraction de signaux simples à interpréter, ainsi, les algorithmes de fin de chaîne (la dernière couche d'algorithme permettant l'interprétation vis-à-vis de l'application visée) sont simplifiés et aucun posttraitement n'est nécessaire. L'intérêt de l'utilisation de ces algorithmes dans deux applications concrètes (détecteur d'hypovigilance et système d'apprentissage de la langue des signes) a été démontré.

Chapitre IV: Suivi et identification d'objets

Dans ce chapitre, nous allons voir comment construire un système de suivi d'objets en mouvement ainsi qu'une méthode d'identification et de catégorisation d'objets. Nous présenterons tout d'abord le système de suivi, basé sur les informations issues des modélisations de la rétine et du cortex visuel primaire. En seconde partie de ce chapitre, nous proposerons, une méthode d'identification d'objets basée sur la modélisation de l'analyse fréquentielle réalisée au niveau du cortex visuel primaire. A terme, ces deux systèmes (suivi et identification) pourront être fusionnés de façon à créer un système d'analyse d'objets pouvant à la fois, suivre, classer et identifier tout objet en mouvement, voire le reconnaître de nouveau après une occultation totale.

IV.1. Suivi d'objet

IV.1.1. Algorithme proposé

Le système de suivi proposé est basé sur une approche contours et contours en mouvement. Nous partons du principe que les objets en mouvement ont été préalablement segmentés à chaque image de la séquence vidéo analysée et nous greffons alors un système de suivi "bio-inspiré". Ce système met en relation les zones segmentées entre deux images successives et permet de s'affranchir des zones éventuelles de segmentation créées par le bruit d'acquisition.

Comme nous nous intéressons aux contours en mouvement, nous filtrons la scène visuelle avec les filtres Parvo contours et MagnoY contours mobiles. A partir de ces deux informations et des zones de segmentation préalablement extraites et étiquetées en composantes connexes [Haralick92], nous créons le système de suivi décrit sur la figure IV.1. Sur cette architecture, on trouve d'une part une analyse du flot optique de chaque zone segmentée qui permet le suivi des objets avec préservation de leur identifiant (ID). L'ID de chaque objet existe tant que l'objet est suivi. D'autre part, une analyse spectrale log polaire permet la caractérisation de la texture des zones segmentées dans le but de reconnaître des zones segmentées associées au bruit. Ce système d'identification du bruit met en évidence le fait qu'en cas de bruit, il n'y a pas de corrélation entre les contours extraits par le filtre Parvo contours et ceux extraits par le filtre MagnoY contours mobiles.

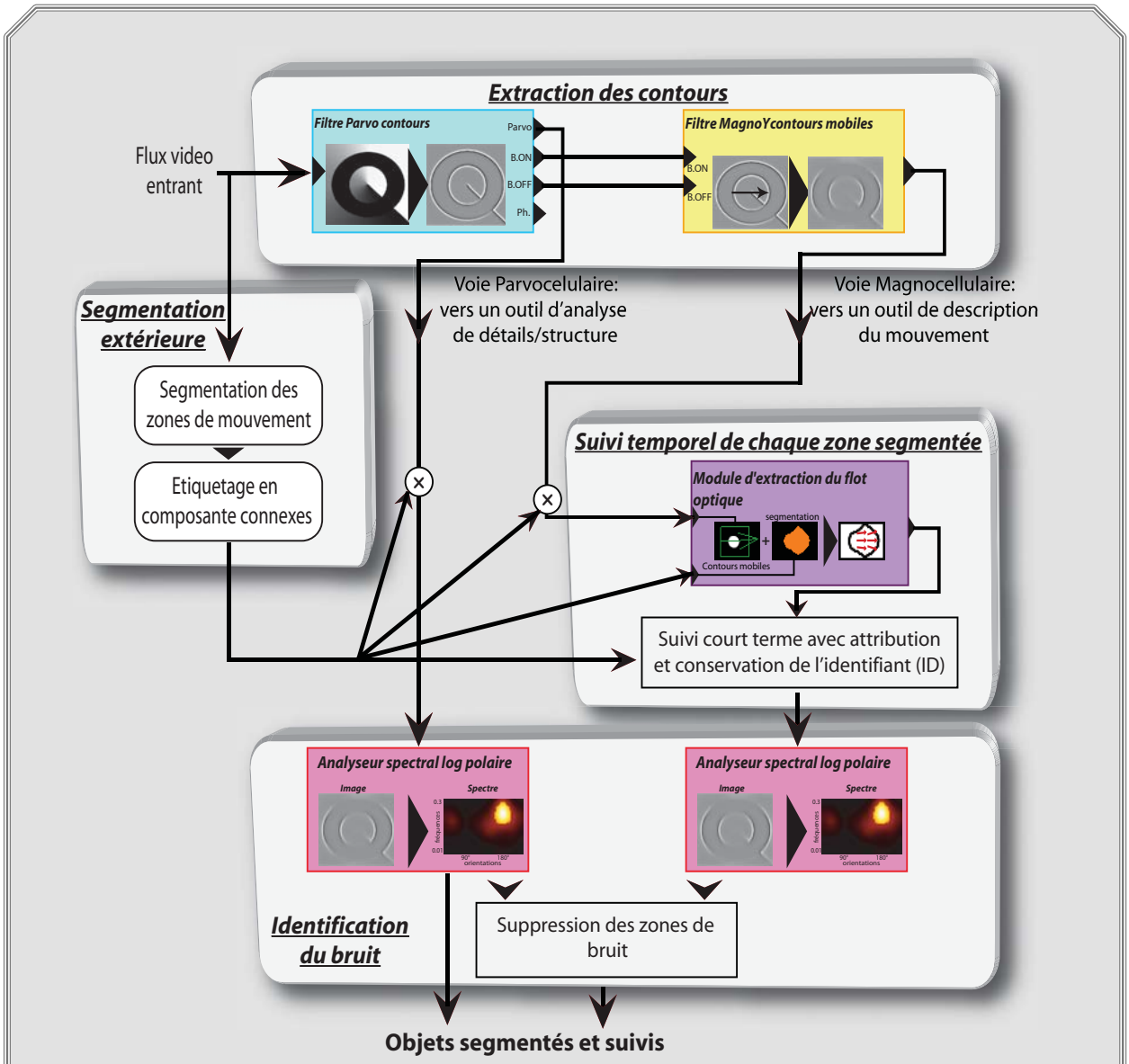


Figure IV.1: système de suivi, après l'étape de suivi proprement dite, une analyse spectrale permet d'identifier les zones segmentées du fait du bruit.

IV.1.2. Suivi court terme avec attribution et conservation de l'identifiant

IV.1.2.1 Principe

Le but est de créer un système de suivi permettant de mettre en correspondance les zones segmentées d'une image à la suivante. Comme décrit sur la figure IV.1, cet algorithme exploite l'information de flot optique des contours en mouvement. La figure IV.2 décrit plus précisément cette partie du système. Plusieurs informations sont à notre disposition sur chaque zone segmentée à chaque image:

→ taille de chaque zone.

→ position spatiale de son centre de gravité.

→ vecteur vitesse moyen de chaque zone.

On construit alors un système de mise en correspondance des zones segmentées aux instants *t* et *t-1* (cf. fig. IV.2). Cette mise en correspondance utilise un jeu de règles appelées «fonctions score» basées sur les caractéristiques de mouvement moyen, de taille et de position de chaque zone segmentée. Le paragraphe suivant décrit ces fonctions score considérées ainsi que le système de fusion permettant la mise en correspondance.

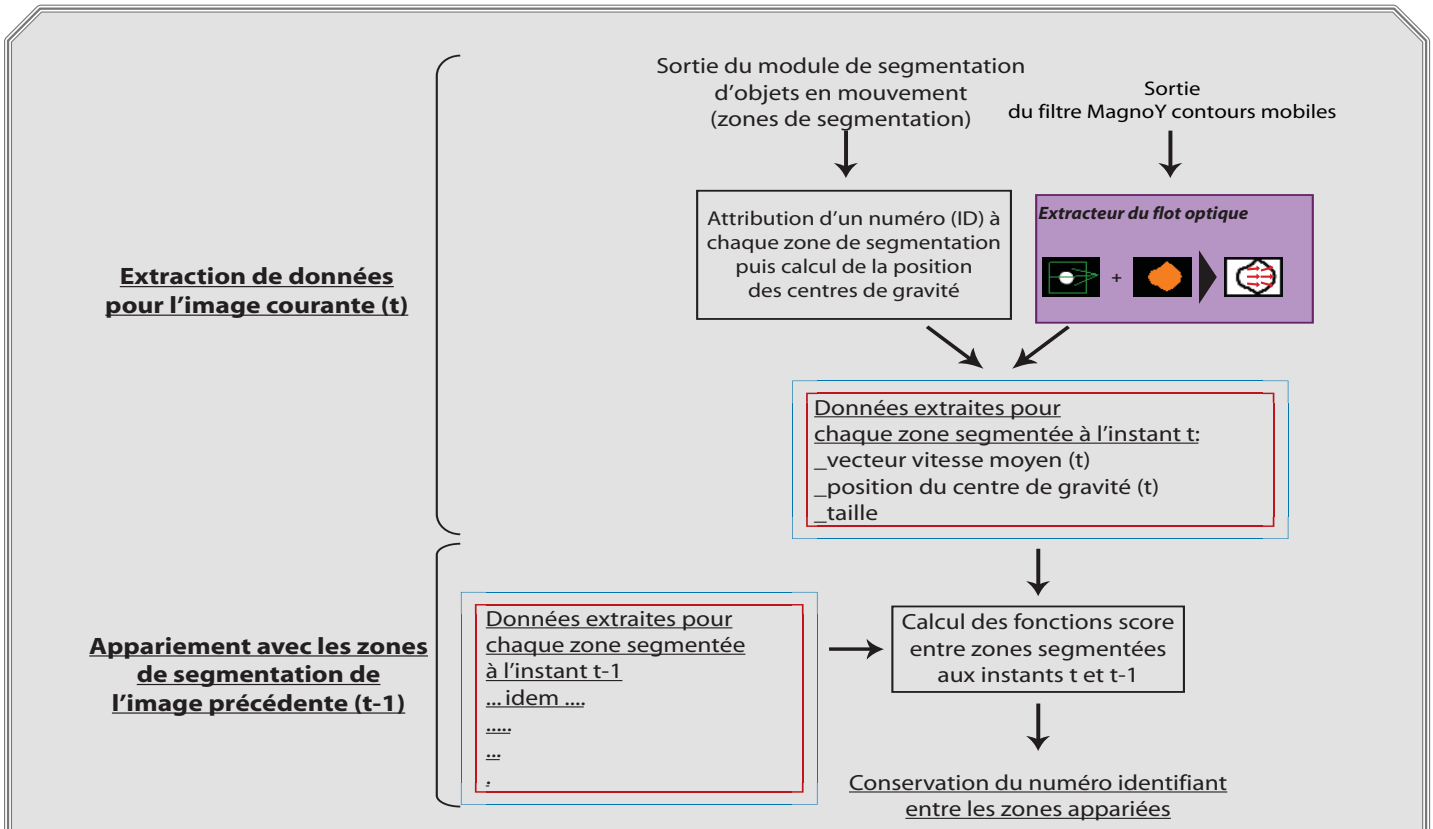


Figure IV.2: méthode d'extraction d'informations sur les zones segmentées de chaque image et mise en place d'un système d'appariement de chaque zone courante avec son correspondant dans l'image précédente. Lorsqu'une zone de l'image précédente est appariée à une de celles de l'image courante, son identifiant (ID) est conservé. Si une des zones de l'image courante n'est pas appariée, elle se voit attribuer un nouvel identifiant.

IV.1.2.2 Règles pour le suivi: définition des fonctions score

Régularité de la taille de la zone de segmentation

Etant donnée la cadence d'acquisition des images, on fait l'hypothèse que la taille des zones segmentées varie lentement d'une image à l'autre. Ce premier critère est tout de même à considérer avec précaution selon le type de mouvement et/ou les occultations. On élabore une fonction score de régularité de taille en se basant sur la quantité de pixels dans la zone segmentée. Soit $n_t(k)$ le nombre de pixels de la zone de segmentation *k* dans l'image *t* et $n_{t-1}(l)$, le nombre de pixels de la zone segmentée *l* de l'image *t-1*. On définit pour chaque

couple de zones segmentées *(k, l)* la fonction score $S^{t/t-1}_{taille}(k, l)$ de correspondance entre les deux zones *k* et *l* selon la relation:

$$S^{t/t-1}_{taille}(k,\ l) = \frac{\min\left(n_t(k), n_{t-1}(l)\right)}{\max\left(n_t(k), n_{t-1}(l)\right)} \qquad \textit{(Eq. IV.1)}$$

Cette fonction prend des valeurs entre 0 et 1, elle est maximale lorsque les deux zones segmentées comparées sont de taille identique.

Régularité du mouvement

Nous considérons de même que le mouvement est régulier entre deux images successives. Une zone de segmentation *k* à l'instant *t* et sa zone correspondante *l* à l'instant *t-1* doivent avoir des vecteurs vitesses proches en orientation et en amplitude. Pour cela, on définit les fonctions scores suivantes:

→ Fonction score de régularité de l'amplitude du vecteur vitesse moyen: on définit la fonction score $S^{t/t-1}_{a}(k, l)$ comprise entre 0 et 1 qui compare l'amplitude du vecteur vitesse moyen $v_t(k)$ de la zone de segmentation *k* à l'instant *t* à $v_{t-1}(l)$, celui de la zone de segmentation *l* à l'instant *t-1*.

$$S^{t/t-1}_{a}(k,\ l) = \frac{\min\left(v_t(k), v_{t-1}(l)\right)}{\max\left(v_t(k), v_{t-1}(l)\right)} \qquad \textit{(Eq. IV.2)}$$

→ Fonction score de régularité de l'orientation du vecteur vitesse moyen: $S^{t/t-1}_{\theta}(k, l)$ comprise entre 0 et 1 compare l'orientation $\theta_t(k)$ du vecteur vitesse moyen de la zone segmentée *k* à l'instant *t*, à $\theta_{t-1}(l)$, celle de la zone de segmentation *l* à l'instant *t-1*. Si les deux vecteurs sont de même orientation, alors $S^{t/t-1}_{\theta}(k, l)$ tend vers 1, sinon $S^{t/t-1}_{\theta}(k, l)$ tend vers 0. Dans le cas de vecteurs d'orientation opposée (l'angle que ces deux vecteurs forment serait supérieur à 90° en valeur absolue), $S^{t/t-1}_{\theta}(k, l)$ est maintenu à 0.

$$S^{t/t-1}_{\theta}(k,\ l) = \max\left(0, \cos\left(\theta_t(k) - \theta_{t-1}(l)\right)\right) \qquad \textit{(Eq. IV.3)}$$

Hypothèse de faible déplacement d'un objet dans la scène

En estimation de mouvement, on considère souvent l'hypothèse d'un déplacement faible de l'objet analysé d'une image à la suivante. Dans notre cas, cela revient à supposer que deux zones spatialement proches, segmentées dans deux images successives, sont plus susceptibles de représenter le même objet que deux zones éloignées. On ajoute donc la fonction score $S^{t/t-1}_{d}(k, l)$, qui tient compte de $\Delta^{t/t-1}(k, l)$, la distance entre les centres de gravité de la zone de segmentation *k* à l'instant *t* et celui de la zone segmentée *l* à l'instant *t-1*. Le choix s'est orienté vers une fonction $S^{t/t-1}_{d}(k,l)$ exponentielle pour s'affranchir des problèmes de calibration. $S^{t/t-1}_{d}(k,l)$ prend sa valeur maximale *1.0* lorsque les deux centres de gravité sont confondus et tend vers *0* lorsqu'ils s'éloignent l'un de l'autre. Le paramètre τ_{Sdist} permet de faire varier la sensibilité de la fonction, il s'agit de sa constante d'espace, une valeur élevée de τ_{Sdist} permet d'obtenir une fonction décroissant plus lentement vers *0* à mesure que $\Delta^{t/t-1}(k, l)$ augmente (nous fixons expérimentalement τ à *5* pixels).

$$S_d^{t/t-1}(k,\ l) = \exp\left(-\Delta_{t/t-1}(k,\ l)/\tau_{Sdist}\right) \qquad \textit{(Eq. IV.4)}$$

Relation entre vecteurs vitesse et direction du déplacement

Le vecteur vitesse moyen estimé par le module de flot optique $v_{t-1}(l)$ de l'objet l à l'image t-1 doit avoir une orientation proche de celle du vecteur déplacement $d_{t/t-1}(k, l)$ reliant les positions des centres de gravité de deux zones appariées entre deux images successives (*t-1* et *t)*. Pour cela nous calculons une fonction score: $S^{t/t-1}_v(k, l)$.

Soient $\theta_v(l)$, l'orientation du vecteur vitesse moyen $v_{t-1}(l)$ de l'objet segmenté l à l'instant *t-1* et $\theta^{t/t-1}_d(k, l)$, l'orientation du vecteur $d_{t/t-1}(k, l)$ reliant les centres de gravité des zones l à l'instant *t-1* et k à l'instant *t*. On définit alors $S^{t/t-1}_v(k, l)$ selon la relation:

$$S_v^{t/t-1}(k,\ l) = \max\left(0, \cos\left(\theta_v(l) - \theta_d^{t/t-1}(k,l)\right)\right) \qquad \textit{(Eq. IV.5)}$$

IV.1.2.3. Fusion des cinq fonctions score

Fusion des fonctions de comparaison entre deux objets segmentés

Les différentes fonctions score sont assimilables à des probabilités. Il serait néanmoins imprudent d'effectuer directement le produit entre ces différentes probabilités. Certaines fonctions score sont plus fiables dans certaines conditions, notamment S_θ, S_a et S_v qui n'ont plus vraiment de sens lorsque l'objet devient statique (lorsque le déplacement devient très faible, voire nul).

Nous proposons de fusionner les fonctions score entre deux zones segmentées k et l en utilisant la relation suivante:

$$S_f^{t/t-1}(k,\ l) = S_{taille}^{t/t-1}(k,\ l).S_d^{t/t-1}(k,\ l).\frac{1+\left[1-S_d^{t/t-1}(k,\ l)\right].\left[S_v^{t/t-1}(k,\ l)+S_a^{t/t-1}(k,\ l).S_\theta^{t/t-1}(k,\ l)\right]}{1+2.\left(1-S_d^{t/t-1}(k,\ l)\right)} \qquad \textit{(Eq. IV.6)}$$

Il s'agit d'une fonction qui favorise avant tout la mise en correspondance de zones de taille similaires et faiblement éloignées et se déplaçant de façon cohérente. Plus précisément, son comportement est descriptible comme suit :

D'une façon générale, on impose que la taille de l'objet suivi soit temporellement régulière, et que la distance parcourue entre deux images soit faible (inférieure à τ_{Sdist}). Les fonction score S_d et S_{taille} conservent un poids fort dans le calcul final. En revanche, les fonctions S_a, S_θ et S_v dépendent de plusieurs facteurs. On distingue deux régimes de fonctionnement:

→ Si les fonctions $S^{t/t-1}_{\theta}(k, l)$ et $S^{t/t-1}_{v}(k, l)$ sont non nulles, alors l'objet est supposé en déplacement, en conséquence, $S^{t/t-1}_{d}(k, l)$ doit a priori tendre vers 0. Afin de vérifier ces hypothèse, lorsque des zones segmentées éloignées sont comparées, $S^{t/t-1}_{d}(k, l)$ devenant faible, ceci a pour effet d'augmenter le poids des fonctions S_a, S_θ et S_v dans le calcul et donc de valider ou non la cohérence du mouvement (régularité de vitesse et d'orientation du mouvement). A l'opposé, si la distance entre zones comparées est faible (c.-à-d. S_d tendant vers 1.0), cela revient à faire l'hypothèse d'un objet statique, alors les contraintes de régularité de vitesse et d'orientation sont relâchées. Cela signifie aussi qu'un objet est susceptible de changer de direction lorsqu'il est lent ou immobile plutôt que lorsqu'il se déplace rapidement.

→ Si les fonctions $S^{t/t-1}_{\theta}(k, l)$ et/ou $S^{t/t-1}_{v}(k, l)$ sont nulles, cela signifie que les zones mises en correspondance n'ont pas de relation cohérente entre leurs vecteurs vitesses ou avec leur déplacement. On vérifie alors si l'objet est statique en ne prenant en compte que les fonctions S_d et S_{taille}.

Cette méthode permet au final de comparer deux zones de segmentation en mouvement en imposant une certaine régularité de son déplacement (orientation et amplitude) et permet aussi d'apparier les zones de segmentation temporairement immobiles (tant que la zone segmentée existe).

Méthode d'appariement

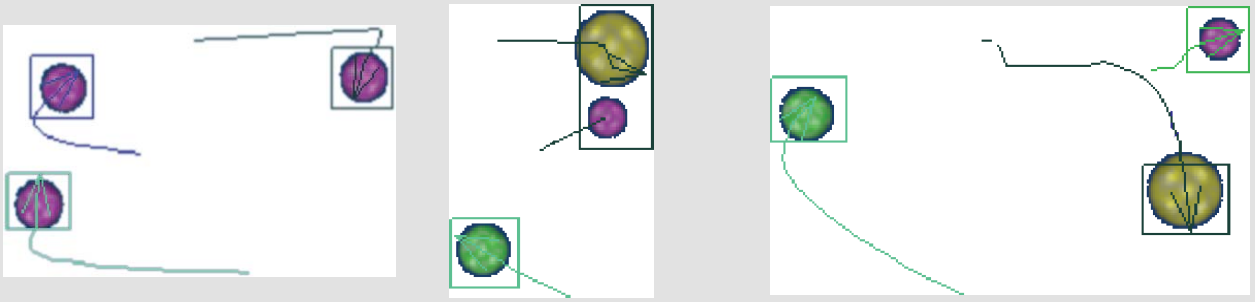
La fusion des fonctions score que nous venons de décrire permet de donner un indicateur qui traduit la probabilité que deux zones segmentées dans deux images successives correspondent à un même objet. Raisonnons maintenant sur l'ensemble des zones segmentées entre deux images successives. Il faut définir un système de vote qui mette en correspondance chacune des zones segmentées de chaque objet d'une image à la suivante. Par ailleurs, la mise en correspondance ne doit pas être forcée si deux zones ne correspondent pas suffisamment.

La zone k de l'image t est mise en correspondance avec la zone l de l'image $t-1$ donnant la valeur $S^{t/t-1}_{f}(k, l)$ la plus forte. Mais, nous introduisons également un critère de "ressemblance minimale" avec le seuil m_S fixé de façon experte à 0.1 pour lequel aucune mise en correspondance n'est faite si la valeur maximum $S^{t/t-1}_{f}(k, l)$ lui est inférieure.

```
// Calcul des fonctions score entre les zones segmentées de deux images successives
Pour chaque zone l à l'image t-1:
        Pour chaque zone k à l'image t
                Calcul de la fonction score finale S^{t/t-1}_f(k, l)
        finPour

        // Mise en correspondance de la zone k avec l donnant S^{t/t-1}_f(k, l) maximum
        Choix du résultat maximum des maxS_f=max[S^{t/t-1}_f(k, l)]
        Si maxS_f>m_S
                Attribution de l'ID de la zone l à la zone correspondante k
        finSi
finPour
```

Cette méthode permet à un objet à l'image *t-1* de n'être mis en correspondance à l'image *t* qu'avec un seul objet (cf. fig. IV.3.a). Elle permet également à plusieurs objets de fusionner à l'image suivante: plusieurs zones *l* de l'image *t-1* peuvent être mises en correspondance avec une même zone de segmentation *k* à l'image suivante *t* lors d'une collision (cf. fig. IV.3.b). Ainsi, deux objets peuvent fusionner leur ID mais s'ils viennent à se séparer à nouveau, le plus grand garde l'ID, l'autre se voit attribuer un nouvel ID (cf. fig. IV.3.c).



a. **b.** **c.**

Figure IV.3: exemples de suivi de 3 balles avec la méthode proposée. Chaque couleur de trajectoire et de boite englobante représente un ID spécifique. a, exemple de suivi simple d'objets, sans problèmes de collisions/occultations . b, exemple d'union des zones de segmentations entre 2 objets proches, l'ID devient unique. c, séparation entre 2 objets préalablement confondus, un nouvel ID est attribué au plus petit.

IV.1.3. Filtrage des zones segmentées dues au bruit

IV.1.3.1. Principe

Nous venons de voir comment suivre un objet durant son déplacement en appariant ses zones de segmentation successives dans une séquence vidéo à l'aide des informations de position, de taille et de vitesse. Ce système permet de suivre aussi bien un objet en mouvement qu'un objet immobile (ou en déplacement faible), tant que la zone de segmentation existe. Dans le cas de vidéos bruitées, des zones de segmentation peuvent être créées par le bruit et il est alors nécessaire de les éliminer. Pour cela, nous proposons une approche par analyse spectrale.

Il est possible de savoir si une zone segmentée provient d'un bruit ou d'un véritable objet en mouvement. Pour cela, nous faisons appel aux capacités d'analyse spectrale du cortex V1: l'idée est de vérifier s'il y a un lien entre les caractéristiques des contours contenus dans la zone de segmentation et les contours correspondant à des changements spatio-temporels détectés dans cette même zone. Si les changements spatio-temporels correspondent à des contours en mouvement qui existent parmi les contours de la zone alors il s'agit véritablement d'une zone de mouvement (cf. fig. IV.4.a), sinon, la zone segmentée correspond à une zone de bruit (cf. fig. IV.4.b). Ceci revient à comparer les caractéristiques spatiales des changements spatio-temporels dans une zone aux caractéristiques spatiales des contours présents. Pour cela, nous nous intéressons aux spectres log polaire des sorties Parvo ON-OFF du filtre Parvo contours et du filtre MagnoY contours mobiles de chaque zone segmentée. Le spectre de l'image de la sortie du filtre Parvo contours caractérise tous les contours de la zone segmentée, le spectre en sortie du filtre MagnoY contours mobiles caractérise quant à lui les changements spatio-temporels: les contours en mouvement et le bruit.

Si la zone segmentée correspond effectivement à un objet en mouvement, alors ses contours mobiles sont présents sur la sortie Parvo ON-OFF **ET** sur la sortie MagnoY contours mobiles (cf. fig. IV.4.a). Dans le cas d'une zone segmentée associée à un bruit temporel, il n'y a pas de corrélation entre les sorties de chacun des deux filtres (cf. fig. IV.4.b).

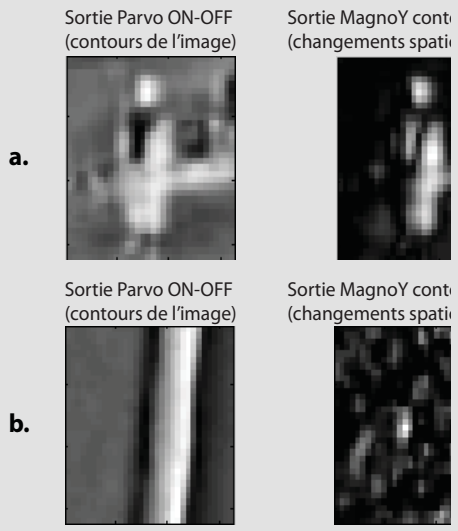


Figure IV.4: comparaison entre les sorties des filtres Parvo ON-OFF et MagnoY contours mobiles d'une zone de mouvement (a) ou d'une zone de bruit (b).

IV.1.3.2. Méthode de différentiation du mouvement et du bruit

Le principe énoncé peut être reformulé de la façon suivante: si les contours en mouvement d'une zone de segmentation ont des orientations qui coïncident avec certaines des orientations de tous les contours de la zone alors il s'agit certainement d'un objet en mouvement, sinon, il s'agit de bruit. On va donc comparer les spectres log-polaires des sorties de chacun des deux filtres.

On construit un système de différenciation basé sur l'analyse des orientations dominantes des sorties Parvo ON-OFF et MagnoY contours mobiles pour chacune des zones segmentées (cf. fig. IV.5). Si la zone segmentée considérée correspond à un objet mobile, le produit des courbes d'énergie cumulée par orientation des sorties Parvo ON-OFF et MagnoY contours mobiles doit avoir un profil proche de la courbe d'énergie cumulée par orientation de la sortie MagnoY seule. Dans le cas contraire, l'orientation des énergies en sortie MagnoY est «aléatoire » et ne correspond pas à la sortie Parvo contours (ON-OFF).

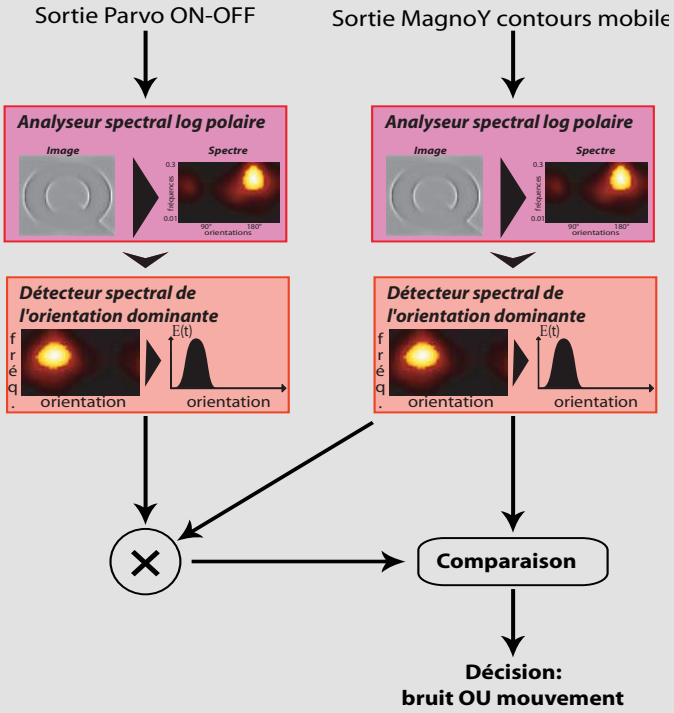


Figure IV.5: système de différenciation entre bruit et mouvement. Pour une zone de segmentation donnée, le système d'analyse spectrale analyse indépendamment les sorties de filtre Parvo ON-OFF et MagnoY et prend une décision sur sa nature: zone de mouvement ou zone de bruit d'après une analyse de la coïncidence entre les orientations des deux spectres.

Afin de comparer les deux courbes d'énergie cumulée par orientation (du produit Parvo*MagnoY et de la sortie MagnoY seule), nous analysons l'intercorrélation entre ces deux signaux. Plus précisément, nous normalisons ces courbes d'énergie cumulée par orientation entre 0 et 1, et retirons la valeur moyenne, puis en calculons l'intercorrélation. De cette façon, la normalisation permet de comparer des signaux d'amplitude proche, et le retrait de la moyenne élimine la composante continue. Ceci a pour conséquence que la valeur maximum de l'intercorrélation est liée au décalage en orientation entre les deux signaux. Si la zone de segmentation englobe une région en mouvement, alors le décalage en orientation est nul, car les deux signaux comparés sont corrélés. Sinon, le décalage est non nul ce qui montre la non-corrélation ou non-cohérence entre les signaux Parvo et MagnoY et traduit la présence d'un bruit.

Ce type de traitement est illustré sur la figure IV.6. On observe en IV.6.a le cas d'une zone associée à un bruit. La réponse des orientations des contours est différente entre la sortie Parvo et la sortie MagnoY, le produit de ces deux sorties n'est pas corrélé avec les orientations de la sortie MagnoY, le maximum d'intercorrélation n'est pas centré en 0. Le système considère la zone segmentée comme une zone de bruit et arrête son suivi. A l'opposé, la figure IV.6.b montre une zone segmentée qui englobe une personne en déplacement de faible amplitude. On observe de grandes similitudes en orientation entre les sorties Parvo ON-OFF et MagnoY contours mobiles. Le calcul de corrélation montre un maximum en 0 ce qui indique cette cohérence entre les deux signaux, cette méthode considère alors cette zone segmentée comme une zone de mouvement et autorise le système à poursuivre le suivi de l'objet.

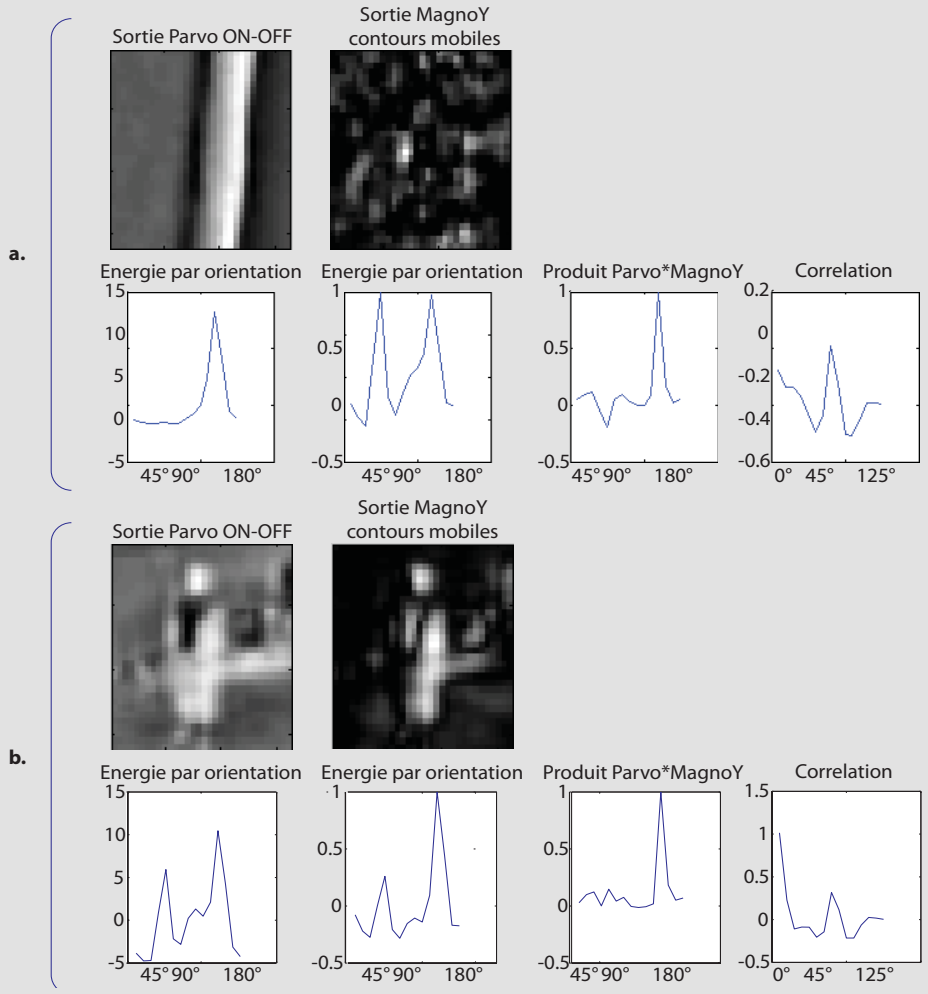


Figure IV.6: illustration de la méthode de différenciation bruit et mouvement: a, la zone de segmentation correspond à du bruit, les courbes d'énergie cumulée par orientation des sorties des filtres Parvo et MagnoY ne sont pas corrélées. b, la zone de segmentation englobant une zone de mouvement (ici un piéton) montre un lien entre les sorties Parvo et MagnoY : les contours en mouvement font partie des orientations présentes dans la zone, le maximum de corrélation est centré en zéro.

IV.1.3.3. Performances et discussion

Les tests de performance de cette méthode ont été réalisés sur la base de test PETS 2001 [PETS]. A partir de la séquence vidéo test 2-camera2, nous avons défini 400 zones de segmentation de bruit et 400 zones de mouvement (piétons, voitures, vélo) (cf. fig. IV.7). Les capacités d'identification des zones de bruit ont alors été évaluées sur ces deux types de zone. Les bonnes reconnaissances, les fausses alarmes (c.-à-d. lorsque le système considère un objet en mouvement comme une zone de bruit) ainsi que le taux d'oubli sont relevés. Les résultats sont présentés sur la table IV.1.

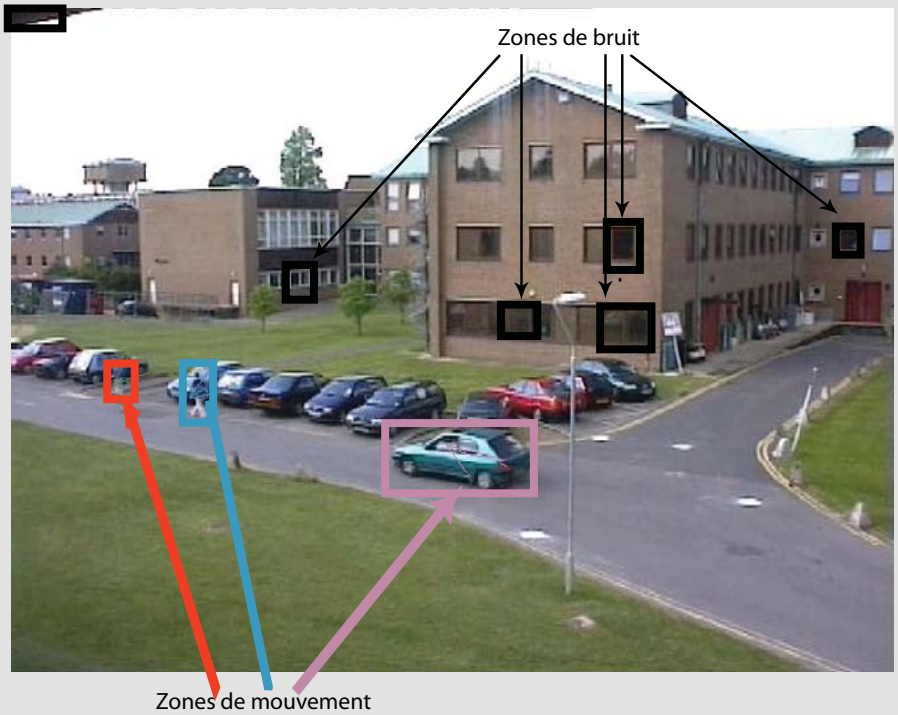


Figure IV.7: extrait de la séquence de test pour l'identification des zones de mouvement et de bruit.

Table IV.1: évaluation des performances du système de reconnaissance du bruit

Taux de succès	*Taux de fausse alarme*	*Taux d'oubli*
89%	*3%*	*9%*

On constate que cet algorithme permet d'identifier les zones de bruit avec efficacité. Le taux de fausse alarme est bas ce qui est primordial, car ce système ne doit pas considérer les objets en mouvement comme des zones de bruit. Enfin, le taux d'oubli de l'ordre de 10% montre que cette méthode présente des limites. Des problèmes peuvent survenir lorsque les zones segmentées présentent de nombreuses orientations, celles-ci pouvant être confondues avec les orientations temporelles créées par le bruit.

Le but de cet algorithme est comparable aux post-traitements à base d'opérateurs morphologiques. Ces deux méthodes ont le même but: éliminer les zones de segmentation de bruit. Mais chacune agit dans un cadre différent: les opérateurs morphologiques éliminent les zones segmentées de taille réduite et ne tiennent pas compte du mouvement, la méthode proposée est capable d'éliminer des zones segmentées de taille importante en se basant sur une analyse de l'information de mouvement contenue dans celles-ci.

Du point de vue du coût de calcul, cette méthode de différenciation fait appel à deux analyses spectrales ce qui peut être coûteux en temps de calcul si la zone à analyser est de taille importante. D'une manière

générale, cette méthode est appliquée ponctuellement dans le système de suivi, toutes les 10 images.

Cette méthode de différenciation ne se base que sur une analyse des orientations dominantes dans les zones segmentées, elle est donc relativement qualitative. Elle peut néanmoins être améliorée si l'on compare les images spectrales log polaire plutôt que leur courbe d'énergie cumulée par orientation, mais le coût de calcul est alors augmenté.

IV.1.4. Performances de l'algorithme de suivi avec identification des zones de bruit

IV.1.4.1. Qualité du suivi

Nous avons testé cet algorithme sur différents types de vidéo de manière à montrer la généricité du système. Sans modifier les paramètres des filtres utilisés, nous testons les performances de suivi dans le cas d'objets en déplacement. Le système de segmentation préalable utilisé pour ces tests est celui présenté au chapitre II (mais tout autre algorithme de segmentation pourrait a priori faire l'affaire), il s'insère dans le système de suivi complet (cf. fig. IV.1) et permet de tester un système de suivi entièrement basé sur un système "bio-inspiré".

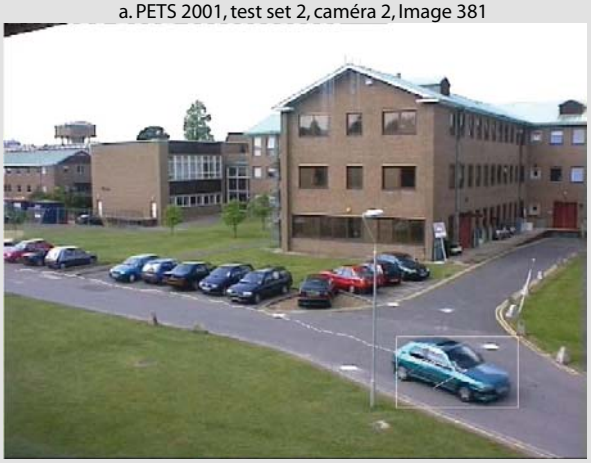
Les figures IV.8.a-c montrent différents résultats de suivi dans une séquence PETS2001 [PETS], pour laquelle la caméra est fixe et il n'y a pas d'occultations. Les zones de segmentation sont encadrées avec une couleur spécifique et la trajectoire de leur centre de gravité est tracée avec la même couleur. Sur la figure IV.8.a, une voiture traverse la scène et est suivie tout le long de son parcours. La figure IV.8.b montre le suivi de trois personnes traversant la scène. Le suivi simultané de différents types de zones en mouvement: "personnes" et "vélo" est montré sur la figure IV.8.c. Pour ces exemples, d'une part, le suivi est réalisé correctement sans perte du suivi et d'autre part, les zones de segmentation crées par le bruit sont isolées.

Les figures IV.8.d et IV.8.e montrent deux extraits de séquences plus difficiles. On teste les performances du système de suivi dans le cas d'une caméra mobile en basse résolution et fortement compressée (vidéo mpeg1). La caméra est mobile et suit la trajectoire du ballon. La tâche de suivi est plus difficile, mais reste efficace, les problèmes apparaissent principalement lors des inversions brusques dans la direction de déplacement de la caméra. Cela montre que le système proposé est également tolérant vis-à-vis des systèmes à caméra mobile tant que la caméra bouge de façon régulière.

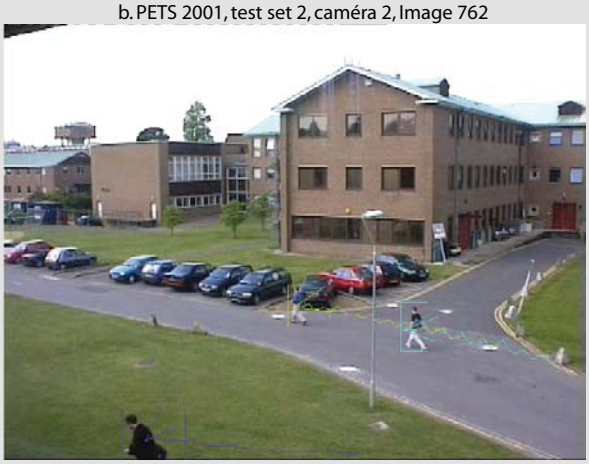
On observe que quelle que soit la zone en mouvement (voiture, piéton, vélo) la segmentation les englobe correctement. Par ailleurs, le suivi des objets est correct tout le long de leur apparition, tant qu'aucune occultation ne se produit (pour rappel, nous ne proposons pas un système de suivi long terme dans cette partie).

Plus précisément, la table IV.2 donne les écarts de suivi par rapport à la vérité terrain pour les vidéos de référence Pets2001 [PETS] en résolution originale (768*576 pixels). Nous évaluons également les performances de conservation de l'ID tout au long de la trajectoire. Pour cela, nous ne considérons que les trajectoires pour lesquelles les objets ne sont pas occultés. On note ici TF le nombre de trajectoires fragmentées par rapport au nombre d'objets à suivre établi en vérité terrain. Un rapport supérieur à 1 montre qu'un objet est suivi durant un certain temps puis est "perdu", puis son suivi est repris avec un ID différent et ainsi de suite.

a. PETS 2001, test set 2, caméra 2, Image 381

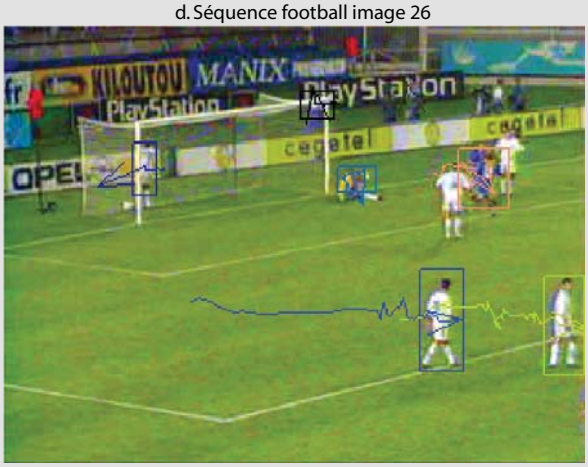


b. PETS 2001, test set 2, caméra 2, Image 762



c. PETS 2001, test set 2, caméra 2, Image 1873

d. Séquence football image 26



e. Séquence football image 165

Figure IV.8: test du système de suivi temporel d'objets en mouvement. a-b-c, test en caméra fixe dans une séquence test PETS 2001 (test set 2, caméra 2) (Résolution de l'image: 576*768 pixels). d-e, Test en caméra mobile sur une séquence vidéo compressée de football (résolution image: 320*240)

Table IV.2: évaluation des performances du calcul de trajectoire des zones en mouvement.

Ecart moyen avec la trajectoire de référence	*Ecart type*	*TF de notre méthode*
6 pixels	*3 pixels*	*32 trajectoires /27 objets*

Ces résultats montrent un suivi spatialement précis de la trajectoire et une fragmentation très faible de la trajectoire lorsque les objets ne sont pas occultés.

IV.1.4.2. Coût de calcul

Cet algorithme est entièrement basé sur une approche contour utilisant les outils présentés aux chapitres I et II. Il fonctionne à 12 images par seconde pour des images de taille 320*240 pixels sur un ordinateur équipé d'un processeur de type Intel Pentium 4 3.0Ghz avec un code C++/Matlab non optimisé.

IV.1.5. Conclusion sur le système de suivi

Dans cette partie, nous avons montré qu'il est possible de construire un système de suivi court terme d'objets en mouvement basé sur une approche contours utilisant les outils de vision par ordinateur inspiré par des modélisations du système visuel humain. Il exploite les deux voies d'information de détail (Parvo) et de mouvement (Magno). Une phase de segmentation est requise au préalable pour permettre d'isoler les réponses de ces deux voies d'informations pour chaque objet en mouvement (par exemple, le système de segmentation présenté au chapitre II). Par suite, ces zones sont suivies temporellement d'une image à l'autre et sont identifiées (par un ID spécifique). Ce système est également capable de s'affranchir des zones de segmentation liées au bruit.

Ce système ne réalise néanmoins qu'un suivi à court terme c.-à-d. d'une image à la suivante. Une méthode d'analyse plus haut niveau est nécessaire pour résoudre les problèmes d'occultation amenant à la perte temporaire d'un objet dans la scène.

IV.2. Reconnaissance d'objets

L'objet de ce paragraphe est de faire d'une part de l'identification d'objet (retrouver un objet précis dans une base) et d'autre part de la catégorisation/classification d'objet (retrouver le type d'un objet: voiture, vélo, etc. à partir d'une base représentative des types objets à classer) comme cela se fait dans notre cortex visuel dans les aires supérieures à l'aire V1. Comme il l'a été évoqué dans [Guyader04, LeBorgne04] la modélisation du cortex V1 par une décomposition en bandes d'orientations et bandes de fréquences permet une catégorisation efficace de la scène analysée. Nous allons également utiliser ce type d'approche avec nos outils d'analyse bio-inspirés. L'objectif est à terme de proposer un outil de reconnaissance à insérer dans le système de suivi présenté au paragraphe précédent dans le but de pouvoir retrouver les objets suivis, mais ayant disparu durant quelques images pour raison d'occultations.

IV.2.1. Algorithme proposé

Afin de caractériser les objets, nous proposons de leur associer une signature spectrale qui soit invariante aux effets de zoom, de translation et de rotation dans le plan de la caméra. La figure IV.9 présente l'architecture de l'algorithme proposé. La structure spatiale des objets est caractérisée par l'organisation de ses contours d'où l'utilisation du filtre Parvo contours et l'analyseur spectral log polaire. A ces deux modules déjà amplement décrits précédemment, nous ajoutons un module d'extraction de la signature spectrale proprement dite.

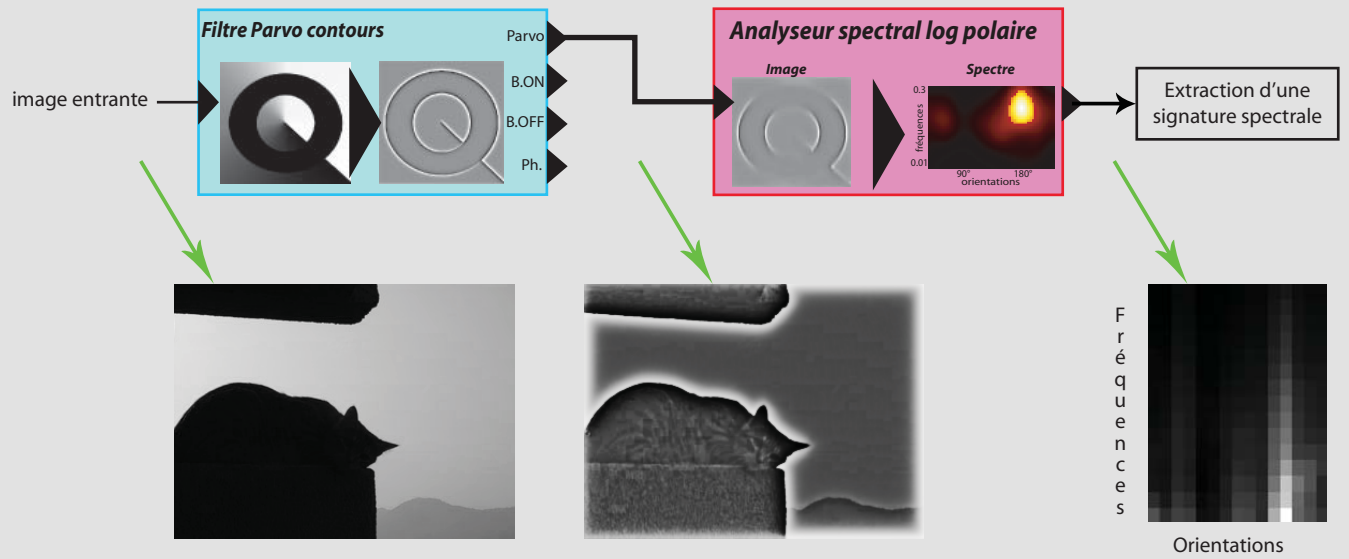


Figure IV.9: principe de fonctionnement de l'analyseur/identificateur d'objets: la scène est traitée à l'aide du filtre Parvo contours puis de l'image de contours résultante est calculé le spectre d'amplitude log polaire donnant une information de structure et de texture sur la zone analysée. Enfin, la signature spectrale de chaque objet est calculée

IV.2.1.1. Définition de la signature spectrale

Rappels sur les propriétés du spectre log polaire

→ Le spectre d'amplitude échantillonné de façon log polaire a pour propriété de porter l'information de structure et de texture des éléments de la scène visuelle.

→ Le spectre log polaire est invariant (dans une certaine mesure) à la translation (cf. fig. IV.10.a). Cette propriété découle de la transformation de Fourier. En effet, un objet ou motif analysé en différentes positions spatiales présente un spectre identique s'il ne subit pas d'autres transformations qu'une translation plane dans le plan image.

→ Lors d'un zoom ou d'une rotation plane, le spectre log polaire se translate : lors d'une rotation, l'énergie se translate selon l'axe des orientations et une partie "sort " de l'image spectrale pour réapparaître de l'autre côté par effet cyclique (cf. fig. IV.10.b). Lors d'un zoom, l'énergie se translate le long de l'axe des fréquences (cf. fig. IV.10.c).

L'idée est donc de définir à partir du spectre log polaire caractéristique de la structure et de la texture des objets et naturellement invariant aux translations planes, une représentation également invariante aux zooms et rotations planes.

Remarque: dans le cas de zoom, on perd ou on gagne de l'information visuelle du fait de la résolution et de la zone d'analyse englobant plus ou moins d'informations: on perd de l'information en basse fréquence et on gagne de nouvelles informations en haute fréquence lors d'un agrandissement. Au contraire, on gagne de l'information en basse fréquence et on en perd en haute fréquence lors d'un rétrécissement. L'information qui perdure dans le temps est donc plutôt au centre de la bande de fréquence admise par notre banc de filtres.

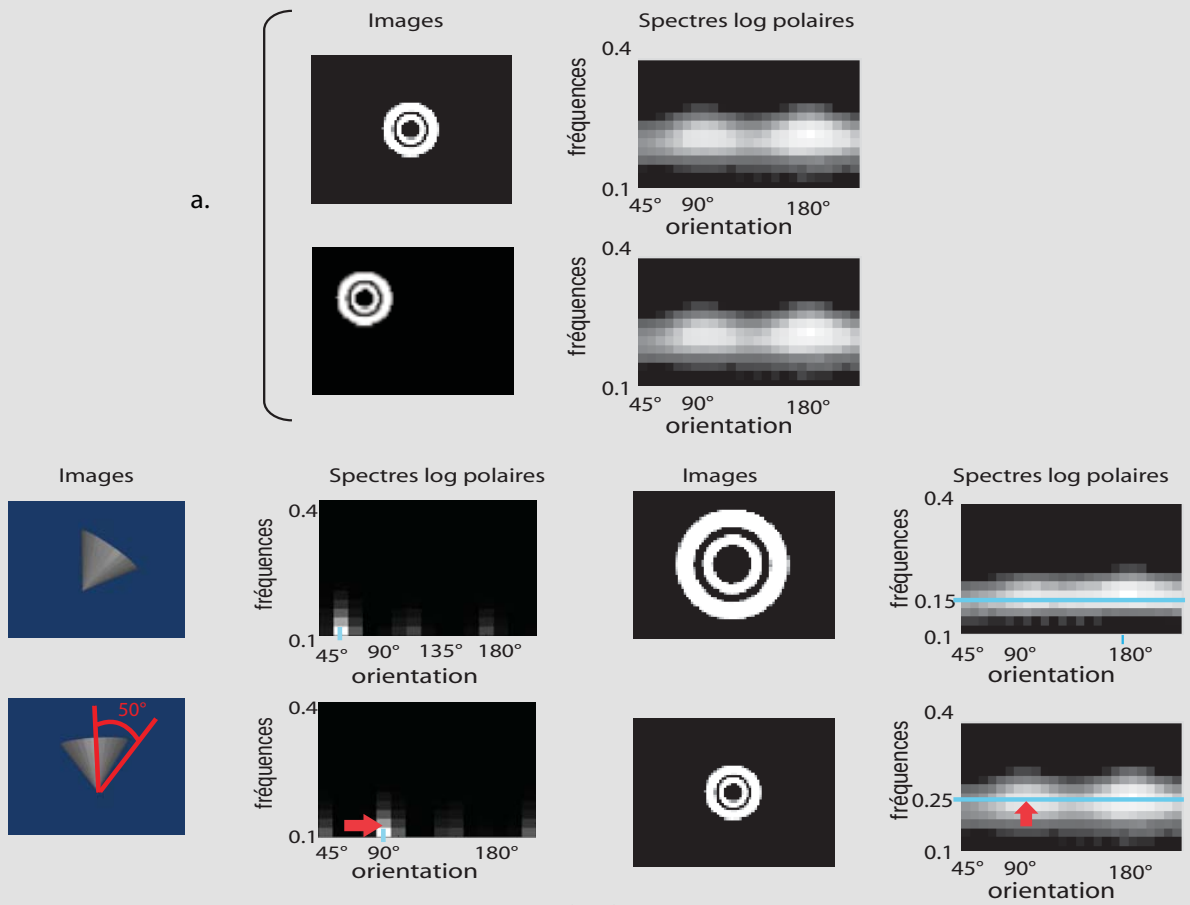


Figure IV.10: comportement du spectre log polaire lors d'une translation (a), d'une rotation dans le plan de la caméra (b) ou lors d'un zoom (c)

Définition de la signature spectrale

Afin de conserver l'invariance à la translation et d'ajouter une invariance à la rotation plane (roll) et au zoom, nous proposons d'utiliser l'autocorrélation de ces spectres log polaires. L'autocorrélation d'une image *S* de taille *M*N* est définie comme suit :

$$C(i,j) = \sum_{1}^{M}\sum_{1}^{N} S(m,n).S(m+i,n+j))$$

$$avec$$

$$1 \le i \le 2M \text{ et } 1 \le j \le 2N$$

(Eq. IV.7)

Dans notre cas, l'image dont on calcule l'autocorrélation est un spectre échantillonné de façon log polaire. Son autocorrélation rend le spectre log polaire invariant aux zooms et rotations car l'autocorrélation possède la propriété d'être invariante aux translations de l'image analysée.

Comme le spectre est invariant aux translations de l'objet et comme l'autocorrélation est invariante aux translations du spectre log polaire de l'objet, alors, l'image d'autocorrélation est invariante aux translations,

rotations planes et zooms de l'objet. Cette image constitue la signature spectrale (ou carte d'identité) d'un objet, indépendamment de sa position, de sa taille et de son orientation dans le plan de la caméra. A titre d'illustration, sur la figure IV.11, toutes les représentations de l'objet conduisent à la même signature spectrale.

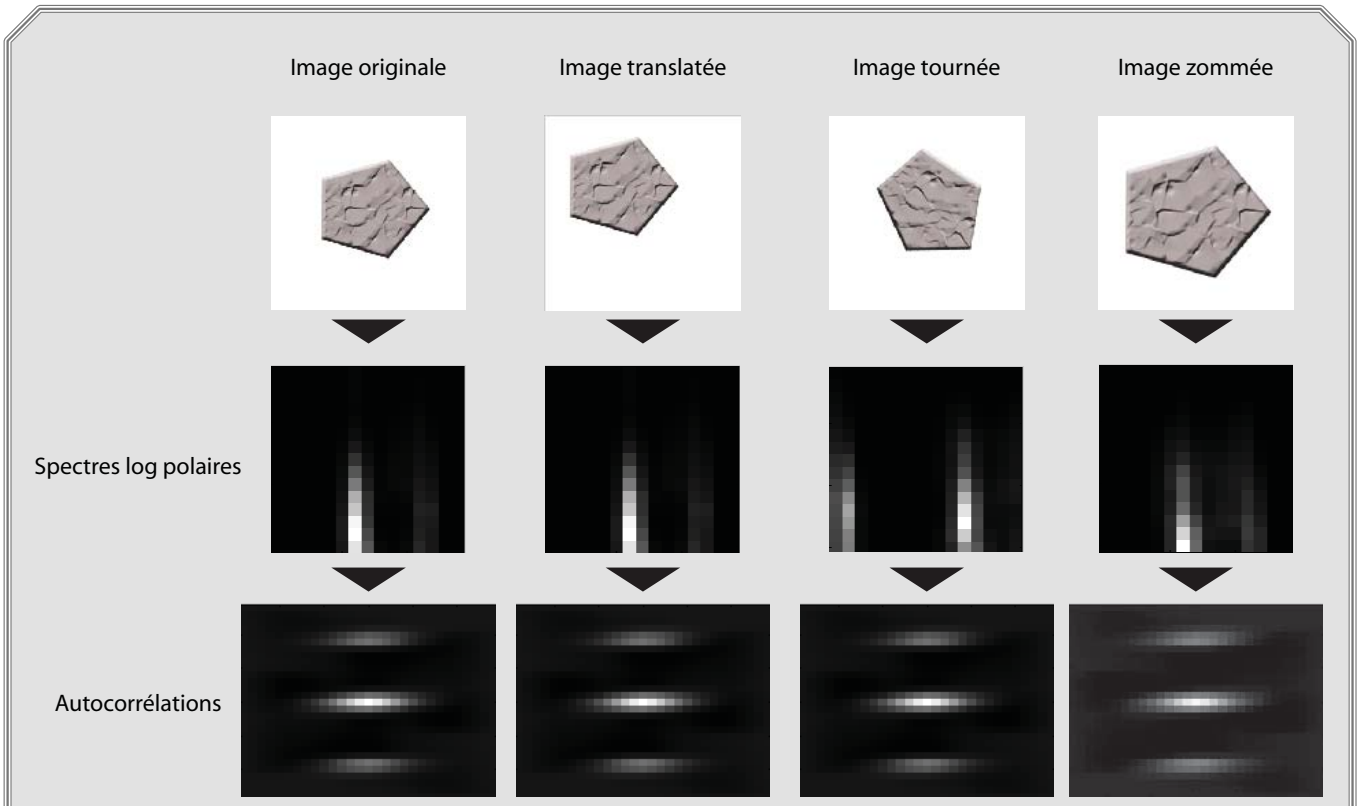


Figure IV.11: autocorrélation du spectre log polaire lors d'une translation (a), d'une rotation dans le plan de la caméra (b) ou lors d'un zoom (c)

IV.2.1.2. Calcul de distance entre signatures spectrales

Dans le but d'identifier ou de classer des objets par rapport à une base de référence, nous proposons une méthode de mise en correspondance basée sur les signatures spectrales. L'idée est de comparer deux objets *a* et *b* par le calcul de la distance euclidienne entre les autocorrélations $C_a(i, j)$ et $C_b(i, j)$ de taille M*N de leurs spectres log polaires (leurs signatures spectrales) selon la relation suivante:

$$d(C_a, C_b) = \sum_{1}^{2M-1} \sum_{1}^{2N-1} \left(C_a(i,j) - C_b(i,j)\right)^2$$

(Eq. IV.8)

Une distance faible indique une forte ressemblance entre les objets. Au contraire, une distance importante montre leur différence.

Nous choisissons de travailler avec un spectre log polaire de précision comparable à celle du système

visuel. Le spectre est échantillonné en 15 bandes d'orientations et 15 bandes de fréquences, ce qui aboutit à une image 15*15=225 pixels. Sachant que l'autocorrélation d'une image de taille N*M est de taille (2N-1)*(2M-1), nous travaillons avec des autocorrélations de taille 29*29 pixels.

IV.2.2. Analyse de performances

Nous avons testé cette méthode sur deux bases, la première permet de tester les performances de la méthode, la seconde en teste les limites. Cette première base est constituée de 72 images de scènes naturelles ayant subi des rotations planes et des effets de zoom. Ceci va nous permettre de tester la méthode proposée dans un cadre idéal. La seconde base (COIL100 [CoilSite]) contient 100 objets subissant des rotations non planes (changements de prise de vue) et des effets de zoom. Cette seconde base va permettre de tester cette méthode dans un cadre d'étude plus complexe ne respectant pas les hypothèses de rotation plane, mais en ajoutant des effets introduits par la prise de vue 3D.

IV.2.2.1. Analyse des résultats sur la base de scènes naturelles

Cette première base contient 72 images réparties en six classes: villes, montagnes, plages, forêts, intérieurs d'habitation et villages (cf. fig. IV.12.a). A partir de chaque image, on crée (cf. fig. IV.12.b):

→ 5 images tournées dans le plan de la caméra (roll) de 10°, 45°, 90°, 135° et 170°.

→ 5 images zoomées d'un facteur respectif: 1.2, 1.5, 1.8, 2 et 2.5 (les zooms sont positifs de manière à minimiser les effets de bords).

Cette base est donc constituée de 792 images.

Figure IV.12: images extraites de la base de scènes naturelles: a, exemple d'images pour chacune des 6 classes: villes, montagnes, plages, forêts, scènes d'intérieurs et villages. b, exemple d'une image ayant subi zooms et rotations.

Identification et tolérance aux effets de zoom et de rotation dans le plan de la caméra

On étudie l'évolution de la distance entre une image en position de référence et ses versions zoomées et tournées. La figure IV.13.a montre l'évolution des spectres log-polaires et des autocorrélations lors de zooms et rotations d'une image de plage. La figure IV.13.b montre l'évolution de la distance entre la signature spectrale de l'image de référence et celle de ses versions transformées. A titre de comparaison, nous représentons la distance moyenne entre la signature spectrale de l'image d'exemple et la signature spectrale de chacune des images des autres classes.

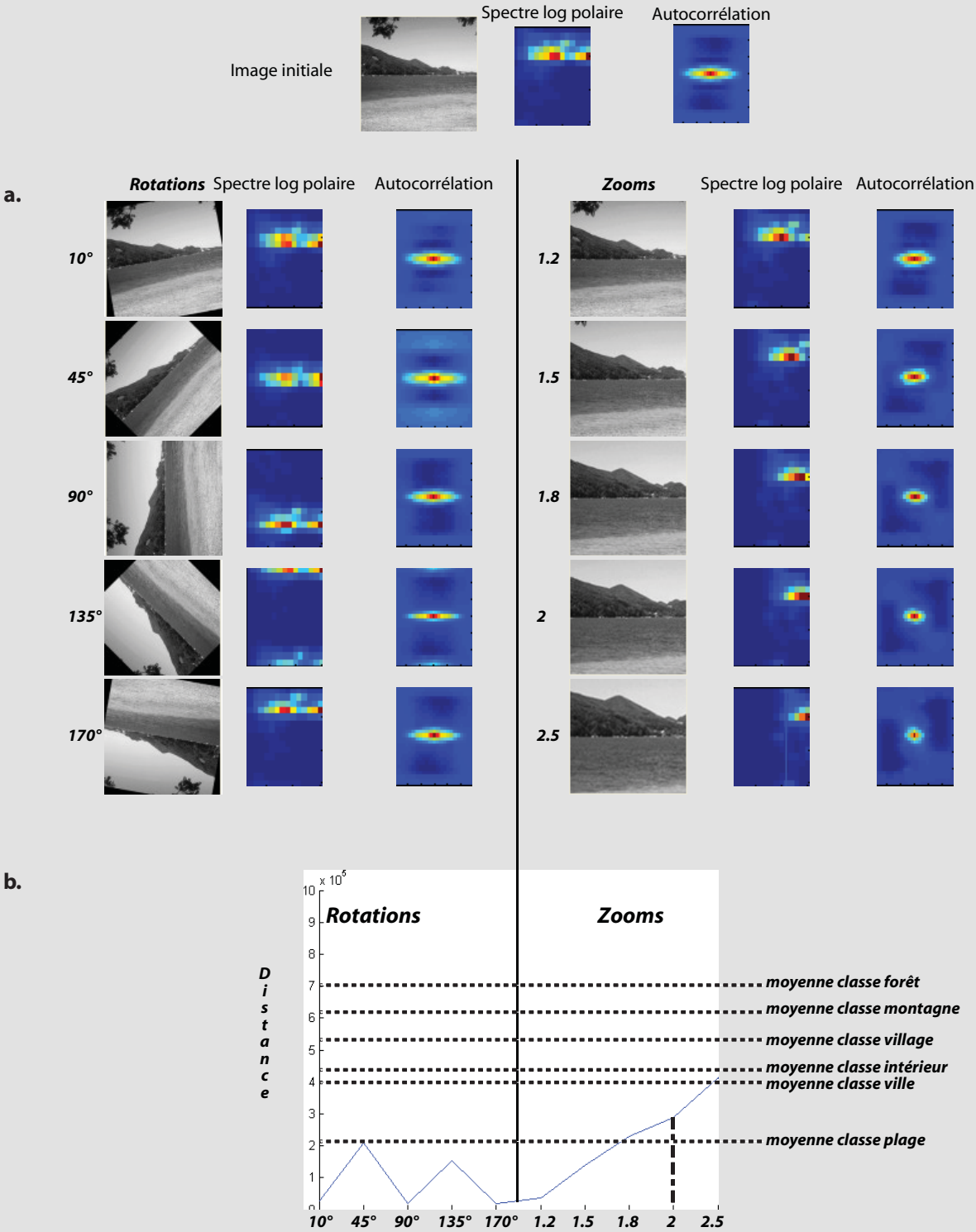


Figure IV.13: a, évolution des spectres log polaires et des autocorrélations en fonction de la rotation et de l'effet de zoom pour une image donnée. b, évolution de la distance entre les signatures spectrales de l'image initiale et des images transformées et comparaison avec les moyennes des distances vis-à-vis des autres images de chaque classe.

On constate que les autocorrélations des spectres varient peu lors d'une rotation, même si le spectre se translate. Les changements les plus remarquables (rotations de 45° et 75°) sont dus aux effets de bord importants sur les images tournées. Ces effets de bords sont néanmoins atténués du fait de l'utilisation d'un fenêtrage de Hanning (cf. chapitre I.5.2) et on constate quoi qu'il en soit que la distance mesurée reste inférieure à la distance vis-à-vis des autres classes. Ainsi, l'évolution de la distance en fonction de la rotation est stable et est toujours inférieure à la distance moyenne aux autres classes.

En ce qui concerne les effets de zoom, la translation du spectre vers les basses fréquences entraîne de manière inévitable une perte d'informations. Ainsi, les autocorrélations évoluent à mesure que l'effet de zoom se renforce. Cela se traduit par un accroissement progressif de la distance en fonction de l'effet de zoom. Il existe donc une limite au-delà de laquelle les autocorrélations n'ont plus de liens suffisants (ici, on fixe la distance limite à 3.10^5, soit un zoom de l'ordre de 2 donnant une distance légèrement supérieure à la classe d'appartenance de l'image ce qui permet une certaine tolérance).

Par ailleurs, la moyenne des distances calculées avec des images issues de la même classe que l'image requête (plage) est inférieure à la moyenne des distances calculées avec des images issues des autres classes. Il est donc possible d'envisager une classification en se basant sur un calcul de distance entre autocorrélations des spectres log polaires.

Classification de scènes

On définit la classification comme la capacité à apparier une image requête avec des images appartenant à la même classe. Afin de tester les performances de classification sur cette base d'images de scènes naturelles, on procède selon une première méthode qui consiste à sélectionner une image parmi les 792 de la base et chercher parmi les 791 restantes laquelle donne la distance la plus faible. On regarde alors si les deux images de distance proche appartiennent à la même classe. Le taux de bonne classification est de 85%. Les figures IV.14.a-b illustrent ces résultats : pour une image requête donnée, les images donnant une distance faible appartiennent bien à la même classe, les autres classes sont beaucoup plus éloignées en terme de distance. Ce premier test montre que la méthode est efficace dans le cadre d'analyse de scènes qui respectent les types de transformation d'image supportés par la méthode (rotation plane et zoom). Des erreurs se produisent seulement pour des images de structure et de texture très proches : dans l'exemple de la figure IV.14.c, les murs verticaux des immeubles sont les plus visibles et leur texture est moins marquée ce qui se rapproche d'une image de plage particulière, également marquée par un contour très fort. Dans ce cas-là, les informations de texture et de structure sont insuffisantes et seule l'information de contexte permettrait de les différentier.

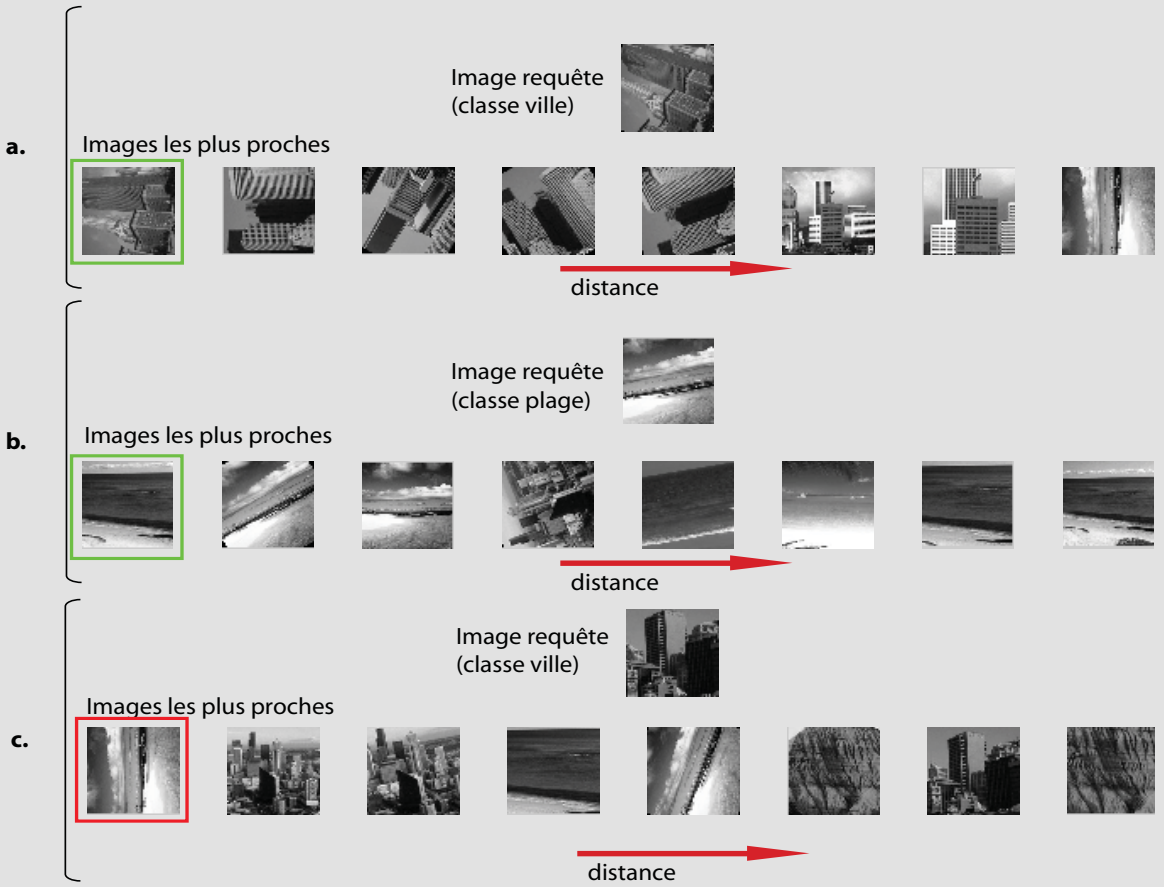


Figure IV.14: exemples de résultats de classification avec la méthode proposée. a et b, exemples de bonne classification. c, exemple d'erreur de classification du fait de la ressemblance des orientations entre certaines images de classes différentes.

Classification par utilisation d'une image moyenne

On évalue les performances de classification par une autre méthode couramment utilisée en classification. Elle consiste à calculer l'autocorrélation moyenne C_m des autocorrélations C_i des différentes images contenues dans une même classe. Chaque classe est donc représentée par une seule autocorrélation caractéristique (cf. fig. IV.15). La classification consiste alors à comparer l'autocorrélation du spectre d'une image requête avec chacune des autocorrélations moyennes représentatives de chaque classe. L'autocorrélation moyenne donnant la distance la plus faible indique alors la classe qui a été reconnue. On obtient par cette méthode un taux de reconnaissance de 82% ce qui est inférieur au résultat précédent du fait que chaque classe est décrite par une seule signature spectrale moyenne (cette signature ne pouvant en effet pas représenter chaque cas de figure au sein de la classe). Ce taux est comparable aux méthodes de classification basée sur les spectres log polaires [Guyader04].

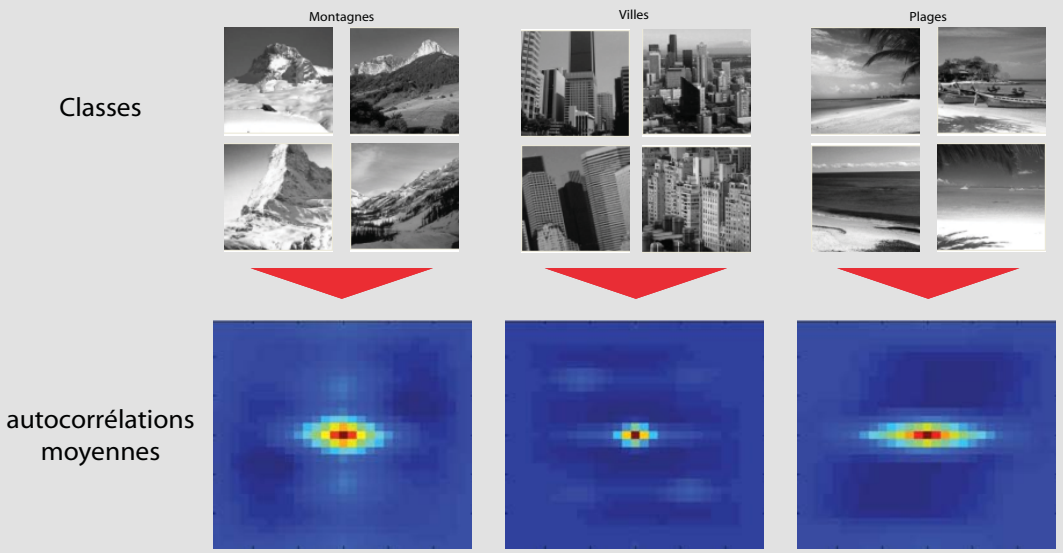


Figure IV.15: autocorrélations moyennes de trois classes d'images

A l'issue des tests effectués sur cette base d'images "idéales", on conclut que la méthode de classification proposée fonctionne de façon satisfaisante pour des images subissant des effets de translation, de rotations dans le plan image et de zoom de faible amplitude (zoom de facteur 2).

IV.2.2.2. Analyse des résultats sur la base d'images COIL100

La base COIL100 contient 7200 images de 100 objets photographiés en couleur sur fond noir, en résolution 128*128. Chacun de ces objets est pris sous 72 angles de vue différents. La figure IV.16 donne quelques exemples d'objets contenus dans cette base.

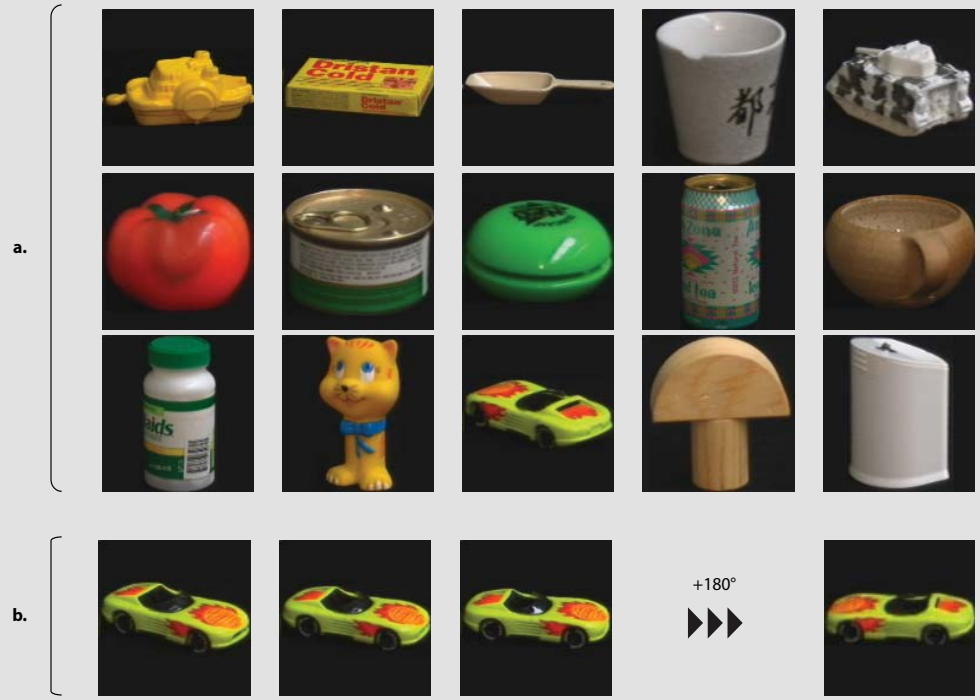


Figure IV.16: images extraites de la base COIL100: a, exemple d'objets dans la base. b, angles de vue différents pour une objet donné.

Cette base d'images est intéressante pour évaluer les performances de notre algorithme, car on y trouve des objets de forme proche (différentes voitures par exemple) et des objets de forme et de textures très différentes (boite, agrafeuse, etc). Les prises de vues sont identiques pour chaque objet et il est donc possible de voir quelle est la tolérance du système pour la reconnaissance d'objets de même orientation, de même type, etc.

Comme dans cette base de nombreux objets ont de fortes ressemblances, nous la décomposons en différentes classes dans le but d'évaluer les performances de classification de la méthode proposée. Une classification de référence a été réalisée lors d'un entretien entre 2 experts. La consigne donnée est de regrouper les objets selon des critères de forme. Il leur était présenté les 100 objets en position de référence (prise de vue en angle 0°) et en niveau de gris. Une phase de débat a permis de construire de façon plus précise les classes. Cette phase d'expertise a permis de dégager les 12 classes présentées sur le tableau IV.3. On obtient une moyenne de 8.3 objets par classe et un écart type de 3.5.

Notons qu'en prenant les mêmes critères, d'autres classifications sont possibles (par exemple, les classes canettes et mugs auraient pu être regroupées) il a néanmoins fallu faire un choix ce qui explique que nous nous basons sur cette classification ci.

Table IV.3: classification manuelle des objets de la base COIL100

Classes	*Critères retenus*	*Nombre d'objets*	*Liste des numéros des objets*	*Exemple d'objets*
Mugs et tasses	*cylindres texturés*	*8*	*10 11 16 18 43 45 59 81*	
Récipients	*cylindres coniques non texturés*	*5*	*30 58 86 89 97*	
Petites figures	*objets verticaux texturés*	*7*	*14 17 20 28 48 52 74*	
Formes rondes	*ovoïdes faiblement texturés*	*6*	*34 35 47 56 73 94*	
Utilitaires	*objets peu texturés de forme variables*	*9*	*21 36 40 44 57 60 66 68 85*	
Véhicules	*orientations et structures associées aux voitures*	*15*	*3 6 8 15 19 23 27 37 38 42 69 76 78 91 100*	
Boites rondes	*boites rondes texturées*	*9*	*25 26 29 32 70 71 72 87 95*	
Tubes	*cylindres minces et texturés*	*15*	*5 9 13 22 24 33 39 50 55 61 64 65 88 90 92*	
Boites rectangulaires	*formes rectangulaires horizontales texturées*	*10*	*1 31 46 53 54 67 79 84 96 98*	
Fruits&légumes	*formes rondes faiblement texturées*	*6*	*2 4 63 75 82 83*	
Formes en bois	*formes cylindriques et carrées faiblement texturées*	*5*	*12 41 51 77 80*	
Canettes	*cylindres texturés de dimensions constantes*	*5*	*7 49 62 93 99*	

Evolution de la signature spectrale en fonction de l'angle de vue

Nous avons mené une évaluation de la tolérance au changement de l'angle de vue des objets. L'objectif est d'analyser pour un objet donné, l'évolution de la distance entre l'autocorrélation du spectre de l'objet vue de face (image de référence à 0°) et l'autocorrélation du spectre du même objet vue sous des angles différents. Ces rotations plus complexes que de simples rotations dans le plan de la caméra entraînent un changement d'apparence de l'objet. Ceci crée au niveau de son spectre log polaire des translations locales des énergies (par effet de compression ou de dilatation des contours selon des orientations particulières) pour les rotations faibles et l'apparition de nouvelles caractéristiques spectrales lors de grandes rotations (cf. fig. IV.17). Ces transformations du spectre log polaire étant plus complexes que des translations, l'autocorrélation n'est plus stable pour les transformations importantes. Nous allons donc évaluer les limites de la méthode.

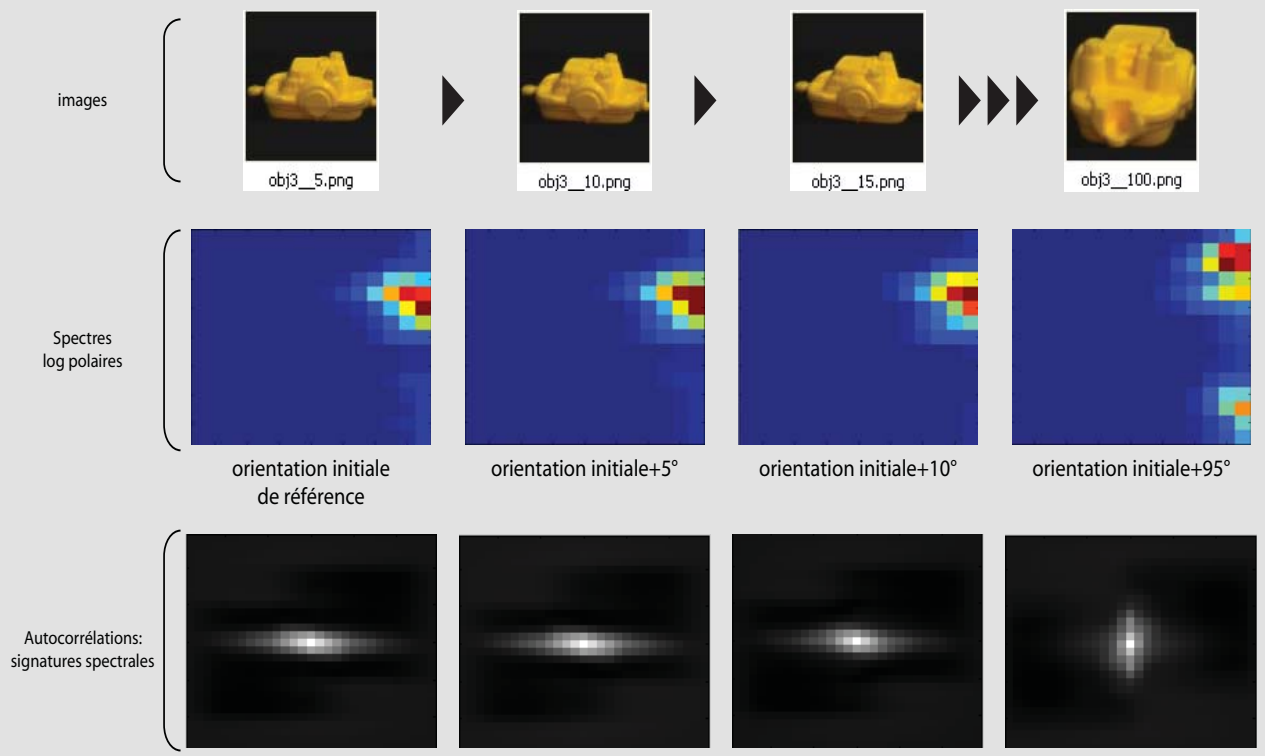


Figure IV.17: évolution du spectre log polaire et de l'autocorrélation d'un même objet lors de changement de prise de vue entraînant une apparence différente du point de vue de la caméra.

L'évaluation de la tolérance aux changements d'angle de vue de l'objet a été réalisée de la manière suivante. Pour chaque objet, on définit l'attitude de référence (prise de vue à 0°) et son autocorrélation correspondante C_0. On calcule ensuite la distance $d(C_0, C_i)$ entre la signature spectrale de référence et la signature spectrale C_i de la prise de vue d'angle i de ce même objet. Ces calculs donnent des courbes d'évolution de la distance entre signatures spectrales en fonction de la rotation de l'objet (cf. fig. IV.18). Notons que les objets ne sont pas toujours symétriques pour une rotation de 180° ce qui explique que les courbes de distances ne le soient pas non plus rigoureusement.

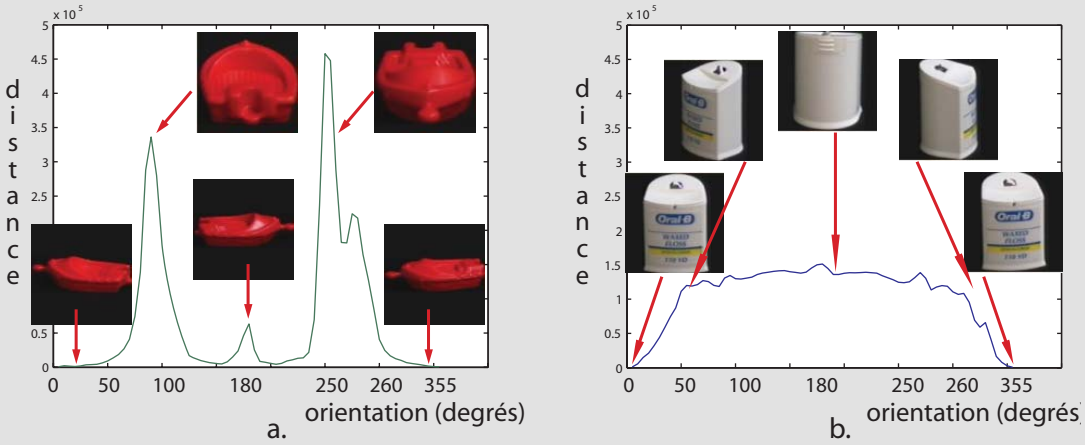


Figure IV.18: évolution de la distance, pour un même objet, entre une position de référence et les autres points de vue. a, objet de la classe "figurine horizontale". b, objet de la classe "tubes"

On observe que d'une manière générale, pour un même objet, la distance en fonction de la rotation est faible pour une rotation faible, mais croit avec l'angle de rotation. Ceci s'explique par un changement d'apparence important pour les grandes rotations, ainsi la signature spectrale devient de plus en plus différente.

La distance évolue donc fortement avec le changement de prise de vue, cette méthode ne peut donc a priori pas permettre d'identifier ou de classer un objet en le comparant à une seule prise de vue de référence. Selon les objets, on observe en effet une forte augmentation de la distance pour un faible changement d'orientation de la prise de vue (cf. fig. IV.16.b), ceci est dû à l'apparence qui évolue très rapidement même pour des rotations faibles. A l'opposé, certains objets donnent une mesure de distance faible pour une large gamme de rotation du fait de leur structure plane et faiblement texturée. On observe néanmoins qu'en moyenne, on mesure les distances les plus faibles pour des rotations inférieures à 15°. Ce qui correspond au pas d'échantillonnage des filtres GloP utilisés pour la transformation log polaire. Si l'on ne se base que sur cette tolérance de 15°, alors, identifier un objet sera possible en le comparant à un ensemble de prises de vue de référence représentant l'objet au maximum tous les 15°. Il faudra donc 24 prises de vue de référence pour identifier l'objet quelque soit sa prise de vue sur 360°.

Par ailleurs, ce changement des caractéristiques spectrales des objets lors du changement de prise de vue entraîne qu'il ne sera pas non plus possible de classer les objets en comparant une image requête avec une seule image moyenne représentative de chaque classe comme nous l'avons fait avec la base précédente.

Identification d'objets

On définit l'identification comme la capacité à apparier une image requête avec une des images représentant exactement le même objet. Nous avons effectué un premier test dont le but est de vérifier les capacités de l'algorithme à retrouver un objet dont on dispose d'une description riche (ses 72 prises de vue) au milieu d'une description également riche d'autres objets (les 99 objets restant avec 72 prises de vue pour chacun d'eux). Il s'agit donc de voir dans quelle mesure notre algorithme distingue des objets entre eux. La procédure de test est la suivante. On sélectionne une image parmi les 7200 et l'on cherche parmi les 7199 restantes celle qui donne la distance la plus faible. On regarde alors si l'objet le plus proche en terme de distance est exactement le même objet, à une rotation près.

On obtient finalement 83.6% de bonne reconnaissance, les images donnant les distances les plus faibles étant généralement celles représentant l'objet sous un point de vue avec une rotation inférieure à 15° par rapport à celui de l'image requête (cf. fig. IV.19.a-b). Ce premier test montre que cette méthode permet d'identifier un objet avec un taux de réussite correct, dans le cas où l'on a une description riche de toutes les poses de cet objet même si celles-ci sont noyées dans la description de toutes les poses d'un grand nombre d'autres objets.

Figure IV.19: exemples de résultats d'identification avec la méthode proposée. a-b, exemples de bonne classification. c, exemple d'erreur d'identification du fait de la ressemblance des orientations entre certaines images d'objets différents.

Le taux d'erreur (16.4%) s'explique par une confusion entre objets de forme et de texture proche. En effet, notre méthode ne prend pas en compte d'information de couleur ou autre. La figure IV.19.c montre une erreur d'identification pour laquelle l'image requête présente un contour extérieur proche de l'image donnant la distance la plus proche. Par ailleurs, la texture des deux objets étant relativement pauvre, les deux éléments ne sont pas facilement différenciés, ces deux facteurs limitent l'efficacité de la méthode. La figure IV.20 illustre un autre cas de confusion possible : deux objets pris sous le même angle de vue et qui ne diffèrent que par leur information de couleur.

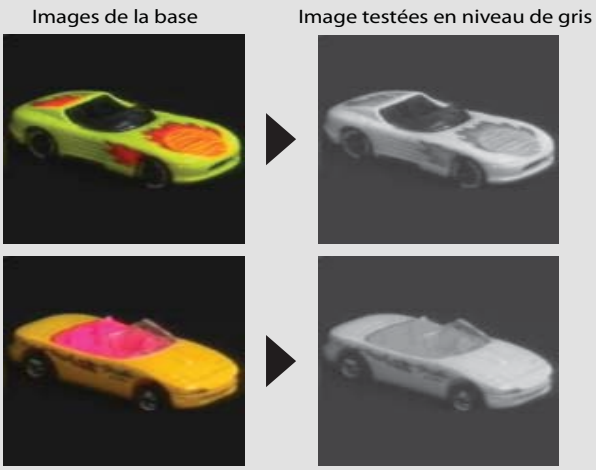


Figure IV.20: deux objets de forme proche et d'orientation identique dans la base de test. En niveau de gris, la ressemblance peut porter à confusion.

Classification

Nous testons maintenant la capacité de cette méthode à classer un objet dans les catégories définies dans la table IV.3. Pour cela, nous procédons selon deux approches différentes.

Classification par comparaison avec toutes les images de la base

Dans ce premier test, nous suivons le même principe que celui utilisé pour l'identification, mais en relâchant cette fois une contrainte. On sélectionne une image d'un objet et on la compare aux 7199 restantes. On regarde ensuite si l'image donnant la distance la plus faible appartient à la même classe que l'image test.

Cette méthode donne 87.4% de bonne reconnaissance. Comme on pouvait s'y attendre ce résultat est supérieur à l'identification réalisée sur le même principe, car nous avons relâché une contrainte autorisant ainsi certains objets à être confondus avec ceux appartenant à la même classe (cf. fig. IV.21.a). Néanmoins, l'augmentation n'est pas très importante et s'explique par la ressemblance de forme et de texture entre certains objets, même issus de classes différentes (cf. fig. IV.21.b).

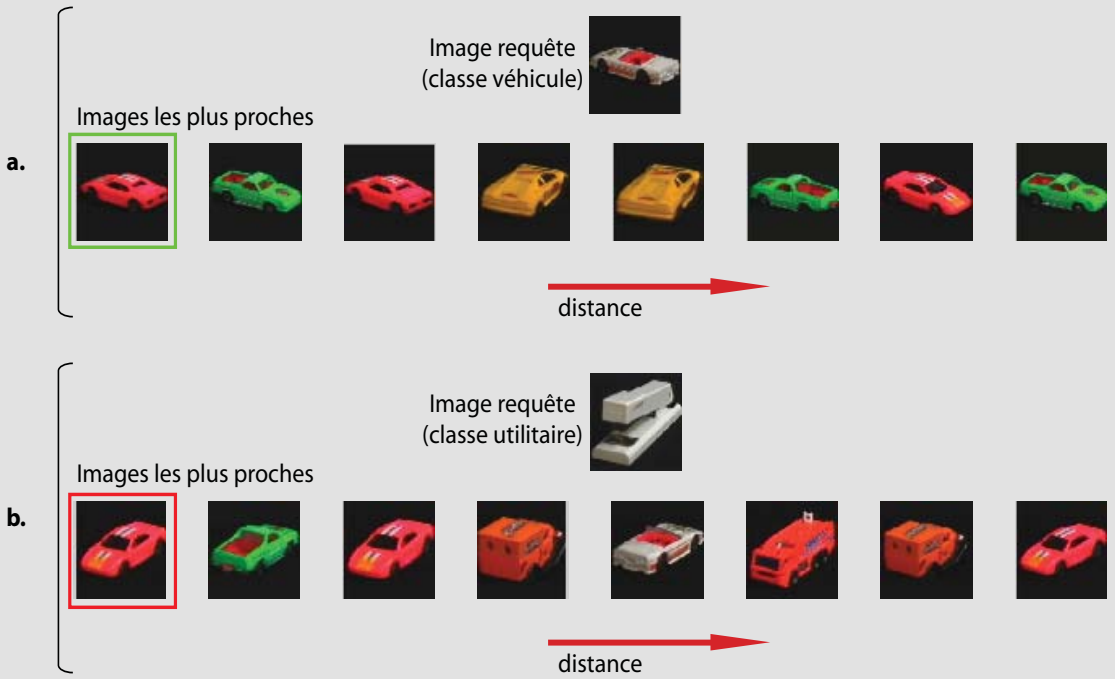


Figure IV.21: exemples de résultats de classification. a, exemple de bonne classification. b, exemple d'erreur de classification du fait de la ressemblance de forme et des orientations des textures entre certaines images d'objets différents.

Classification par comparaison à une image de référence

On analyse les capacités de classification de notre algorithme selon une seconde méthode. On calcule désormais la distance entre chaque objet pris sous un unique angle de vue de référence (0° par exemple), la base de tests consiste alors en 100 images représentant les 100 objets de la base avec le même angle de vue. Pour chaque image requête, on regarde si l'image donnant la distance la plus faible appartient à la même classe.

On obtient un taux de classification relativement faible de 73 %. Les figures IV.22.a-b montrent des exemples de classification correcte. La figure IV.22.c montre quant à elle un exemple de mauvaise classification. Les erreurs proviennent du fait que les objets au sein d'une même classe peuvent présenter des spectres et donc des autocorrélations très différentes (cf. fig. IV.23). Ceci s'explique par le fait que la classification réalisée manuellement se base sur des considérations de forme alors que dans l'exemple de la figure IV.23, bien que la forme soit identique, la texture est très différente. Il apparaît donc comme nécessaire de se baser sur un autre critère pour une classification de ce type ou alors il serait nécessaire d'enrichir la base de test de manière à décrire chaque classe avec un grand nombre d'objets représentant toutes les configurations possibles propres à chaque classe. On peut également remettre en cause la classification experte qui dans cet exemple (cf. fig. IV.23) n'était basée que sur un critère de forme, la texture ayant néanmoins une importance majeure.

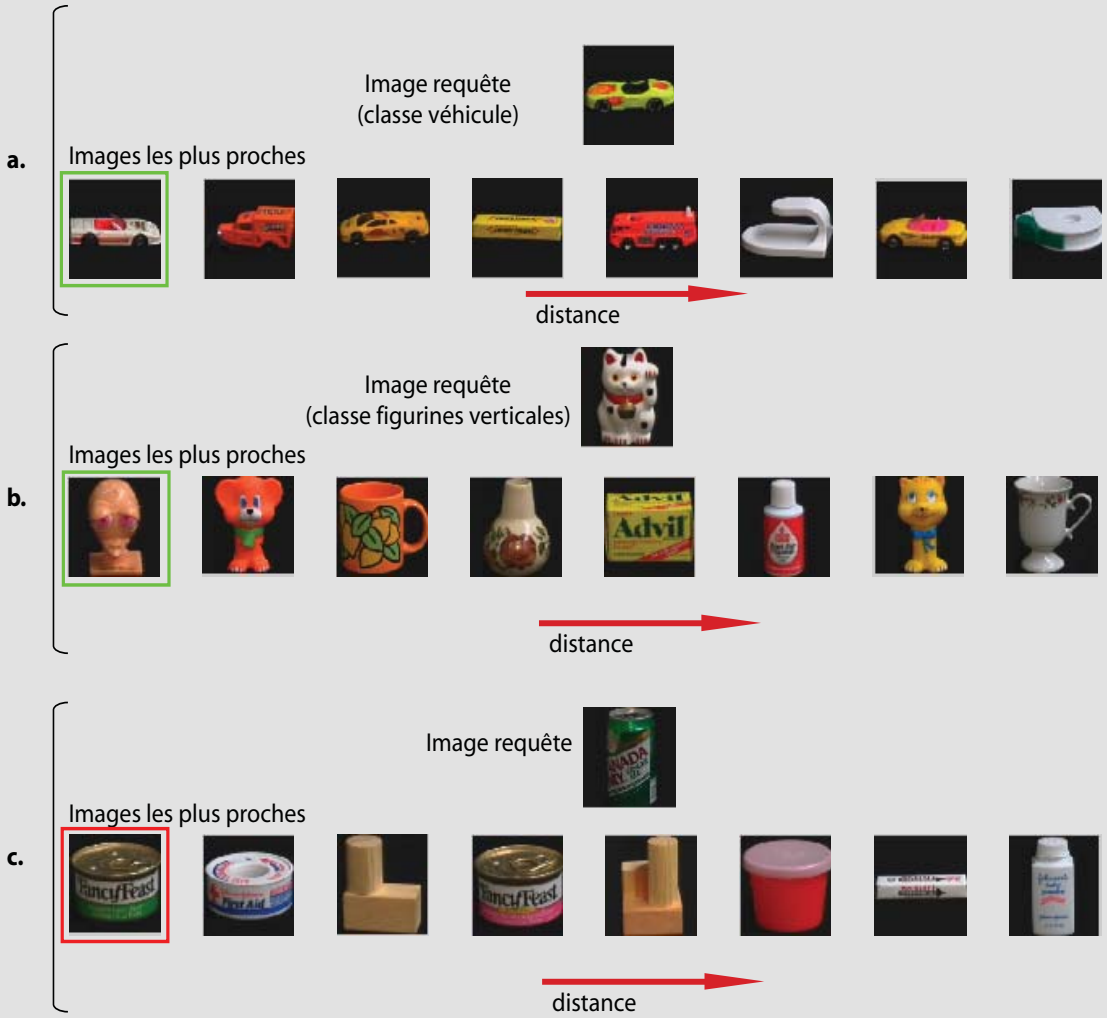


Figure IV.22: exemples de résultats de classification par comparaison entre prise de vue identique. a-b, exemples de bonne classification. c, exemple d'erreur de classification.

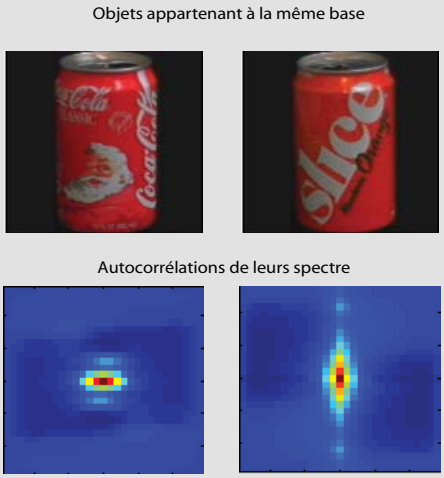


Figure IV.23: exemple d'objets appartenant à la même classe mais présentant des autocorrélations différentes

Nous venons de voir que sur une base d'images dans laquelle les objets subissent des rotations 3D, la classification est plus difficile avec la méthode proposée. Il est néanmoins possible d'obtenir de bonnes performances de classification et d'identification si l'on effectue des comparaisons entre objets en utilisant sur une grande quantité d'images représentant chaque objet sous différents angles de vue.

IV.2.3. Conclusion de la méthode d'identification et de classification

Nous avons proposé dans cette partie une méthode d'identification et de classification d'objets basée sur une analyse de la signature spectrale d'un objet. Le critère d'identification est invariant aux effets de translation, de zoom et de rotation de l'objet dans le plan de la caméra. Il n'est néanmoins pas invariant aux changements d'angles de prise de vue 3D de l'objet. Il est alors nécessaire d'avoir une description riche des apparences possibles de l'objet pour mener à bien son identification ou sa classification. Cette étude n'est basée que sur une analyse des contours, la prise en compte de la couleur serait un atout non négligeable pour un renforcement des performances.

VI.3. Perspectives: système de suivi long terme d'objets

Nous avons vu dans ce chapitre des méthodes respectivement de suivi court terme d'objets en mouvement et d'identification/classification. L'une est capable de suivre les objets en mouvement et de décrire leur évolution temporelle à court terme. L'autre est capable d'identifier des objets et de les classer. Leur utilisation conjointe pourrait permettre de construire un système de suivi très complet, l'intérêt principal étant de résoudre le problème des occultations lors du suivi grâce à la méthode d'identification (cf. fig.IV.24).

Les objets segmentés et suivis seraient d'une part classés et d'autre part ré-identifiés lorsqu'ils réapparaissent après une occultation. Plus précisément, on conserverait en mémoire, pour chaque objet suivi ses dernières signatures spectrales. La liste des objets dont le suivi a été perdu durant les *n* dernières images serait conservée et contiendrait elle aussi la carte d'identité spectrale de chacun d'eux. Lors de chaque nouvelle

apparition d'objet, leur signature spectrale serait comparée à celles des derniers objets «disparus» et un lien (reprise du suivi) serait mis en place en cas de correspondance.

Ce système est en cours de développement, et fait l'objet d'actives recherches.

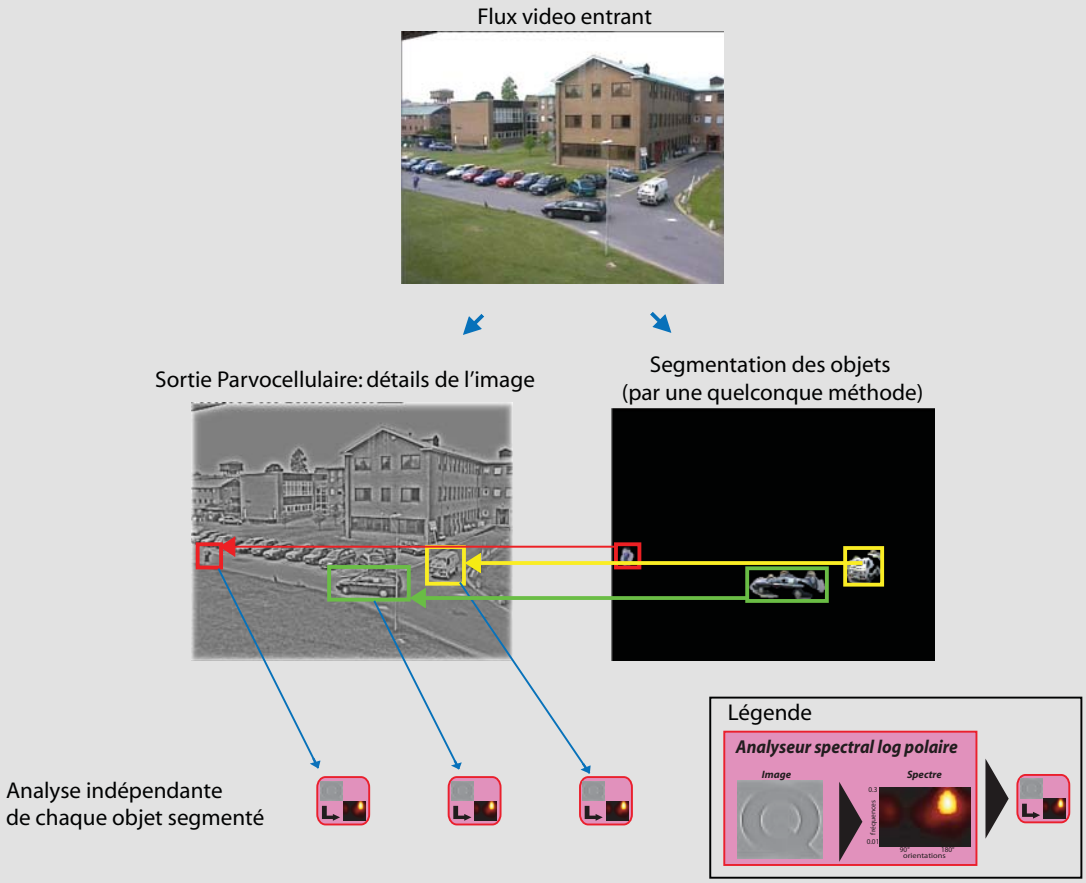


Figure IV.24: ajout de l'analyseur/identificateur d'objets au système de suivi d'objets : chaque objet est pris en charge par un analyseur spectral local permettant l'analyse de sa signature spectrale.

Chapitre V: Conclusion et perspectives

I. Conclusion

Ce travail de thèse a permis de rassembler des modélisations du système visuel humain de façon à synthétiser des outils de vision par ordinateur génériques et performants. Nous avons vu de quelle façon ils peuvent être assemblés de manière à créer des systèmes d'analyse tirant partie des différents avantages apportés par la modélisation du modèle humain.

Plus précisément, nous avons décrit les différents traitements se produisant au niveau de la rétine et au niveau du cortex visuel primaire. Ces traitements appliqués à l'information visuelle captée par l'oeil ont alors été modélisés et ont permis de construire les premiers éléments bas niveau de traitement d'image : un filtre passe-bas spatio-temporel, un module de compression logarithmique, un filtre passe-haut temporel et un analyseur spectral dans le domaine log polaire. Aussi, nous avons décrit les deux voies d'informations issues de la rétine: la voie Parvo porteuse de l'information de contour et de détail et la voie Magno porteuse de l'information de mouvement. Ces deux voies effectuent des traitements visant à renforcer la qualité des signaux extraits avant de les transmettre au cortex visuel primaire V1 qui effectue une analyse par bandes d'orientations et bandes de fréquences.

Nous avons ensuite proposé des modules de traitement d'image exclusivement basés sur les "briques" issues de la modélisation du système visuel humain: un module d'extraction de contours, un module d'extraction des contours en mouvement, un module d'analyse des orientations, un module de détection d'alertes de mouvement, un module de segmentation d'objets en mouvement et nous avons décrit un système existant de calcul de vitesse basé sur des approches similaires. Ces modules de vision par ordinateur mettent à profit la qualité des traitements "bio-inspirés" et permettent d'obtenir une panoplie intéressante d'outils de traitement d'images et de vidéos. Ces modules sont de plus capables de s'adapter facilement à la plupart des scènes analysées (conditions normales, pénombre, bruit, changements de luminosité, etc.) et fonctionnent en temps réel.

Nous avons ensuite montré comment agencer cet ensemble d'outils d'analyse pour effectuer des interprétations de gestes humains. Nous avons abouti à deux applications concrètes: un système de détection des risques d'accident pour les conducteurs (risque d'hypovigilance et de stress) ainsi qu'un système d'apprentissage de la langue des signes.

Enfin, nous avons montré comment élaborer un système de suivi de mouvement et d'identification/classification des objets. A terme ces deux systèmes seront assemblés pour créer un système de suivi complet gérant les problèmes d'occultations tout en étant capable de donner une description (identification, caractérisation) des objets segmentés et suivis.

Nous nous sommes efforcés de mettre en évidence l'intérêt de s'appuyer sur une modélisation du système visuel humain pour développer des algorithmes de traitement d'images. La partie 2 de ce mémoire dédiée à des exemples d'utilisation des modules bas niveau du système visuel humain est loin d'être exhaustive

en terme d'applications possibles. En effet, la généricité des modules bas niveau offre la possibilité d'envisager la conception de systèmes permettant de traiter d'autres cas comme par exemple des détecteurs de visage, un système d'analyse des émotions, le contrôle de qualité de surfaces d'objets, etc.

Cependant, nous sommes conscients que la rétine et l'aire V1 ne constituent pas l'ensemble des traitements subis par l'information visuelle. De ce fait, il n'est pas encore envisageable de construire un système algorithmique qui aurait toutes les aptitudes du système visuel humain.

II. Perspectives

Du point de vue des analyses qu'il serait encore possible de réaliser avec les modules actuels, de nombreux développements sont encore à explorer. En ce qui concerne l'analyse du visage, il serait intéressant de voir dans quelle mesure l'analyse du spectre log polaire pourrait permettre de reconnaître les formes de la bouche. La classification des spectres qui sont liés aux différentes formes de bouche pourrait conduire à terme à des systèmes de reconnaissance de la parole tout comme des systèmes d'analyse d'émotions.

Une analyse plus poussée de l'utilisation des spectres log polaires pour l'identification et la classification d'objets est également à envisager. Nous avons vu dans ce manuscrit une méthode de classification qui présente des propriétés intéressantes, mais présente certaines limites notamment pour la reconnaissance d'objets 3D. Envisager des méthodes de classifications basées sur les mêmes informations (de spectres log polaires) auxquelles on ajouterait des informations telles que la couleur serait intéressant.

Les perspectives à long terme concernent avant tout la prise en compte de modélisations plus fines du système visuel humain. Concernant la rétine, nos connaissances sont aujourd'hui encore limitées. Nous savons que la rétine n'est pas qu'un simple capteur et que nombre de traitements sont réalisés à ce niveau sans que nous les connaissions tous. A la vue des bénéfices apportés par les modélisations partielles que nous avons utilisées, on peut imaginer un formidable potentiel de ressources et de traitements effectués au niveau de la rétine dans son ensemble. Par ailleurs, nous n'avons exposé dans ce manuscrit que des traitements en niveau de gris. Les recherches actuelles sur la perception de la couleur [Alleyson05] montrent des possibilités de traitement d'information applicables au monde de la vision par ordinateur. Il serait par exemple envisageable de considérer un système de segmentation basé sur les modèles perceptifs de la couleur ce qui permettrait d'accéder à une plus grande richesse d'information. Enfin, la rétine présente contrairement aux capteurs CCD actuels une répartition aléatoire et non régulière des photorécepteurs qu'il serait intéressant de prendre en considération. Les avantages sont multiples pour le traitement d'image: limitation des problèmes de repliement spectral, réduction de la quantité de données à traiter, etc.

Au niveau des aires du cortex visuel, seule l'aire V1 a été modélisée. Les recherches montrent que les aires supérieures permettent un niveau d'analyse plus fin permettant d'interpréter des formes et des actions (mouvements, stimulus temporels) spécifiques. Ainsi, il serait intéressant de modéliser les aires pour lesquelles les connaissances sont suffisantes de manière à créer des modules d'analyse de plus haut niveau.

Annexe: synthèse des paramètres

Dans cette annexe est présentée une synthèse du paramétrage utilisé pour chaque module exposé dans ce document. Les valeurs numériques proposées ont été choisies pour des analyses d'images de taille de l'ordre de 320*240 pixels et des vidéos acquises à 25 images par seconde. Ces paramètres sont néanmoins adaptés à des résolutions et vitesses d'acquisition proches.

Il est explicité, en quoi le changement de chacun des paramètres affecte la réponse de chaque module. Ceci permet de comprendre le principe du paramétrage et autorise une amélioration des performances au cas par cas.

1. Paramètres associés aux filtres Parvo et MagnoY.

1.1. Filtre Parvo contours

1.1.1 Correction de luminance: paramètres associés à l'adaptation locale de luminance des photorécepteurs

L'étage d'adaptation locale de gain que l'on trouve au niveau des photorécepteurs nécessite deux paramètres: la luminance locale L qui provient de la sortie des cellules horizontales et le paramètre de compression V_0.

→ Le paramètre L dépend du filtre passe-bas associé aux cellules horizontales, on le modifie en changeant la constante d'espace α_h du filtre des cellules horizontales (couramment fixé à 5 pixels). Ceci modifie la taille du voisinage pris en compte pour le calcul de L.

→ Le paramètre V_0 ajuste le taux de compression. Sa valeur est fixée par défaut à 230, une valeur moindre donnerait une compression moindre. Cette valeur donne un comportement général satisfaisant. A titre informatif, V_0 donne des résultats intéressants du point de vue visuel pour un intervalle de valeurs comprises entre 160 et 250 [Durette05].

1.1.2. Filtrage passe-bas des photorécepteurs: minimisation du bruit spatio-temporel

Ce filtre a pour but de minimiser le bruit haute fréquence introduit lors de l'acquisition de l'image. Il permet aussi d'homogénéiser la réponse de chaque pixel (les capteurs associés à chaque pixel n'ont pas toujours les mêmes caractéristiques intrinsèques).

→ Ce bruit haute fréquence est filtré de façon efficace avec une constante d'espace de 1 pixel. Une valeur plus faible augmente la fréquence de coupure spatiale limitant le filtrage du bruit haute fréquence. Une valeur plus forte renforce le filtrage du bruit, mais risque d'atténuer la réponse des contours dans l'image.

→ Le filtrage temporel passe-bas est paramétré avec une constante de temps de 1/25 seconde ce qui permet un filtrage efficace des très hautes fréquences temporelles tout en laissant le système suffisamment réactif. Une valeur plus faible augmente la réactivité, mais favorise le passage du bruit haute fréquence. Une valeur plus forte force un moyennage temporel de l'information, ceci minimise le bruit, mais atténue la réponse des contours en mouvement rapide.

1.1.3. Filtrage passe-bas des cellules horizontales: calcul de la luminance locale moyenne

Ce filtre a pour but d'estimer la luminance locale moyenne pour chaque pixel de l'image.

→ Une constante d'espace forte permet d'estimer la luminance locale sur un voisinage large, une valeur faible rend l'estimation très localisée. Le paramètre couramment utilisé est de 5 pixels.

→ La constante de temps est fixée à 1/25 seconde ce qui laisse le système temporellement assez réactif. Une valeur plus forte donnera une estimation de luminance locale moyennée temporellement ce qui entraînera un effet retard entre l'image courante et le calcul de la luminance moyenne. Une valeur plus faible permet de se rapprocher d'une estimation image par image.

1.1.4. Adaptation locale des cellules ganglionnaires

Le dernier élément de traitement de la sortie du filtre Parvo est l'adaptation des cellules ganglionnaires. Il s'agit du même module que celui de l'adaptation locale des photorécepteurs.

→ Le paramètre L est la réponse locale moyenne des cellules ganglionnaires, pour cela, on peut utiliser un filtre passe bas purement spatial de constante d'espace fixée à 5 pixels. Une valeur plus haute permet une adaptation locale de plus grande étendue spatiale. Une valeur plus faible tendra à minimiser l'adaptation locale jusqu'à court-circuiter cet étage.

→ Le paramètre V_0 ajuste le taux de compression. Sa valeur est fixée par défaut à 230, une valeur moindre donnerait une compression moindre. Cette valeur donne un comportement général satisfaisant. A titre informatif, V_0 donne des résultats intéressants du point de vue visuel pour un intervalle de valeurs comprises entre 160 et 250 [Durette05].

1.2. Filtre MagnoY contours mobiles

1.2.1. Filtrage passe-haut temporel des cellules amacrines

Nous fixons la constante de temps du filtrage passe-haut à 1/12 seconde. Ceci permet d'extraire les changements temporels supérieurs à 12Hz. Une valeur de constante de temps plus haute permet d'extraire les changements plus lents (de fréquence plus faible) (c.-à-d. on analyse le mouvement de façon plus globale temporellement). Au contraire, une constante plus faible permet d'extraire les changements très rapides seulement, plus particulièrement, le bruit résiduel haute fréquence.

1.2.2. Filtrage spatio-temporel passe-bas des cellules ganglionnaires

Nous fixons la constante d'espace à 5 pixels de façon à estimer une moyenne locale de mouvement dans un voisinage faible. Une valeur plus forte augmente l'étendue du calcul de moyenne sur la surface de l'image (c.-à-d. on analyse le mouvement de façon plus globale spatialement).

1.2.3. Adaptation locale des cellules ganglionnaires

Se référer au 1.1.4, les mêmes paramètres sont utilisés (le principe reste le même).

2. Paramètres de l'analyseur spectral

La bande de fréquence «intéressante» est considérée comme la portion de fréquences du spectre contenant les informations propres à la scène visuelle (les objets, les formes...c.-à-d. le sujet d'étude) par opposition aux informations liées à la capture (bruit etc). Dans notre étude, nous nous intéressons aux contours des objets dans la scène. Le bruit haute fréquence et la composante continue (luminance moyenne dans l'image) ne nous intéressent pas.

Pour cela, nous fixons la fréquence centrale normalisée maximale à $f_{max}=0.35$. Les plus hautes fréquences (entre 0.4 et 0.5) ne sont alors que faiblement voire pas du tout sélectionnées.

Le spectre log polaire est composé de 15 orientations couvrant les 180° et de 15 bandes de fréquences (précision de 12°). Avec un pas d'échantillonnage fréquentiel de 1.2 octave, la fréquence centrale minimum normalisée est de l'ordre de 0.023.

3. Paramètres associés au détecteur d'évènements

3.1. Réglage de la réponse temporelle

Le paramétrage consiste à ajuster le paramètre τ_{E1} régissant le comportement d'atténuation temporelle du signal $E_1(t)$: plus τ_{E1} est grand, plus l'indicateur de changements temporels $\alpha(t)$ tiendra compte des événements temporellement éloignés (effet d'oubli atténué).

τ_{E1} est fixé par défaut à 0.5 seconde. Cela signifie qu'un mouvement de même amplitude qu'un mouvement exécuté 0.5 seconde auparavant sera considéré comme au moins supérieur de 30% à ce mouvement précédent.

3.2. Réglage des seuils de déclenchement pour la segmentation d'évènements

L'ajustement du seuil de déclenchement $Vd=E_{noise}$ permet de maintenir un niveau de sensibilité minimisant les problèmes de bruit. Pour cela, on estime ce seuil durant une période de non-mouvement, par calcul de la moyenne μ_E de l'amplitude du bruit et de son écart type σ_E. En fixant $Vd=\mu_E+3\sigma_E$ nous garantissons de ne pas déclencher de fausse alarme due au bruit avec un intervalle de confiance de trois écarts types si le bruit considéré est gaussien.

3.3. Seuil de déclenchement du détecteur binaire d'alerte de mouvement

Le seuil de déclenchement du détecteur binaire est flexible. Les problèmes de fausse alarme due au bruit ont été minimisés en amont. Ce seuil va permettre de sélectionner le mouvement de façon qualitative: un seuil bas, proche de zéro permet de segmenter le moindre mouvement détecté. Un seuil haut tendant vers 1 permet de sélectionner les mouvements de plus grande amplitude. Nous fixons le seuil de déclenchement à 0.2 de manière à sélectionner une grande majorité des mouvements détectés.

4. Paramètres de l'algorithme de segmentation de mouvement

4.1. Estimation très localisée de la quantité de mouvement: filtrage Seg_p

Le premier filtrage passe-bas a pour but d'homogénéiser très localement l'énergie de la sortie MagnoY. Ainsi, suivant la taille des objets à segmenter, la fréquence de coupure spatiale du filtre devra être ajustée. Une valeur élevée de cette fréquence de coupure ne minimise que faiblement les fluctuations locales de l'énergie. Cela signifie que la sortie de ce filtre (*Seg_p*) conservera des fluctuations d'énergie sur les zones en mouvement, que ce soit pour des objets réellement en mouvement ou pour le bruit résiduel. On risque alors de segmenter le bruit et de décomposer un objet en plusieurs régions. Au contraire, une fréquence de coupure trop faible risque d'étendre les zones de segmentation, trop largement autour de l'objet en mouvement, le bruit lui sera beaucoup moins facilement segmenté. Par défaut, nous utilisons la même fréquence de coupure spatiale que celle utilisée pour les cellules horizontales de la rétine (constante d'espace de 5 pixels).

D'une manière générale, nous recommandons de fixer la fréquence de coupure spatiale de l'ordre de la taille minimum des objets à segmenter.

Aussi, on donne une réponse temporelle à ce filtre ce qui permet de lisser temporellement l'énergie et surtout d'étendre la réponse des contours en mouvement sur les surfaces homogènes des objets en mouvement. Une valeur faible de la constante de temps permet au système de répondre de façon rapide, mais segmente surtout les contours de l'objet sans ses surfaces homogènes. Au contraire, une constante de forte valeur rend le système moins réactif mais permet d'étendre fortement la zone de segmentation sur toute la surface de l'objet.

Nous fixons la constante de temps à 1/25 seconde. Néanmoins, si le bruit est trop important, une augmentation de cette constante est recommandée.

4.2. Estimation de la quantité de mouvement moyenne locale: filtrage Seg_h

Le second filtrage a lui pour fonction d'estimer le mouvement moyen dans un voisinage étendu. Ainsi, sa fréquence de coupure spatiale devra être inférieure à celle du premier filtrage. Aussi, une fréquence de coupure trop haute conduira à un risque plus élevé de segmentation du bruit et, plus préoccupante, une décomposition des objets en plusieurs parties suivant la concentration locale d'énergie sur leur surface. Nous avons donc fixé la fréquence de coupure spatiale à une valeur très basse, 5 fois plus faible que celle des cellules horizontales de la rétine (constante d'espace de 25 pixels).

La constante de temps est elle moins critique, son influence est moindre. Elle est fixée par défaut à une valeur faible (1/25 seconde). L'augmenter permet d'étendre quelque peu la zone de segmentation par effet de moyenne temporelle.

4.3. Ajustement de la pondération de Seg_h pour la segmentation

Le paramètre δ_b est directement lié au bruit maximum admissible en sortie magnoY. Il représente la limite entre le signal utile de mouvement et le bruit résiduel au niveau de la sortie *Seg_p*. Une étude préalable du niveau de bruit en sortie MagnoY est requise pour une segmentation optimale. Si l'on considère le niveau de bruit comme constant, le calcul de sa moyenne μ_b et de son écart type σ_b sur une série d'images statiques de la scène à analyser permettra d'ajuster δ_b. Néanmoins, les qualités de filtrage du bruit au niveau des filtres Parvo et MagnoY permettent d'obtenir un bruit d'amplitude faible quelque soit le type de scène vidéo analysé.

Par défaut, nous fixons ce seuil à $\mu_b + 3\sigma_b$ ce qui permet de s'assurer que le bruit reste en dessous du seuil de segmentation avec un intervalle de confiance de trois écarts types.

La différence de gain *a* entre les deux filtrages *Seg_p* et *Seg_h* est fixée par défaut à *0.5*. Une valeur plus faible a pour effet une baisse de l'efficacité de l'adaptation locale. Cela tend à effectuer un seuillage à valeur fixe sur toute l'image d'où un risque de segmentation large autour des objets en mouvement fort et un risque de non segmentation des objets à mouvement lent. A l'inverse, une valeur proche de *1.0* tendra à nous ramener à un comportement avec sur-segmentation du bruit et sous-segmentation des objets texturés de grande taille.

5. Paramètres du système d'estimation de vitesse

5.1. Paramètre des filtres de vitesse

Nous décrivons ici rapidement les paramètres associés aux filtres de vitesse. Se référer à la thèse de Torralba pour une description plus fine des paramètres.

→ La constante d'espace de ces filtres (k^2) est fixée à 1 de façon à respecter $a+b=2$ [Torralba99].

→ La constante de temps est fixée à 1/25 cycle par seconde et le paramètre *a* (le même que celui énoncé précédemment) est fixé à 0.99. Ce qui permet au système d'être sensible aux grandes vitesses.

5.2. Paramètre des filtres passes bas spatiaux placés en aval des filtres de vitesse

Ces filtres doivent intégrer les valeurs de vitesse de chaque pixel au sein d'une zone en mouvement. Ils permettent d'homogénéiser la vitesse sur tout l'objet. Si le but est d'estimer le vecteur vitesse moyen, alors, il est nécessaire d'effectuer un lissage très fort de l'information. Pour cela, nous fixons la constante d'espace à environ trois fois la taille de l'objet.

Remarque: si l'on veut au contraire décrire plus précisément les mouvements internes à l'objet segmenté, cette constante d'espace doit être réduite. Néanmoins apparaissent rapidement les problèmes d'ouverture qui risquent de fausser l'estimation si la contrainte de spectre blanchi n'est pas respectée.

BIBLIOGRAPHIE

[Adelson85] Adelson E. H, Bergen J. R, (1985) *"Spatiotemporal energy models for the perception of motion"*. J. Opt. Soc. Am. A/vol. 2, no. 2, 284-299.

[Alleysson05] Alleysson D. Chaix de Lavarenne, B. Hérault, J, (2005) *"Non linear and uniform filtering for estimating spatial information in the cone mosaic"*, 18th Symposium of the International Colour Vision Society, ICVS'2005, Lyon, July 2005.

[Aran05] Aran O, Keskin C, Akarun L, (2005)*"Sign Language Tutoring Tool"* EUSIPCO'05, Antalya, Turkey.

[Aran06] Aran O, Akarun L, (2006) *"Recognizing Two Handed Gestures with Generative, Discriminative and Ensemble Methods via Fisher Kernels"* International Workshop on Multimedia Content Representation, Classification and Security (MRCS'06), Istanbul, Turkey.

[Atick92] Atick J, Redlich A, (1992) *"What does the retina know about natural scenes?"* Neural Comput. vol.4:196-210.

[Bailly01] Elisei F, Bailly G. Odisio M, et Badin P, (2001) *"Clones parlants 3D vidéo-réalistes : Application à l'interprétation de FAP MPEG-4"*, Journées CORESA 2001, Dijon.

[Bartlett03] Bartlett M.S, Littlewort G, Fasel I, Movellan J.R, (2003) *"Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction",* CVPR2003, Madison, Wisconsin.

[Barlow53] Barlow H. B, (1953), *"Summation and inhibition in the frog's retina"*. Journal of Physiology, 119, 69-88.

[Barron94] Barron J.L, Fleet D.J, Beauchemin S.S, (1994), *"Performance of Optical Flow Techniques"* International Journal of Computer Vision, vol. 12, no. 1, 43-77.

[Beaudot94] 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.

[BioIdSite] BioID Face Database www.bioid.com

[BiopacSite] Data acquisition and analysis systems for research and education: www.biopac.com

[Blakemore69] Blakemore C, Campbell F.W, (1969), *"On the existance of neurones in the human visual system selectively sensitive to the orientation and size of retinal images"* Journal of Physiology, 203, 237-260.

[Blasdel92a] Blasdel G.G, (1992), *"Differential imaging of ocular dominance and orientation selectivity in monkey striate cortex"* Journal of Neuroscience, vol. 12, 3115-3138.

[Blasdel92b] Blasdel G.G, (1992), *"Orientation selectivity, preference and continuity in monkey striate cortex"* Journal of Neuroscience, vol. 12, 3139-3161.

[Bogdan06] Bogdan K. (2006), *"Model Based Facial Pose Tracking Using a Particle Filter"* Geometric Modeling and Imaging-New Trends (GMAI'06), 203-208, London, England.

[Bouchet04] Bouchet J, Nigay L, Ganille T, (2004), *"ICARE Software Components for Rapidly Developing*

Multimodal Interfaces" Conference Proceedings of ICMI 2004, ACM Press, 251-258, Pennsylvania, USA.

[Braathen01] Braathen B, Bartlett M.S, Littlewort-Ford G, Movellan J.R, (2001), *"3-D head pose estimation from video by nonlinear stochastic particle filtering"* Machine Perception Lab. Technical, UC San Diego, Report 5.

[Bullier98] Bullier J, (1998), *"Architecture fonctionnelle du système visuel"* dans : Boucart, Hennaff & Belin (eds) *"La vision : aspects perceptifs et cognitifs"* Edition SOLAL Neuropsychologie.

[Bullier01] Bullier J, (2001),*" Integrated model of visual processing"* Brain Research, vol. 36, no. 2, 96-107.

[Buser87] Buser P, Imbert M, (1987). *"Vision"* Hermann Edition, Paris, ISBN: 2-7056-6032-1.

[Caffier03] Caffier PP, Erdmann U, Ullsperger P, (2003), *"Experimental evaluation of eye-blink parameters as a drowsiness measure"* Eur J Appl Physiol vol. 89, 319-325.

[Chauvin03] 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.

[CoilSite] Labelled images for object recognition tests : http://tev.itc.it/DATABASES/objects.html

[Collins04] Collins T, (2004), *"Analysing Video Sequences using the Spatio-temporal Volume"* MSc Informatics Research Review, November 2004.

[Daugman88] Daugman J.D, (1988). *"Complete discrete 2-d Gabor transforms by neural networks for image analysis and compression"* IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 36, 1169-1179.

[DeValois82] De Valois R, et al. (1982). *"The orientation and direction selectivity of cells in macaque visual cortex"* Vision Research, vol. 22, 531-544.

[DeValois85] Webster M.A, De Valois R.L, (1985), *"Relationship between spatial-frequency and orientation tuning of striate-cortex cells"* Journal of Optical Society of America A, vol. 2, no. 7, 1124-1132.

[DeValois88] De Valois R. L, De Valois K, (1988), *"Spatial Vision"* Oxford: Oxford University Press.

[Dorea05] Dorea C, PardàsM. Marqués F. (2005), *"A Motion-based Binary Partition Tree Approach to Video Object Segmentation"* Proceedings of the ICIP-05, Genova, Italia.

[Douglas95] Douglas R, Mahowald M, Mead C, (1995), *"Neuromorphic analogue VLSI"* Annu. Rev. Neurosci. vol. 18, 255-81.

[DriverSimulator05] Benoit A, Bonnaud L, Caplier A, Ngo P, Lawson L, Treviesan D, Levacic V, Mancas C, Chanel G, (2005), *"Multimodal Focus Attention Detection in an Augmented Driver Simulator"* eNTERFACE'05 workshop, Mons, Belgium.

[DriverSimulator06] Benoit A, Bonnaud L, Caplier A, Damousis Y, Jourde F, Nigay L, Serranos M, (2006), *"Multimodal Signal Processing and Interaction for a Driving Simulator: Component-based Architecture"* eNTERFACE'06 workshop, Dubrovnik, Croatia.

[Durette05] Durette, B, Heraul, J. (2005) *"Traitement visuels biomimétiques pour la suppléance perceptive"*, rapport interne LIS.

[EnterfaceSite] Official website of the eNTERFACE workshops, the SIMILAR NoE Summer School on Mul-

timodal Interfaces: www.enterface.net

[Eveno03] Eveno N, (2003), *"Segmentation des lèvres par un modèle déformable analytique"* Thèse de doctorat, Institut National Polytechnique, Grenoble.

[FeretSite] The FERET Database http://www.itl.nist.gov/iad/humanid/feret/

[Freixenet02] Freixenet J, Munoz X, Raba D, Marti J, Cu X, (2002)*"Yet another survey on image segmentation:Region and boundary information integration"* ECCV, Springer- Berlin Heidelberg, 408-422.

[Garcia05] Duffner S., Garcia C. (2005), *"A Hierarchical Approach for Precise Facial Feature Detection"* CORESA 2005, p. 29-34, Rennes, France, 7-8 novembre 2005.

[GsbmeSite] www.lx040-001.gsbme.unsw.edu.au/

[Guyader04] Guyader N, (2004), *"Perception visuelle et analyse des scènes : modèles de structures corticales"*.Thèse de Doctorat de l'Université Joseph Fourier, Grenoble, France.

[Hammal06] Hammal Z, (2006), *"Segmentation des Traits du Visage, Analyse et Reconnaissance d'Expressions Faciales par le Modèle de Croyance Transférable"* Thèse de Doctorat de l'Université Joseph Fourier, Grenoble, France.

[Haralick92] Haralick R.M, Shapiro L.G, (1992), *"Computer and Robot Vision"* Volume I, Addison-Wesley, 28-48.

[Hartline38] Hartline H. K, (1938), *"The response of single optic nerve fibers of the vertebrate eye to illumination of the retina"* American Journal of Physiology, vol. 121, 400–415.

[Harvey90] Harvey L.O, Doan, V.V, (1990), *"Visual masking at different polar angles in the two dimensional Fourier plane"* Journal of Optical Society of America A, vol. 7, 116-127.

[Hawken87] Hawken M.J, Parker AJ, (1987), *"Spatial properties of neurons in the monkey striate cortex"* Proc R Soc Lond B Biol Sci. ; vol. 231, 251-88.

[Heeger87] Heeger D. J, (1987), *" Model for the extraction of image flow"* Journal Optical Society of America, vol. 4, no. 8, 1455-71.

[Hérault01] Herault J, (2001), *"De la rétine biologique aux circuits neuromorphiques"* Traité IC2, Les Systemes de Vision, J-M Jolion ed. Hermès.

[HeraultTeach] Herault J, http://www.lis.inpg.fr/pages_perso/herault/ Enseignement en perception visuelle.

[Hubel62] Hubel D.H, Wiesel T.N, (1962), *"Receptive fields, binocular interaction and functional architecture in the cat's visual cortex"* Journal of Physiology, vol. 160, 106–154.

[Hubel74] Hubel D.H, Wiesel T.N, (1974), *"Sequence regularity and geometry of orientation columns in monkey striate cortex"* J. Comp. Neurol, vol. 158, 267-293.

[Jepson90] Fleet D.J, Jepson A.D, (1990), *"Computation of component image velocity from local phase information"* International Journal of Computer Vision, vol. 5, 77–10.

[Kaplan86] Kaplan E, Shapley R. M, (1986), *"The primate retina contains two groups of ganglion cells, with high and low contrast sensitivity"* Proc. Natn. Acad. Sci. USA 83, 2755-2757.

[Kolb96] Kolb H, Fernandez E, Nelson R, (1996), *"Webvision : The Organization of the Retina and the Visual*

System" From http://webvision.med.utah.edu/

[Kuffler53] Kuffler S.W, (1953), *"Discharge patterns and functional organization of mammalian retina"* Journal of Neurophysiology, vol. 16, 37-68.

[Kühne01] Kühne G, Richter S, Beier M, (2001), *"Motion-based segmentation and contour based classification of video objects"*, in Proceedings of the ninth ACM international conference on Multimedia, 41-50.

[LeBorgne04] Le Borgne H. (2004) "*Analyse de données par réseau de neurones auto-organisés"* Thèse de doctorat de l'INPG, Grenoble, France.

[Mallat99] Mallat S, (1999), *"A Wavelet Tour of Signal Processing"* Academic Press; 2nd edition (September 15, 1999) ISBN: 978-0124666061.

[Marcelja80] Marcelja S, (1980), *"Mathematical description of the response of simple cortical cells"* Journal of Optical Society of America, vol. 70, 1297-1300.

[Massot06] Massot C, (2006), *"Perception 3D dans le cortex visuel"*. Thèse de Doctorat de l'Université Joseph Fourier, Grenoble, France.

[McGillSite] http://www.lecerveau.mcgill.ca/

[Mead88] Mead C.A, Mahowald M. A, (1988),*"A silicon model of early visual processing"* Neural Networks, 1:91--97.

[MpisearchSite] The Machine Perception Toolbox (MPT): http://mplab.ucsd.edu/grants/project1/free-software/mptwebsite/API/index.html

[Nabet92] Nabet B, (1992), *"Nonlinear vision: determination of neural receptive fields, function and networks"* Electronic hardware for vision modeling, chapter 18, 463-474 CRC Press. In R. B. Pinter and B. Nabet editors

[OpenCVSite] Open Source Computer Vision Library: www.intel.com/technology/computing/opencv/

[OpenInterfaceSite] OpenInterface is the SIMILAR Software Platform that allows integration of software components dedicated to multimodal interaction and multimodal data fusion: http://www.openinterface.org/

[Palagi92] Palagi P.M, (1992), *"Décomposition fréquentielle des textures: caractérisation et segmentation"* Thèse de Doctorat, Institut National Polytechnique, Grenoble.

[PETS] Performance Evaluation of Tracking and Surveillance: http://www.cvg.cs.rdg.ac.uk/

[Richefeu04] Richefeu J, Manzanera A, (2004), *"A new hybrid differential filter for motion detection"* ICCVG'04 ,Warsaw, Poland, 22-24.

[Rivet06] Rivet B, (2006), *"La bimodalité de la parole au secours de la séparation de sources"* thèse de doctorat INPG, Grenoble, France.

[Saric04] Šarić, Z., Jovičić, S., (2004) *"Adaptive microphone array based on pause detection"* Acoustics Research Letters Online, Volume 5, Issue 2, 68-74..

[Sherman02] Guillery RW, Sherman SM (2002) *"The thalamus as a monitor of motor outputs"* Philosophical Transactions of the Royal Society of London vol. 357: 1809-1821

[SignLanguageTutoringTool06] Aran O, Ari I, Benoit A, Huerta Carrillo A, Fanard F.X, Campr P, Akarun L,

Caplier A, Rombaut M, Sankur B, (2006), *"Sign Language Tutoring Tool"* eNTERFACE'06 workshop, Dubrovnik, Croatia.

[SimilarSite] Réseau d'excellence européen www.similar.cc 2003-2007.

[Spinei98] Spinei A. (1998) "Estimation du mouvement par triades de filtres de Gabor. Application au mouvement d'objets transparents" Thèse de doctorat INPG, Grenoble, France.

[Smirnakis97] Smirnakis S.M, Berry M.J, Warland D.K, Bialek W, Meister M, (1997), *"Adaptation of Retinal Processing to Image Contrast and Spatial Scale*" Nature vol. 386, 69-73.

[Szmajda05] Szmajda B.A, Jusuf P.R, Grunert U. Martin P.R, (2005), *"Projection of wide-field ganglion cells to the lateral geniculate nucleus of the marmoset"* Proceedings of the Australian Neuroscience Society. vol. 16, 91.

[Tao03] Tao L, Qi L, Shenghuo Z, and Mitsunori O, (2003), *"Survey on Wavelet Applications in Data Mining"* SIGKDD EXplorations, January 2003. Volume 4, Issue 2, Pages 49-68

[TorcsSite] TORCS - The Open Racing Car Simulator: http://torcs.sourceforge.net/

[Torralba97] Torralba A.B, Hérault J, (1997), *"From retinal circuits to motion processing: a neuromorphic approach to velocity estimation"* 5th Eurorean Symposium on Artificial Neural Networks (ESANN '97), Bruges, Belgium, 47-54.

[Torralba99] Torralba A.B.(1999) *"Analogue Architectures for Vision Cellular Neural Networks and Neuromorphic Circuits"*. Thèse de Doctorat de l'Université Joseph Fourier, Grenoble, France.

[Tsaig02] Tsaig Y, Averbuch A, (2002), *"Automatic segmentation of moving objects in video sequences: a region labeling approach"* IEEE Trans. on Circuits Syst. Video vol. 12, pp. 597--612.

[Viola02] Viola P. and Jones M. (2002) "Robust real-time object detection". Int'l. J. Computer Vision, 57(2):137–154.

[Watson85] Watson A. B, Ahumada A.J, (1985), *"Model of human visual-motion sensing"*. Journal Optical Society of America, vol. 2, no. 2, 322-42

[Werblin88] Werblin F, Maguire G, Lukasiewicz P, Eliasof S, Wu S.M, (1988), *"Neural interactions mediating the detection of motion in the retina of the tiger salamander"*. Vis. Neurosci. vol. 1, 317-29.

[Xiao03] Xiao J, Moriyama T, Kanade T, Cohn J.F, (2003), *"Robust Full-Motion Recovery of Head by Dynamic Templates and Re-Registration Techniques,"* International Journal of Imaging Systems and Technology, vol. 13, pp. 85-94, Sept. 2003.

[Yellott82] Yellott J.I, (1982), *"Spectral analysis of spatial sampling by photoreceptors: Topological disorder prevents aliasing"* Vision Research, vol. 22, 1205-1210.

[Zhang01] Zhang D, Lu G, (2001), *"Segmentation of Moving Objects in Image Sequence: A Review",* Circuits, systems, and signal processing : CSSP vol.20, no.2, 143 - 184.

[Zivkovic01] Zivkovic Z, Van der Heijden F, (2001), *"A stabilized adaptive appearance changes model for 3D head tracking"*. In IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (RATFG-RTS'01), 175-182.

Publications

ARTICLES EN REVUE INTERNATIONALE

◦ Benoit A, Bonnaud L, Caplier A, Ngo P, Lawson L, Trevisan D.G, (2006), “Multimodal Focus Attention Detection in an Augmented Driver Simulator”, Personal and Ubiquitous Computing (accepted, not yet published).

◦ A. Benoit, L. Bonnaud, A. Caplier, Y. Damousis, F. Jourde, L. Lawson, L. Nigay, M. Serrano, D. Tzovaras, (2007), “Multimodal Signal Processing and Interaction for a Driving Simulator: Component-based Architecture”, Journal of Multimodal User Interface (JMUI).

◦ Aran O. Ari I. Benoit A. Carrillo A.H. Fanard F.X. Campr P. Akarun L, Caplier A. Rombaut M. Sankur B. (2006), “Sign Language Tutoring Tool”, eNTERFACE 2006, Dubrovnik, Croatia (extended version).

En cours de relecture:

◦ Benoit A, Caplier A, (2006), “Using Visual Human System Properties For The Interpretation Of Rigid And Non Head Gestures Involved In Verbal Communication Process”, International Journal of Image and Video Processing (IJIVP).

CONFERENCES INTERNATIONALES

◦ Aran O. Ari I. Benoit A. Carrillo A.H. Fanard F.X. Campr P. Akarun L, Caplier A. Rombaut M. Sankur B. (2006), “Sign Language Tutoring Tool”, eNTERFACE 2006, Dubrovnik, Croatia.

◦ Benoit A. Bonnaud L. Caplier A. Damousis Y. Tzovaras D. Jourde F. Nigay L. Serrano M. Lawson L. (2006), “Multimodal Signal Processing and Interaction for a Driving Simulator: Component-based Architecture”, eNTERFACE 2006, Dubrovnik, Croatia.

◦ Burger T, Benoit A, Caplier A, (2006), “Intercepting static hand gestures in dynamic context”, IEEE ICIP, Atlanta, USA, September 2006.

◦ Benoit A, Bonnaud L, Caplier A, Ngo P, Lawson L, Trevisan D.G, (2006), “Multimodal Focus Attention Detection in an Augmented Driver Simulator”, AIAI, Athens, Greece.

◦ Benoit A, Caplier A, (2005), “Motion Estimator Inspired From Biological Model For Head Motion Interpretation” IEE WIAMIS, Montreux, Switzrland.

◦ Benoit A, Caplier A, (2005), “Head Nods Analysis : Interpretation Of Non Verbal Communication Gestures” IEEE ICIP, Genova, Italy.

◦ Benoit A, Caplier A, (2005), “Biological Approach For Head Motion Detection And Analysis”, EUSIPCO, Antalya, Turkey.

◦ Benoit A, Caplier A, (2005), “Hypovigilence Analysis: Open or Closed Eye or Mouth ? Blinking or Yawning Frequency?” IEEE AVSS, Como, Italy.

CONFERENCE NATIONALE

◦ Benoit , A. Caplier, A. (2005), “Interprétation des mouvements de tête impliqués dans le processus de communication non verbale” Orasis, Fournol, France.

The human visual system as a complete solution for image processing

The aim of this work is to demonstrate that the human visual modelling yields to efficient tools dedicated to image processing. The benefits are directly related to the human visual system properties and give the ability to solve common major problems encountered in image analysis such as noise reduction, details enhancement, back light correction. We propose a set of low level image processing tools which realize specific processing such as contours extraction, spectrum analysis, event detection. These tools are combined in order to create high level image analysis and we propose in this manuscript two examples: a face analyser which extracts eye blinks, yawnings, head motion, this application being applied to hypovigilance detection for drivers. The second considered application concerns general motion analysis and we propose a moving object tracker able to eliminate noise segmentation problems. Finally, we propose a bio-inspired algorithm for object identification and classification.

Le système visuel humain au secours de la vision par ordinateur

L'objectif de ce travail est de mettre en évidence l'intérêt d'utiliser des modélisations du système visuel humain pour développer des outils de traitement d'images. En effet, il est reconnu que le SVH peut s'affranchir d'un certain nombre de difficultés couramment rencontrées en vision par ordinateur: les problèmes d'éclairage, les contre-jours, le bruit, etc. Nous proposons un ensemble de modules de traitement d'image bas niveau qui permettent de réaliser des tâches spécifiques telles que l'extraction de contours, l'analyse spectrale et la détection de mouvement. Ces outils sont assemblés dans le but d'effectuer des tâches d'analyse de plus haut niveau. Nous présentons deux applications possibles: un système d'analyse du visage capable de détecter les clignements d'yeux, les bâillements, inséré dans un système de détection de l'hypovigilance chez le conducteur. Une seconde application traite de l'analyse du mouvement en général avec un système de suivi et de reconnaissance d'objets.