DE LA PERCEPTION LOCALE DES DISTORSIONS DE CODAGE A L'APPRECIATION GLOBALE DE LA QUALITE VISUELLE DES IMAGES ET VIDEOS. APPORT DE L'ATTENTION VISUELLE DANS LE JUGEMENT DE QUALITE

Alexandre Ninassi

▶ **To cite this version:**

Alexandre Ninassi. DE LA PERCEPTION LOCALE DES DISTORSIONS DE CODAGE A L'APPRECIATION GLOBALE DE LA QUALITE VISUELLE DES IMAGES ET VIDEOS. APPORT DE L'ATTENTION VISUELLE DANS LE JUGEMENT DE QUALITE. Interface homme-machine [cs.HC]. Université de Nantes, 2009. Français. NNT : . tel-00426909

HAL Id: tel-00426909

https://theses.hal.science/tel-00426909

Submitted on 28 Oct 2009

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.

UNIVERSITÉ DE NANTES

ÉCOLE DOCTORALE

« SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE MATHEMATIQUES »

Année : 2009

Thèse de Doctorat de l'Université de Nantes

Spécialité : TRAITEMENT DU SIGNAL ET INFORMATIQUE APPLIQUEE

Présentée et soutenue publiquement par

Alexandre NINASSI

le 17 mars 2009

à l'École polytechnique de l'université de Nantes

DE LA PERCEPTION LOCALE DES DISTORSIONS DE CODAGE A L'APPRECIATION GLOBALE DE LA QUALITE VISUELLE DES IMAGES ET VIDEOS. APPORT DE L'ATTENTION VISUELLE DANS LE JUGEMENT DE QUALITE

Jury

Présidente	:	Mme GUERIN-DUGUE Anne	Professeur des Universités, INPG, Grenoble
Rapporteurs	:	M. COUDOUX Francois-Xavier	Professeur des Universités, Université de Valenciennes
		M. LABIT Claude	Directeur de Recherche, INRIA, Rennes
Examinateurs	:	M. BARBA Dominique	Professeur des Universités, Polytech'Nantes
		M. EBRAHIMI Touradj	Professeur, EPFL, Lausanne
		M. LE CALLET Patrick	Professeur des Universités, Polytech'Nantes
		M. LE MEUR Olivier	Ingénieur de recherche, THOMSON R&D, Cesson-Sévigné

Directeur de Thèse : Dominique BARBA
Laboratoire : IRCCyN
Co-encadrant : Patrick LE CALLET
Laboratoire : IRCCyN
Composante de rattachement du directeur de thèse : École polytechnique de l'université de Nantes

N° ED 503-041

À Sophie.

Remerciements

Je tiens à remercier

Madame Anne Guérin-Dugué, Professeur des Universités à l'INPG de Grenoble, qui m'a fait l'honneur de présider le jury de cette thèse ;

Monsieur François-Xavier Coudoux, Professeur des Universités à l'Université de Valenciennes, et Monsieur Claude Labit, Directeur de Recherche à l'INRIA de Rennes, d'avoir bien voulu accepter la charge de rapporteur ;

Monsieur Touradj Ebrahimi, Professeur à l'EPFL de Lausanne, d'avoir bien voulu juger ce travail ;

Dominique Barba, directeur de cette thèse, Patrick Le Callet et Olivier Le Meur pour m'avoir conseillé et guidé tout au long de mes travaux ;

Philippe Guillotel, responsable du laboratoire Compression de Thomson R&D France, pour m'avoir accueilli dans son équipe ;

Tous les membres du laboratoire Compression de Thomson R&D France et de l'équipe Image Vidéo Communication de l'IRCCyN pour leur accueil, leur sympathie et leur contribution à ce travail ;

Guillaume et Ana pour leur relecture finale ;

Ma famille, mes amis.

Table des matières

Introduction générale **1**

I Distorsions visuelles en images et vidéos **5**

Introduction **7**

1 État de l'art **9**

1.1 Introduction 9

1.2 Vidéo numérique et facteurs humains 9

1.2.1 Représentation et codage 10

1.2.2 Les distorsions 14

1.3 Modélisation du système visuel humain 16

1.3.1 La perception de la luminance 17

1.3.2 La perception des couleurs 18

1.3.3 La sensibilité au contraste 19

1.3.4 L'organisation multi-canal 24

1.3.5 Les effets de masquage 26

1.4 Conclusion 30

2 Imagerie des distorsions perçues : conception de nouvelles méthodes et validation comparative **33**

2.1 Introduction 33

2.2 Revue des méthodes existantes 34

2.2.1 Les approches purement de type signal 35

2.2.2 Les approches structurelles 35

2.2.3 Les approches modélisant le système visuel humain 38

2.2.4 Discussion 40

2.3 Principe général des deux modèles proposés 40

2.4 Espace de couleur et adaptation en luminance 42

2.5 Modélisation du comportement multi-canal 44

2.5.1 Décomposition en canaux perceptuels (DCP) . . . 44
2.5.2 Adaptation de la transformée en ondelettes . . . 44
2.6 Modélisation de la sensibilité aux contrastes . . . 49
2.6.1 Modèle basé Fourier . . . 49
2.6.2 Modèle basé ondelettes . . . 50
2.7 Modélisation de l'effet de masquage spatial . . . 51
2.7.1 Masquage de contraste . . . 51
2.7.2 Masquage semi-local . . . 52
2.8 Normalisation et cumul inter sous-bandes des erreurs . . . 55
2.9 Résultats qualitatifs . . . 56
2.9.1 Comparaison de cartes de distorsions perceptuelles . . . 56
2.9.2 Perspectives d'amélioration des cartes de distorsions perceptuelles . . . 61
2.10 Conclusion . . . 63

3 Conception de séquences de distorsions visuelles de vidéos 65

3.1 Introduction . . . 65
3.2 Revue des méthodes existantes . . . 66
3.2.1 Les approches purement de type signal . . . 66
3.2.2 Les approches structurelles . . . 67
3.2.3 Les approches modélisant le système visuel humain . . . 67
3.2.4 Discussion . . . 68
3.3 Principe général du modèle proposé . . . 70
3.4 Cumul temporel court terme . . . 72
3.5 Les tubes spatio-temporels . . . 72
3.5.1 Estimation de l'information de mouvement . . . 74
3.5.2 Construction des tubes spatio-temporels . . . 74
3.6 Filtrage temporel des distorsions spatiales dans les tubes . . . 77
3.7 Évaluation temporelle des distorsions dans les tubes . . . 77
3.8 Résultats qualitatifs . . . 79
3.9 Conclusion . . . 82

Conclusion **83**

II Critères objectifs de qualité visuelle d'images et de vidéos **85**

Introduction **87**

4 État de l'art sur l'évaluation subjective et objective de la qualité visuelle d'images et de vidéos **89**

4.1 Introduction . . . 89
4.2 Tests subjectifs d'évaluation de qualité . . . 90
4.2.1 Tests subjectifs : maîtriser l'environnement . . . 90
4.2.2 Tests subjectifs : les sources de biais . . . 91
4.2.3 Les différents protocoles de tests subjectifs . . . 92
4.2.4 Traitement des résultats obtenus lors de tests . . . 96
4.2.5 Conclusion . . . 99
4.3 Indicateurs de performance de critères objectifs de qualité visuelle . . . 99
4.3.1 Coefficient de corrélation linéaire : indicateur de précision . . . 100
4.3.2 Coefficient de corrélation de rang : indicateur de monotonicité . . . 101
4.3.3 Outlier ratio : indicateur de cohérence . . . 101
4.3.4 Erreur de prédiction de la qualité . . . 102
4.3.5 Tests de significativité ou tests statistiques de différence significative . . . 102
4.4 Métriques de qualité visuelle avec référence complète . . . 103
4.4.1 Approches reposant sur le calcul de cartes de distorsions : cumul spatial des distorsions . 104
4.4.2 Approches reposant sur le calcul de séquences temporelles de cartes de distorsions : cumul spatial et temporel des distorsions . . . 106
4.4.3 Autres approches . . . 108
4.4.4 Conclusion . . . 109
4.5 Conclusion . . . 110

5 Critères objectifs de qualité visuelle d'images **111**

5.1 Introduction . . . 111
5.2 Cumul spatial . . . 111
5.3 Expérimentations . . . 112
5.3.1 Bases d'évaluation subjective . . . 112
5.3.2 Évaluation des performances . . . 116
5.4 Conclusion . . . 123

6 Critères objectifs de qualité visuelle de vidéos **125**

6.1 Introduction . . . 125

6.2 Cumul spatial et cumul temporel . . . 126

6.2.1 Cumul spatial . . . 127

6.2.2 Cumul temporel . . . 128

6.3 Expérimentations . . . 131

6.3.1 Base d'évaluation subjective . . . 131

6.3.2 Évaluation des performances . . . 132

6.4 Conclusion . . . 140

Conclusion **141**

III Attention visuelle et construction du jugement de qualité visuelle **143**

Introduction **145**

7 État de l'art sur l'attention visuelle **147**

7.1 Introduction . . . 147

7.2 Les mouvements oculaires et l'attention visuelle . . . 147

7.2.1 Les mouvements oculaires . . . 147

7.2.2 L'attention visuelle sélective . . . 149

7.3 Attention visuelle et évaluation de qualité . . . 152

7.3.1 Attention visuelle et évaluation subjective de la qualité . . . 152

7.3.2 Attention visuelle et évaluation objective de la qualité . . . 152

7.4 Conclusion . . . 154

8 Attention visuelle et construction du jugement de qualité d'images **155**

8.1 Introduction . . . 155

8.2 Expérimentations oculométriques et tests subjectifs de qualité . . . 155

8.2.1 Dispositif oculométrique : l'oculomètre . . . 156

8.2.2 Exploration libre . . . 158

8.2.3 Tâche d'évaluation de qualité . . . 158

8.2.4 Déroulement de l'ensemble des tests oculométriques . . . 159

8.2.5 Construction d'une saillance spatiale . . . 160

8.3 Impact de la tâche d'évaluation de la qualité sur l'attention visuelle . . . 161

8.3.1 Tâche et durée des fixations . . . 162

8.3.2 Tâche et cartes de saillance . . . 163

8.3.3 Discussion . . . 169
8.4 Impact de l'attention visuelle sur les performances de métriques de qualité . . . 172
8.4.1 Métriques de qualité basées saillance . . . 172
8.4.2 Analyse quantitative . . . 173
8.4.3 Discussion . . . 177
8.5 Conclusion . . . 179

9 Attention visuelle et construction du jugement de qualité de vidéos 181
9.1 Introduction . . . 181
9.2 Expérimentations oculométriques et tests subjectifs de qualité . . . 182
9.2.1 Exploration libre . . . 182
9.2.2 Tâche d'évaluation de qualité . . . 183
9.2.3 Déroulement de l'ensemble des tests oculométriques . . . 184
9.2.4 Construction d'une saillance spatio-temporelle . . . 184
9.3 Impact de la tâche d'évaluation de la qualité sur l'attention visuelle . . . 187
9.3.1 La tâche et la durée des fixations/poursuites . . . 187
9.3.2 La tâche et les séquences de saillance . . . 189
9.3.3 Discussion . . . 194
9.4 Impact de l'attention visuelle sur les performances de métriques de qualité . . . 195
9.4.1 Métriques de qualité fondées sur la saillance . . . 196
9.4.2 Analyse quantitative . . . 197
9.4.3 Discussion . . . 198
9.5 Conclusion . . . 199

Conclusion 200

Conclusion et perspectives 201

Annexes 204

A Biologie du système visuel humain 207
A.1 L'oeil : organe de la vision . . . 207
A.2 La rétine . . . 208
A.2.1 La rétine : Une structure multicouche . . . 209
A.2.2 Les champs récepteurs . . . 210
A.3 De la rétine au cortex . . . 211

B Implémentation de la DCP **215**

C Résultats par observateur des métriques de qualité d'images basées saillance **219**

Bibliographie **224**

Publications liées à la thèse **235**

Introduction générale

A peine plus d'un siècle nous sépare de la naissance du cinéma. Une invention fantastique qui fut rendue possible grâce au développement de la photographie. Depuis cette époque, la technologie de l'image n'a cessé d'évoluer et de se perfectionner. Alors que la télévision analogique vit ses dernières années en France, les vidéos numériques sont accessibles depuis nos téléviseurs, nos ordinateurs et même depuis nos téléphones portables. La télévision numérique migre déjà vers la télévision numérique haute définition. La vidéo numérique en général devient un contenu presque aussi commun sur internet que les images numériques. De nos jours, les images et les vidéos numériques sont omniprésentes et la quantité de données associées est gigantesque.

Ces évolutions ont nécessité le développement d'un bon nombre de techniques de traitement de l'image et la vidéo. Au centre de toutes ces techniques se trouve un spectateur que l'on cherche à satisfaire.

L'être humain, dont le spectateur fait partie, perçoit son environnement à l'aide de ses systèmes sensoriels, lesquels ne sont pas parfaits. Le système visuel humain n'échappe pas cette imperfection. C'est pourquoi les techniques de traitement d'images et de vidéos ont tout intérêt à prendre en compte et à exploiter la manière dont l'homme perçoit son environnement visuel. Cependant, la réalité est toute autre. Les différents maillons de la chaîne de traitement d'images ou de vidéos sont autant de sources de distorsions pouvant altérer leur qualité visuelle. Il est donc rapidement apparu nécessaire d'évaluer la qualité visuelle produite par un système de traitement d'images ou de vidéos. Dans ce domaine, on distingue deux catégories de méthodes : les méthodes subjectives et les méthodes objectives. Les méthodes subjectives sont les plus proches de la réalité, car elles consistent à faire évaluer la qualité par un groupe d'observateurs. Cependant, ces méthodes sont gourmandes en temps et demandent des conditions expérimentales de visualisation bien précises. Par conséquent, elles représentent un coût trop important pour être utilisées de façon systématique dans l'industrie du traitement d'images ou de vidéos. D'où la nécessité de développer des méthodes objectives permettant d'évaluer la qualité visuelle de façon automatique. Ces méthodes objectives sont appelées des métriques de qualité visuelle. Les métriques de qualité sont classées en trois catégories en fonction de la nature des informations nécessaires pour effectuer l'évaluation. Ces trois catégories sont :

- Les métriques de qualité avec référence complète, notées FR (*Full Reference*) [Le Callet 01], qui comparent la version de l'image ou de la vidéo à évaluer avec une version de référence de celle-ci. Pour cette catégorie de métrique, la version originale et la version à évaluer doivent être disponibles.
- Les métriques de qualité avec référence réduite, notées RR (*Reduced Reference*) [Carnec 04], qui comparent

une description de l'image ou de la vidéo à évaluer avec une description d'une version de référence de celle-ci. Une description est un ensemble de paramètres mesurés sur l'image ou la vidéo. Pour cette catégorie de métrique, la version à évaluer et une description de la version originale doivent être disponibles.

– Les métriques sans référence, notées NR (*No Reference*) [Wang 02b], qui caractérisent les distorsions de l'image ou de la vidéo à évaluer uniquement à partir de celle-ci et, éventuellement, à partir de connaissances a priori sur celles-ci. Pour cette catégorie, seule la version à évaluer est requise.

La performance des métriques de qualité ainsi que leur caractère généraliste dépendent de la quantité d'information disponible. Les métriques de qualité avec référence complète sont censées être les plus précises et les plus robustes. Les deux autres catégories nécessitent souvent, mais pas toujours, des connaissances a priori sur les distorsions présentes dans l'image ou la vidéo à évaluer, elles sont donc plutôt adaptées à des situations spécifiques.

L'évaluation de la qualité globale des images et des vidéos n'est pas le seul le besoin des concepteurs de systèmes de traitement d'images ou de vidéos. L'évaluation des distorsions visuelles perçues localement est aussi une information très recherchée pour la mise au point des systèmes de traitement d'images ou de vidéos. Cette évaluation locale des distorsions visuelles est aussi beaucoup plus difficile à mettre en oeuvre. Dans ce cas de figure, il n'existe pas véritablement de méthodes subjectives permettant d'atteindre la « vérité terrain ». Par conséquent, il n'est pas non plus possible d'évaluer directement les performances d'une méthode objective d'évaluation locale des distorsions perceptuelles. Cette réalité participe sans doute au fait que les recherches dans le domaine se dirigent davantage vers l'optimisation des méthodes de construction d'une note de qualité visuelle globale, que vers des méthodes objectives d'évaluation locale des distorsions visuelles. Ceci explique peut-être que les méthodes actuelles d'évaluation locale des distorsions visuelles ne permettent pas d'obtenir des résultats satisfaisants.

Si la problématique de l'évaluation locale des distorsions perceptuelles des images a été abordée à de nombreuses reprises dans la littérature, la problématique de l'évaluation locale des distorsions perceptuelles de séquences d'images animées l'a été beaucoup moins du fait de sa complexité. Pourtant cette problématique s'avère fondamentale pour de nombreuses applications. Par exemple une adaptation locale des techniques de codage et de compression pilotée par une évaluation locale des distorsions perceptuelles pourrait permettre d'améliorer significativement la qualité visuelle des vidéos encodées. La question récurrente est celle de la répartition du débit disponible dans les différentes zones spatiales des images, ou spatio-temporelles des vidéos, afin d'obtenir la meilleure qualité visuelle possible.

Notre travail

Dans cette thèse, nous nous intéressons à la fois à l'évaluation de la qualité d'images et de vidéos avec référence complète, et à la fois à l'évaluation locale des distorsions perceptuelles dans les images et les vidéos.

D'une façon générale, l'évaluation objective de la qualité visuelle avec référence complète doit reposer sur une modélisation du système visuel humain. On peut penser a priori que plus cette modélisation est complète,

meilleures seront les performances. Les métriques de qualité d'images s'appuyant sur les modélisations les plus poussées du système visuel humain obtiennent les meilleures performances, mais cela au prix d'une complexité opératoire élevée. Cette complexité est souvent un frein à leur utilisation par les concepteurs de systèmes de traitement d'images et surtout de vidéos. C'est pourquoi nous nous sommes intéressés à comment les simplifier.

Ces modélisations ont souffert de l'absence de « vérité terrain » concernant l'évaluation locale des distorsions visuelles, et certaines propriétés intéressantes n'ont pas été modélisées faute, peut-être, de ne pouvoir évaluer leur impact. Il nous a alors semblé pertinent de tenter de prendre en compte certaines de ces propriétés, comme le masquage semi-local.

L'évaluation objective des distorsions perceptuelles spatio-temporelles que l'on rencontre dans les séquences d'images animées ainsi que l'évaluation objective de la qualité globale des vidéos sont des problématiques qui ont été encore peu abordées. C'est pourquoi nous avons orienté une partie de nos recherches dans cette direction. Pour cela nous nous intéressons aux variations temporelles des distorsions perceptuelles spatiales.

Nous nous sommes aussi intéressés à d'autres mécanismes du système visuel humain qui avaient été très peu étudiés dans un contexte d'évaluation de qualité quand nous avons débuté nos travaux. Il s'agit de l'attention visuelle et de ses mécanismes de sélection de l'information. De nombreuses questions se posent quant à son influence dans la construction du jugement de qualité. Il semble, en effet, assez plausible qu'un artefact de codage apparaissant sur une zone de forte saillance soit plus gênant qu'un artefact apparaissant sur une zone peu ou pas intéressante visuellement. Comprendre le fonctionnement de ses mécanismes dans un contexte d'évaluation de la qualité visuelle peut permettre d'améliorer les performances des métriques de qualité.

Ces travaux s'inscrivent dans la continuité de deux autres travaux de thèse effectués au sein de l'équipe Image Vidéo-Communication de l'IRCCyN :

- Patrick Le Callet [Le Callet 01] a développé dans sa thèse des métriques de qualité d'images couleurs reposant sur des modélisations élaborées du système visuel humain. Ces métriques ont été évaluées à partir de tests subjectifs d'évaluation de la qualité.
- Olivier Le Meur [Le Meur 05] a proposé dans sa thèse des modèles d'attention visuelle, pour images fixes et images animées, reposant sur des modélisations élaborées du système visuel humain. Ces modèles d'attention visuelle ont été testés à partir de tests oculométriques.

Structure de ce mémoire

L'évaluation subjective de la qualité visuelle globale passe par l'évaluation subjective des distorsions au niveau local. De même, la première étape des méthodes d'évaluation objective de la qualité, consiste généralement à évaluer localement les distorsions perceptuelles avant de les cumuler en une note de qualité globale. Comme nous l'avons évoqué précédemment chacune de ces deux étapes fait écho à un besoin particulier de l'industrie de l'image et de la vidéo. C'est pourquoi nous avons consacré une partie de ce mémoire à chacune d'elles. Ensuite, les aspects novateurs que présentent les interactions entre la construction du jugement de qualité et l'attention visuelle en font un sujet d'étude à part entière. Ce mémoire est donc divisé en trois parties :

- La première partie (chapitre 1 à 3) traite de l'évaluation locale des distorsions visuelles en proposant des méthodes objectives d'évaluation pour les images et pour les vidéos.
- La seconde partie (chapitre 4 à 6) est dédiée à l'évaluation objective de la qualité en proposant des métriques de qualité visuelle pour les images et pour les vidéos.
- La troisième partie (chapitre 7 à 9) revient sur la construction du jugement de qualité en étudiant l'attention visuelle et son impact sur l'évaluation de la qualité des images et des vidéos.

Première partie

Distorsions visuelles en images et vidéos

Introduction

La première partie de cette thèse est consacrée à l'évaluation locale des distorsions. Schématiquement et sans entrer dans le détail, la qualité d'images ou de vidéos peut être définie comme une fonction d'un ensemble de gênes visuelles. La gêne, elle, peut être définie comme une fonction de la perception des dégradations et de la nature de la zone qui les porte. La perception des distorsions correspond à la sensation provoquée par la visibilité d'erreurs. La visibilité des erreurs est considérée au niveau local, et elle est établie par rapport à un seuil appelé seuil de visibilité. Si une erreur est inférieure au seuil de visibilité, elle est invisible, sinon elle est visible. On peut donc considérer quatre niveaux d'analyse : la visibilité, la perception, la gêne et enfin la qualité. Cette première partie se situe aux niveaux de la visibilité et de la perception.

Dans cette partie, nous tentons d'apporter des éléments de réponse à un besoin des concepteurs de systèmes de traitement d'images ou de vidéos en proposant des méthodes objectives évaluant localement les distorsions. Les résultats de ces méthodes se présentent sous la forme de cartes de distorsions visuelles pour les images, et sous la forme de séquences de distorsions visuelles pour les vidéos.

Les méthodes proposées sont des méthodes d'évaluation locale des distorsions avec référence complète. Ceci sous-entend d'une part que la version à évaluer est comparée à une version de référence qui doit être disponible, d'autre part qu'il n'est pas nécessaire d'avoir des connaissances a priori sur la nature des distorsions. Cependant, il peut être intéressant de connaître les sources et les types de distorsions que nous pouvons rencontrer et qui sont principalement liés dans ce qui nous intéresse aux systèmes de compression et de codage de l'information.

Évaluer la qualité visuelle, c'est évaluer la qualité perçue au travers du système visuel humain. Une modélisation de celui-ci est donc indispensable pour élaborer une méthode d'évaluation objective.

Cette partie se décompose en trois chapitres.

Dans le chapitre 1 nous présentons d'une part la vidéo numérique, dont l'image numérique peut être considérée comme un cas particulier, au travers de ses grands concepts et des distorsions liées au codage numérique. D'autre part, nous détaillons les éléments de modélisation du système visuel humain ayant un intérêt particulier dans un contexte d'évaluation de qualité.

Le chapitre 2 est dédié à l'évaluation locale des distorsions dans les images. Nous y présentons les méthodes existantes, pour ensuite proposer de nouvelles méthodes permettant d'améliorer certains aspects de la modélisation, ou d'en simplifier d'autres sans diminuer les performances.

Le chapitre 3 est consacré à l'évaluation locale des distorsions dans les vidéos. Nous y présentons les méthodes existantes, pour ensuite proposer une méthode innovante basée sur l'évaluation des distorsions à l'échelle des fixations oculaires.

Chapitre 1

État de l'art

1.1 Introduction

Dans cette première partie nous nous intéressons à l'évaluation locale des distorsions dans les images et les vidéos numériques. Derrière le terme « numérique » se trouve un certain nombre de concepts importants pour comprendre les sources et la nature des distorsions que l'on peut rencontrer. C'est pourquoi, la première partie de ce chapitre est consacrée à la vidéo numérique et aux distorsions qui lui sont associées.

Comme nous l'avons évoqué précédemment, une méthode objective d'évaluation de la qualité ou d'évaluation locale des distorsions doit reposer sur un modèle du système visuel humain. La modélisation d'un système biologique quel qu'il soit, en l'occurrence ici le système visuel humain, nécessite la connaissance des étapes biologiques du traitement de l'information. La biologie du système visuel ayant déjà fait l'objet d'études détaillées dans de nombreuses thèses dans notre domaine, nous nous concentrerons dans ce chapitre sur la modélisation de celui-ci. Cependant, le lecteur pourra se reporter à l'annexe A pour plus de détails sur les mécanismes et le fonctionnement du système visuel humain. La seconde partie de ce chapitre sera consacrée à la présentation des modélisations existantes de certaines propriétés du système visuel humain. Ces propriétés sont choisies pour leur intérêt dans l'élaboration de méthodes d'évaluation des distorsions perceptuelles.

1.2 Vidéo numérique et facteurs humains

Les images animées dans toutes leurs déclinaisons (cinéma, télévision, vidéo, etc.) sont l'une des inventions du vingtième siècle ayant le plus de succès. Plus récemment, le développement d'algorithmes de compression performants a facilité le passage de l'analogique au numérique. De nos jours le numérique est présent pratiquement partout dans la chaîne : de la production à la distribution. Les principaux objectifs dans l'élaboration des équipements vidéo numériques étaient de réduire les besoins en bande passante et en espace de stockage tout en proposant un niveau de qualité visuelle au moins égal à celui des équipements analogiques. La qualité visuelle est donc arrivée au centre des préoccupations des professionnels. C'est pourquoi la prise en compte des facteurs humains dans le codage et la représentation des vidéos numériques est rapidement devenue incontournable.

Cette section débute avec un bref rappel sur le signal vidéo numérique, suivi des méthodes de compression

et des normes. Puis dans un contexte d'évaluation de la qualité visuelle, il semble pertinent de s'intéresser aux différentes distorsions que l'on peut rencontrer, en particulier les plus courantes.

1.2.1 Représentation et codage

1.2.1.1 Le codage de la couleur

De nombreuses méthodes de compression et de normes vidéo comme le PAL[ITU-R Rec. BT.470-6 98], NTSC[ITU-R Rec. BT.470-7 98], ou MPEG [1], sont déjà basées sur la vision humaine dans la façon où l'information couleur est traitée. En particulier, ils prennent en compte la non linéarité de la perception de la luminance, l'organisation en canaux de couleur, et la faible sensibilité chromatique du système visuel humain.

Dans un écran de télévision cathodique (CRT : *Cathod Ray Tube*), la relation entre les valeurs RGB (*Red-Green-Blue*) des pixels des images et l'intensité lumineuse émise par l'écran est non linéaire.

La théorie des signaux antagonistes de la perception des couleurs indique que le système visuel humain décorrèle l'information en trois signaux de différences blanc-noir, rouge-vert et bleu-jaune, qui sont traités par des canaux visuels différents (cf. section 1.3.2). De plus, l'acuité visuelle chromatique est significativement inférieure à l'acuité visuelle achromatique. Afin d'exploiter ce comportement, les couleurs primaires rouge, vert et bleu sont rarement utilisées pour le codage de l'information couleur. A la place, les signaux de différence de couleurs (chrominance) similaires à ceux mentionnés précédemment seront utilisés. Dans la vidéo à composantes séparées, par exemple, l'espace de couleurs utilisé est appelé YUV ou YCbCr, où Y code la luminance, U ou Cb code la différence entre le bleu primaire et la luminance, et V ou Cr code la différence entre le rouge primaire et la luminance. La faible acuité visuelle chromatique permet une réduction importante de la quantité d'information des signaux de différence de couleurs. En vidéo numérique, ceci est réalisé par le sous-échantillonnage chromatique. La notation communément utilisée est la suivante :

- 4 : 4 : 4 indique qu'aucun sous-échantillonnage chromatique n'a été effectué.
- 4 : 2 : 2 indique un sous-échantillonnage d'un facteur deux horizontalement ; ce format est utilisé dans la norme sur l'encodage de la télévision numérique pour studio défini par la recommandation BT601-5 [ITU-R Rec. BT.601-5 95] de l'IUT-R, par exemple.
- 4 : 2 : 0 indique un sous-échantillonnage d'un facteur deux à la fois horizontalement et verticalement ; c'est la plus proche approximation de l'acuité chromatique humaine réalisée seulement par sous-échantillonnage chromatique. Ce format est le plus couramment utilisé en Motion-JPEG ou MPEG, autrement dit pour la distribution de vidéo.
- 4 : 1 : 1 indique un sous-échantillonnage d'un facteur 4 horizontalement.

1.2.1.2 L'entrelacement

Durant le développement de la télévision analogique, il fut noté qu'un papillotement (*flicker*) pouvait être perçu à certaines cadences de rafraîchissement de l'image, et que l'importance de ce papillotement était fonction

1. `http://www.chiariglione.org/mpeg/`

de la luminosité de l'écran et des conditions d'éclairage de l'environnement. Un film projeté dans un cinéma avec des niveaux lumineux relativement bas peut être projeté à une cadence de rafraîchissement de 24 Hz sans gêne. Par contre un écran CRT nécessite une cadence de rafraîchissement supérieure à 50 Hz pour que le papillotement disparaisse. L'inconvénient d'une telle cadence de rafraîchissement de l'image est l'augmentation considérable de la bande passante nécessaire. D'un autre côté la sensibilité spatiale du système visuel humain décroît à une telle fréquence temporelle (cf. CSF dans hautes fréquences spatio-temporelles, section 1.3.3). La combinaison de ces deux propriétés a motivé la technique d'entrelacement. L'entrelacement est un compromis entre la résolution spatiale verticale et la résolution temporelle. Au lieu d'échantillonner la vidéo à 25 (PAL) ou 30 (NTSC) images par seconde, la séquence est filmée à 50 ou 60 trames (*field*) entrelacées par seconde. Une trame correspond soit aux lignes paires, soit aux lignes impaires d'une d'image, lesquelles étant échantillonnées à des instants temporels différents et affichées alternativement (la trame contenant les lignes paires est dite trame du haut ou supérieure, et la trame contenant les lignes impaires est dite trame du bas ou inférieure). Par conséquent la bande passante du signal peut être réduite par un facteur 2, alors que les résolutions horizontale et verticale sont conservées pour les zones fixes des images, et que la cadence de rafraîchissement est suffisamment élevée pour les objets dont la taille est supérieure à une ligne. La plupart des codec multimédias pour ordinateur ne supporte que le mode progressif, qui est mieux adapté pour les écrans d'ordinateur. Les normes de compression pour la vidéo numérique ont eux à supporter à la fois le mode progressif et le mode entrelacé (cf. section 1.2.1.4).

1.2.1.3 La compression vidéo numérique

Les données numériques représentant des informations visuelles comme la vidéo nécessitent des capacités de stockage et de transmission très importantes. Par exemple, une vidéo au format télévisuel classique (NTSC par exemple), c'est-à-dire non comprimée, a un débit de plusieurs centaines de Mb/s. Des techniques de compression efficaces sont nécessaires pour manipuler des quantités de données aussi importantes. La compression vidéo est une méthode de compression de données, qui consiste à réduire la quantité de données, en limitant au maximum l'impact sur la qualité visuelle de la vidéo. Des méthodes génériques de compression de données sans perte pourraient être utilisées pour compresser des images ou des vidéos. Ces méthodes auraient l'avantage d'assurer une parfaite reconstruction des données initiales, cependant elles ne sont pas vraiment adéquates car elles ne prennent en compte que les caractéristiques du flux binaire. En compression vidéo, différents types de redondances sont exploités :

- La redondance spatio-temporelle : typiquement les valeurs des pixels sont corrélées avec les valeurs des pixels dans leur voisinage. Cette corrélation est à la fois spatiale, les pixels adjacents de l'image courante sont similaires, et temporelle, les pixels des images passées et futures ont des valeurs aussi très proches de celle du pixel courant.
- La redondance psychovisuelle : la sensibilité du système visuel humain, ainsi que ces variations peuvent être exploitées en compression vidéo. L'idée est de supprimer les informations qui ne sont pas visibles pour un observateur humain. Ces méthodes de compression sont dites avec perte.

En vidéo analogique, ces deux types de redondance sont exploités au travers du codage des couleurs basé sur la vision, ainsi qu'au travers des techniques d'entrelacement. La compression de vidéo numérique offre d'autres possibilités via des techniques adéquates. La plupart des méthodes de compression vidéo numériques actuelles comportent les étapes suivantes :

- une transformation : le signal numérique initial (pixels) est transformé en un ensemble d'éléments (coefficients) dont beaucoup ont une valeur très faible. Par exemple, cette transformation peut être réalisée par une transformée discrète en cosinus (DCT), ou par une transformée en ondelettes.
- une quantification : après la transformation, la précision numérique des coefficients transformés est réduite afin de diminuer l'information transmise. Le degré de quantification appliqué à chaque coefficient peut être choisi en considérant la sensibilité du système visuel humain ; les coefficients hautes fréquences peuvent être représentés plus approximativement que les coefficients basses fréquences, par exemple. La quantification est l'étape responsable de la perte d'informations et donc une source importante de distorsions.
- le codage : après que les données aient été quantifiées en un ensemble fini de valeurs, ces valeurs quantifiées peuvent être codées sans perte en utilisant leur redondance dans le flux binaire. Un codage entropique, qui s'appuie sur le fait que certains symboles apparaissent plus fréquemment que d'autres, est souvent utilisé ici. Les schémas de codage entropique les plus connus sont le codage de Huffman et le codage arithmétique.

Un aspect essentiel de la compression vidéo numérique est d'exploiter la similarité entre images successives plutôt que de les coder indépendamment. Une méthode simple pour exploiter la redondance temporelle est de coder la différence pixel à pixel des images successives. Une compression plus importante peut être obtenue en utilisant l'estimation et la compensation de mouvement, qui est une technique pour décrire une image en se basant sur le contenu des images d'un voisinage temporel proche au moyen de vecteurs de mouvement. En compensant le mouvement des objets de cette manière, la différence entre les images, appelée l'erreur de prédiction, peut être encore réduite. Plus l'erreur de prédiction est faible, plus la compression est importante.

Le développement de la vidéo numérique a nécessité le développement de normes techniques afin de permettre l'interopérabilité du codage. La section suivante présente l'évolution de la principale famille de normes : MPEG.

1.2.1.4 Les normes

MPEG [2], acronyme de *Moving Picture Experts Group*, est le groupe de travail de l'ISO (*International Organization for Standardization*) et de la CEI (Commission Electrotechnique Internationale) pour les technologies de l'information. Ce groupe d'experts est chargé du développement de normes internationales pour la compression, la décompression, le traitement et le codage de la vidéo, de l'audio et de leur combinaison. L'activité de ce groupe de travail a commencé en 1988, et depuis il a produit :

- MPEG-1, une norme pour le stockage et l'extraction de la vidéo et de l'audio, qui a été approuvé en Novembre 1992. Elle a été prévue pour être générique, c'est-à-dire que la normalisation n'a concerné que la syntaxe de codage et le schéma de décodage. MPEG-1 définit une technique de codage hybride

2. `http://www.chiariglione.org/mpeg/`

DCT/DPCM basé bloc avec prédiction et compensation de mouvement.

- MPEG-2, une norme pour la télévision numérique, approuvée en Novembre 1994. La technique de codage proposée dans MPEG-2 est aussi générique, c'est un raffinement de MPEG-1. Des considérations particulières sont données pour les sources entrelacées. De plus de nouvelles fonctionnalités comme la scalabilité sont introduites. De façon à conserver une faible complexité pour les équipements ne nécessitant pas de supporter tous les formats vidéo, des profils et des niveaux ont été créés. Les profils décrivent les fonctionnalités, et les niveaux décrivent les paramètres comme la résolution ou la cadence de rafraîchissement. Les profils et les niveaux permettent donc de définir différentes classes de conformité MPEG-2.
- MPEG-4, une norme pour les applications multimédia, dont la première version a été approuvé en octobre 1998. MPEG-4 aborde les besoins en robustesse dans les environnements où les risques d'erreurs sont importants, les fonctionnalités d'interactivité, l'accès, et la manipulation des données basée sur le contenu, ainsi que la compression pour les bas et très bas débits. En 2001, les instances de normalisation ISO/CIE (MPEG) et l'IUT ont conjugué leurs efforts au sein du groupe de travail JVT (*Joint Video Team*) pour développer le système de codage AVC (*Advanced Video Coding*). En 2003, le système AVC est intégré en tant que partie 10 à la norme MPEG-4. MPEG-4 AVC (ou H264), permet des gains importants en compression par rapport à MPEG-2 et a déjà été retenu comme le successeur de celui-ci pour la TV haute définition, la TV sur ADSL et la TNT. Une extension de cette partie, appelée *Scalable Video Coding* (SVC), a été finalisée en 2007, et permet de proposer différents niveaux de qualité à partir d'un seul encodage (scalabilité spatiale, temporelle et en qualité)
- MPEG-7 : une norme pour décrire et chercher du contenu multimédia.
- MPEG-21 : une norme proposant une architecture pour l'interopérabilité et l'utilisation simple de tous les contenus multimédia.

La norme la plus utilisée ces dernières années à des fins commerciales a été MPEG-2, avec en particulier les DVD, la télévision numérique en définition standard (SD), et les versions SD de la télévision sur ADSL et de la TNT. Cependant, les produits reposant sur la norme MPEG-4 commencent à se développer, et en particulier pour la haute définition. MPEG-4 est présent dans les HD-DVD et les Blu-Ray, ainsi que pour la télévision numérique en haute définition (HD), et les versions HD de la télévision du ADSL et de la TNT.

Les flux vidéo MPEG-2 et MPEG-4/AVC ont une structure hiérarchique comme illustrée figure 1.1. La séquence est composée de trois types d'images codées, les images Intra (I), les images prédites unidirectionnelles (P), et les images prédites bidirectionnelles (B). Chaque image est découpée en bandes, qui sont une collection de macroblocs consécutifs ou non. Chaque macrobloc contient plusieurs blocs de 8x8 pixels. Une transformation DCT est calculée sur ces blocs, tandis que l'estimation de mouvement est calculée sur les macroblocs. Les coefficients DCT sont ensuite quantifiés et codés avec un code à longueur variable.

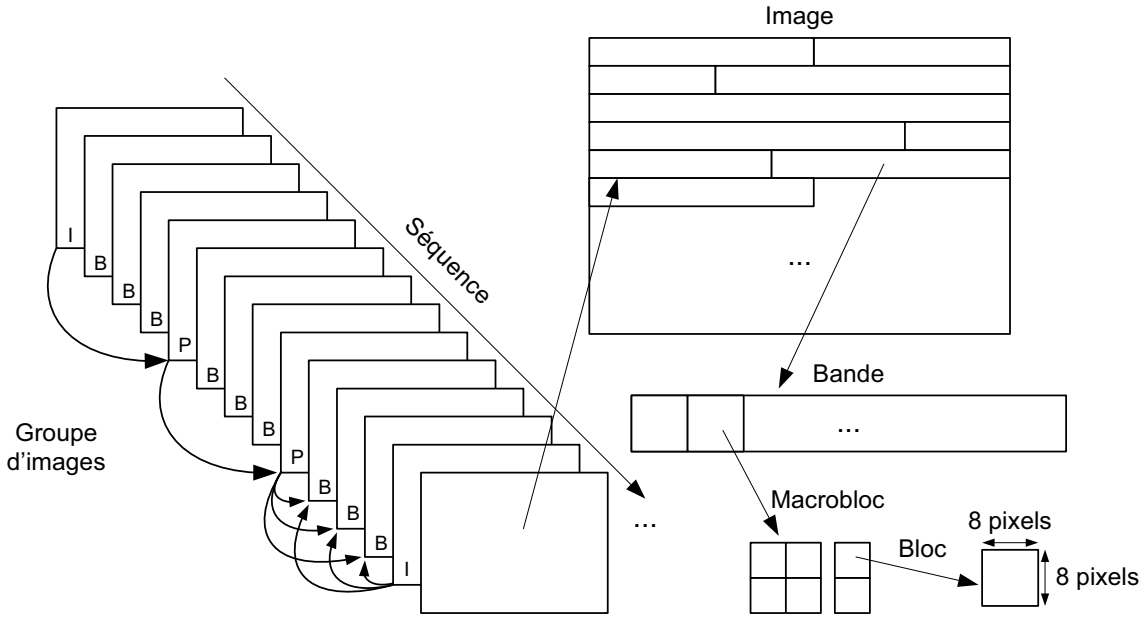


Figure 1.1 – Structure hiérarchique MPEG.

1.2.2 Les distorsions

1.2.2.1 Les distorsions liées à la compression

Comme indiqué dans la section précédente, les techniques de compression utilisées dans les diverses normes présentent des caractéristiques similaires. La plupart d'entre eux utilisent une transformée DCT par bloc, avec compensation de mouvement et quantification des coefficients. Dans ce procédé de codage, les distorsions introduites par la compression sont liées à une seule opération : la quantification. Bien que d'autres facteurs affectent les distorsions produites, comme la compensation de mouvement ou la taille de la mémoire tampon de décodage, ils n'introduisent pas intrinsèquement de distorsions, mais ils affectent indirectement le codage au travers du facteur d'échelle de la quantification. Différents types de distorsions peuvent être identifiés dans une vidéo numérique décompressée :

- L'effet de bloc (*blocking effect*), comme son nom l'indique, est l'apparition de motifs en forme de blocs dans la séquence décompressée. Cet effet est dû à la différence de quantification d'un bloc à l'autre (en général 8x8 pixels) dans les schémas de compression utilisant une DCT par bloc, qui introduit des discontinuités à la frontière des blocs adjacents. L'effet de bloc est souvent la distorsion la plus frappante dans une vidéo décompressée à cause de la régularité et de l'étendu des motifs insérés.
- Le flou (*blurring*) se manifeste comme une perte des détails spatiaux, et comme une réduction de la finesse des contours. Il est dû à la suppression des coefficients hautes fréquences, via une quantification grossière.
- Le *color bleeding* est une bavure des couleurs entre deux zones de chrominance fortement différentes. Il est le résultat de la suppression des coefficients hautes fréquences dans les composantes chromatiques. A cause du sous-échantillonnage de la chrominance, le « color bleeding » s'étend sur la totalité du macrobloc.

- L'« effet des fonctions de base DCT », est important lorsqu'un seul coefficient DCT est dominant sur un bloc. A des niveaux de quantifications élevés, le résultat est l'accentuation de l'image de la fonction de base DCT dominante, et la réduction de toutes les autres fonctions de base.
- Un effet « escalier » peut apparaître sur les contours obliques. Cet effet est dû au fait que les fonctions de base de la DCT sont plus adaptées à la représentation des contours horizontaux et verticaux. Les contours, dont l'orientation n'est ni horizontale ni verticale, nécessitent davantage de coefficients hautes fréquences pour obtenir une représentation précise. La quantification trop importante de ces coefficients introduit des irrégularités sur ces contours obliques (*jagged*).
- L'effet d'ondulation (*ringing*) est fondamentalement associé au phénomène de Gibbs (oscillation de reconstruction d'un signal discontinu avec une somme de signaux continus). Il est donc plus marqué sur les contours à fort contraste que sur les zones uniformes. Cet effet est un résultat direct de la quantification provoquant des irrégularités dans la reconstruction. L'effet de *ringing* se produit à la fois sur la luminance et la chrominance.
- Les contours fantômes sont une conséquence du transfert par compensation de mouvement de la discontinuité de la frontière d'un bloc, induite par un effet de bloc, depuis une image de référence vers une image prédite.
- Le mouvement saccadé (*jagged motion*) peut être dû à une mauvaise estimation de mouvement. Les estimateurs de mouvement fonctionnant par bloc sont plus efficaces lorsque les mouvements de tous les pixels d'un macrobloc sont identiques. Lorsque l'erreur résiduelle d'estimation de mouvement est importante, elle peut être quantifiée grossièrement.
- La quantification de mouvement est souvent réalisée seulement sur la luminance et le même vecteur est utilisé pour les composantes de chrominance. Il peut en résulter une « *chominance mismatch* » sur un macrobloc.
- L'effet « mosquito », est une distorsion temporelle observée principalement sur les zones faiblement texturées comme des fluctuations de la luminance (ou de la chrominance) autour des contours à fort contraste ou des objets en mouvement. Il est la conséquence de la variation du choix des paramètres de codage d'une même zone d'une scène dans les images successives d'une séquence.
- Le papillotement (*flickering*) apparaît principalement dans les zones texturées. La texture des blocs de ces zones est compressée avec des pas de quantifications variant au cours du temps, ce qui a pour effet de créer des papillotements dans ces zones.

Alors que plusieurs de ces distorsions sont liées à une technique de codage fonctionnant par bloc, quelques-unes d'entre elles sont observées avec d'autres méthodes de compression aussi. Dans la compression utilisant la transformée en ondelettes, par exemple, la transformation est réalisée sur l'image entière, il n'y a pas donc de distorsions liées à l'utilisation des blocs. En contrepartie, le flou et le *ringing* sont les distorsions les plus frappantes.

1.2.2.2 Les autres dégradations

A part les distorsions liées à la compression, le rendu des séquences vidéo numériques peut être perturbé d'une part par des erreurs de transmission, d'autre part modifié par l'existence de prétraitements ou post-traitements.

Les erreurs de transmission du flux binaire dans un canal de transmission bruité sont une source importante et souvent négligée de distorsions. Une variété de protocoles est utilisée pour transporter l'information audio-visuelle, synchroniser le média et ajouter les informations temporelles. La plupart des applications nécessitent que les vidéos soient envoyées en flux continu, c'est-à-dire qu'il soit possible de décoder et d'afficher les vidéos en temps réel. Deux différents types d'erreurs peuvent se produire pendant le transport sur des canaux bruités. Les paquets peuvent être perdus, ou ils peuvent être retardés suffisamment longtemps pour qu'ils ne soient pas reçus à temps pour le décodage. Le retard est lié aux techniques de routage et d'ordonnancement des routeurs et des « *switchs* » du réseau. An niveau applicatif, ces deux types d'erreurs ont le même effet : un morceau du flux n'est pas disponible, des paquets sont manquants au moment où ils devraient être décodés. Les effets visuels de ces pertes varient beaucoup d'un décodeur à l'autre, en fonction de leur capacité à gérer des flux corrompus. Des décodeurs ne peuvent fonctionner en présence de certaines erreurs, tandis que d'autres utilisent des techniques adaptatives de dissimulation d'erreurs comme l'interpolation spatiale ou temporelle afin de réduire leurs effets.

1.2.2.3 Discussion

Nous venons de décrire les principaux types de distorsions susceptibles de dégrader la qualité d'une vidéo numérique. Une partie de ces distorsions sont d'ailleurs communes avec les distorsions susceptibles de dégrader la qualité d'une image numérique. L'autre partie de ces distorsions est donc propre à la vidéo numérique, et elles ne doivent leur existence qu'à la dimension temporelle intrinsèque à la vidéo. Cette catégorie de distorsions que l'on peut qualifier de distorsions temporelles ne doit en fait leur existence qu'à l'instabilité locale des distorsions spatiales entre les images successives de la vidéo. Une distorsion temporelle considérée localement peut donc être décrite comme une variation temporelle d'une distorsion spatiale. Cette constatation offre un angle d'étude intéressant pour l'évaluation locale des distorsions dans une vidéo.

1.3 Modélisation du système visuel humain

La modélisation du système visuel humain découle principalement de sa structure biologique et fonctionnelle, ainsi que d'expérimentations psychophysiques. La biologie du système visuel humain est détaillée en annexe A et le lecteur pourra s'y reporter si besoin est. Cette section est focalisée sur la modélisation des propriétés du système visuel humain qui nous semblent les plus importantes pour élaborer des méthodes d'évaluation locale des distorsions, ainsi que des méthodes d'évaluation de la qualité visuelle.

La perception d'une zone de l'image engendre trois types de sensations. Les sensations de teinte et de saturation sont liées à la perception de la chromacité de la zone observée, alors que la sensation de luminosité reflétera la luminance perçue.

1.3.1 La perception de la luminance

Le système visuel humain est naturellement confronté à une dynamique importante de l'intensité lumineuse. Face à cette dynamique, des mécanismes d'adaptation se sont mis en place lui permettant de maintenir sa sensibilité aussi bien dans des conditions d'illumination importante, conditions photopiques, que dans des conditions d'illumination faibles, conditions stocopiques. Les trois principaux mécanismes d'adaptation à la luminance sont :

- La variation de l'ouverture de la pupille (entre 1,5 et 8 mm) qui laisse passer plus ou moins de lumière en fonction des conditions d'illumination. Le temps de réaction est de l'ordre de la seconde.
- La modification des concentrations photochimiques des photorécepteurs qui permet de modifier leur sensibilité. Plus l'intensité lumineuse est importante, et plus la concentration diminue entraînant une réduction de la sensibilité, et vice versa. Le temps d'adaptation est plutôt lent puisqu'il faut compter une heure pour une complète adaptation à l'obscurité.
- La modification de la réponse neuronale de toutes les couches cellulaires de la rétine. Cette adaptation est moins efficace que la précédente, mais beaucoup plus rapide.

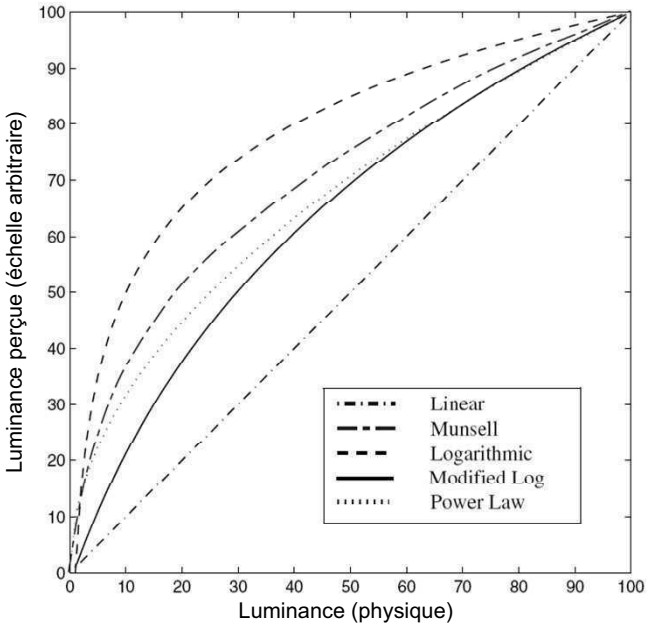


Figure 1.2 – Relation liant la luminance perçue et la luminance réelle et comparaison avec des modèles.

Outre la capacité d'adaptation, la relation entre la luminance perçue (brillance) et la luminance réelle (luminance physique) n'est pas linéaire. De nombreuses expérimentations ont été menées dans le but de déterminer la nature de cette relation. Ces expérimentations consistant le plus souvent à faire classer des nuances de gris par des observateurs. Les expérimentations les plus nombreuses sont celles associées au système de Munsell [Newhall 43]. La figure 1.2 illustre la relation obtenue par ces expérimentations (notée Munsell) ainsi que différentes modélisations mathématiques à titre de comparaison. Sur la figure 1.2, on observe que la fonction

logarithmique surestime la luminance perçue (Munsell), alors que la fonction linéaire la sous-estime. La fonction logarithme modifiée, quant à elle, a tendance à sous-estimer la luminance perçue pour les basses valeurs de luminance.

C'est une fonction puissance qui semble être la relation la plus adaptée pour modéliser cette non linéarité dans la perception de la luminance. La dynamique des valeurs des zones sombres est augmentée, alors que celle des zones claires est réduite. La luminance perçue (brillance), notée L_p, est déduite de la luminance (physique) L_o :

$$L_p = a \times L_o^e - L_0 \tag{1.1}$$

H. Bodmann et al. [Bodmann 80] définissent l'exposant e égal à 0.31 ± 0.03. Les deux autres coefficients a et L_0, permettent de s'adapter à différentes échelles. Dans le cas de la figure 1.2, $a = 1$ et $L_0 = 0$.

1.3.2 La perception des couleurs

Les modèles les plus plausibles de la perception des couleurs sont basés sur la théorie des signaux antagonistes. Selon ces modèles, l'information lumineuse reçue sur la rétine est séparée en trois composantes perceptives : une composante achromatique A, et 2 composantes purement chromatiques Cr1 et Cr2. Ces composantes résultent de la combinaison des signaux issus des trois types de cônes L, M, S. En général, cette combinaison est considérée comme linéaire :

$$\begin{pmatrix} A \\ Cr1 \\ Cr2 \end{pmatrix} = [T] \times \begin{pmatrix} L \\ M \\ S \end{pmatrix} \tag{1.2}$$

R. De Valois [De Valois 92] et O. Faugeras [Faugeras 76] ont tous les deux proposé un modèles physiologique de construction des trois composantes perceptives. Celui de R. De Valois se base sur les signaux arrivant sur les zones excitatrices et inhibitrices des champs récepteurs :

$$[T]_{DeValois} = \begin{pmatrix} 0.375 & 0.6875 & 0.00625 \\ 0.5625 & -0.7187 & 0.1562 \\ -0.8125 & 0.5938 & 0.2187 \end{pmatrix} \tag{1.3}$$

Le modèle de O. Faugeras utilise les différentes courbes d'absorption des différents types de cônes :

$$[T]_{Faugeras} = \begin{pmatrix} 13.63 & 8.33 & 0.42 \\ 64 & -64 & 0 \\ -5 & -5 & -10 \end{pmatrix} \tag{1.4}$$

D'autres modèles sont basés sur des expérimentations psychophysiques, comme par exemple les travaux de M. Webster [Webster 90], de P. Flanagan [Flanagan 90], et de J. Krauskopf [Krauskopf 82]. La matrice $[T]$ définie par J. Krauskopf nous intéresse particulièrement car nous utiliserons une composante de cet espace dans notre étude.

$$[T]_{Krauskopf} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ -0.5 & -0.5 & 1 \end{pmatrix} \tag{1.5}$$

1.3.3 La sensibilité au contraste

La réponse du système visuel humain dépend plus des variations locales de luminance (ΔL) par rapport à la luminance avoisinante (L), que des valeurs absolues de luminance. Cette propriété est connue sous de loi de Weber-Fechner. Cette variation relative de luminance est mesurée au travers de ce qui est appelé *contraste*. Mathématiquement, le contraste de Weber peut s'exprimer par la relation :

$$\frac{\Delta L}{L} = C^{te} \tag{1.6}$$

Cette loi indique que si sur un fond uniforme de luminance L, dite d'adaptation, on superpose un stimulus type médaillon de luminance $\Delta L + L$, le rapport ($\frac{\Delta L}{L}$) est pratiquement constant dans un large domaine de luminance. Une légère incohérence en basses luminances a été corrigée par Moon et Spencer [Moon 44], ce qui donne l'expression suivante :

$$\frac{\Delta L}{L} = \left(\frac{C_\infty}{L}\right)\left(0,456 + \sqrt{L}\right)^2 \tag{1.7}$$

Au seuil de détection, et lorsque la luminance augmente, le rapport $\frac{\Delta L}{L}$ tend vers la constante de Weber C_∞. Cette constante dépend de la géométrie et de la taille du stimulus.

On appelle contraste seuil, ou seuil de détection, la valeur minimale de contraste nécessaire pour qu'un observateur détecte un changement d'intensité lumineuse. L'étude de la sensibilité du système visuel humain est réalisée à l'aide d'expérimentations psychophysiques mesurant la variation de ce contraste seuil face à divers facteurs. Les résultats de ces expérimentations sont exprimés en général en terme de seuil de sensibilité au contraste, appelé aussi seuil différentiel de visibilité. La sensibilité est définie comme l'inverse du contraste seuil. Cette sensibilité sera donc d'autant plus élevée que le contraste seuil sera faible, et inversement.

Le seuil de sensibilité est dépendant des nombreuses caractéristiques des stimuli. Outre la luminance, parmi les caractéristiques qui ont fait l'objet d'études, on peut citer la fréquence spatiale, l'orientation, la fréquence temporelle, la couleur, la vélocité, l'excentricité, etc. Ces dépendances sont ensuite modélisées par des fonctions de sensibilité au contraste (FSC, plus connue sous l'abréviation anglaise CSF pour *Contrast Sensitivity Function*). La plupart du temps, les modèles sont élaborés à partir de résultats expérimentaux sur la détection de signaux sinusoïdaux qui utilisent la définition du contraste de Michelson :

$$C = \frac{L_{max} - L_{min}}{L_{max} + L_{min}} \tag{1.8}$$

où L_{min} et L_{max} correspondent aux valeurs respectivement de luminance minimale et maximale.

1.3.3.1 Sensibilité aux fréquences spatiales

La figure 1.3a reprend l'illustration de Campbell-Robson qui présente la forme de la CSF de façon assez intuitive. La luminance des pixels est modulée de façon sinusoïdale suivant l'axe horizontal. La fréquence spatiale augmente exponentiellement de la gauche vers la droite, tandis que le contraste diminue exponentiellement de 100% à 0.5% du bas vers le haut. Le minimum et le maximum de luminance restent constant sur chaque ligne horizontale. Si le seuil différentiel de visibilité ne dépendait que du contraste, les bandes verticales alternativement claires et foncées devraient toutes apparaître avec la même hauteur. Cependant, les bandes apparaissent plus hautes au milieu de l'image que sur les bords. Cette enveloppe de visibilité représente la CSF pour un signal sinusoïdal. La forme et la position du maximum de cette enveloppe dépendent de la distance d'observation.

Dans la littérature, E. Peli et al. [Peli 93] détaillent un ensemble assez complet de CSFs pour des signaux achromatiques, pour différentes configurations de stimuli. Dans [Barten 99] puis dans [Barten 04], P. Barten propose une formulation plus complète des CSFs, en terme de dépendance à de nombreux paramètres. Un exemple de CSF classique isotrope, proposée par J. Mannos et D. Sakrison [Mannos 74] est représenté à la figure 1.3b. Sa formulation est la suivante :

$$CSF(f) = 2.6 \times (0.0192 + 0.114f)e^{(-(0.114f)^{1.1})} \tag{1.9}$$

où, f est exprimé en cycle par degré (cpd).

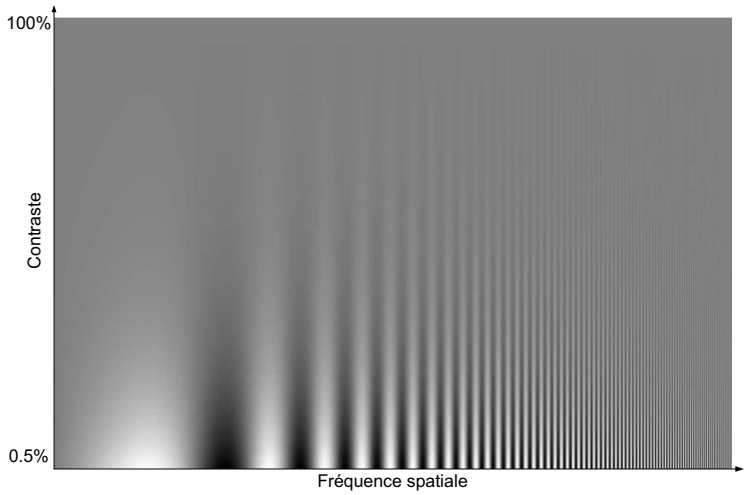
Cette courbe illustre le comportement passe-bande du SVH vis-à-vis des fréquences spatiales de la composante luminance, où les fréquences spatiales sont exprimées en cycle par degré (cpd). La sensibilité du SVH est maximale pour les fréquences spatiales intermédiaires (entre 4 et 13 cpd), et diminue pour les basses et aux hautes fréquences spatiales. Au-delà, de 50 cpd, l'oeil ne perçoit plus rien. Autrement dit, une modification de contraste d'un signal, comme une dégradation, sera plus facilement visible par un observateur dans une zone de fréquences spatiales intermédiaires (autour de 6 à 8 cpd par exemple) que dans une zone de très hautes fréquences spatiales (supérieures 20 cpd par exemple). Une autre fonction de sensibilité au contraste de luminance est présentée ici car elle sera utilisée dans la suite de nos travaux. Il s'agit de la fonction anisotrope de S. Daly [Daly 93]. Cette fonction exprime la sensibilité en fonction de la fréquence radiale ω en cy/deg, l'orientation θ en degrés, le niveau d'adaptation en luminance l en cd/m^2, la surface de l'image s en degrés^2, la distance d'observation d en mètres et l'excentricité e en degrés. Le modèle de Daly est le suivant :

$$S_A(\omega, \theta, l, s, d, e) = P \times min\big[S(\frac{\omega}{bw_a, bw_e, bw_\theta}, l, s), S(\omega, l, s)\big], \tag{1.10}$$

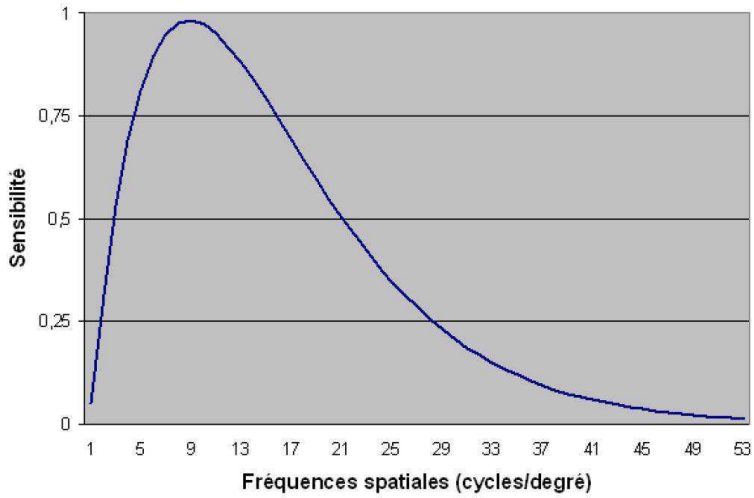
dans laquelle P désigne la sensibilité maximale. Les paramètres bw_a,bw_e,bw_θ assurent la prise en compte des changements de la largeur de bande en fonction respectivement de la distance, de l'excentricité et de l'orientation. Leurs expressions sont données par :

$$bw_a = 0.856 \times d^{0.14}, \tag{1.11}$$

$$bw_e = \frac{1}{1 + 0.24 \times e}, \tag{1.12}$$



(a)



(b)

Figure 1.3 – (a) Illustration de la sensibilité au contraste de Campbell-Robson [Campbell 68]. La CSF apparaît comme étant l'enveloppe de visibilité du signal modulé. (b) Fonction normalisée de sensibilité au contraste proposée par J. Mannos et D. Sakrison [Mannos 74].

$$bw_\theta = 0.15 \times cos(4\theta) + 0.85. \tag{1.13}$$

$S(\omega, l, s)$ est défini par :

$$S(\omega, l, s) = ((3.23 \times (\omega^2 s)^{-0.3})^5 + 1)^{-1/5} \times A_l \times 0.9 \times \omega \times e^{-(B_l \times 0.9 \times \omega)} \times \sqrt{1 + 0.06 \times e^{B_l \times 0.9 \times \omega}}, \tag{1.14}$$

avec

$$A_l = 0.801 \times (1 + 0.7/l)^{-0.2}, \tag{1.15}$$

$$B_l = 0.3 \times (1 + 1000/l)^{0.15}. \tag{1.16}$$

En général, les valeurs des paramètres P, l et e sont respectivement 250, 100 et 0. La valeur nulle pour ce dernier paramètre vient du fait que l'on considère que toute l'image est vue et inspectée dans la zone fovéale de la rétine (excentricité nulle).

En ce qui concerne la couleur, la sélectivité du SVH est plus importante, et la CSF est plus proche d'un filtre passe-bas en fréquences spatiales. Les travaux de P. Le Callet [Le Callet 01], ont permis une modélisation des CSF associées aux composantes chromatiques Cr1 et Cr2, dont les fréquences de coupure sont respectivement 5.5 cpd et 4.1 cpd. Pour la composante Cr1, la CSF est modélisée par la fonction suivante :

$$S_{Cr1}(w, \theta) = \frac{33}{1 + (\frac{w}{5.52})^{1.72}}(1 + 0.27sin(2\theta)), \tag{1.17}$$

Pour la composante Cr2, la CSF est modélisée par la fonction suivante :

$$S_{Cr2}(w, \theta) = \frac{5}{1 + (\frac{w}{4.12})^{1.64}}(1 + 0.24sin(2\theta)), \tag{1.18}$$

Les courbes de sensibilité des CSF sont illustrées figure 1.4 pour une orientation nulle ($\theta = 0$).

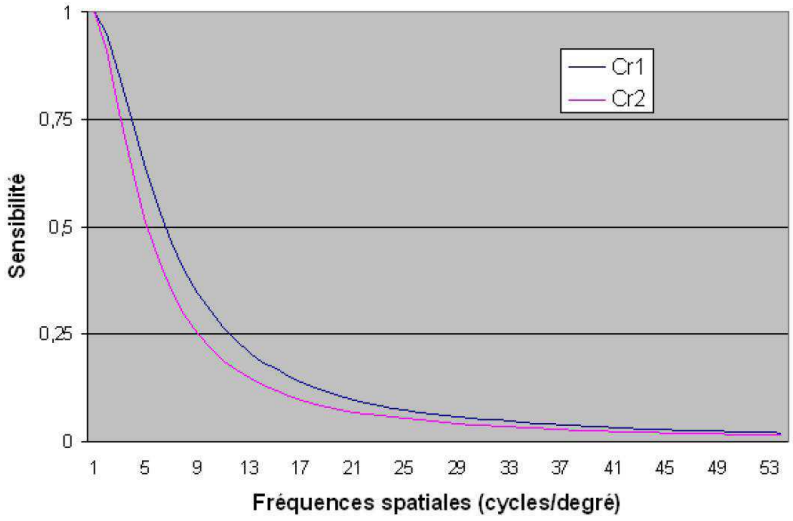


Figure 1.4 – Fonction de sensibilité au contraste [Le Callet 01].

1.3.3.2 Sensibilité aux fréquences temporelles

De même que la sensibilité du SVH varie en fonction des fréquences spatiales, la sensibilité du SVH varie en fonction des fréquences temporelles. Autrement dit, si on considère une zone d'une image dont le contraste varie temporellement de façon sinusoïdale et avec une amplitude constante, à certaines fréquences temporelles les variations seront visibles alors qu'à d'autres les variations ne seront pas perçues par le système visuel. H. De Lange fut l'un des premiers à mener des expérimentations de manière à caractériser des fonctions temporelles de sensibilité au contraste (TCSF). Il montre [De Lange 58] que la sensibilité est maximale à environ 8Hz. Au-dessus de cette fréquence la sensibilité décroît très rapidement, et atteint une valeur de 1 pour une fréquence comprise entre 50Hz et 70Hz. Cette fréquence est appelée CFF (*Critical Flicker Frequency*), et représente la transition entre un scintillement de lumière et une lumière continue. La sensibilité décroît aussi pour les basses fréquences temporelles, mais de façon plus modérée.

1.3.3.3 Interactions spatio-temporelles

Il existe des débats sur la séparabilité espace-temps des CSF spatio-temporelles. Une telle propriété serait intéressante en terme de modélisation, puisque cela permettrait d'exprimer la sensibilité spatio-temporelle tout simplement comme le produit d'une composante spatiale, et d'une composante temporelle. Dans ce cas, la CSF serait définie par la relation :

$$S(f_s, f_t) = S_S(f_s, \theta) \cdot S_T(f_t), \tag{1.19}$$

où $S(f_s, f_t)$ décrit la sensibilité au contraste spatio-temporel, f_s, f_t et θ représentent respectivement la fréquence spatiale, la fréquence temporelle et l'orientation, $S_S(f_s, \theta)$ et $S_T(f_t)$ représentent respectivement la fonction de sensibilité spatiale et la fonction de sensibilité temporelle.

Des études [Robson 66, Koenderink 79] ont montré assez tôt que la CSF spatio-temporelle n'était pas séparable en espace-temps pour les basses fréquences. Il a été montré que :

- la taille de la cible influence la sensibilité aux basses fréquences temporelles : une cible de taille importante tend à réduire la sensibilité ;
- le contraste des contours influence la sensibilité aux basses fréquences temporelles : une cible présentant des contours contrastés augmente la sensibilité.

Kelly [Kelly 79a] a mesuré la sensibilité au contraste dans des conditions d'observation stabilisées (c'est-à-dire en stabilisant le stimulus sur la rétine des observateurs par compensation des mouvements oculaires), et a adapté [Kelly 79b] une fonction analytique sur ses mesures. Il a obtenu une très bonne approximation de la CSF spatio-temporelle pour des stimuli de type *counterphase flicker*. Burbeck et Kelly [Burbeck 80] ont montré que cette CSF achromatique pouvait être approximée par la combinaison linéaire de deux composantes séparables en espace-temps appelées CSF excitatrice et inhibitrice. Ils firent de même avec les CSF chromatiques [Kelly 83].

Dans sa thèse [van den Branden Lambrecht 96b] C. J. van den Branden Lambrecht a proposé aussi un modèle excitateur-inhibiteur de la CSF spatio-temporelle, reprenant la formulation de Burbeck et Kelly [Burbeck 80], et dont une illustration est donnée figure 1.5. La sensibilité spatio-temporelle est exprimée comme la différence

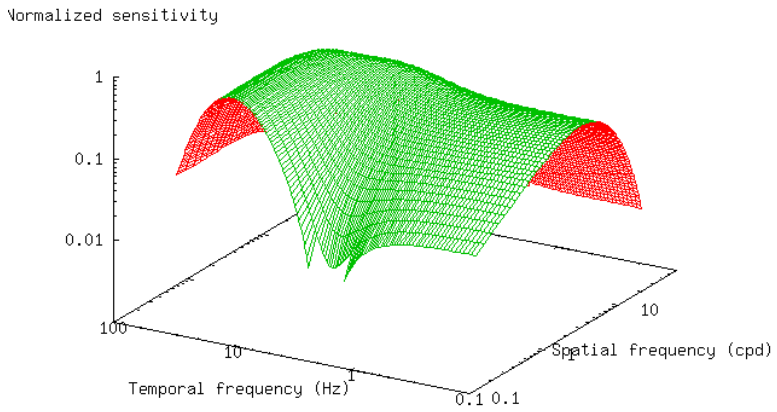


Figure 1.5 – Fonction de sensibilité au contraste [van den Branden Lambrecht 96b].

d'un mécanisme excitateur $E(f_s, f_t)$ et d'un mécanisme inhibiteur $I(f_s, f_t)$:

$$S(f_s, f_t) = E(f_s, f_t) - I(f_s, f_t), \tag{1.20}$$

avec :

$$E(f_s, f_t) = K_1 \cdot S_{S,f_{t_1}}(f_s) \cdot S_{T,f_{s_1}}(f_t), \tag{1.21}$$

$$I(f_s, f_t) = K_2 \cdot (E(f_s, f_{t_2}) - S_{S,f_{t_2}}(f_s)) \cdot (E(f_{s_2}, f_t) - S_{T,f_{s_2}}(f_t)), \tag{1.22}$$

où les fréquences spatiales f_{s_1} et f_{s_2}, et les fréquences temporelles f_{t_1} et f_{t_2} sont choisies pour mesurer les courbes de sensibilité temporelles $S_{T,f_{s_1}}$ et $S_{T,f_{s_2}}$, et les courbes de sensibilité spatiales $S_{S,f_{t_1}}$ et $S_{S,f_{t_2}}$ qui servent à approximer la CSF spatio-temporelle. K_1 et K_2 sont des constantes.

1.3.4 L'organisation multi-canal

Comme évoqué dans l'annexe A, la sensibilité des différentes cellules du système visuel humain à certains types d'informations, comme la couleur, l'orientation ou la fréquence, suggère qu'il existe des regroupements de l'information préalablement à son traitement. Les résultats de plusieurs expérimentations psychophysiques confortent cette idée et présentent le système visuel humain comme un système multi-canal [Braddick 78].

1.3.4.1 Décomposition spatiale de l'information

La décomposition spatiale de l'information du SVH en différents canaux s'effectue selon une sélectivité radiale (de 1 à 2 octaves), et une sélectivité angulaire (de 20 deg à 60 deg). Dans la littérature, on trouve plusieurs décompositions. On peut citer la transformée Cortex de Watson [Watson 87] qui décompose le signal en 5 couronnes de fréquences radiales ayant chacune une largeur de bande d'une octave et présentant une

sélectivité angulaire constante de 45 deg (sauf pour la couronne des plus basses fréquences spatiales). Cette décomposition est réalisée au moyen de filtres dit Cortex. Les paramètres de ces filtres Cortex ont été affinés à partir de nombreuses expérimentations psychophysiques [Sénane 96, Bedat 98, Le Callet 01] à la fois pour la luminance et la couleur. On parle alors de décomposition en canaux perceptuels (DCP). Une partie de nos travaux reposant sur cette décomposition, elle sera détaillée section 2.5.1. On peut citer aussi les filtres de Gabor qui repose sur la ressemblance entre la forme des champs récepteurs corticaux et la transformations bidimensionnelles de Gabor. Cependant, dans la pratique il faudrait considérer un grand nombre de filtres pour couvrir tout le pavage fréquentiel du SVH. On peut aussi citer des approches multi-résolutions, moins fidèles au SVH mais plus directes à implanter, comme par exemple des transformations pyramidales [Burt 83a], ou comme des transformées en ondelettes 2D classiques. L'avantage de ces transformations est la bonne localisation spatiale, mais les inconvénients sont les sélectivités fréquentielles et angulaires. Le cas de la transformée en ondelettes 2D fait l'objet d'une partie de nos travaux et sera revu en détail dans la section 2.5.2.

1.3.4.2 Décomposition temporelle de l'information

Les mécanismes temporels ont aussi fait l'objet des plusieurs études, cependant la littérature révèle que le consensus n'est pas aussi clair sur la décomposition temporelle de l'information que sur la décomposition spatiale. Cependant, la tendance qui en ressort est l'existence de deux mécanismes, l'un passe-bas et l'autre passe-bande [Watson 86, Fredericksen 98]. Il est fait mention respectivement des canaux *sustained* et *transient*. L'existence d'un troisième canal a été évoquée [Mandler 84, Hess 92], et a été remise en question dans d'autres études [Hammett 92, Fredericksen 98]. Dans leurs travaux Fredericksen and Hess [Fredericksen 98] ont pu mettre en adéquation un grand nombre de données psychophysiques avec seulement un canal *sustained* et un canal *transient*. La réponse fréquentielle de ces deux canaux est illustrée figure 1.6.

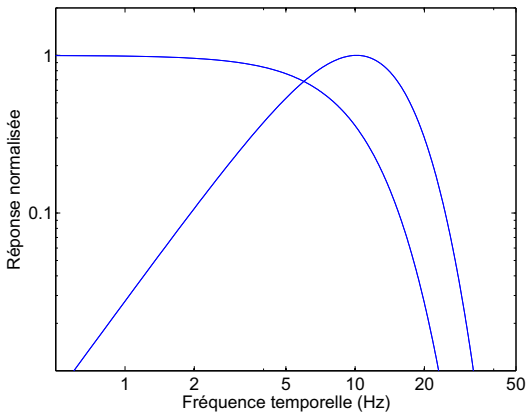


Figure 1.6 – Réponse fréquentielle des canaux visuels sustained (passe-bas) et transient (passe-bande) de la décomposition temporelle de l'information [Fredericksen 98].

L'identification et la caractérisation d'une décomposition spatiale de l'information d'une part, et d'une décomposition temporelle de l'information d'autre part, ne permettent pas de répondre directement à la question

d'une décomposition spatio-temporelle de l'information. Les avis sur la question sont partagés, et deux écoles se distinguent. La première école affirme que la décomposition spatio-temporelle de l'information est séparable en une décomposition temporelle suivie d'une décomposition spatiale. Une décomposition spatio-temporelle peut donc être réalisée, par un enchaînement des deux décompositions, en adaptant simplement les gains des filtres en fonction de la position spatio-temporelle de la sous-bande. Cependant, cet avis ne fait pas consensus, et la seconde école pense que l'interdépendances des décompositions spatiale et temporelle ne permet pas de décrire la décomposition spatio-temporelle de l'information en une décomposition temporelle d'une part, et une décomposition spatiale d'autre part. Il nous semble donc critiquable d'utiliser une telle décomposition spatio-temporelle de l'information.

1.3.5 Les effets de masquage

1.3.5.1 Le masquage spatial

En accord avec la modélisation perceptuelle précédente, les signaux ayant des caractéristiques voisines sont traités par les mêmes canaux visuels et suivent donc le même cheminement de l'oeil jusqu'au cortex. Il existe des interactions aux effets non linéaires entre de tels signaux voisins. Le masquage, ou effet de masquage, est un de ces effets. Il traduit la variation du seuil différentiel de visibilité d'un stimulus due à la présence d'un autre signal dans son voisinage, qualifié de signal masquant, ayant un niveau plus élevé. L'effet de masquage est d'autant plus important que les deux signaux ont des caractéristiques voisines. Par abus de langage, on parle d'effet de masquage aussi bien dans le cas de l'augmentation du seuil différentiel de visibilité, que dans le cas de la diminution de la valeur du seuil. Dans le premier cas il s'agit réellement de masquage (*masking effect*), alors que dans le second cas il s'agit de ce qu'on appelle la facilitation (*pedestal effect*) ou un signal va augmenter la visibilité d'un autre. S'intéresser à l'effet de masquage dans le SVH revient à modéliser la variation du seuil de détection en fonction des caractéristiques du signal masquant. L'effet de masquage a fait l'objet de nombreuses études en raison de son importance dans les différents axes du traitement d'image et vidéo (toute variation du signal d'image en deçà du seuil différentiel de visibilité est non percue). Différents modèles ont été proposés dans la littérature [Foley 94, Legge 80, Heeger 92, Teo 94, Le Callet 01, Daly 93]. Relativement à l'organisation multi-canal du SVH, trois grands types de masquage ont été identifiés :

- le plus important est le masquage intra-canal se traduisant par une interaction entre stimuli et signal masquant de caractéristiques voisines (fréquence, orientation, composante) ;
- le masquage entre stimuli et signal masquant de caractéristiques différentes c'est-à-dire n'appartenant pas au même canal, mais appartenant à la même composante. Cet effet est appelé masquage inter-canal ;
- le masquage entre différentes composantes qui est appelé masquage inter-composante ;

L'effet de masquage est souvent représenté par une courbe caractéristique comme celle présentée figure 1.7. L'axe horizontal représentant le contraste du signal masquant C_M, et l'axe vertical représentant le contraste du stimulus (la cible) au seuil différentiel de visibilité C_T, appelé aussi contraste seuil. Le seuil C_T en l'absence de signal masquant est noté C_{T_0}, dans ce cas cela veut dire en fait que le signal masquant correspond à un

signal constant (luminance uniforme). Pour des valeurs de contraste du signal masquant supérieures à C_{M_0}, le contraste seuil augmente en même temps que le contraste du signal masquant. On retrouve sur cette courbe l'effet de masquage (A), l'effet de facilitation (B).

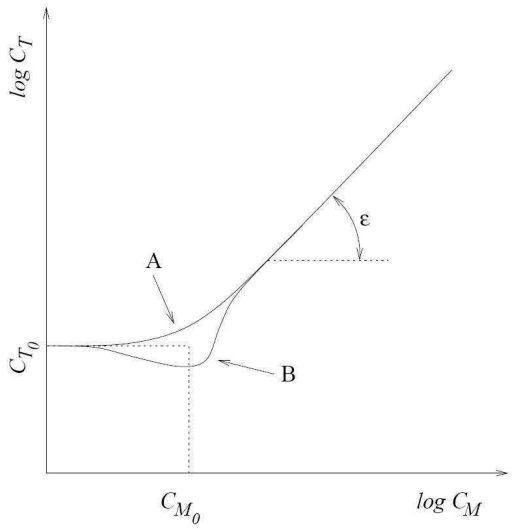
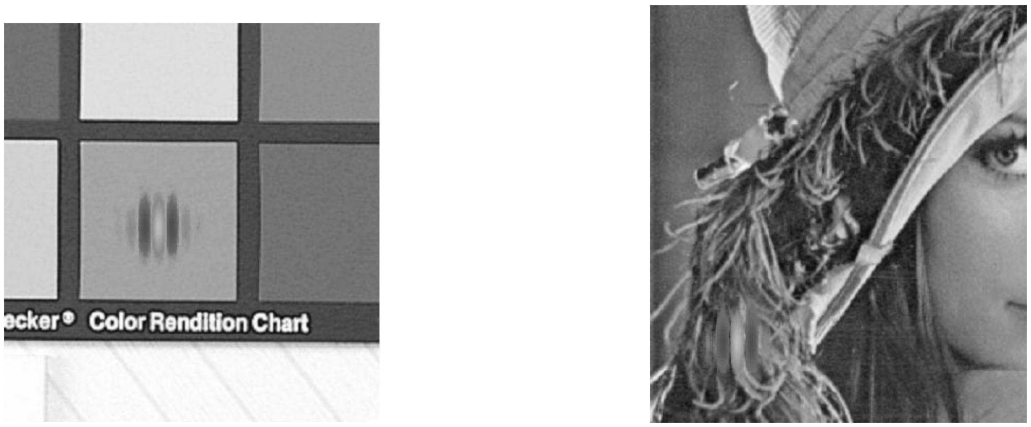


Figure 1.7 – Illustration de courbes de masquage classiques. Les valeurs de contraste seuil de la cible C_T sont donnés en fonction des valeurs de contraste du signal masquant C_M. Pour des stimuli dont les caractéristiques ne sont pas trop proches l'effet principal est le masquage (A). Dans le cas où les stimuli sont très proches un effet de facilitation (B) peut apparaître pour des valeurs de contraste C_M pas très grandes.

Tous les effets de masquage qui ont été abordés jusqu'ici sont souvent qualifiés d'effets de masquage dus au contraste, cependant une autre forme de masquage existe. En effet, il y a des situations où le masquage de contraste ne permet pas d'expliquer complètement l'élévation du seuil de visibilité. La figure 1.8 illustre un cas simpliste d'une telle situation. Dans les deux images la même distorsion a été introduite. Autant cette distorsion est facilement perçue lorsqu'elle est positionnée dans une zone homogène de l'image de gauche, autant il est plus difficile de la distinguer dans l'image de droite. En effet, les plumes du chapeau créent une zone dont la complexité locale est importante, c'est une zone dite très « active » où le SVH a besoin de plus de temps pour détecter la distorsion. Si les caractéristiques du signal de distorsion sont proches de la texture qui l'entoure, il est parfois impossible de trouver la distorsion sans connaître l'image originale.

L'effet de masquage impliqué dans ce type de situation est qualifié de masquage entropique [Watson 97b], de masquage de texture [Gaubatz 05], ou encore de masquage d'activité [Nadenau 00]. Le masquage entropique est lié au masquage de contraste, mais la différence entre les deux réside dans le support spatial pris en compte. Les deux expliquent la modification de la sensibilité due à la présence de fort contraste. Cependant, le masquage de contraste est typiquement appliqué très localement quasiment point à point. Cela signifie que c'est le contraste en un point qui détermine la capacité de masquage en ce point. Même s'il est vrai que dans une modélisation multi-canal, un même point peut générer du masquage dans plusieurs sous-bandes de fréquences spatiales et d'orientations différentes. Le masquage entropique, quant à lui, considère explicitement une semi-localité autour

d'un point, c'est-à-dire son proche voisinage, pour déterminer la capacité de masquage en ce point.



(a) (b)

Figure 1.8 – Illustration du masquage entropique sur un cas simpliste [Nadenau 00].

1.3.5.2 Le masquage temporel

De même que dans le cas du masquage spatial, le masquage temporel traduit une modification du seuil de visibilité d'un signal due à la présence d'un autre signal. Cette modification du seuil de visibilité est due à l'interaction de stimuli temporellement adjacents. Ces effets de masquage sont moins bien connus que ceux rencontrés dans le domaine spatial. Dans sa forme la plus générale, la question du masquage temporel s'intéresse à la façon dont interagissent deux stimuli proches temporellement. La réponse est complexe et dépend de nombreux facteurs comme la structure spatiale du signal masquant, la similarité entre les caractéristiques spatiales entre la cible et le signal masquant, l'intervalle de temps séparant la cible et le signal masquant, la différence de luminance entre la cible et le signal masquant, etc. Dans les études sur le masquage temporel, on distingue une situation particulière d'étude qui s'intéresse aux effets de masquage dus à de fortes discontinuités temporelles, comme les changements de plan ou des transitions rapides (sombre-clair, clair-sombre) [Seyler 59, Seyler 65, Tam 95, Ahumada 93].

Dans la terminologie sur le sujet, on distingue le masquage « avant » (*forward masking*), et le masquage « arrière » (*backward masking*). Le masquage avant est obtenu lorsque que le signal masquant est présenté avant la cible, alors que le masquage arrière est obtenu lorsque le signal masquant est présenté après la cible. Le degré de masquage dépend dans les deux cas de la nature du masque (masque homogène, masque de type bruit aléatoire, etc.). Le masquage avant est plus intuitif. La réduction de la perception après des discontinuités temporelles comme des transitions sombre-clair, clair-sombre, a d'abord été évaluée à quelques centaines de millisecondes [Seyler 59, Seyler 65]. Puis, plus récemment, en étudiant la visibilité d'artéfact de codage de type MPEG-2 après un changement de plan, Tam et al [Tam 95] n'ont trouvé des effets de masquage significatifs que sur la première image suivant le changement de plan. Le masquage arrière est plus éphémère mais tout de même existant. On évalue sa durée à une dizaine de millisecondes. Il peut être expliqué par la variation de la latence

des signaux neuronaux en fonction de leur intensité [Ahumada 93].

Turvey [Turvey 73], affirme qu'il existe deux mécanismes différents qui aboutissent à deux types de masquage : le masquage d'intégration (*integration masking*) et le masquage d'interruption (*interruption masking*). Le masquage d'intégration est le processus par lequel la cible et le masque sont ajoutés pour ne former qu'une seule image composite. Un tel ajout s'opère physiquement lorsque la cible et le masque sont superposés par une présentation simultanée. Cette intégration peut avoir lieu aussi, si le masque est présenté suffisamment tôt après que la cible ait disparu, ou avant que la cible n'apparaisse. Dans les deux cas, la superposition n'a pas lieu physiquement, mais plutôt dans ce qui est appelé la mémoire iconique. L'intégration est d'autant plus importante que la cible et le masque sont proches temporellement. Ce type de masquage est illustré sur la figure 1.9. La courbe en pointillé de cette figure montre les résultats d'expérimentations dans lesquelles il était demandé aux observateurs de reconnaître le plus possible de lettres parmi trois lettres présentées en présence d'un masque constitué de différents motifs dont la luminosité était deux fois supérieure à celle de la cible. Le masque et la cible étaient présentés tous les deux pendant 10 ms. Le délai entre le début de présentation de la cible et le début de présentation du masque (*SOA : stimulus onset asynchrony*) variait de 0 à 184 ms. Lorsque le SOA vaut 0 ms,

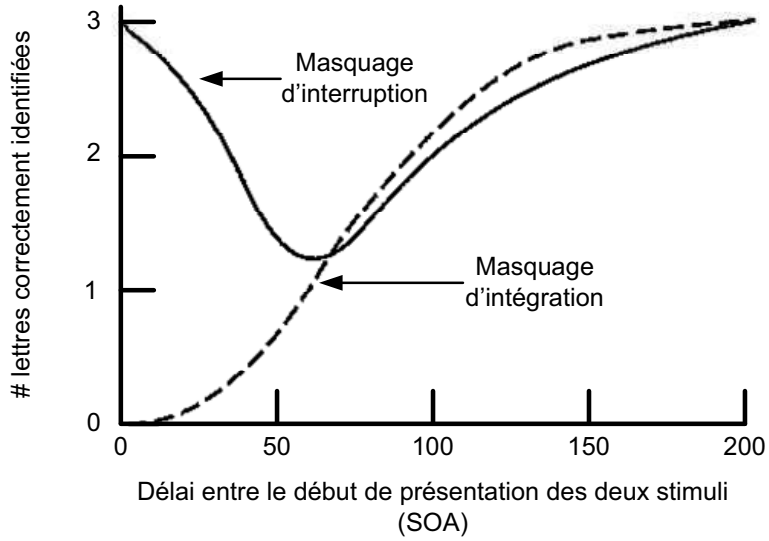


Figure 1.9 – Illustration du masquage d'intégration et du masquage d'interruption [Turvey 73]. Le masquage d'intégration se produit avec un masque constitué de motifs clairs. Le masquage d'interruption se produit avec un masque constitué de motifs sombres.

il y a une réellement superposition optique, mais lorsque le SOA devient supérieur à 10 ms, l'effet de masquage a lieu dans la mémoire iconique. Les résultats montrent que lorsque le masque est intégré optiquement (SOA = 0 ms) avec la cible l'effet est dévastateur sur l'identification. Le même résultat est constaté lorsque le SOA vaut 16 ms. Lorsque le SOA augmente, l'effet de masquage diminue et il devient pratiquement nul vers 200 ms.

L'existence du masquage d'interruption est illustrée par la courbe continue de la figure 1.9. Cette courbe

montre une fonction de masquage très différente dans sa réponse temporelle de celle du masquage d'intégration : pas d'effet de masquage au début, un effet de masquage maximal vers un SOA de 50 ms, puis une diminution de l'effet de masquage. Ces résultats sont obtenus avec les mêmes motifs et les mêmes SOA que précédemment. La différence réside uniquement dans la luminosité du masque qui est deux fois inférieure à celle de la cible. Les résultats de ces deux expérimentations tendent à montrer l'existence de deux mécanismes différents.

Les connaissances en masquage temporel sont bien moindres que pour le masquage spatial. D'après les connaissances auxquelles nous avons eu accès, il n'existe pas de modèle de masquage temporel s'appuyant sur un nombre important de données psychophysiques. Le masquage temporel semble dépendre de nombreux facteurs pour lesquels il faudrait mettre en oeuvre de nombreuses expérimentations psychophysiques, avant de pouvoir proposer une modélisation. La caractérisation des effets de masquage temporel est un sujet de travail en soi qui, malgré son intérêt certain, ne rentre pas dans le cadre de nos travaux.

1.4 Conclusion

La vocation de ce chapitre était de rassembler et de présenter des connaissances sur les images et les vidéos numériques, ainsi que sur les modélisations existantes du système visuel humain. La première partie de ce chapitre nous a permis d'aborder les concepts importants de la vidéo numérique, en particulier la compression. Nous avons également fait une synthèse sur les différents types de distorsions qui peuvent dégrader la qualité qu'une vidéo numérique. Nous avons observé que les distorsions temporelles pouvaient être décrites comme une variation temporelle de distorsions spatiales, ce qui nous ouvre un axe de recherche intéressant pour l'évaluation locale des distorsions dans les vidéos, mais aussi pour l'évaluation de la qualité visuelle des vidéos. La problématique pourrait être par exemple : comment passer des variations de distorsions spatiales à la perception d'une distorsion temporelle ? C'est une question que nous aborderons dans le chapitre 3.

La seconde partie de chapitre était consacrée au système visuel humain et plus particulièrement aux modélisations existantes de certaines de ses propriétés. Les propriétés qui nous intéressent ici sont celles présentant un intérêt dans un contexte d'évaluation objective de la qualité : la perception de la luminance, la perception des couleurs, la sensibilité au contraste, l'organisation multi-canal, et les effets de masquage. Une modélisation du système visuel doit prendre en compte ces différentes propriétés. Dans nos travaux, et afin de cloisonner notre étude, nous avons fait le choix de nous focaliser sur la perception de la luminance à cause de son rôle prépondérant dans la perception.

L'organisation multi-canal est une propriété importante du système visuel humain. La décomposition spatiale de l'information est incontournable et les études sur le sujet permettent une modélisation proche de la réalité. Par contre la décomposition temporelle de l'information reste problématique à mettre en oeuvre, car ses interactions avec la décomposition spatiale ne font pas consensus. Les méthodes que nous proposerons dans la suite de ce mémoire s'appuieront sur une décomposition spatiale de l'information, tout à fait adaptée pour les méthodes concernant les images. Par contre pour les méthodes concernant les vidéos, nous ne modéliserons pas la décomposition temporelle décrite par les canaux *sustained* et *transient*, mais nous proposerons une autre

approche reposant sur l'étude des variations temporelles des distorsions spatiales. Les décompositions spatiales existantes présentant une complexité de calcul parfois rédhibitoire en vue d'une utilisation sur de la vidéo, nous proposerons également une décomposition spatiale à complexité plus réduite.

La modélisation du comportement multi-canal du système visuel humain est nécessaire à la bonne prise en compte des effets de masquage. Dans les méthodes que nous proposerons, nous prendrons en compte les effets de masquage spatial et nous laisserons de côté les effets de masquage temporel, car les connaissances auxquelles nous avons eu accès sur le sujet nous semblent trop limitées pour que nous en proposions une modélisation pertinente. Cependant, les effets de masquage spatial ne seront pas limités au masquage de contraste comme c'est souvent le cas dans la littérature, mais nous prendrons aussi en compte le masquage entropique.

Chapitre 2

Imagerie des distorsions perçues : conception de nouvelles méthodes et validation comparative

2.1 Introduction

L'objet de ce chapitre est la conception de cartes de distorsions visuelles d'images fixes. Il s'agit de construire une carte de distorsions perceptuelles représentant les distorsions qu'auraient perçues des observateurs humains entre une image dite *originale* et une version dite *dégradée* de cette même image. Pour atteindre cet objectif, il est nécessaire de modéliser la perception visuelle humaine. Dans ce chapitre nous nous intéressons à deux aspects particuliers de la modélisation du système visuel humain : la décomposition en sous-bandes de fréquences spatiales et les effets de masquages.

Le comportement multi-canal du système visuel humain peut être modélisé par une décomposition en canaux perceptuels. Idéalement, cette décomposition peut être effectuée dans le domaine de Fourier, mais au prix d'une complexité de calcul importante. Une transformée spatiale comme la transformée en ondelettes pourrait être une alternative intéressante en terme de complexité. Même si la correspondance entre les sous-bandes perceptuelles et les sous-bandes ondelettes n'est pas directe, cette alternative mérite d'être étudiée de près.

La modélisation des effets de masquage en évaluation objective de la qualité se limite presque toujours à la modélisation du masquage de contraste. Cependant, le masquage de contraste ne permet pas d'expliquer toujours l'élévation du seuil de visibilité en particulier sur les images naturelles lorsque que la complexité spatiale devient importante. Dans ces situations l'effet de masquage est qualifié de masquage semi-local. Celui-ci est aussi appelé par d'autres : masquage entropique [Watson 97b], masquage de texture [Gaubatz 05], ou encore masquage d'activité [Nadenau 00].

Dans ce chapitre, nous proposons plusieurs méthodes de conception de cartes de distorsions visuelles d'images. D'une part, nous introduisons le masquage entropique dans un modèle du système visuel humain reposant sur le domaine de Fourier. Et d'autre part, nous proposons un modèle du système visuel humain reposant sur le domaine des ondelettes et prenant en compte le masquage entropique.

Dans une première partie nous faisons une revue de la littérature sur les méthodes d'évaluation locale des distorsions des images fixes. Nous décrivons des approches mathématiques, des approches structurelles et des approches modélisant le système visuel humain. La suite du chapitre est consacrée à la description des différents aspects des modèles proposés : le comportement multi-canal, la sensibilité au contraste et les effets de masquage.

2.2 Revue des méthodes existantes

Dans la littérature sur l'évaluation objective de la qualité visuelle, la plupart des approches évaluent la qualité de manière globale, c'est-à-dire que l'ensemble des distorsions perçues dans l'image est résumé sous la forme d'une grandeur unique appelée note de qualité. Comme nous l'avons évoqué en introduction générale, ces approches répondent à un besoin des concepteurs de systèmes de traitement d'images et l'évaluation de leurs performances est facilitée par l'existence de tests subjectifs d'évaluation de qualité. Cependant, l'évaluation locale des distorsions est aussi un besoin fort des concepteurs de systèmes de traitement d'images, notamment en codage, car elle permet d'évaluer et de comparer localement les distorsions introduites par différents systèmes. L'information générée par cette évaluation locale est plus riche qu'une note globale, et peut permettre une analyse plus fine d'un système de traitement d'images. La problématique de la conception de cartes d'erreurs perceptuelles d'images a été abordée dans la littérature sur l'évaluation objective de la qualité. D'ailleurs, il s'agit souvent d'une étape précédant l'élaboration de la note de qualité. Le figure 2.1 présente la structure générale d'une métrique de qualité d'images reposant sur la construction de cartes d'erreurs. L'étape de création de cartes d'erreurs y est encadrée en rouge. Cependant, l'évaluation quantitative de la pertinence des cartes d'erreurs reste un problème tant qu'il n'existera pas de données subjectives.

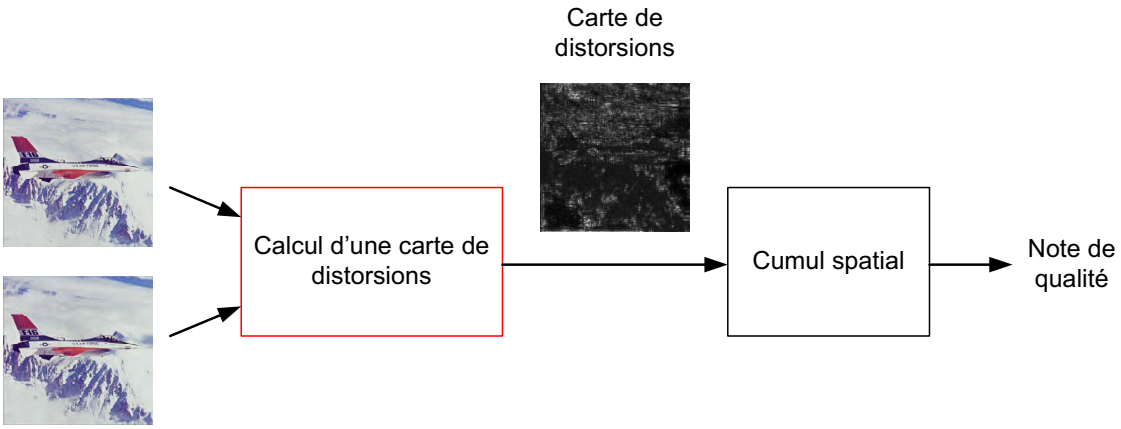


Figure 2.1 – Structure générale d'une métrique de qualité d'images reposant sur le calcul de cartes d'erreurs

Dans cette section, nous allons passer en revue les principales approches qui se dégagent dans la littérature sur le sujet.

2.2.1 Les approches purement de type signal

Les méthodes les plus directes pour construire des cartes de distorsions sont les méthodes basées sur une différence mathématique $D(m,n)$ entre l'image originale $I_o(m,n)$ et l'image dont la qualité est à évaluer $I_d(m,n)$:

$$D(m,n) = |I_o(m,n) - I_d(m,n)|^p \tag{2.1}$$

Lorsque $p = 2$, cette carte d'erreurs est utilisée pour le calcul de mesure de distorsions comme la MSE (*Mean Square Error*) ou le PSNR (*Peak Signal to Noise Ratio*) sur lesquelles nous reviendrons dans la section 4.4.1.1. L'avantage de ces cartes d'erreurs mathématiques réside dans leur simplicité d'implémentation, ainsi que dans leur rapidité de calcul. Par contre, elles présentent l'inconvénient majeur de n'être basées que sur le signal ; elles ne prennent donc pas en compte les propriétés du système visuel humain. Le niveau de perception des distorsions dues à des erreurs dans le signal n'est malheureusement pas simplement lié au niveau de ces erreurs. Pour s'en convaincre, il suffit de regarder l'exemple de la figure 2.2, dans lequel une image originale est présentée avec une version dégradée, ainsi que la carte d'erreurs mathématique associée. L'interprétation de cette carte laisserait penser que les distorsions dans la zone correspondant aux montagnes sont plus visibles que dans la zone correspondant au ciel, ce qui n'est pas le cas.

2.2.2 Les approches structurelles

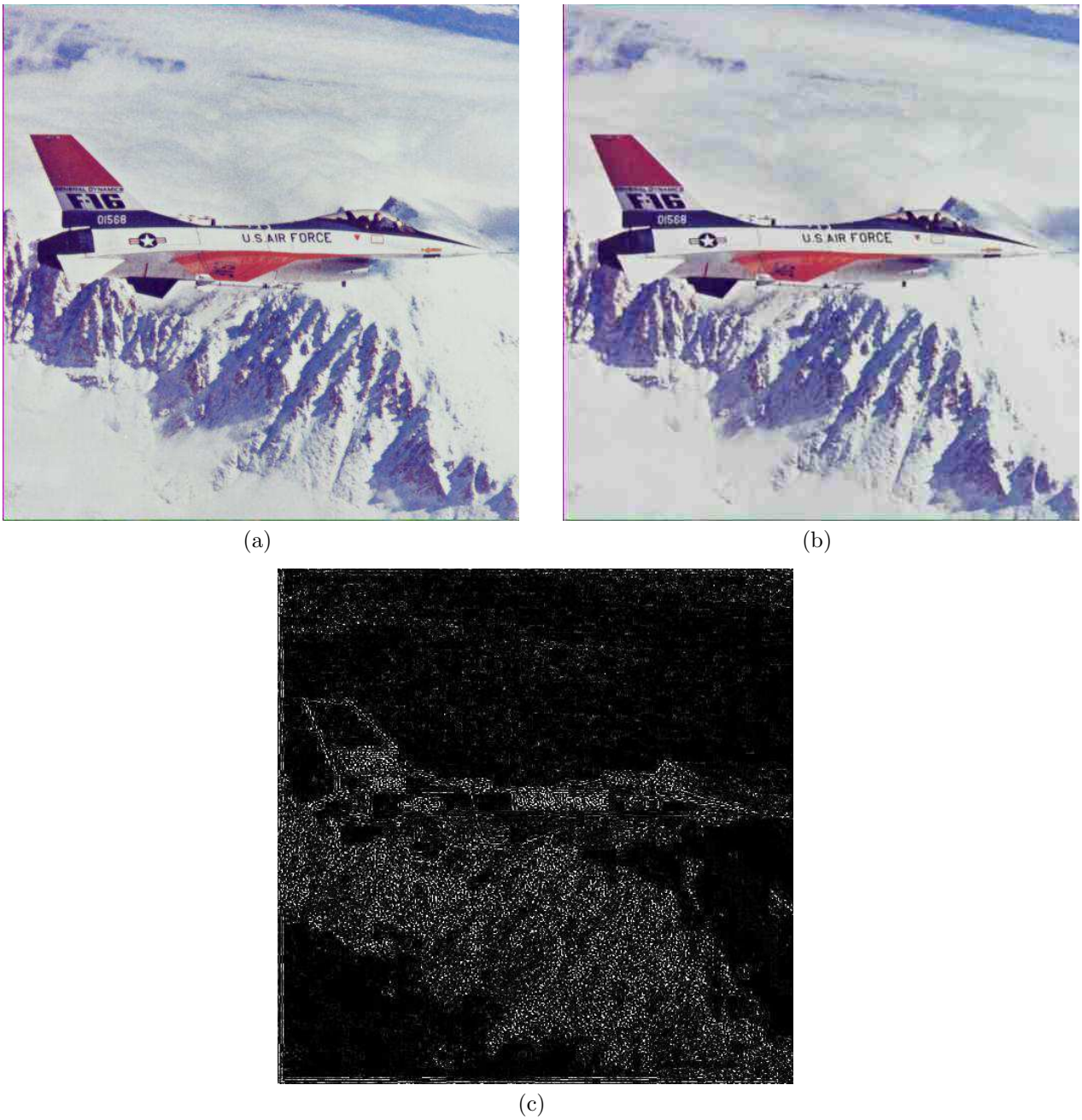
Une autre approche est celle de Wang et *al.* avec la SSIM (Structural SIMilarity) [Wang 04a]. L'idée principale de la SSIM est de mesurer la « similarité de structure » entre deux images, plutôt qu'une différence pixel à pixel comme le font les approches purement de type signal. L'hypothèse sous-jacente est que l'oeil humain est plus sensible aux changements dans la structure de l'image. Cette approche ne repose pas sur une modélisation du système visuel humain, mais elle prend en compte des spécificités des images auxquelles il est sensible. Les images naturelles sont fortement structurées, c'est-à-dire que les pixels d'une image sont très dépendants les uns des autres, et en particulier lorsqu'ils sont proches les uns des autres. Ces structures jouent un rôle important dans la perception de la scène. Par conséquent, une modification de la structure de l'image impacte la perception que l'on a de cette image. Toutefois, le calcul de similarité ne se limite pas seulement à la comparaison des structures entre les images ; les différences de luminance et de contraste entre les deux images sont également évaluées. Comme nous l'avons évoqué dans le chapitre 1, la luminance et le contraste jouent effectivement un rôle important dans la perception.

Le calcul de similarité s'effectue entre une fenêtre f_o de l'image originale et la fenêtre f_d correspondante de la version dégradée, en combinant trois mesures : une mesure de similarité de luminance $l(f_o, f_d)$, une mesure de similarité de contraste $c(f_o, f_d)$ et une mesure de similarité de structure $s(f_o, f_d)$:

$$SSIM(f_o, f_d) = [l(f_o, f_d)]^\alpha . [c(f_o, f_d)]^\beta . [s(f_o, f_d)]^\gamma \tag{2.2}$$

avec :

$$l(f_o, f_d) = \frac{2\mu_{f_o}\mu_{f_d} + C_1}{\mu_{f_o}^2 + \mu_{f_d}^2 + C_1} \tag{2.3}$$



(a) (b)

(c)

Figure 2.2 – Carte d'erreurs mathématiques (c) entre l'image *Avion* originale (a), et une version dégradée (b) par un schéma de compression de type JPEG. Plus les valeurs sont sombres, et plus l'erreur mathématique est faible.

$$c(f_o, f_d) = \frac{2\sigma_{f_o}\sigma_{f_d} + C_2}{\sigma_{f_o}^2 + \sigma_{f_d}^2 + C_2} \tag{2.4}$$

$$s(f_o, f_d) = \frac{\sigma_{f_o f_d} + C_3}{\sigma_{f_o}\sigma_{f_d} + C_3} \tag{2.5}$$

Par simplification et en posant $\alpha = \beta = \gamma = 1$, on obtient la relation suivante :

$$SSIM(f_o, f_d) = \frac{(2\mu_{f_o}\mu_{f_d} + C_1)(\sigma_{f_o f_d} + C_2)}{(\mu_{f_o}^2 + \mu_{f_d}^2 + C_1)(\sigma_{f_o}^2\sigma_{f_d}^2 + C_2)} \tag{2.6}$$

avec :

- μ_{f_o}, μ_{f_d} : respectivement les moyennes de f_o et f_d (indicateur de luminance) ;
- $\sigma_{f_o}^2$, $\sigma_{f_d}^2$: respectivement les variances de f_o et f_d (indicateur de contraste) ;
- $\sigma_{f_o f_d}$: la covariance entre f_o et f_d ;
- $C1 = (K_1.L)^2$, $C2 = (K_2.L)^2$: deux variables destinées à stabiliser la division quand le dénominateur est de valeur très faible ;
- L étant la dynamique des valeurs des pixels (soit 255 pour des images codées sur 8 bits) ;
- $K_1 = 0,01$ et $K_2 = 0,03$ par défaut.

La SSIM est une généralisation de la mesure UQI (Universal Quality Index) définit par Wang et Bovick dans [Wang 01] et [Wang 02a] et dans laquelle $C_1 = C_2 = 0$. Cette généralisation permet de stabiliser la mesure pour des valeurs de $(\mu_{f_o}^2 + \mu_{f_d}^2)$ et $(\sigma_{f_o}^2\sigma_{f_d}^2)$ proches de zéro.

En appliquant la SSIM sur une fenêtre glissante centrée sur chaque pixel (m, n) des images à comparer, il est possible de créer une carte des erreurs structurelles. La figure 2.3 présente la carte SSIM calculée entre les images (a) et (b) de la figure 2.2. On peut observer sur cette carte que les contours à fort contraste, comme les contours de l'avion, ont un effet perturbateur sur les valeurs de SSIM calculées dans les fenêtres les contenant. Le fort contraste du contour entraîne une trop forte corrélation entre les deux fenêtres, ce qui provoque une sous-estimation des erreurs par la SSIM. Par contre, on peut observer la détection des frontières des effets de blocs dans les zones de ciel et de neige, ce qui est un point intéressant. Cependant, comme pour la carte des d'erreurs mathématiques de la figure 2.2, les valeurs relatives des erreurs mesurées entre les zones de ciel et de montagne sont discutables, car elles laissent penser que les montagnes sont visuellement plus dégradées que le ciel.

Dans la continuité de la SSIM, d'autres méthodes basées sur les erreurs structurelles ont été proposées. On peut citer la SSIM multi-échelle (*MS-SSIM multi-scale SSIM*) également proposée par Wang et *al.* dans [Wang 03]. Cette méthode reprend les concepts de la SSIM mais les applique à une approche multi-résolution. Les niveaux de résolutions sont calculés à partir des images de départ par filtrage passe-bas et sous-échantillonnage. Les mesures $l(m, n)$, $c(m, n)$ et $s(m, n)$ sont calculées à différentes résolutions puis combinées selon la relation :

$$MS - SSIM(f_o, f_d) = [l(f_o, f_d)]^{\alpha_M} . \prod_{j=1}^{M} [c(f_o, f_d)]^{\beta_j} . [s(f_o, f_d)]^{\gamma_j} \tag{2.7}$$

avec j représentant les différents niveaux de résolutions, M étant le nombre de niveaux de résolution utilisé, et les paramètres α_M, β_j, γ_j permettant d'ajuster l'importante des différentes composantes. Même si cette

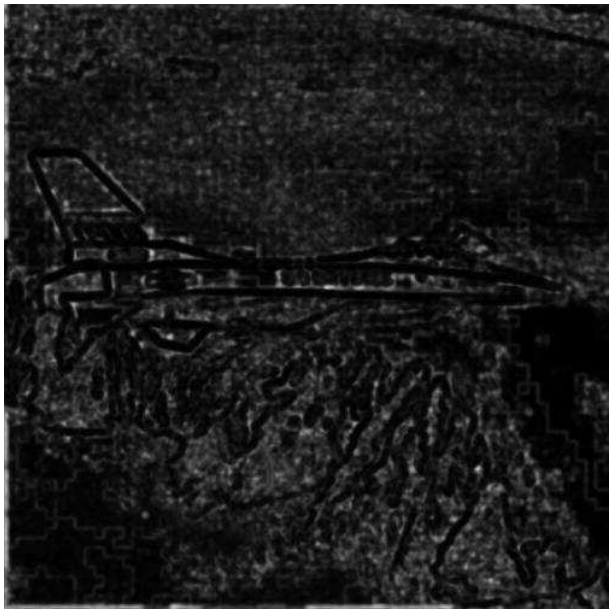


Figure 2.3 – Carte d'erreurs structurelles SSIM, entre l'image *Avion* originale (Fig.2.2a), et une version dégradée (Fig.2.2b) par un schéma de compression de type JPEG. Plus les valeurs sont sombres, et plus les erreurs structurelles sont faibles.

approche est plus flexible que la SSIM, et tente d'inclure des notions de variations de la sensibilité en fonction des fréquences spatiales (cf. CSF) au travers des différents niveaux de résolution, aucune de ces deux méthodes ne permet de prendre en compte les conditions d'observation dont l'importance est primordiale.

2.2.3 Les approches modélisant le système visuel humain

Les approches les plus intéressantes sont celles qui tentent de modéliser les mécanismes de la perception humaine. Ces approches peuvent être plus ou moins complètes et complexes, selon les aspects qu'elles prennent en compte. Elles ont toutes pour objectif d'évaluer la visibilité des erreurs.

Le VDP (*Visible Difference Predictor*) de Daly [Daly 92, Daly 93] a pour but de créer des cartes représentant la probabilité de détection des différences entre l'image d'origine et l'image dégradée. Chaque site (m, n) de la carte représente la probabilité qu'un observateur humain perçoive la différence entre l'image de référence et l'image à évaluer au site considéré. L'image originale et l'image dégradée sont exprimées en valeur de luminance avant de passer par un ensemble de traitement : filtrage par une CSF, décomposition en canaux, calcul du contraste, modélisation des effets de masquage, et calcul des probabilités de détection. Une décomposition Cortex [Watson 87] modifiée est utilisée pour la décomposition en canaux, qui transforme l'image en 31 sous-bandes indépendantes (cinq bandes de fréquences radiales chacune avec six orientations plus une sous-bande basses fréquences). Pour chaque sous-bande, une carte d'élévation de seuil est calculée à partir des valeurs de contraste. Puis les erreurs dans chaque sous-bande sont normalisées par les valeurs d'élévations de seuil associées, avant d'être transformées en probabilité de détection par une fonction « psychométrique ». Les cartes de probabilité

de détection de chaque sous-bande sont ensuite cumulées afin d'obtenir la carte finale.

La méthode de Lubin [Lubin 93, Lubin 95] tente aussi d'estimer une probabilité de détection des différences entre une image originale et une image dégradée. Cette méthode est connue sous le nom de VDM (*Sarnoff Visual Discrimination Model*). Un flou est d'abord appliqué sur les images pour simuler la PSF (*Point Spread Function*) du système optique de l'oeil. Les images sont ensuite ré-échantillonnées afin de refléter l'échantillonnage des photo-récepteurs de la rétine. La décomposition utilisée est une pyramide Laplacienne [Burt 83b] à sept niveaux. Elle est suivie par le calcul d'un contraste à bande limitée [Peli 90]. La sélectivité angulaire est réalisée au moyen de filtres orientés de Freeman et Adelson [Freeman 91] pour quatre orientations. La CSF est simulée en normalisant la sortie de chaque filtre par une valeur de sensibilité approximant celle de la sous-bande correspondante. L'effet de masquage est réalisé au moyen d'une fonction sigmoïde. Finalement, la carte de distorsions est calculée par cumul utilisant une sommation de Minkowski entre les sous-bandes. Les valeurs encodées par cette carte sont appelées JND (*Just Noticeable Difference*). Cette méthode a été modifiée plus tard pour obtenir la métrique Sarnoff JND pour des vidéos couleurs [Lubin 97].

Teo et Heeger [Teo 94, Heeger 95] proposent une modélisation prenant en compte la PSF, l'effet de masquage de luminance (ou adaptation à la luminance), la décomposition multi-canal, la normalisation du contraste. La décomposition est effectuée selon une pyramide hexagonale avec des filtres QMF (*Quadrature Miror Filter*) selon quatre résolutions spatiales et six orientations. L'effet de masquage est modélisé par une normalisation du contraste et une saturation de la réponse. La normalisation du contraste est différente de celle de Daly et Lubin, car toutes les sorties de toutes les orientations d'un même niveau de résolution sont utilisées pour calculer la normalisation. Il ne considère donc pas que les orientations d'une même résolution soient indépendantes, mais seulement que les différents niveaux de résolutions sont indépendants.

Bradley [Bradley 99] décrit un modèle appelé WVDP (*Wavelet Visible Difference Predictor*) qui est une simplification du VDP de Daly. Il utilise les résultats de Watson [Watson 97c] sur la détection de bruit de quantification après transformation par les ondelettes 9/7, et les combine avec une élévation de seuil et une fonction psychométrique de probabilité de détection, d'une façon similaire à celle de Daly.

Watson a proposé un modèle dans le domaine DCT (*Discrete Cosine Transform*) [Watson 93]. Même si ce modèle ne permet pas de sortir directement des cartes de distorsions, une simple modification de l'ordre des cumuls d'erreurs permet d'obtenir une carte de distorsions au niveau bloc. Ce modèle repose sur la transformée DCT 8×8 couramment utilisée en traitement d'image et en compression vidéo. Contrairement aux autres méthodes citées, cette méthode décompose le spectre en 64 sous-bandes uniformes. Après la transformée DCT par bloc, des valeurs de contraste sont calculées par sous-bande, un seuil de visibilité est construit pour chaque coefficient de chaque sous-bande et cela dans chaque bloc en utilisant la sensibilité de base de la sous-bande. Les sensibilités de base de chaque sous-bande sont déduites empiriquement. Les seuils sont corrigés en fonction du masquage de luminance et du masquage de texture. Les erreurs dans chaque sous-bande sont pondérées par les seuils de visibilité correspondants, puis cumulées par des sommations de Minkowski.

Le Callet [Le Callet 01] a proposé une approche s'inscrivant dans la lignée du VDP de Daly, mais prenant

en compte l'information chromatique. Le modèle proposé décompose les images couleur en trois composantes perceptuelles. Chaque composante couleur est décomposée en canaux perceptuels, puis la visibilité des différences entre les images est définie à partir de fonction de masquage. Un des intérêts de ce modèle est qu'il est basé sur de nombreuses expérimentations psychophysiques. Nos travaux s'inspirant de ce modèle, celui-ci est détaillé dans la suite de ce chapitre.

2.2.4 Discussion

Les approches purement de type signal ne sont manifestement pas adaptées à l'évaluation de la qualité visuelle car elles ne prennent pas en compte la perception humaine. Les approches structurelles sont plus intéressantes car elles se basent sur des considérations perceptuelles. Toutefois, ne reposant pas sur une modélisation du système visuel, elles ne peuvent pas prétendre prédire convenablement les distorsions dans le cas général. Les approches s'appuyant sur un modèle du système visuel humain sont les plus intéressantes. On constate que toutes les approches présentées possèdent des points communs :

- utilisation d'une décomposition en sous-bandes du signal d'image,
- modélisation de la sensibilité au contraste,
- modélisation des effets de masquage.

Ces points communs touchent des propriétés essentielles du système visuel humain permettant de prédire la visibilité des erreurs entre l'image originale et l'image dégradée. Une méthode d'évaluation de la qualité ou une méthode d'évaluation locale des distorsions doit s'appuyer sur une modélisation de ces propriétés. Cette modélisation varie d'une approche à l'autre.

Les décompositions en sous-bandes, qui permettent de modéliser correctement les effets de masquage, sont réalisées de manières très différentes selon les approches. Les paramètres guidant le choix de la décomposition ne sont pas seulement liés à une modélisation fidèle du système visuel, mais sont aussi liés à des considérations de complexité algorithmique ou à des considérations de compatibilité avec les algorithmes de codage. Il est important de trouver un bon compromis entre toutes ces considérations, pour que la complexité de l'approche soit compatible avec son utilisation, sans pour autant que la décomposition utilisée perde son intérêt à cause d'une modélisation trop imparfaite.

Les modèles de masquage utilisés dans les approches présentées portent plus spécifiquement sur les effets de masquage de contraste. Cependant, le masquage de contraste ne constitue pas le seul masquage à prendre en compte, et comme le souligne Watson dans [Watson 97b], une métrique de qualité devrait prendre en compte non seulement le masquage de contraste, mais aussi le masquage semi-local.

2.3 Principe général des deux modèles proposés

Nous proposons deux approches multi-canal du système visuel humain, nous permettant de construire des cartes de distorsions perceptuelles. Ces deux approches exploitent seulement le signal achromatique. Elles modélisent la sensibilité au contraste (CSF), le comportement multi-canal et à la fois le masquage de contraste et

le masquage semi-local. Le premier modèle est largement inspiré des travaux de Daly [Daly 93] et de Le Callet [Le Callet 01], et repose sur une décomposition en sous-bandes réalisée dans le domaine de Fourier, il sera noté FQA (*Fourier based Quality Assessment*). Les modèles utilisant une décomposition dans le domaine de Fourier produisent de bonnes performances, mais leur complexité est élevée. Nous avons cherché à réduire la complexité en proposant un second modèle, qui est une adaptation au domaine des ondelettes du premier modèle. Il sera noté WQA (*Wavelet based Quality Assessment*). Les décompositions utilisées dans chacun des modèles proposés présentent des caractéristiques très différentes. Le modèle FQA repose sur une décomposition en canaux perceptuels caractérisée à partir d'expérimentations psychovisuelles. Cette décomposition est donc très proche de celle réalisée par le système visuel humain. De son côté, le modèle WQA repose sur une décomposition en ondelettes adaptée pour se rapprocher de la décomposition en canaux perceptuels. Bien qu'adaptée, cette décomposition est plus éloignée du système visuel humain que la précédente. La correspondance entre le système visuel humain et le domaine des ondelettes est d'ailleurs connue pour être seulement approximative [Bradley 99, Zeng 02]. Cependant il est intéressant d'étudier dans quelle mesure cette décomposition est handicapante dans un contexte d'évaluation de qualité.

Ces deux modèles présentent une structure plutôt classique au regard des approches existantes de critères objectifs de qualité avec référence complète s'appuyant sur une modélisation du système visuel humain. La structure de ces deux modèles est illustrée respectivement figure 2.4 pour le modèle FQA et figure 2.5 pour le modèle WQA.

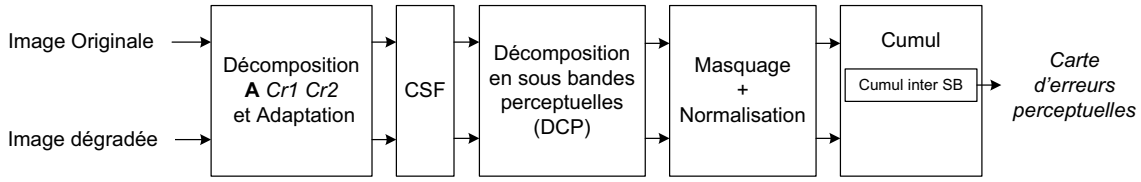


Figure 2.4 – Structure du modèle basé Fourier (FQA).

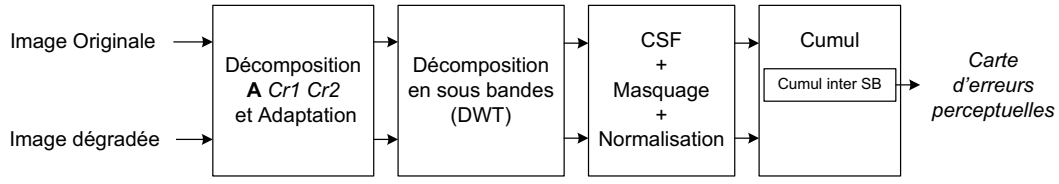


Figure 2.5 – Structure du modèle basé ondelettes (WQA).

Les deux modèles proposés se décomposent en quatre étapes. Dans les sections suivantes nous allons décrire en détails ces différentes étapes de la construction de cartes d'erreurs perceptuelles. La première consiste à se projeter dans un espace perceptuel afin d'obtenir la composante achromatique des images après d'adaptation en luminance du système visuel.

2.4 Espace de couleur et adaptation en luminance

Les images numériques couleur sont généralement représentées par les trois composantes couleurs correspondant aux couleurs primaires de la synthèse additive : Rouge, Vert et Bleu. Ce codage appelé RVB (ou *RGB* en anglais) correspond à la manière dont les écrans à tube (TV ou d'ordinateur) représentent les couleurs. Pour représenter perceptuellement une image numérique, il est d'abord nécessaire de transformer les composantes RVB en signaux lumineux en prenant en compte les caractéristiques des trois fonctions gamma de l'écran d'affichage. Les caractéristiques de ces fonctions sont propres à chaque écran et peuvent être mesurées au moyen d'un colorimètre dont une illustration est présentée figure 2.6.

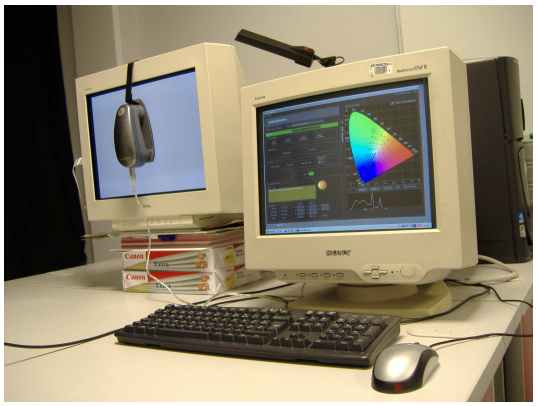


Figure 2.6 – Mesures des gammas d'un écran CRT à l'aide d'un colorimètre et d'un logiciel adapté.

La fonction non linéaire propre à chaque composante est ensuite modélisée à l'aide des relations suivantes :

$$L_R = Offset_R + L_{R,max} \times (\frac{R}{R_{max}})^{\gamma_R} \tag{2.8}$$

$$L_V = Offset_V + L_{V,max} \times (\frac{V}{V_{max}})^{\gamma_V} \tag{2.9}$$

$$L_B = Offset_B + L_{B,max} \times (\frac{B}{B_{max}})^{\gamma_B} \tag{2.10}$$

avec :

- L_R, L_V et L_B, les luminances des composantes Rouge, Vert ou Bleu respectivement ;
- $L_{R,max}$, $L_{V,max}$ et $L_{B,max}$, respectivement la luminance maximale de la composante Rouge, Vert ou Bleu ;
- γ_R, γ_V et γ_B, des paramètres dépendants du dispositif d'affichage mesurés à l'aide d'un colorimètre, et appelés communément les gammas de l'écran (cf. figure 2.6).
- $Offset_R$, $Offset_V$ et $Offset_B$, les valeurs de luminance pour un niveau nul respectivement de la composante Rouge, Vert ou Bleu ;
- R_{max}, V_{max} et B_{max}, les valeurs maximales respectivement de la composante Rouge, Vert ou Bleu, et égale à 255 pour un codage sur 8 bits.

Les signaux lumineux (L_R, L_V, L_B) doivent ensuite être projetés dans un espace perceptuel. Nous avons choisi l'espace colorimétrique de Krauskopf (A, Cr1, Cr2) car celui-ci a été validé comme espace perceptuel couleur [Bedat 98], et c'est aussi l'espace utilisé dans [Le Callet 01]. Les signaux lumineux sont convertis à l'aide de la relation suivante :

$$\begin{pmatrix} A \\ Cr1 \\ Cr2 \end{pmatrix} = L_{max} \begin{pmatrix} \frac{0.2244}{L_{R,max}} & \frac{0.6811}{L_{V,max}} & \frac{0.0942}{L_{B,max}} \\ \frac{0.0891}{L_{R,max}} & \frac{-0.0617}{L_{V,max}} & \frac{-0.0275}{L_{B,max}} \\ \frac{-0.1029}{L_{R,max}} & \frac{-0.2874}{L_{V,max}} & \frac{0.3903}{L_{B,max}} \end{pmatrix} \begin{pmatrix} L_R \\ L_V \\ L_B \end{pmatrix} \tag{2.11}$$

avec :

- L_R, L_V et L_B, les luminances des composantes Rouge, Vert ou Bleu respectivement ;
- $L_{R,max}$, $L_{V,max}$ et $L_{B,max}$, les luminances maximales de la composante Rouge, Vert ou Bleu respectivement ;
- $L_{max} = L_R + L_V + L_B$, la luminance maximale.

La composante A représente la composante achromatique et les composantes Cr1 et Cr2 sont deux composantes chromatiques reposant sur la théorie des signaux antagonistes. La figure 2.7 illustre sur un exemple la décomposition d'une image dans l'espace colorimétrique de J. Krauskopf (A, Cr1, Cr2).

(A) (Cr1) (Cr2)

Figure 2.7 – Illustration sur l'image *Isabelle* des trois composantes de l'espace colorimètrique perceptuel de J. Krauskopf [Krauskopf 82] : composante A (achromatique), Cr1 (axe rouge-vert) et Cr2 (axe bleu-jaune).

Notre étude est limitée à une approche achromatique, donc seule la composante A sera utilisée. Comme introduit section 1.3.1, la perception de la luminance n'est pas linéaire. Cette non linéarité dans la perception de la luminance est modélisée par la relation 1.1.

L'étape suivante est la modélisation du comportement multi-canal du système visuel humain. La décomposition en sous-bandes de chacun des deux modèles proposés est décrite dans la section suivante.

2.5 Modélisation du comportement multi-canal

2.5.1 Décomposition en canaux perceptuels (DCP)

La décomposition en canaux perceptuels traduit la sélectivité spatio-fréquentielle du système visuel humain. Elle met en oeuvre un découpage du plan spatio-fréquentiel tel qu'obtenu par une transformée de Fourier. Dans nos travaux, nous utilisons une transformée de Fourier rapide FFT (*Fast Fourier Transform*) afin de calculer la transformée de Fourier discrète (TFD) des images. L'élaboration de la décomposition en canaux perceptuels est réalisée à partir d'un ensemble de filtres : les filtres DoM et les filtres Fan. Pour des détails d'implémentation, le lecteur pourra se reporter à l'annexe B.

La décomposition en canaux perceptuels est issue des travaux de Sénane et de Le Callet [Sénane 96, Le Callet 01]. Cette décomposition propose un pavage fréquentiel en dix-sept canaux comme l'illustre la figure 2.8. Cette décomposition est fortement inspirée de la transformée Cortex de Watson, cependant le nombre et les caractéristiques des canaux sont différents. Parmi ces différences on peut noter que la DCP ne présente pas un découpage dyadique en fréquences radiales et que la sélectivité angulaire varie en fonction des fréquences radiales. Contrairement à la décomposition de Watson, la DCP s'appuie sur un ensemble d'expérimentations psychophysiques.

Dans les conditions normalisées de visualisation, on distingue quatre domaines de fréquences spatiales radiales (BF, II, III, IV). Le domaine BF (basses fréquences) correspond aux fréquences spatiales comprises entre 0 et 1.5 cpd (cycles par degré), le domaine II correspond aux fréquences comprises entre 1.5 et 5.7 cpd, le domaine III correspond aux fréquences comprises entre 5.7 et 14.2 cpd et le domaine IV correspond aux fréquences comprises entre 14.2 et 28.2 cpd. La sélectivité angulaire dépend du domaine de fréquence spatiale considéré. Pour le domaine BF, il n'y a pas de sélectivité angulaire. Pour le domaine II, la sélectivité angulaire est de 45°, ce qui définit quatre canaux orientés. Pour les domaines III et IV, la sélectivité angulaire est de 30°, ce qui définit six canaux orientés pour chaque domaine.

2.5.2 Adaptation de la transformée en ondelettes

2.5.2.1 Transformée en ondelettes

Par transformée en ondelettes, nous désignons la transformée en ondelettes bi-dimensionnelle, ou ondelettes 2D. La transformation en ondelettes est une technique d'analyse multirésolution de l'image qui consiste à décomposer une image en un ensemble de sous-bandes de résolution inférieure.

En dimension 2 (comme en dimension supérieure), on construit classiquement les bases d'ondelettes de $L^2(\mathbb{R}^2)$ par produit séparable de fonctions d'une variable.

Soit ϕ une fonction d'échelle et ψ l'ondelette correspondante, on définit trois ondelettes :

$$\psi^1(x,y) = \phi(x)\psi(y), \quad \psi^2(x,y) = \psi(x)\phi(y), \quad \psi^3(x,y) = \psi(x)\psi(y) \tag{2.12}$$

Ces trois ondelettes extraient des détails de l'image suivant des orientations différentes. A chaque échelle 2^{-j}, la transformée en ondelettes bi-dimensionnelle est caractérisée par les coefficients suivants, pour tout point

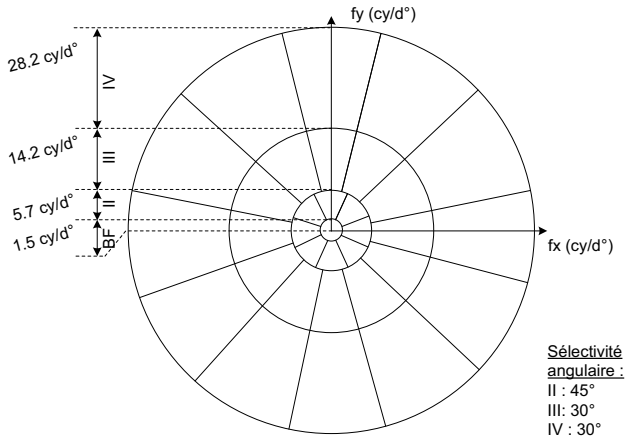


Figure 2.8 – Pavage du plan fréquentiel pour la Décomposition en Canaux Perceptuels (DCP) de la composante A. Les dix sept sous-bandes de la DCP sont réparties sur les quatre couronnes (BF, II, III, IV).

$n = (n1, n2)$:

$$a_j[n] = \langle f, \phi^2_{j,n} \rangle \quad \text{et} \quad d^k_j[n] = \langle f, \psi^k_{j,n} \rangle \quad \text{pour} \quad 1 \leq k \leq 3 \tag{2.13}$$

En pratique, ces coefficients sont obtenus par filtrages mono-dimensionnels successifs sur les lignes puis les colonnes (ou inversement). La fonction d'échelle ϕ peut être perçue comme la réponse impulsionnelle d'un filtre passe-bas demi-bande, et l'ondelette ψ comme celle d'un filtre passe-haut demi-bande. En codage d'image et vidéo, souvent on utilise en pratique le banc de filtres biorthogonal 9-7 de Daubechies [Antonini 92], qui présente un très bon compromis entre séparation spectrale et complexité. C'est d'ailleurs cette transformée que nous utiliserons dans nos travaux. Comme la transformée en ondelettes est obtenue par filtrage, on repère en général les coefficients définis ci-avant par la sous-bande à laquelle ils appartiennent. On note traditionnellement LL_j la bande fréquentielle des coefficients $a_j[n]$, LH_j celle des $d^1_j[n]$, HL_j celle des $d^2_j[n]$, et HH_j celle des $d^3_j[n]$. Autrement dit, la sous-bande LL_j représente les valeurs d'approximation ou les basses fréquences, les sous-bandes LH_j représentent les détails horizontaux, les sous-bandes HL_j représentent les détails verticaux, et les sous-bandes HH_j représentent les détails diagonaux.

La figure 2.9 représente une image source et sa décomposition ondelettes sur deux niveaux.

2.5.2.2 Transformée en ondelettes et décomposition en canaux perceptuels

La transformée en ondelettes présente des similarités avec l'organisation multi-canal du système visuel humain. Comme la décomposition en canaux perceptuels, la transformée en ondelettes décompose l'image en un certain nombre de sous-bandes. Ces sous-bandes correspondent à une bande de fréquences limitée, ainsi qu'à un ensemble limité d'« orientations ». De plus, le contenu de chaque sous-bande correspond à une localisation spatiale particulière. Cependant, des différences entre la DCP et la DWT existent.

Tout d'abord, la DWT est une transformée séparable et dyadique, alors que la DCP n'est ni séparable, ni dyadique. Ensuite les grandes différences se situent au niveau des gammes de fréquences et d'orientations des

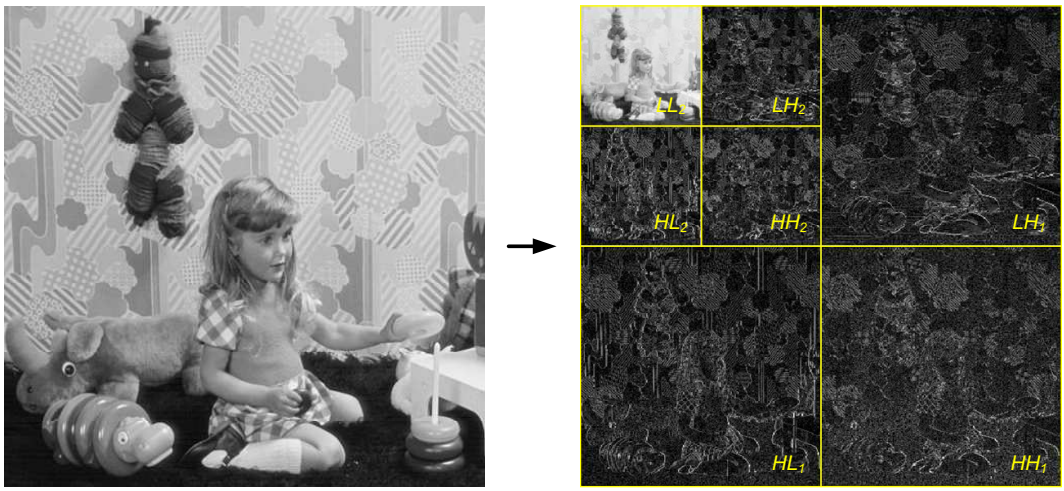


Figure 2.9 – Image source *Isabelle* et sa transformée dyadique en ondelettes sur deux niveaux. Plus un coefficient est sombre, plus sa valeur absolue est faible.

sous-bandes. La sélectivité angulaire de la DCP est de 45° ou de 30° en fonction de la gamme de fréquences radiales de la sous-bande. La sélectivité angulaire de la transformée en ondelettes est seulement de 45°, de plus les sous-bandes diagonales contiennent à la fois les orientations à 45° et les orientations à -45°. Cela peut poser un problème pour la prise en compte des effets de masquage, car les signaux d'orientation proche de 45° ne vont pas masquer significativement ceux d'orientation proche de -45°. Il est donc possible que des interférences entre ces deux orientations se produisent. La figure 2.10 illustre ce phénomène. Cette figure montre que les sous-bandes diagonales (HH_j) de la transformée en ondelettes contiennent à la fois les orientations à 45° et les orientations à -45°. Par contre, les signaux issus de la transformée en ondelettes des contours de lignes horizontales et verticales se retrouvent bien dans les sous-bandes correspondant à leur orientation.

Une autre différence réside dans la forme des filtres utilisés pour la décomposition en ondelettes. Les sous-bandes verticales et horizontales « débordent » sur les sous-bandes diagonales. La transformée en ondelettes de contours diagonaux peut se retrouver à la fois dans les sous-bandes diagonales et dans les sous-bandes horizontales et verticales. Si le contraste d'un contour est suffisamment important, l'amplitude des coefficients ondelettes correspondant à ce contour dans les sous-bandes horizontales et verticales peut être suffisamment importante pour entraîner une surestimation de l'effet de masquage sur ce contour. De plus, l'énergie déplacée des sous-bandes diagonales vers les sous-bandes verticales et horizontales ne sera pas prise en compte dans le calcul des effets de masquage sur les structures diagonales. Ce phénomène est illustré par la figure 2.10. Sur cette figure, les signaux issus de la transformée en ondelettes des contours de lignes orientées à 45° et à -45° se retrouvent principalement dans les sous-bandes diagonales, mais aussi dans les sous-bandes horizontales et verticales. Par ailleurs, on peut observer que la position des lignes a aussi une influence sur le débordement des sous-bandes diagonales vers les sous-bandes horizontales et verticales, ainsi au premier niveau de décomposition,

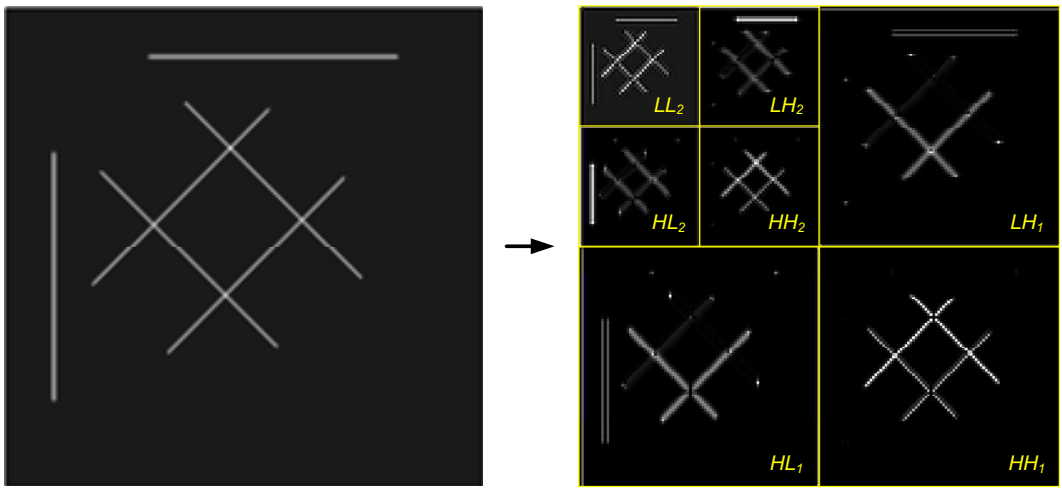


Figure 2.10 – Image source synthétique et sa transformée dyadique en ondelettes sur deux niveaux. Plus un coefficient est sombre, plus sa valeur absolue est faible.

les signaux issus de la transformée en ondelettes des deux lignes diagonales les plus basses ont une amplitude plus importante plus que celle des deux lignes diagonales les plus hautes.

Pour la luminance les gammes de fréquences spatiales radiales de la DCP sont respectivement de $0cy/d^\circ$ à $1.5cy/d^\circ$, de $1.5cy/d^\circ$ à $5.7cy/d^\circ$, de $5.7cy/d^\circ$ à $14.2cy/d^\circ$, et enfin de $14.2cy/d^\circ$ à $28.2cy/d^\circ$. Les gammes de fréquences spatiales horizontales et verticales de la DWT dépendent à la fois de la fréquence spatiale maximale f_{max} visible dans de l'image (i.e. des conditions d'observation), et à la fois du niveau de décomposition. Pour une fréquence f_{max} donnée, la gamme de fréquences d'une sous-bande ondelettes est $[2^{-l}f_{max}; 2^{-(l-1)}f_{max}]$, où l est le niveau de décomposition de la sous-bande ($l = 1$ correspond aux fréquences les plus hautes). Le nombre de niveau de décomposition a un impact sur la différence entre les gammes de fréquences des sous-bandes de la DWT et celles des sous-bandes de la DCP (cf. figure 2.11).

Afin d'augmenter au maximum la correspondance entre la DCP et notre décomposition basée sur la DWT, le nombre de niveau de décomposition de la DWT est déterminé en fonction de la fréquence spatiale maximale f_{max} visible dans l'image (i.e. des conditions d'observation). La fréquence d'échantillonnage spatial d'une image f_s exprimée en nombre de pixels par degré de champ visuel peut s'exprimer par :

$$f_s = 2\tan(0.5^\circ)rH \tag{2.14}$$

dans laquelle r représente la distance d'observation en nombre de fois la hauteur de l'image, et H représente la hauteur de l'image en nombre de pixels. Le nombre minimal de pixels nécessaire pour représenter une période du signal étant de deux, la fréquence f_{max} s'exprime ensuite en cycle par degré (cy/d°) :

$$f_{max} = \frac{f_s}{2} \tag{2.15}$$

Le nombre de niveaux de décomposition est ensuite déterminé en augmentant celui-ci jusqu'à faire coïncider

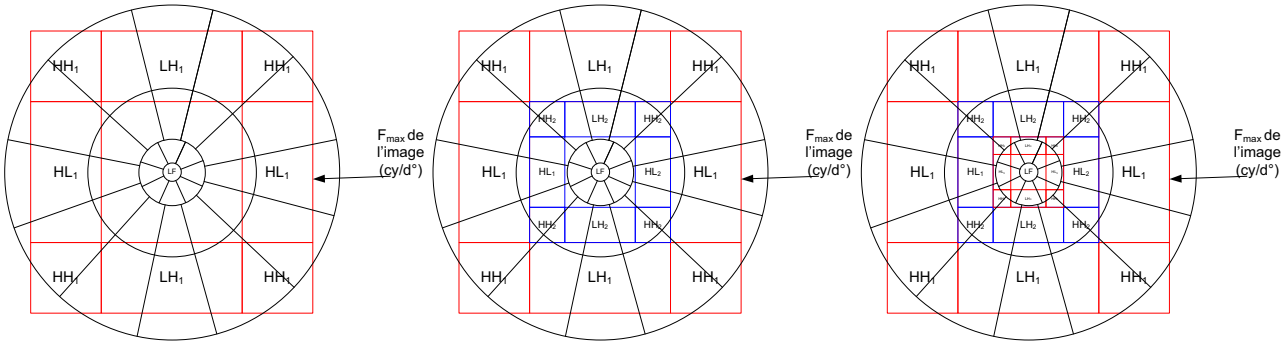


Figure 2.11 – Illustration de la correspondance entre la DCP et la DWT en fonction du niveau de décomposition de la DWT : (a) un niveau, (b) deux niveaux et (c) trois niveaux.

la basse fréquence de la DWT, avec la couronne basse fréquence $(0 - 1.5cy/d°)$ de la DCP comme illustré figure 2.12.

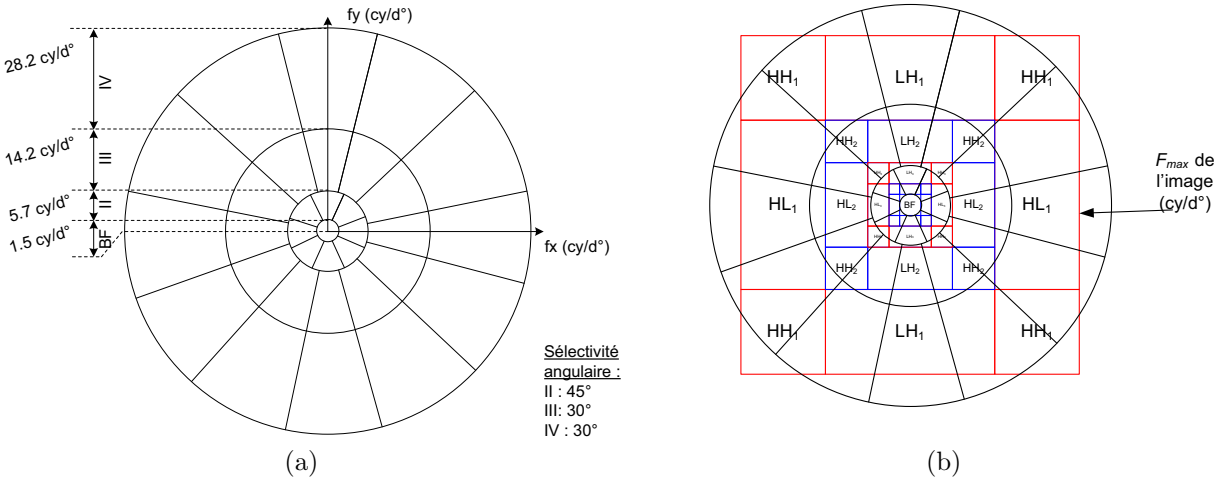


Figure 2.12 – (a) Décomposition en canaux perceptuels (DCP). (b) Décomposition en ondelettes dépendante des fréquences spatiales : mise en correspondance de la sous-bande basse fréquence de la DCP et de la DWT, ici le niveau de décomposition est égal à quatre.

La transformée en ondelettes que nous avons utilisée est la transformée biorthogonale 9/7 de Daubechies [Antonini 92], utilisée dans bon nombre d'applications liées au traitement d'images comme JPEG2000 [Christopoulos 00]. Les valeurs des coefficients des filtres de synthèse et de codage de cette transformée sont données tableaux 2.1 et 2.2.

i	Lowpass Filter hL(i)	Highpass Filter hH(i)
0	0.6029490182363579	1.115087052456994
±1	0.2668641184428723	-0.5912717631142470
±2	-0.07822326652898785	-0.05754352622849957
±3	-0.01686411844287495	0.09127176311424948
±4	0.02674875741080976	

Table 2.1 – Coefficients du filtre d'analyse de la transformée biorthogonale 9/7 de Daubechies.

i	Lowpass Filter gL(i)	Highpass Filter gH(i)
0	1.115087052456994	0.6029490182363579
±1	0.5912717631142470	-0.2668641184428723
±2	-0.05754352622849957	-0.07822326652898785
±3	-0.09127176311424948	0.01686411844287495
±4	0.02674875741080976	

Table 2.2 – Coefficients du filtre de synthèse de la transformée biorthogonale 9/7 de Daubechies.

2.6 Modélisation de la sensibilité aux contrastes

Comme nous l'avons vu précédemment, les fonctions de sensibilités au contraste permettent de décrire la sensibilité du système visuel humain en fonction de nombreux paramètres. Dans nos travaux nous avons utilisé la CSF anisotropique issue des travaux de Daly. Cette CSF permet de décrire la sensibilité pour la composante achromatique en fonction de deux paramètres : la fréquence spatiale radiale ω et l'orientation θ.

2.6.1 Modèle basé Fourier

Dans le modèle fondé sur la transformation de Fourier, la CSF est directement appliquée dans le plan fréquentiel après transformation de la composante A de l'image par une FFT 2D, et avant la mise en oeuvre de la décomposition en canaux perceptuels. Les coefficients de Fourier $c(\omega, \theta)$ de la composante achromatique sont normalisés par les valeurs de la CSF pour des conditions de visualisation fixées :

$$\tilde{c}(\omega, \theta) = c(\omega, \theta) \cdot S_A(\omega, \theta) \tag{2.16}$$

où $\tilde{c}(\omega, \theta)$ représente le coefficient (ω, θ) normalisée par la CSF, et $S_A(\omega, \theta)$ représente la valeur de la CSF pour la fréquence spatiale (ω, θ). Les conditions de visualisation étant fixées, les valeurs $S_A(\omega, \theta)$ de la CSF 2D de Daly sont calculées selon la relation suivante (détaillée dans la section 1.3.3.1) :

$$S_A(\omega, \theta, l, s, d, e) = P \times min\big[S(\frac{\omega}{bw_a, bw_e, bw_\theta}, l, s), S(\omega, l, s)\big], \tag{2.17}$$

Les paramètres l, s, d, e sont fixés par les conditions d'observation. La figure 2.13 illustre une représentation de cette CSF pour des conditions de visualisation données, sur laquelle on peut observer le comportement anisotrope de cette CSF.

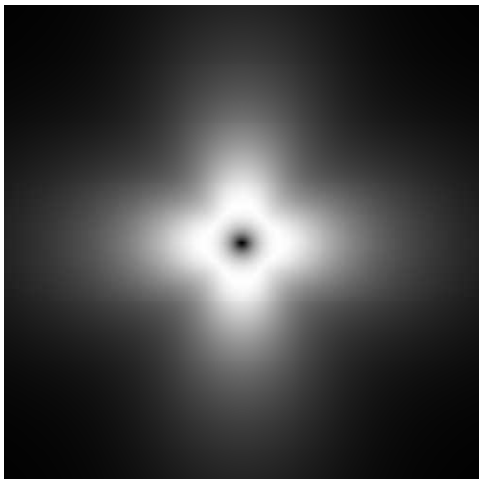


Figure 2.13 – Représentation de la CSF 2D de Daly en fonction de la fréquence spatiale ω et l'orientation θ, et pour des conditions d'observation données. Plus un coefficient est sombre, plus sa valeur la sensibilité est faible.

2.6.2 Modèle basé ondelettes

La transformée en ondelettes ne permet pas d'obtenir une représentation fréquentielle complète de l'image comme le permet la transformée de Fourier (cf. section 2.6.1). Dans la littérature on trouve plusieurs solutions pour appliquer la CSF. Une première solution pourrait consister à appliquer la CSF dans l'espace de Fourier. Pour cela, il faudrait effectuer une transformation directe, appliquer la CSF, puis effectuer une transformation inverse. Cette méthode serait sans doute la plus précise, mais elle augmenterait considérablement la complexité opératoire. Cette méthode n'est donc pas cohérente avec l'utilisation de la transformée en ondelettes dont l'objectif est de réduire justement cette complexité. Une autre solution serait de modifier les filtres de la transformée en ondelettes pour qu'ils prennent directement en compte la variation de la sensibilité due aux fréquences spatiales. Cette solution serait plus cohérente que la précédente, mais elle n'est pas envisageable à cause de la dépendance de la CSF aux conditions d'observation : il serait nécessaire de recalculer de nouveaux filtres en fonction des conditions d'observation. C'est pourquoi nous avons choisi d'utiliser une méthode de faible complexité et adaptée à la représentation fréquentielle de la transformée en ondelettes : la CSF est appliquée au moyen d'une valeur de CSF par sous-bande. L'application de la CSF consiste en une normalisation de coefficients ondelettes $c_{l,o}(m,n)$ en utilisant une valeur unique de CSF par sous-bande :

$$\tilde{c}_{l,o}(m,n) = c_{l,o}(m,n) \cdot N_{l,o}^{CSF} \tag{2.18}$$

où $\tilde{c}_{l,o}(m,n)$ représente la valeur du coefficient ondelette normalisé par la CSF au site (m,n) et dans la sous-bande de niveau l et d'orientation o. Pour chaque sous-bande (l,o), une valeur de CSF $N_{l,o}^{CSF}$ est calculée à partir de la CSF 2D de Daly déjà utilisée dans le modèle basé sur le domaine de Fourier. Ces valeurs sont calculées en moyennant les valeurs de la CSF 2D de Daly sur la gamme de fréquences spatiales de chaque sous-bande, comme

illustré figure 2.14. Les gammes de fréquences $f_{l,o}$ des sous-bandes sont données par les relations suivantes, et

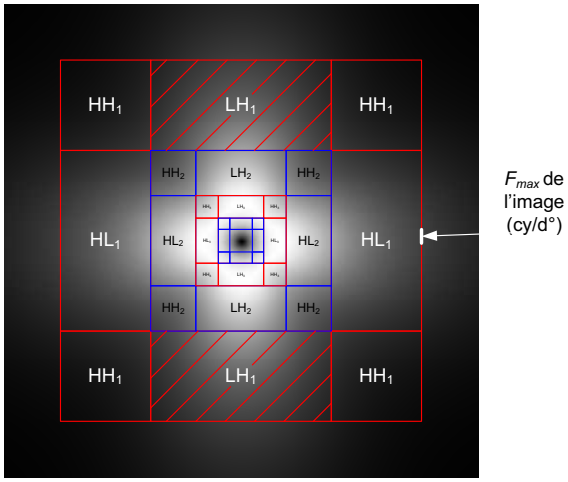


Figure 2.14 – Le calcul des valeurs $N_{l,o}^{CSF}$ de chaque sous-bande s'effectue en moyennant les valeurs de la CSF 2D de Daly sur la gamme de fréquences spatiales de chaque sous-bande. Pour la sous-bande LH_1 il s'agit des valeurs de la zone hachurée.

correspondent à l'intersection de la gamme de fréquences spatiales horizontales et de la gamme de fréquences spatiales verticales de la sous-bande considérée :

$$\begin{aligned} f_{l,LH} &= f_{L_l} \cap f_{H_l} \\ f_{l,HL} &= f_{H_l} \cap f_{L_l} \\ f_{l,HH} &= f_{H_l} \cap f_{H_l} \\ f_{l,BF} &= f_{L_l} \cap f_{L_l} \end{aligned} \tag{2.19}$$

où l représente le niveau de décomposition, LH, HL, HH représentent les différentes orientations, BF la sous-bande basses fréquences. Les gammes de fréquences horizontales ou verticales f_{L_l} et f_{L_l} sont définies par les relations suivantes :

$$\begin{aligned} f_{L_l} &= [0; 2^{-l} f_{max}] \\ f_{H_l} &= [2^{-l} f_{max}; 2^{-(l-1)} f_{max}] \end{aligned} \tag{2.20}$$

où l représente le niveau de décomposition ($l = 1$ correspond aux fréquences les plus hautes), et f_{max} représente la fréquence spatiale maximale possible de l'image dans les conditions d'observation choisies.

2.7 Modélisation de l'effet de masquage spatial

2.7.1 Masquage de contraste

Dans le modèle fondé sur la transformation de Fourier, la modélisation des effets de masquage repose sur le modèle de S. Daly [Daly 93]. Ce modèle ne prend en compte que les interactions de masquage intra-canal (pas d'effet de facilitation). Comme dans les travaux de Le Callet [Le Callet 01], les valeurs de l'élévation de seuil

$T_{\rho,\theta}(m,n)$ provoquée par le masquage sont calculées en chaque site (m,n) dans chaque sous-bande (ρ,θ) de la DCP, selon la relation suivante :

$$T_{\rho,\theta}(m,n) = (1 + (k_1 \cdot (k_2 \cdot |\tilde{c}_{\rho,\theta}(m,n)|)^s)^b)^{\frac{1}{b}} , \tag{2.21}$$

avec :

- $\tilde{c}_{\rho,\theta}(m,n)$ étant la valeur normalisée par la CSF du site (m,n) dans la sous-bande de la couronne ρ et d'orientation θ
- $k_1 = 0.0153$
- $k_2 = 392.5$
- l, b étant des constantes dépendant de la sous-bande (ρ,θ).

Dans le modèle fondé sur la transformation en ondelettes, le masquage est appliqué de la même façon sur les coefficients ondelettes normalisés par la CSF $\tilde{c}_{l,o}(m,n)$. Dans ce modèle, le contraste est décrit par la valeur absolue des coefficients ondelettes. Comme nous l'avons discuté dans la section 2.5.2.2 la correspondance entre la DCP et la DWT n'est pas complète. L'utilisation de la même fonction de masquage est donc une approximation des fonctions de masquage validées dans le cas de la DCP. Il est bien entendu que la modélisation des effets de masquage dans le cas du modèle basé ondelettes ne sera pas aussi fine que dans le cas de la DCP, Cependant il est intéressant d'évaluer son efficacité dans notre contexte. L'élévation de seuil $T_{l,o}(m,n)$ pour le site (m,n) de la sous-bandes ondelettes (l,o) est donnée par la relation :

$$T_{l,o}(m,n) = (1 + (k_1 \cdot (k_2 \cdot |\tilde{c}_{l,o}(m,n)|)^s)^b)^{\frac{1}{b}} , \tag{2.22}$$

où les paramètres k_1 et k_2 ont les mêmes valeurs que dans la relation (2.21). La dépendance des paramètres l, b à la sous-bande considérée, est maintenue en établissement une correspondance entre les sous-bande de la DCP et les sous-bandes ondelettes. Pour une sous-bande ondelette (l,o), les paramètres l et b sélectionnés seront ceux de la sous-bande de la DCP correspondant à la gamme de fréquences et d'orientation la plus proche.

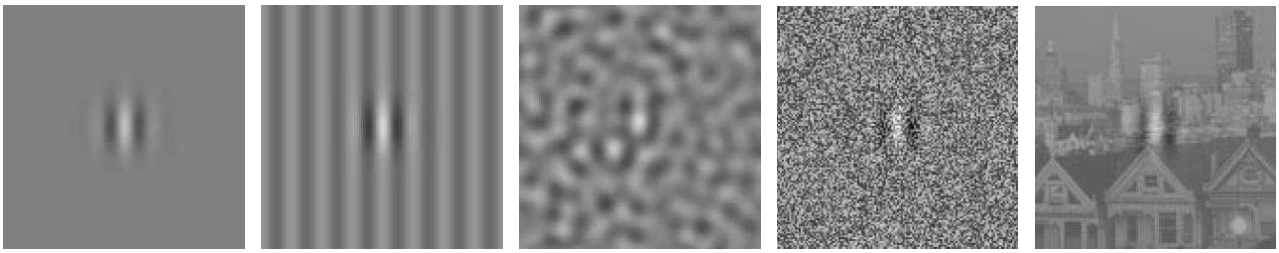
2.7.2 Masquage semi-local

Comme nous l'avons introduit en section 1.3.5.1, la masquage de contraste ne permet pas toujours d'expliquer l'importance de l'élévation du seuil de visibilité. C'est Watson [Watson 97b] qui a introduit le terme de masquage entropique, après avoir réalisé des mesures d'élévation du seuil de visibilité avec des masques de différentes natures :

- masque cosinusoïdal,
- bruit isotrope filtré passe bande,
- bruit blanc de distribution uniforme,
- image naturelle.

Le contraste des différents masques étant choisi de façon à ce que leur énergie de contraste soit équivalente, il a remarqué que plus le masque avait un caractère aléatoire, ou imprédictible, et plus l'élévation de seuil était

importante. La figure 2.15 présente des exemples de masques utilisés par Watson. Il met l'accent sur la notion de capacité d'apprentissage (*learnability*) du masque et de la cible dans les phénomènes de masquage. Il est plus facile de détecter une variation lorsque le masque est simple et facile à se représenter, que lorsque le masque est complexe et difficile à apprendre. L'impact de l'apprentissage sur la mesure de seuil de visibilité avait été noté par Swift et Smith dans [Swift 83], où ils montrèrent que l'apprentissage modifie la pente de la courbe reliant le seuil de détection au contraste du signal masquant. Ils trouvèrent d'ailleurs que cette courbe varie en fonction de l'apprentissage, et qu'elle a une asymptote dont la valeur de la pente se situe autour de 0.65. Dans ces travaux sur le VDP (Visible Difference Predictor) [Daly 93], Daly utilise ce résultat en modifiant la valeur d'un exposant de sa fonction de masquage. Il s'agit du paramètre s dans les relations (2.21) et (2.22). Cet exposant varie de 1 pour un aucun apprentissage, jusqu'à 0.65 pour des masques très bien connus par les observateurs. Différentes valeurs de ce paramètre sont fixées pour les différentes sous-bandes fréquentielles. Cette solution ne permet pas de prendre en compte toute la spécificité des images influençant à la fois la capacité d'apprentissage, ou la connaissance que les observateurs peuvent en avoir.



(a) (b) (c) (d) (e)

Figure 2.15 – Cible de Gabor (a) ajoutée à quatre signaux masquants (b-e).

Dans notre modèle nous proposons d'étendre le modèle de masquage de Daly afin qu'il prenne en compte la spécificité de l'image autour du site pour lequel l'élévation de seuil est calculée. Comme le fait remarquer Watson dans [Watson 97b], le calcul de l'entropie est un bon point de départ pour évaluer la spécificité du voisinage du site en question. Nous appelons ce voisinage une *semi-localité*, la localité étant le site en lui-même. Le masquage prenant en compte cette semi-localité dans le calcul de l'élévation du seuil de visibilité sera donc qualifié de masquage semi-local dans la suite. Le calcul de l'entropie dans cette semi-localité, nous donne une mesure de la quantité d'information contenue dans cette zone. Plus cette zone contient d'information (i.e. valeur d'entropie importante), plus la capacité à l'apprendre sera faible entraînant par le fait un effet de masquage important.

Les fonctions de masquage utilisées dans les deux modèles proposés et décrites par les relations (2.21) et (2.22) deviennent les deux fonctions décrites par les relations (2.23) et (2.24) respectivement pour le modèle basé DCP et pour le modèle basé DWT :

$$T_{\rho,\theta}(m,n) = (1 + (k_1 \cdot (k_2 \cdot |\tilde{c}_{\rho,\theta}(m,n)|)^{s(m,n)})^b)^{\frac{1}{b}} , \tag{2.23}$$

$$T_{l,o}(m,n) = (1 + (k_1 \cdot (k_2 \cdot |\tilde{c}_{l,o}(m,n)|)^{s(m,n)})^b)^{\frac{1}{b}} \tag{2.24}$$

où les paramètres sont les mêmes que pour la relation (2.21), excepté pour le paramètre $s(m,n)$ qui remplace le paramètre s et permet de prendre en compte la spécificité de la semi-localité autour du site (m,n). Ce paramètre varie entre 1 et 0.65 en fonction de la *complexité* de la semi-localité, et selon la relation :

$$s(m,n) = S + \Delta s(m,n), \tag{2.25}$$

où S est une constante dépendant de la sous-bande considérée, et $\Delta s(m,n)$ est une mesure de la complexité semi-locale. La spécificité semi-locale modifie la pente de la fonction de masquage comme l'illustre la figure 2.16. Plus la complexité est importante, et plus la pente l'est aussi. La valeur de $\Delta s(m,n)$ est déterminée par une fonction sigmoïde (de mise à l'échelle) appliquée à l'entropie. L'élévation de seuil étant calculée pour la composante A de l'image, l'entropie l'est aussi.

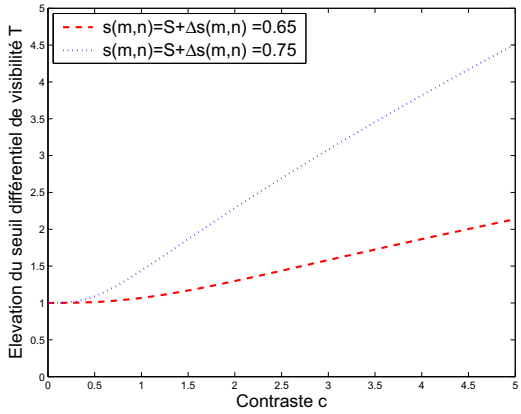


Figure 2.16 – Modification de la pente de la fonction de masquage due à la complexité semi-locale $\Delta s(m,n)$.

L'entropie de la semi-localité est calculé sur un voisinage de taille $N \times N$ autour du site (m,n) selon la relation :

$$E(m,n) = -\sum p(k) \cdot log(p(k)), \tag{2.26}$$

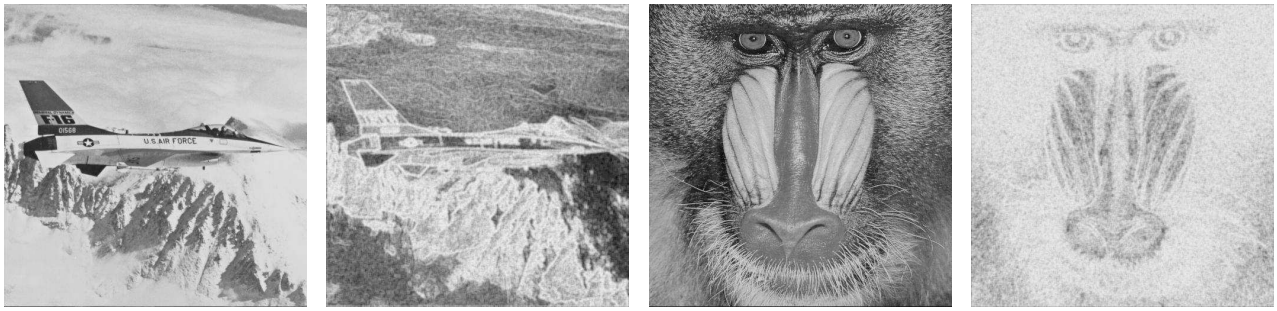
où $p(k)$ est la probabilité calculée à partir de l'histogramme des valeurs de luminance dans le voisinage $V(m,n)$ considéré, et $E(m,n)$ est la carte des valeurs d'entropie obtenue. Deux exemples de cartes d'entropie sont données figure 2.17.

On peut observer que les zones complexes comme les montagnes de l'image *Avion*, ou le pelage sur l'image *Mandrill*, sont bien détectées par la mesure d'entropie. Ensuite, les valeurs d'entropie sont projetées, à l'aide d'une fonction sigmoïde, sur l'intervalle $[0; 1-S]$ afin d'obtenir les valeurs $\Delta s(m,n)$:

$$\Delta s(m,n) = \frac{b1}{1 + e^{-b2 \cdot (E(m,n) - b3)}}, \tag{2.27}$$

où les paramètres $b1$, $b2$, $b3$ sont déduits empiriquement (expérimentations sur différentes textures).

Une fois les élévations de seuil différentiel de visibilité déterminées dans chaque sous-bande, les erreurs entre l'image source et l'image à évaluer doivent être calculées et normalisées par les élévations de seuil au sein de chaque sous-bande, puis cumulées en une carte unique de distorsions.



(a) (b) (c) (d)

Figure 2.17 – Deux exemples (b), (d) de cartes d'entropie $E(m,n)$ pour les images *Avion* (a) et *Mandrill* (c) respectivement. Plus la valeur est sombre, plus l'entropie est faible.

2.8 Normalisation et cumul inter sous-bandes des erreurs

Dans le contexte de la création de cartes de distorsions perceptuelles, le calcul des seuils différentiels de visibilité a pour objectif la détermination de la visibilité des erreurs entre une image originale et une image dégradée. Dans les deux modèles proposés les distorsions sont calculées, après normalisation par la CSF, au moyen d'une différence entre les sous-bandes de l'image originale et de l'image dégradée. Ces valeurs de différence sont ensuite normalisées par les valeurs d'élévation de seuil calculées grâce aux fonctions de masquage, selon les relations suivantes :

$$VE_{\rho,\theta}(m,n) = \frac{|\tilde{c}^O_{\rho,\theta}(m,n) - \tilde{c}^D_{\rho,\theta}(m,n)|}{max(T^O_{\rho,\theta}(m,n), T^D_{\rho,\theta}(m,n))}\,. \tag{2.28}$$

$$VE_{l,o}(m,n) = \frac{|\tilde{c}^O_{l,o}(m,n) - \tilde{c}^D_{l,o}(m,n)|}{max(T^O_{l,o}(m,n), T^D_{l,o}(m,n))}\,. \tag{2.29}$$

où $VE_{\rho,\theta}(m,n)$ et $VE_{l,o}(m,n)$ représentent respectivement les cartes d'erreurs perceptuelles par sous-bande pour le modèle basé Fourier et pour le modèle basé ondelettes. Les exposants O et D représentent respectivement l'image originale et l'image dégradée.

Les cartes d'erreurs perceptuelles de chaque sous-bande sont ensuite cumulées suivant les orientations, puis suivant les fréquences radiales afin d'obtenir la carte d'erreurs perceptuelles finale $VE(m,n)$. Différentes stratégies sont possibles pour réaliser ce cumul. Dans les travaux de sa thèse N. Bekkat [Bekkat 99] a montré que l'ordre entre le cumul fréquentiel angulaire et radial ne constituait pas un élément sensible, et comme dans les travaux de thèse de P. Le Callet [Le Callet 01], nous choisissons d'effectuer le cumul angulaire avant le cumul radial parce que cela est plus simple à mettre en oeuvre. Dans notre étude nous avons choisi une approche simple et souvent utilisée dans la littérature (cf. section 2.2.3) : la sommation de Minkowski.

Pour le modèle basé Fourier, le cumul est réalisé par :

$$VE_{\rho}(m,n) = \left(\frac{1}{M_\theta} \sum_{\theta=1}^{M_\theta} (VE_{\rho,\theta}(m,n))^{\beta_\theta} \right)^{\frac{1}{\beta_\theta}}\,, \tag{2.30}$$

puis par :

$$VE(m,n) = \left(\frac{1}{N_\rho} \sum_{\rho=1}^{N_\rho} (VE_\rho(m,n))^{\beta_\rho} \right)^{\frac{1}{\beta_\rho}} . \tag{2.31}$$

avec :

- $VE_\rho(m,n)$: résultat du cumul des orientations,
- M_θ : le nombre d'orientations θ pour la couronne ρ,
- N_ρ : le nombre de sous-bandes radiales ρ,
- β_θ et β_ρ : les exposants de Minkowski respectivement pour les orientations et les fréquences radiales.

Pour le modèle basé ondelettes, le cumul est réalisé par :

$$VE_l(m,n) = \left(\frac{1}{3} \sum_{d} (VE_{l,o}(m,n))^{\beta_o} \right)^{\frac{1}{\beta_o}}, \forall o \in [LH, HL, HH] . \tag{2.32}$$

puis par :

$$VE(m,n) = \left(\frac{1}{L} \sum_{l=0}^{L} (VE_l(m,n))^{\beta_l} \right)^{\frac{1}{\beta_l}} . \tag{2.33}$$

avec :

- $VE_l(m,n)$: le résultat du cumul des orientations,
- L : le nombre de niveau de la décomposition en ondelettes,
- β_o et β_l : les exposants de Minkowski respectivement pour les orientations o et les niveaux l.

2.9 Résultats qualitatifs

Un certain nombre de mécanismes du système visuel humain étant modélisés, il est maintenant possible de calculer une carte d'erreurs perceptuelles représentant les distorsions visibles entre une image originale et une image dégradée. Dans cette section nous allons évaluer qualitativement les cartes d'erreurs ainsi créées, puis nous proposerons des perspectives d'amélioration.

2.9.1 Comparaison de cartes de distorsions perceptuelles

La figure 2.18 présente les cartes de distorsions obtenues selon différentes méthodes pour une version de l'image *Avion* compressée avec JPEG2000 :

- Notre modèle basé Fourier (appelé FQA) ;
- Notre modèle basé ondelettes (appelé WQA) ;
- Erreur quadratique ;
- SSIM [Wang 04a].

L'image compressée avec JPEG2000 présente des distorsions particulièrement visibles dans le ciel (au-dessus de l'avion) et peu visible dans les montagnes (sous l'avion). On observe que la répartition relative des erreurs perceptuelles entre les différentes zones de l'image est assez proche concernant les cartes FQA et WQA. Par contre les cartes d'erreurs quadratiques et de SSIM présentent une répartition assez différente des erreurs. Sur

cet exemple les cartes FQA et WQA sont plus cohérentes que les deux autres, puisqu'elles donnent des erreurs plus visibles dans le ciel que dans les montagnes. En effet, les cartes d'erreurs quadratiques et de SSIM indiquent que les distorsions sont plus importantes dans les montagnes que dans le ciel.

La validation de cartes de distorsions est un problème complexe et peu abordé dans la littérature. Le problème réside dans le fait qu'on ne dispose pas de vérité terrain pour évaluer quantitativement les cartes de distorsions. La constitution d'une vérité demanderait la réalisation de tests subjectifs permettant d'obtenir une évaluation locale des distorsions. A notre connaissance, de tels tests n'ont jamais été menés, il n'existe donc pas de méthodologie permettant d'obtenir cette vérité. Cependant, la définition d'une telle méthodologie serait un sujet de recherche intéressant avec des retombées scientifiques importantes. En évaluation de qualité, la validation quantitative des modèles se limite à éprouver leur capacité de prédiction de la qualité en terme de note globale sur l'image ou la vidéo. Nous détaillons d'ailleurs ces techniques dans la section 4.2.

Afin de confirmer les tendances évoquer précédemment sur l'image *Avion* et de valider qualitativement nos cartes de distorsions perceptuelles, nous avons mené une expérimentation avec quelques observateurs.

(a) (c) (e)

(b) (d) (f)

Figure 2.18 – (a) est l'image *Avion*, (b) est une version de l'image *Avion* compressée avec JPEG2000, (c) et (d) sont les cartes de distorsions issues de notre modèle basé Fourier et de notre modèle basé ondelettes respectivement, (e) et (f) sont les cartes de distorsions issues de l'erreur quadratique et de la SSIM [Wang 04a] respectivement.

2.9.1.1 Tests de préférences des cartes d'erreurs

Dans ce test, nous avons présenté à des experts en compression vidéo une image de référence et une version dégradée de cette image, en leur demandant entre plusieurs types de cartes de distorsions de choisir laquelle

correspondait le mieux à leur perception des distorsions entre les deux images. Il s'agit d'un test de préférence d'une carte de distorsions parmi trois. Les trois cartes de distorsions proposées étaient :

- Notre modèle basé Fourier ;
- Erreur quadratique ;
- SSIM [Wang 04a].

L'image originale et l'image dégradée étaient présentées côte à côte. Les cartes d'erreurs étaient visualisées sous forme de cartes de chaleur, les teintes bleues correspondant aux zones sans ou avec des dégradations très peu visibles, et les teintes rouges-oranges correspondant aux zones avec des dégradations très visibles. Une illustration de l'interface de test est présentée figure 2.19.

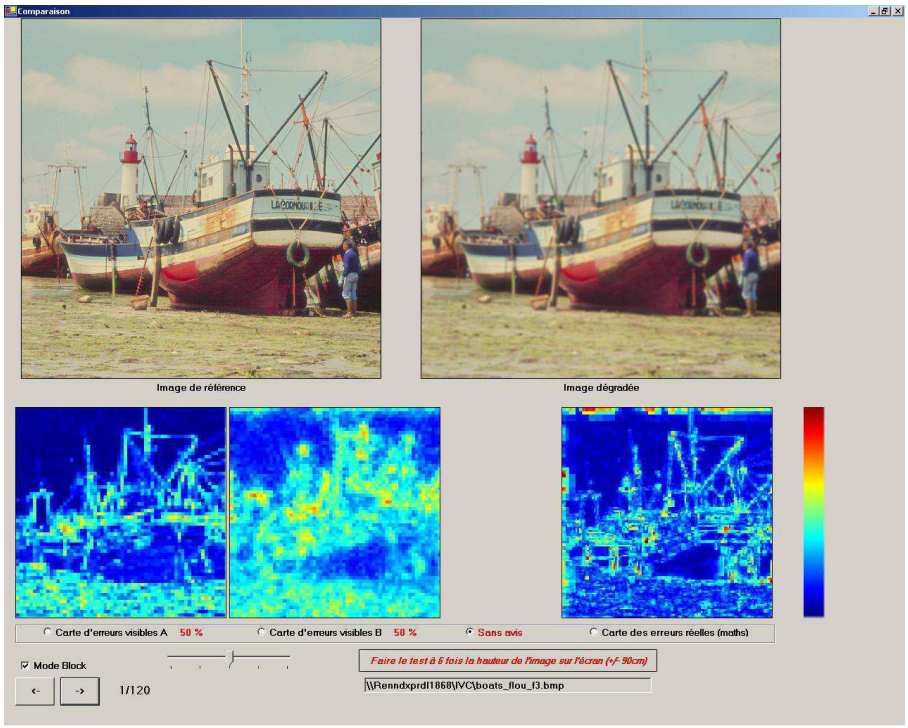


Figure 2.19 – Illustration de l'interface du test de préférence entre les cartes FQA, SSIM et d'erreurs quadratiques.

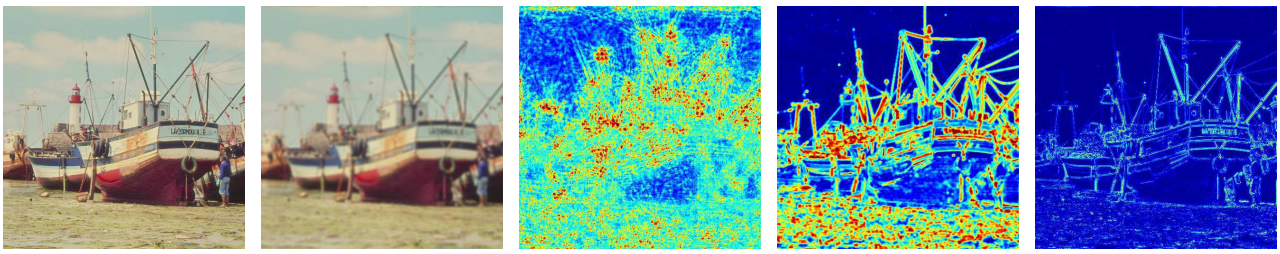
La base de test comprenait 120 images de différents contenus ayant été dégradées de différentes façons (JPEG, JPEG2000, flou). Cette base d'images est décrite en détail dans la section 5.3.1.1. La consigne donnée aux experts était d'effectuer le test à une distance d'observation égale à six fois la hauteur des images affichées. Cependant la distance d'observation n'était pas contrôlée pendant celui-ci. Les images étaient présentées dans un ordre aléatoire pour chaque observateur, mais ils pouvaient revenir sur leur choix précédent jusqu'à la fin du test, la durée de celui-ci n'étant pas limitée. Seule la carte d'erreurs quadratiques était identifiée en tant que telle. Afin de ne pas influencer les observateurs dans leur choix, la méthode ayant permis la création des autres cartes était cachée. Les experts ayant passé le test étaient au nombre de cinq.

Types de dégradations	Cartes de SSIM	Cartes de FQA	Pas de préférence	Cartes d'erreurs quadratiques
Toutes	20.67%	60.00%	17.17%	2.17%
JPEG	12.40%	72.40%	14.40%	0.80%
JPEG2000	21.60%	52.80%	12.00%	3.60%
Flou	31.20%	37.60%	29.60%	1.60%

Table 2.3 – Résultats du test de préférence. Pour chaque choix possible le pourcentage des réponses est donné.

Les résultats de ce test de préférence sont présentés dans le tableau 2.3. On observe que les cartes d'erreurs quadratiques sont rarement choisies par les experts, qui préfèrent les cartes FQA dans 60.00% des cas, et les cartes SSIM dans 20.67% des cas. Ce test montre qualitativement que les cartes données par le modèle FQA sont une meilleure représentation des distorsions perceptuelles que les cartes SSIM, a fortiori que les simples cartes d'erreurs quadratiques. Les cartes WQA et FQA étant très proches, les cartes WQA sont, par extension, une bonne représentation des distorsions perceptuelles. Le score très faible des cartes d'erreurs quadratiques renforce encore l'idée qu'elles ne sont pas une bonne représentation des distorsions visuelles, alors qu'elles sont, malgré tout, encore beaucoup utilisées sous différentes formes en compression vidéo. On peut se poser la question de l'existence d'un biais étant donné que les observateurs savaient quelle était la carte d'erreurs quadratiques. Cependant, les observateurs étant des experts en vidéo, on peut imaginer qu'ils ne se sont pas laissés influencer par cette information.

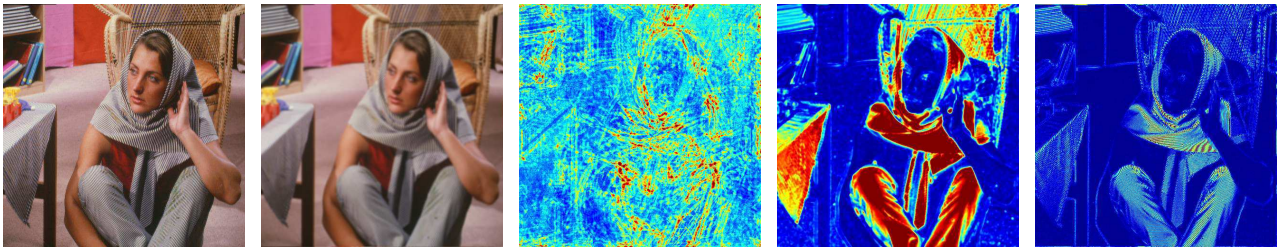
Par ailleurs, les résultats montrent que la répartition des choix entre les différentes cartes de distorsions n'est pas la même suivant le type de dégradations. Il en ressort que les cartes FQA ont toujours été préférées quel que soit le type de dégradations, même si pour le flou la préférence est moins forte. Les figures 2.20 et 2.21 montrent des exemples les cartes d'erreurs présentées pendant le test pour deux images dégradées par du flou.



(a) (b) (c) (d) (e)

Figure 2.20 – Image *Boats* originale (a) et cartes de distorsions pour une version dégradée par du flou (b) : carte d'erreurs FQA (c), carte d'erreurs SSIM (d) et carte d'erreurs quadratiques (e). Le rouge correspond à des erreurs très importantes et le bleu à des erreurs très faibles.

Les cas typiques où les cartes SSIM ont été choisies, sont les images contenant des zones fortement structurées qui ont été dégradées par du flou, entraînant une modification structurelle particulièrement gênante de ces zones. Une image caractéristique est l'image *Barbara* (cf. figure 2.21) qui contient une nappe, un pantalon et un châle comportant des rayures blanches et noires. De plus, on remarque que ce sont pour les dégradations de type flou que les experts ont le plus de mal à exprimer une préférence entre les cartes de distorsions (dans 29.6% des



(a) (b) (c) (d) (e)

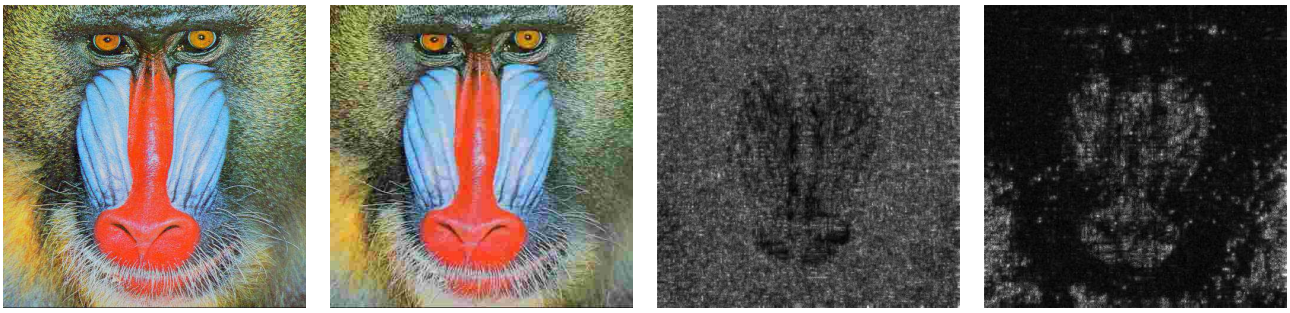
Figure 2.21 – Image *Barbara* originale (a) et cartes de distorsions pour une version dégradée par du flou (b) : carte FQA (c), carte SSIM (d) et carte d'erreurs quadratiques (e). Le rouge correspond à des erreurs très importantes et le bleu à des erreurs très faibles.

cas). Ces observations nous laissent penser que cet aspect pourrait être une piste d'amélioration de nos cartes d'erreurs perceptuelles.

Les deux modèles proposés permettent donc de construire des cartes d'erreurs perceptuelles pertinentes qualitativement. Comme nous l'avons décrit précédemment, ces méthodes modélisent le masquage semi-local contrairement à ce qui est généralement fait dans la littérature. Dans la section suivante, nous évaluons qualitativement l'intérêt de modéliser cet effet de masquage au travers de son impact sur les cartes de distorsions perceptuelles.

2.9.1.2 Impact du masquage semi-local sur les cartes de distorsions perceptuelles

Les images (a) et (b) de la figure 2.22 représentent l'image *Mandrill* originale et une version compressée avec JPEG respectivement. La différence entre les cartes de distorsions perceptuelles avec (cf. Figure 2.22(d)) et sans (cf. Figure 2.22(c)) masquage semi-local est significative. L'effet de masquage sur les zones les plus difficiles à « apprendre » (de complexité locale importante), comme le pelage et les moustaches, est sous-estimé sans le masquage semi-local, mais est plus proche de la réalité avec masquage semi-local.



(a) (b) (c) (d)

Figure 2.22 – (a) est l'image *Mandrill*, (b) est une version de l'image *Mandrill* compressée avec JPEG, (c) et (d) sont les cartes d'erreurs perceptuelles issues du modèle WQA respectivement sans et avec masquage semi-local.

Les images (a) et (b) de la figure 2.23 représentent l'image *Avion* originale et une version compressée avec JPEG2000 respectivement. La différence entre les cartes d'erreurs perceptuelles avec (cf. Figure 2.23(d)) et sans

(cf. Figure 2.23(c)) masquage semi-local est aussi significative. L'effet de masquage sur les zones de complexité locale importante, comme les montagnes, est sous-estimé sans le masquage semi-local, mais est plus cohérent avec le masquage semi-local. Par contre, l'effet de masquage a tendance à être surestimé sur des zones spécifiques comme les contours à fort contraste. Par exemple, on peut observer une surestimation du masquage sur les contours de la queue de l'avion. Ceci est dû à la mesure de l'activité semi-locale. En effet, le calcul d'entropie est un bon estimateur de l'activité semi-locale, mais il a tendance à surestimer l'activité dans le voisinage des contours fortement contrastés, dont le potentiel de masquage est déjà correctement estimé par le masquage de contraste. Nous proposons dans la section suivante une piste d'amélioration à ce problème.

(a) (b) (c) (d)

Figure 2.23 – (a) est l'image *Avion*, (b) est une version de l'image *Avion* compressée avec JPEG2000, (c) et (d) sont les cartes d'erreurs perceptuelles issues du modèle WQA respectivement sans et avec masquage semi-local.

2.9.2 Perspectives d'amélioration des cartes de distorsions perceptuelles

2.9.2.1 Apport d'une branche structurelle

Les résultats des tests de préférence, ainsi que nos observations et celles des experts nous ont fait remarquer que les cartes d'erreurs perceptuelles, pourraient être enrichies par une meilleure prise en compte des erreurs structurelles (cf. section 2.2.2). Il s'agit typiquement des apparitions ou des pertes de contours entre l'image originale et l'image dégradée.

Afin de prendre en compte ces erreurs structurelles nous nous inspirons d'une solution proposée par Le Callet dans [Le Callet 01]. Cependant, nous effectuons une détection de contours (Sobel), non pas sur les images en entrée du modèle, mais dans l'espace perceptuel, c'est-à-dire sur une représentation visuelle des images de référence et dégradée. Pour le modèle WQA, cette représentation visuelle est calculée en normalisant dans chaque sous-bande les coefficients ondelettes par la CSF et par les élévations de seuil dues au masquage, puis en effectuant une transformée en ondelettes inverse. L'image ainsi reconstruite correspond à la représentation visuelle.

A partir des contours détectés sur les représentations visuelles des images de référence et dégradée, nous répartissons les distorsions selon quatre classes C_1, C_2, C_3 et C_4 :

– C_1 : Erreurs correspondant à une perte de contours existant dans l'image originale,

- C_2 : Erreurs correspondant à l'apparition de nouveaux contours dans l'image dégradée (effet de bloc par exemple),
- C_3 : Erreurs localisées sur une zone de contours soit dans l'image originale, soit dans l'image dégradée (cela correspond à l'union de C_1 et de C_2),
- C_4 : Les autres erreurs.

Pour chaque classe C_i, on peut construire une carte $VE_{C_i}(m,n)$ dont la valeur au site (m,n) est égale à la valeur de l'erreur $VE(m,n)$ en ce site si sa classe est C_i, et nulle sinon. Une nouvelle carte d'erreurs perceptuelles $E_S(m,n)$ peut être calculée en combinant linéairement les quatre cartes $E_{C_i}(m,n)$:

$$VE_S(m,n) = \sum_{i=1}^{4} \alpha_i . VE_{C_i}(m,n) \tag{2.34}$$

où les coefficients α_i permettent de modifier le poids donné à chaque classe. Par exemple, en choisissant des valeurs de α_1 et α_2 supérieures aux valeurs de α_3 et α_4, les classes C_1 et C_2 sont favorisées. On donne ainsi plus d'importance aux erreurs structurelles car les classes C_1 et C_2 correspondent aux apparitions et aux pertes de contours.

La figure 2.24 présente des cartes d'erreurs perceptuelles issues du modèle WQA avec et sans la branche structurelle.

(a) (b) (c) (d)

Figure 2.24 – (a) est l'image *Avion*, (b) est une version de l'image *Avion* compressée avec JPEG, (c) et (d) sont les cartes d'erreurs perceptuelles issues du modèle WQA respectivement sans et avec la branche structurelle.

On observe que les erreurs perceptuelles dues aux frontières des effets de blocs (dans le ciel et dans la neige) sont mieux représentées lorsque la branche structurelle est utilisée. La branche structurelle proposée n'a pas vocation à répondre complètement au problème, mais ouvre des perspectives d'amélioration des cartes d'erreurs du modèle WQA. Il apparaît qu'un modèle reposant sur une évaluation de la visibilité des erreurs peut être avantageusement complété par des considérations sur les erreurs structurelles.

2.9.2.2 Adaptation de la mesure d'entropie comme estimateur de la complexité semi-locale

Nous avons observé dans la section 2.9.1.2, que la mesure d'entropie utilisée pour estimer le masquage semi-local pouvait conduire à surestimer le masquage sur les contours à fort contraste. Cette surestimation est due

au calcul de l'entropie dont les valeurs sont importantes sur les contours contrastés. Or, en terme de masquage semi-local, un contour même fortement contrasté a une capacité de masquage faible en comparaison d'une zone plus complexe.

Pour remédier à cette surestimation, nous proposons une correction de la mesure d'entropie. Cette correction consiste à raffiner la carte d'entropie en supprimant les forts contours, comme illustré figure 2.25.

(a) (b) (c)

Figure 2.25 – (a) est la carte d'entropie de l'image *Avion*,(b) est la carte de détection des contours de l'image *Avion*, et (c) la carte d'entropie corrigée de l'image *Avion*.

Cette détection de contours peut être réalisée au moyen d'un filtrage de Sobel par exemple et d'un seuillage adapté. La nouvelle carte d'entropie $E'(m,n)$ est donnée par l'équation :

$$E'(m,n) = \begin{cases} E'(m,n) & \text{si} \quad Sobel(m,n) > s_{sobel} \\ 0 & \text{sinon} \end{cases} , \tag{2.35}$$

où $Sobel(m,n)$ est la carte issue d'un filtrage de Sobel et s_{sobel} un seuil déterminé empiriquement sur un certain nombre d'images.

L'entropie n'est sans doute pas la mesure parfaite pour modéliser convenablement le masquage semi-local, cependant elle est une première approximation nous permettant dans cette étude d'en illustrer l'intérêt. Il existe donc sur ce point des perspectives d'amélioration de la modélisation du masquage semi-local. Un axe de recherche intéressant serait de définir une « mesure semi-locale » permettant de modéliser plus finement le masquage semi-local, et cela en s'appuyant sur des expérimentations psychovisuelles, comme par exemple des mesures de seuil différentiel de visibilité en présence de signaux masquants correspondant à de nombreuses textures différentes.

2.10 Conclusion

Ce chapitre était consacré à la conception de cartes de distorsions visuelles d'images. Après avoir présenté les différentes approches de la littérature, nous avons proposé essentiellement deux approches s'appuyant sur une modélisation du système visuel humain et permettant la construction de carte d'erreurs perceptuelles.

Nous nous sommes intéressés à deux aspects particulièrement important de la modélisation du système visuel humain, à savoir la modélisation du comportement multi-canal du système visuel humain, et la modélisation

des effets de masquages. La modélisation du comportement multi-canal est un point primordial, mais son implantation peut s'avérer de coût opératoire important. Dans ce cadre, nous avons montré que le passage d'une modélisation réalisée dans le domaine de Fourier à une modélisation réalisée dans le domaine des ondelettes nous permettait d'obtenir des cartes d'erreurs perceptuelles pertinentes et qualitativement proches, et cela en réduisant la complexité de calcul. La modélisation des effets de masquage est un autre point crucial que nous avons abordé. Nous avons montré l'importance de prendre en compte le masquage semi-local, en plus du masquage de contraste. Une validation expérimentale (tests de préférence) nous a permis de montrer que les cartes perceptuelles issues des modèles proposés étaient plus pertinentes que les cartes de SSIM ou que celles d'erreurs quadratiques. Par ailleurs, des perspectives d'amélioration des cartes d'erreurs perceptuelles ont aussi été proposées.

Chapitre 3

Conception de séquences de distorsions visuelles de vidéos

3.1 Introduction

L'objet de ce chapitre est la conception de séquences de cartes de distorsions visuelles de vidéos. Il s'agit de construire une séquence temporelle de cartes d'erreurs perceptuelles représentant les erreurs perçues par des observateurs humains entre une vidéo dite *originale* et une version dite *dégradée* de cette même vidéo. Pour atteindre cet objectif, il est nécessaire de modéliser la perception visuelle humaine, aussi bien d'un point de vue spatial, que d'un point de vue temporel. Une façon simple de traiter ce problème serait de construire la séquence des cartes des distorsions purement spatiales image par image entre la vidéo originale et la vidéo dégradée. Cependant, cette approche simpliste ne prendrait pas en compte correctement les aspects temporels. Les distorsions temporelles telles que le papillotement (*flickering*) ou l'effet « mosquito » jouent un rôle fondamental dans l'évaluation locale des distorsions d'une vidéo. Une distorsion temporelle est généralement définie comme une évolution temporelle, ou une fluctuation des distorsions spatiales d'une zone particulière. La perception à un instant donné des distorsions spatiales peut être en grande partie modifiée par leur évolution temporelle, comme par exemple une augmentation ou une diminution des distorsions, ou encore comme des changements périodiques des distorsions.

Dans ce chapitre, nous proposons une méthode de conception de séquences temporelles de distorsions visuelles qui repose sur l'étude des évolutions temporelles des distorsions spatiales. La perception temporelle de l'information visuelle, dont les distorsions temporelles font partie, étant étroitement liée aux mécanismes de l'attention visuelle, nous avons choisi d'évaluer d'abord les distorsions temporelles au niveau des fixations oculaires et des mouvements de poursuite. Pour le système visuel humain, la dimension temporelle est une contrainte beaucoup plus importante lors de l'exploration d'une vidéo que lors de l'exploration d'une image fixe. En évaluation de qualité, une image fixe peut être considérée comme un cas particulier de vidéo sans mouvement de caméra, sans objet en mouvement et dont les distorsions sont purement spatiales. L'image fixe n'évoluant par temporellement, il est possible d'explorer presque toute l'image, suivant sa durée de présentation. Par contre, la

dimension temporelle intrinsèque de la vidéo empêche l'observateur de l'explorer entièrement et c'est au travers des fixations et des mouvements de poursuite qu'un observateur explore une vidéo. L'évaluation des fluctuations temporelles des distorsions spatiales est réalisée dans des segments spatio-temporels de la vidéo, lesquels correspondent aux fixations ou aux mouvements de poursuite pouvant se produire sur chaque site des images de la séquence. Le résultat de cette évaluation est une séquence de distorsions visuelles.

Dans une première partie nous faisons une revue de la littérature sur sujet. Nous décrirons des approches purement de type signal, des approches structurelles et des approches modélisant le système visuel humain. La suite du chapitre est dédiée à la description de l'approche proposée.

3.2 Revue des méthodes existantes

Le problème de la conception de séquences de distorsions visuelles prenant en compte l'aspect temporel a déjà été abordé dans la littérature sur l'évaluation objective de la qualité. D'ailleurs, il s'agit souvent d'une étape précédant l'élaboration de la note de qualité globale d'une vidéo. La figure 3.1 représente la structure générale d'une métrique de qualité de vidéos reposant sur le calcul de séquences de cartes d'erreurs. L'étape de création de séquences de cartes de distorsions y est encadrée en rouge.

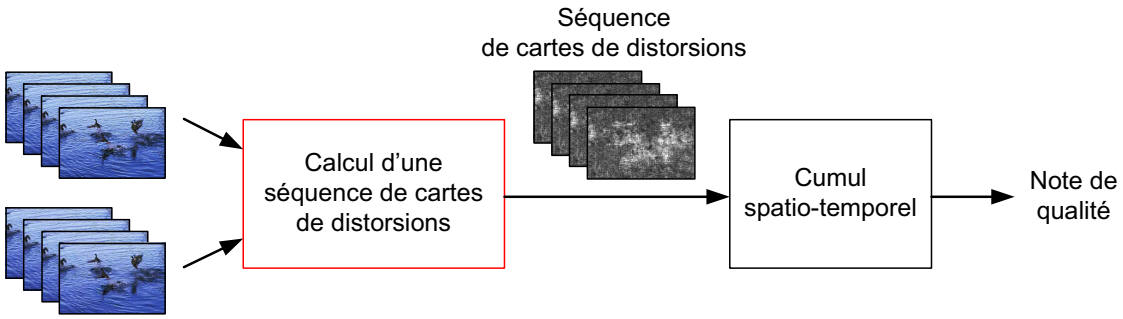


Figure 3.1 – Structure générale d'une métrique de qualité de vidéos reposant sur le calcul de séquences de cartes de distorsions.

Dans cette section, nous allons passer en revue les grandes approches qui se dégagent de littérature.

3.2.1 Les approches purement de type signal

Les approches purement de type signal permettant de construire des séquences de cartes d'erreurs sont de simples applications image par image des méthodes présentées section 2.2.1. Elles ne prennent donc pas en compte les aspects temporels. Les avantages et les inconvénients restent les mêmes que dans le calcul d'une carte d'erreurs pour une image. L'avantage des séquences de cartes d'erreurs mathématiques réside dans leur faible complexité d'implantation et de calcul. Par contre, elles ont l'inconvénient majeur de n'être basées que sur le signal et pas sur la perception humaine. Le PSNR (*Peak Signal to noise Ratio*) (cf. section 4.4.2.1) est une métrique largement utilisée en compression de la vidéo, alors qu'il ne prend pourtant pas en compte la dimension temporelle.

3.2.2 Les approches structurelles

Une extension temporelle de la SSIM a été proposée par Wang et *al.* dans [Wang 04b]. Les auteurs introduisent deux ajustements temporels. Dans leur implantation, ces ajustements ne modifient pas les cartes d'erreurs à proprement parler, car ils sont utilisés dans la phase de cumul permettant la construction de la note de qualité. Cependant, ils peuvent être interprétés comme une pondération des cartes d'erreurs précédent cette phase de cumul. Le premier ajustement consiste à réduire l'importance des zones sombres par rapport aux zones plus claires. Ils considèrent que les zones sombres attirent moins le regard. Par conséquent pour chaque image d'une vidéo, les zones sombres ont une probabilité plus faible d'être regardées. Cette considération est discutable, car l'attention visuelle est davantage dirigée par les contrastes que par les niveaux de luminance eux-mêmes. En effet, une petite zone sombre au milieu d'une grande zone claire attirera aussi bien le regard, qu'une petite zone claire au milieu d'une grande zone sombre. L'effet sur les cartes de distorsions se traduit par une pondération de chaque valeur $SSIM(f_o, f_d)$ par un facteur w_{f_o} dépendant de la luminance moyenne locale :

$$w_{f_o} = \begin{cases} 0 & \mu_{f_o} \leq 40 \\ (\mu_{f_o} - 40)/10 & 40 < \mu_{f_o} \leq 50 \\ 1 & \mu_{f_o} > 50 \end{cases} \tag{3.1}$$

où la valeur μ_{f_o} correspond à la luminance moyenne sur f_o ; f_o et f_d étant respectivement les fenêtres de l'image originale et de l'image dégradée, sur lesquelles est calculée la mesure structurelle.

Le second ajustement temporel est lié au mouvement entre deux images. Il consiste à diminuer l'importance des distorsions présentes dans une image lorsque le mouvement global moyen entre cette image et la précédente est élevé. Ils justifient cet ajustement par le fait que certaines distorsions sont moins bien perçues lors de mouvements rapides de la scène (mouvement de caméra par exemple). Ces considérations sont, pour partie, proches des CSF spatio-temporelles présentées dans le chapitre 1, où la sensibilité au contraste varie en fonction des fréquences temporelles. Cet ajustement peut être considéré comme une pondération de chaque carte par une valeur dépendant du mouvement global moyen entre l'image considérée et l'image précédente. Plus le mouvement est important, plus la pondération tend vers zéro. Dans cette approche, la prise en compte des mécanismes temporels est limitée. La prise en compte du mouvement est très globale et ne permet pas d'améliorer localement les séquences de distorsions obtenues.

3.2.3 Les approches modélisant le système visuel humain

Plusieurs auteurs ont développé des métriques de qualité reposant sur une modélisation du système visuel humain. Nous nous intéressons aux approches dont la création de séquences de distorsions est une étape dans l'élaboration de la note de qualité.

Van den Branden Lambrecht a proposé plusieurs métriques de qualité. Ces métriques sont basées sur des modèles multi-canaux du SVH [van den Branden Lambrecht 96b]. La métrique, appelée MPQM (*Moving Picture Quality Metric*) [van den Branden Lambrecht 96c], est basée sur :

- une définition locale du contraste,
- une décomposition spatiale utilisant des filtres de Gabor,
- deux canaux liés à l'aspect temporel (*transient* et *sustained*),
- une CSF spatio-temporelle,
- un modèle de masquage de contraste intra-canal.

Une version couleur du MPQM utilisant un espace couleur basé sur la théorie des signaux antagonistes a été proposée dans [van den Branden Lambrecht 96a]. Une méthode moins complexe a aussi été proposée sous le nom NVFM (Normalization Video Fidelity Metric) dans [Lindh 96]. Cette méthode utilise, entre autres, une décomposition pyramidale orientée plutôt que des filtres de Gabor pour la décomposition spatiale. C'est une extension spatio-temporelle de la métrique pour image de Teo et Heeger présentée précédemment, et qui exploite le masquage inter-canal. Ces métriques ont l'avantage de reposer sur une modélisation avancée du système visuel. Outre la complexité, un inconvénient réside dans l'application de la CSF spatio-temporel qui est une simple pondération des sous-bandes spatio-temporelles. De plus, des questions se posent sur la séparabilité des domaines spatial et temporel (cf. section 1.3.3.3) de la CSF utilisée. Par ailleurs, le fait que la littérature ne s'accorde pas sur le nombre de canaux temporels est aussi problématique.

Winkler [Winkler 99] a proposé une méthode pour évaluer la qualité des vidéos couleur, appelé PDM (*Perceptual Distortion metric*). Cette méthode utilise une transformation de l'espace colorimétrique et évalue les distorsions sur chacune des trois composantes couleur. Deux flux temporels, correspondant aux canaux *sustained* et *transient*, sont également calculés en utilisant des filtres IIR. La décomposition spatiale comprend 5 niveaux de résolution et 4 orientations. Chaque canal est pondéré en fonction de la CSF et le masquage repose sur un modèle de masquage *excitateur-inhibiteur* proposé par Watson et Solomon [Watson 97a]. Les avantages et les inconvénients sont les mêmes que pour les métriques proposées par Van den Branden Lambrecht.

La métrique DVQ (*Digital Video Quality*) de Watson [Watson 98, Watson 01] est une méthode d'évaluation des vidéos couleur qui opère dans le domaine transformé (DCT). Le domaine DCT présente un avantage certain du point de vue calculatoire, parce que la DCT est implantée de façon efficace et que la plupart des codeurs vidéo sont basés sur la DCT. Une modélisation en trois dimensions des seuils différentiels de visibilité pour les sous-bandes DCT spatio-temporelles est proposée. Son principe est le suivant : calcul de la DCT de l'image originale et de l'image dégradée, calcul d'un contraste local, application une CSF temporelle, normalisation des résultats par les seuils différentiels de visibilité, enfin calcul du signal d'erreur. La méthode est appliquée à chaque composante après une transformation de l'espace colorimétrique. Dans cette métrique, un seul canal temporel est considéré. De plus, la question de la séparabilité espace-temps est de nouveau posée.

3.2.4 Discussion

Aucune des approches de la littérature, ne s'intéresse directement à l'évolution temporelle des distorsions. Les approches purement de type signal ne considèrent pas les aspects temporels et les approches structurelles ne les modélisent qu'assez grossièrement. Les modélisations les plus poussées tentent plutôt de modéliser les mécanismes

transient et *sustained*, ainsi que la sensibilité au contraste spatio-temporelle (CSF) avec les problèmes, évoqués précédemment, que cela pose. Dans ces modélisations, on fait l'hypothèse que l'observateur se trouve toujours en vision stabilisée, de manière à pondérer visuellement les erreurs en terme de visibilité.

L'étude de l'évolution temporelle des distorsions nous semble une approche intéressante à examiner. Comme nous l'avons évoqué dans la section 1.2.2.3, une distorsion temporelle considérée localement peut être décrite comme une variation temporelle d'une distorsion spatiale. Par conséquent, l'étude locale de l'évolution temporelle des distorsions spatiales devrait permettre d'évaluer les distorsions spatio-temporelles. Même si cette approche ne fait partie de celles des méthodes existantes, on trouve dans la littérature connexe, des travaux [Tan 98, Masry 04] présentant des similitudes avec celle-ci. Ces travaux ne s'intéressent pas directement aux variations temporelles de distorsions mais plutôt aux variations temporelles de la qualité. Dans ces travaux, les variations temporelles de la qualité ont été étudiées dans le contexte de l'évaluation continue de la qualité. Dans ce contexte, des métriques de qualité essayent de reproduire la notation continue de la qualité telle qu'elle est enregistrée lors de tests subjectifs d'évaluation continue de la qualité, comme par exemple avec le protocole SSCQE (*Single Stimulus Continuous Quality Evaluation*). L'évolution de la qualité est donc évaluée globalement au travers d'une note par image. Cette évaluation globale par image ne permet donc pas d'évaluer localement les distorsions spatio-temporelles. Cependant, Masry et Hemami [Masry 04] ont introduit l'existence de deux mécanismes intéressants pour le cumul temporel des distorsions : un mécanisme court terme et un mécanisme long terme. Le mécanisme court terme est simulé comme une étape de lissage des notes de qualité par image, autrement dit des distorsions purement spatiales cumulées en notes de qualité. Le mécanisme long terme est, quant à lui, simulé par un traitement récursif opéré sur les notes par image lissées par le mécanisme court terme. La construction de séquences de distorsions spatio-temporelles peut s'apparenter au mécanisme court terme introduit par Masry et Hemami, à la différence que dans ce cas, il n'y a pas de cumul spatial des distorsions. Le mécanisme long terme, quant à lui, concerne l'élaboration de la note globale de qualité pour une séquence vidéo. Il sera l'objet du chapitre 6.

Différents aspects sont à considérer dans l'élaboration de ce mécanisme court terme. Un premier aspect concerne la façon dont une vidéo est explorée par le système visuel humain. Ce sont les mécanismes de l'attention visuelle qui permettent au système visuel humain d'explorer une séquence vidéo. C'est au cours des fixations et des mouvements de poursuite qu'un observateur peut évaluer les distorsions d'une vidéo. La séquence vidéo est donc « fragmentée » par le système visuel en une multitude de segments spatio-temporels. Notre approche doit tenter de reproduire ce mécanisme.

Un autre aspect à considérer concerne les fluctuations temporelles des distorsions spatiales elles-mêmes. Évaluer, subjectivement ou objectivement, la qualité d'une image, se limite à évaluer la gêne provoquée par des distorsions purement spatiales : pas d'évolution temporelle des distorsions sur la durée de présentation de l'image. Par contre, l'évaluation subjective ou objective de la qualité de vidéos implique de prendre en compte la dimension temporelle. Les distorsions conservent évidemment leur support spatial, mais leur évolution temporelle joue un rôle perceptuel fondamental. On identifie un certain nombre de distorsions temporelles comme par

exemple le papillotement (*flickering*), le *jerkiness* ou encore l'effet « mosquito » (cf. section 1.2.2). De façon générale, une distorsion temporelle peut être décrite comme une variation ou une fluctuation temporelle des distorsions spatiales sur une zone particulière de la vidéo. La perception d'une distorsion spatiale peut être largement modifiée (accentuée ou diminuée) par son évolution temporelle. Si cette évolution n'a pas de caractère périodique, alors sa vitesse, c'est-à-dire le temps nécessaire pour passer d'un niveau de distorsion à un autre plus ou moins élevé, influencera la perception que l'on aura de cette distorsion temporelle. Si cette évolution temporelle a un caractère périodique, alors sa fréquence temporelle influencera significativement la perception que l'on en aura, de même que le caractère répétitif de cette évolution en lui-même.

L'approche que nous proposons se base sur l'élaboration d'un mécanisme court terme d'évaluation des distorsions spatio-temporelles, ainsi que sur la prise en compte des différents aspects évoqués. Dans la section suivante nous allons décrire le principe général de l'approche proposée.

3.3 Principe général du modèle proposé

Dans ce travail, nous avons choisi d'évaluer les distorsions temporelles d'une vidéo au travers de l'étude de l'évolution temporelle des distorsions spatiales. La perception des distorsions temporelles est liée aux mécanismes de l'attention visuelle. Le système visuel humain est intrinsèquement un système limité dans le sens où il n'est pas capable de percevoir instantanément, et avec une grande précision, l'ensemble de son champ visuel. Pour remédier à ce problème, l'inspection visuelle du champ visuel est réalisée au travers des différents mécanismes de sélection liés à l'attention visuelle. L'attention visuelle est décrite dans son ensemble dans la section 7.2. Cependant, nous allons présenter rapidement quelques notions permettant d'expliquer les fondements de notre méthode.

L'attention visuelle se manifeste en grande partie au travers de différents mouvements oculaires qui peuvent être décomposés en trois grands types de mouvements [Hoffman 98] : les saccades, les fixations et les mouvements de poursuite. Les saccades sont des mouvements très rapides qui permettent à l'homme d'explorer son champ visuel. Les fixations sont des mouvements résiduels qui se produisent lorsque le regard est focalisé sur une zone particulière du champ visuel. Les mouvements de poursuite sont des mouvements qui permettent à l'oeil de suivre une zone (objet ou partie d'objet) en mouvement dans son champ visuel. Les saccades mobilisent les ressources visuelles sur les différentes parties d'une scène. Entre deux saccades se produit une fixation, ou un mouvement de poursuite. Au-delà des mouvements oculaires on distingue deux formes de focalisation de l'attention visuelle : une focalisation dite *overt* ou une focalisation dite *covert*. La première forme de focalisation se traduit directement par un mouvement de l'oeil, contrairement à la seconde qui utilise la vision périphérique, comme lorsqu'on regarde du « coin de l'oeil ». En focalisation *overt* la précision d'analyse est supérieure à celle en focalisation *overt*, car dans le premier cas la zone fovéale de la rétine est exploitée et dans le second cas c'est la zone péri-fovéale de la rétine qui l'est. Même si la parafovéa joue probablement un rôle dans la perception des distorsions temporelles, nous avons choisi de simplifier le problème en faisant l'hypothèse que les distorsions sont principalement perçues par la zone fovéale de la rétine. Le modèle que nous proposons est donc

purement fovéal, et la focalisation de l'attention visuelle est de type *overt*. Un observateur évaluant la qualité d'une vidéo explore cette vidéo en effectuant un ensemble de fixations. Si on ne considère que la zone fovéale de la rétine, l'exploration de la vidéo entraîne sa « fragmentation » en segments spatio-temporels qui sont évalués successivement par l'observateur. Chaque segment spatio-temporel correspond à l'information perçue par le système visuel pendant une fixation, ou un mouvement de poursuite. Ces segments sont limités spatialement par la projection de la vidéo sur la zone fovéale de la rétine. Cette approche n'est évidemment valable que pour la fovéa. Elle n'est plus valable si on considère aussi la parafovéa.

Lors de l'évaluation des distorsions d'un objet en mouvement par le système visuel, l'oeil doit être en mouvement de poursuite de façon à stabiliser l'image de cet objet sur la fovéa. Dans ce cas, l'évaluation locale des distorsions par une méthode objective doit reproduire ce phénomène. Les objets en mouvement doivent donc être évalués en prenant en compte leur mouvement, et par conséquent la construction des segments spatio-temporels aussi. Par ailleurs, les segments spatio-temporels sont limités temporellement soit par la durée de la fixation, soit par la durée du mouvement de poursuite. La durée d'un mouvement de poursuite est difficile à déterminer, c'est pourquoi, étant donné que la durée d'une fixation est inférieure à la durée d'un mouvement de poursuite, nous avons choisi d'évaluer les distorsions temporelles à l'échelle de temps des fixations. La durée moyenne d'une fixation est de l'ordre 400ms (cf. section 9.3.1).

Cette évaluation court terme correspond au modèle proposé pour la construction de séquences de distorsions perceptuelles. La structure générale de ce modèle est présentée figure 3.2. Dans ce modèle, les distorsions spatio-

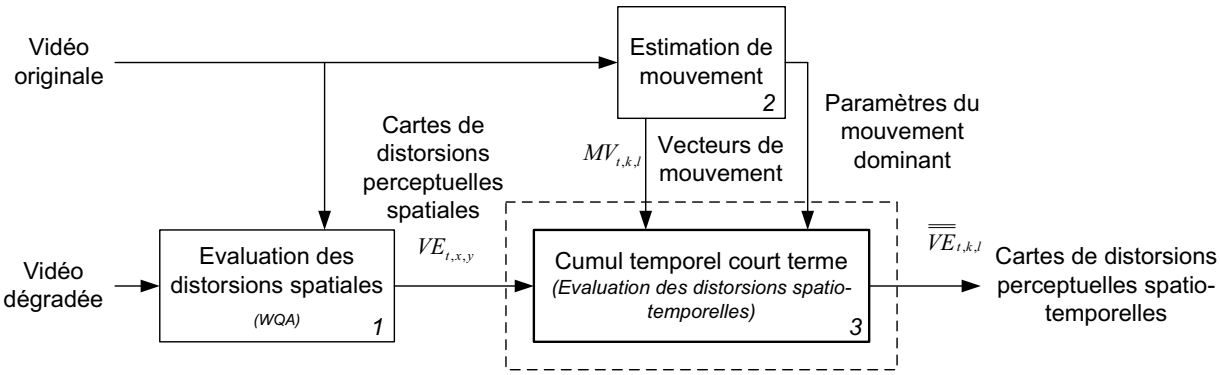


Figure 3.2 – Structure générale du modèle de construction de séquences de cartes d'erreurs perceptuelles.

temporelles sont évaluées au travers d'une analyse temporelle des cartes de distorsions perceptuelles purement spatiales $VE_{t,x,y}$. Les cartes de distorsions perceptuelles spatiales (notées $VE_{t,x,y}$) sont calculées en utilisant le modèle WQA, basé ondelettes, présenté dans le chapitre 2, cette première étape est notée 1 sur la figure 3.2. La seconde étape est constituée de l'estimation de l'information de mouvement apparent (noté $MV_{t,k,l}$) nécessaire à l'évaluation des objets en mouvement. La dernière étape, notée 3 sur la figure 3.2, correspond à l'évaluation des distorsions spatio-temporelles. Les résultats de cette étape sont les cartes $\overline{\overline{VE}}_{t,k,l}$ des distorsions spatio-temporelles, qui sont calculées à partir des cartes $VE_{t,x,y}$ de distorsions perceptuelles purement spatiales et des

informations sur le mouvement. Pour chaque image d'une séquence, une carte des distorsions spatio-temporelles est calculée. Cette carte encode en chaque site (x, y) le niveau de distorsions spatio-temporelles perceptibles.

3.4 Cumul temporel court terme

L'évaluation spatio-temporelle des distorsions est un problème complexe. Dans le modèle proposé, nous réalisons une évaluation à court terme des distorsions temporelles, en simulant les distorsions perçues au niveau des fixations oculaires. Comme nous l'avons évoqué précédemment, la séquence vidéo doit être divisée en segments spatio-temporels correspondant à chaque fixation possible. Cela signifie que la fixation peut commencer à tout instant t, et sur tout site (x, y) de la séquence. Lors d'une fixation, la perception des distorsions dépend à la fois du niveau moyen des distorsions et des fluctuations temporelles de celles-ci. En effet, le niveau moyen des distorsions joue un rôle important dans l'évaluation locale des vidéos qui est comparable à celui des distorsions purement spatiales des images fixes. Cependant, ce sont les fluctuations temporelles de celles-ci qui sont à l'origine des distorsions temporelles (et donc spatio-temporelles). Il est donc impératif d'en tenir compte dans l'évaluation objective des distorsions spatio-temporelles. Pour tenir compte de ces deux aspects, deux traitements sont à considérer. D'une part les variations temporelles des distorsions doivent être lissées afin d'obtenir le niveau moyen de distorsions perceptible au cours d'une fixation. D'autre part, les faibles variations temporelles de distorsions doivent être éliminées. Seules les variations temporelles de distorsions les plus importantes perceptuellement doivent être prises en compte. La figure 3.3 donne les principaux traitements impliqués dans cette évaluation. Le premier traitement (noté 3.1 sur la figure) est consacré à la création des structures spatio-temporelles nécessaires à l'analyse des variations de distorsion au cours d'une fixation. Ces structures sont appelées « des tubes spatio-temporels ». Le procédé est ensuite séparé en deux branches parallèles. L'objectif de la première branche est d'évaluer le niveau moyen des distorsions au cours d'une fixation. L'objectif de la seconde branche est d'évaluer les variations de distorsion survenues sur la durée de la fixation et auxquelles les humains sont les plus sensibles. Ensuite, ces deux branches sont fusionnées pour obtenir les cartes des distorsions spatio-temporelles.

Les différentes étapes de l'évaluation des distorsions perceptuelles spatio-temporelles réalisée par le cumul temporel court terme des cartes de distorsions spatiales, sont détaillées dans les sections suivantes.

3.5 Les tubes spatio-temporels

Les tubes spatio-temporels sont des structures $2D + t$ modélisant les segments spatio-temporels d'une vidéo collectés par le système visuel humain au cours de ses fixations. L'information sur le mouvement apparent est nécessaire à la création des tubes, c'est pourquoi nous allons d'abord décrire comment cette information est estimée. Nous décrirons ensuite la construction des tubes.

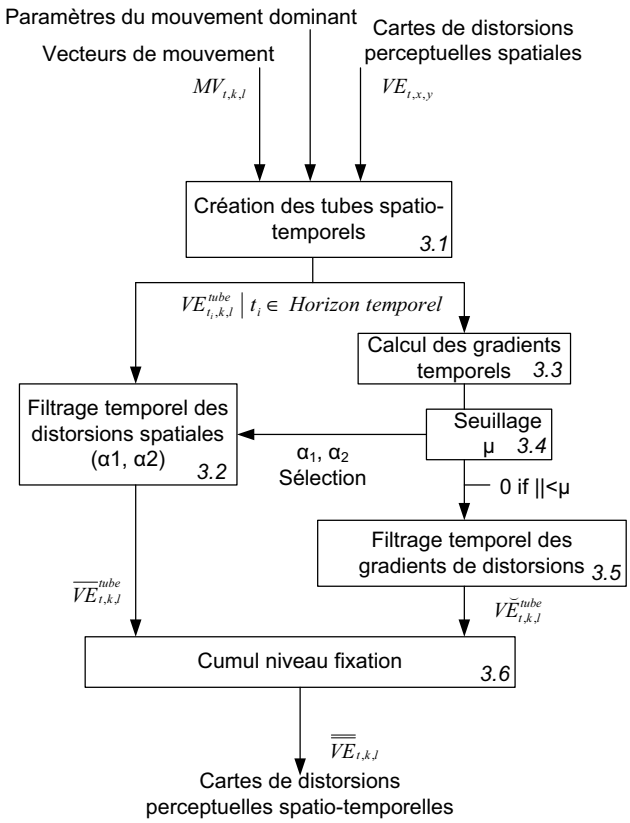


Figure 3.3 – Structure du cumul temporel court terme, ou autrement dit de l'étape d'évaluation des distorsions perceptuelles spatio-temporelles.

3.5.1 Estimation de l'information de mouvement

Dans cette étape, notée 2 sur la figure 3.2, le mouvement local entre deux images successives est estimé, ainsi que le mouvement dominant. Cette étape est réalisée au moyen d'un estimateur de mouvement hiérarchique classique (HME : *Hierarchical Motion Estimator*). Le mouvement local est estimé sur des blocs 8×8. Le mouvement local et le mouvement dominant sont tous les deux utilisés pour construire les structures spatio-temporelles dans lesquelles les distorsions vont être évaluées. Ces structures spatio-temporelles correspondent aux segments spatio-temporels perçus au cours des fixations ou des mouvements de poursuite.

Lors d'un mouvement de poursuite les segments spatio-temporels doivent suivre l'objet en mouvement afin de simuler la stabilisation de celui-ci sur la fovéa. Le mouvement apparent local est donc utilisé pour reconstruire la trajectoire passée de cet objet en mouvement (pour simplifier l'expression par la suite on le notera simplement comme mouvement local). Le mouvement local $\overrightarrow{V}_{local}$ (le vecteur de mouvement) qu'on considère de type translation plane purement, est calculé pour chaque bloc (k, l) d'une image au moyen d'un algorithme de *block-matching* hiérarchique. L'estimation de mouvement est donc réalisée sur différents niveaux (différentes résolutions), et les résultats de chaque niveau sont utilisés comme initialisation par le niveau suivant.

Le mouvement dominant est utilisé pour déterminer l'horizon temporel sur lequel on peut suivre un objet, cet horizon temporel dépend entre autres de l'apparition et de la disparition de l'objet dans la scène comme nous le détaillons dans la section suivante. Afin d'estimer le mouvement dominant, la transformation globale qui se produit entre deux images successives est estimée à partir des vecteurs de mouvement locaux précédents. Nous considérons que le déplacement $\overrightarrow{V}_{\Theta}(x, y)$, au site (x, y) lié à un modèle de mouvement paramétrique Θ est donné par un modèle de mouvement affine 2D :

$$\overrightarrow{V}_{\Theta}(s) = \begin{pmatrix} a_1 + a_2 x + a_3 y \\ a_4 + a_5 x + a_6 y \end{pmatrix}, \tag{3.2}$$

où $\Theta = [a_1, a_2, a_3, a_4, a_5, a_6]$ représentent les paramètres affines du modèle. Les paramètres affines sont calculés avec une technique statistique robuste de M-estimation [Odobez 95].

Le mouvement local et le mouvement dominant sont ensuite utilisés pour construire les tubes spatio-temporels.

3.5.2 Construction des tubes spatio-temporels

Les tubes spatio-temporels des distorsions sont créés dans l'étape 3.1 de la figure 3.3. Le but de cette étape est de diviser la séquence vidéo en segments spatio-temporels correspondant à chaque fixation possible. Un tube spatio-temporel est calculé pour chaque bloc d'une image t. Un tube spatio-temporel est une structure spatio-temporelle contenant le passé d'un bloc en termes de distorsions spatiales (cf. Fig. 3.4). Cela signifie que cette structure contient les différentes valeurs de distorsion de ce bloc sur un horizon temporel spécifique. La valeur de distorsion de ce bloc à l'instant t est obtenue en moyennant les valeurs de distorsion des pixels de ce bloc à l'instant t. Les vecteurs de mouvement $MV_{t,k,l}$ sont utilisés pour compenser en mouvement le bloc (k, l). Cette

compensation en mouvement permet de reconstituer la trajectoire du bloc considéré comme sur la figure.

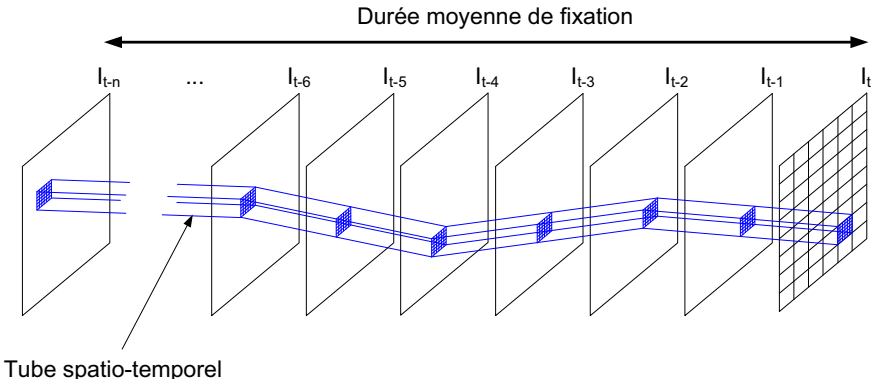


Figure 3.4 – Illustration d'un tube spatio-temporel. La trajectoire passée d'un bloc de l'image I_t est reconstituée à partir de l'historique des vecteurs de mouvement de ce bloc.

L'horizon temporel est limité à 400ms (durée moyenne d'une fixation). Cependant, il peut être réduit en fonction d'événements se produisant dans le passé, comme par exemples l'apparition ou la disparition d'un objet, le découvrement d'une zone par un objet en mouvement.

Pour détecter ces événements, chaque bloc est classé en comparant son mouvement avec la représentation paramétrique du mouvement dominant : chaque bloc est soit conforme (*inlier*), soit non-conforme (*outlier*), au mouvement dominant. Ensuite la classification entre deux images consécutives des blocs compensés en mouvement est comparée. Si un bloc change de classification (*inlier*/*outlier*) cela signifie qu'un événement dans le passé nécessite que le tube spatio-temporel se termine. Le tableau 3.1 présente les changements de classifications et les événements associés.

Changement de Classification t→t-1	Mouvement dominant (MD) nul	Termine tube	Mouvement dominant (MD) non nul	Termine tube
Inlier → *Outlier*	Objet s'arrête	*	Objet rejoint MD	*
Inlier → *Outlier*	Objet disparaît	Oui	Objet disparaît	Oui
Inlier → *Outlier*	Zone découverte	Oui	Zone découverte	Oui
Outlier → *Inlier*	Objet apparaît	Oui	Objet apparaît	Oui
Outlier → *Inlier*	Objet se met en mouvement	*	Objet quitte MD	*

Table 3.1 – Changements de classification et événements associés. Le symbole * signifie que l'événement correspond à un changement de mouvement oculaire (fixation/poursuite). La terminaison du tube associée à cet événement est discutable.

La figure 3.5 illustre le calcul de l'horizon temporel dans un cas simple : un objet est en translation sur un fond fixe, c'est-à-dire avec un mouvement dominant nul.

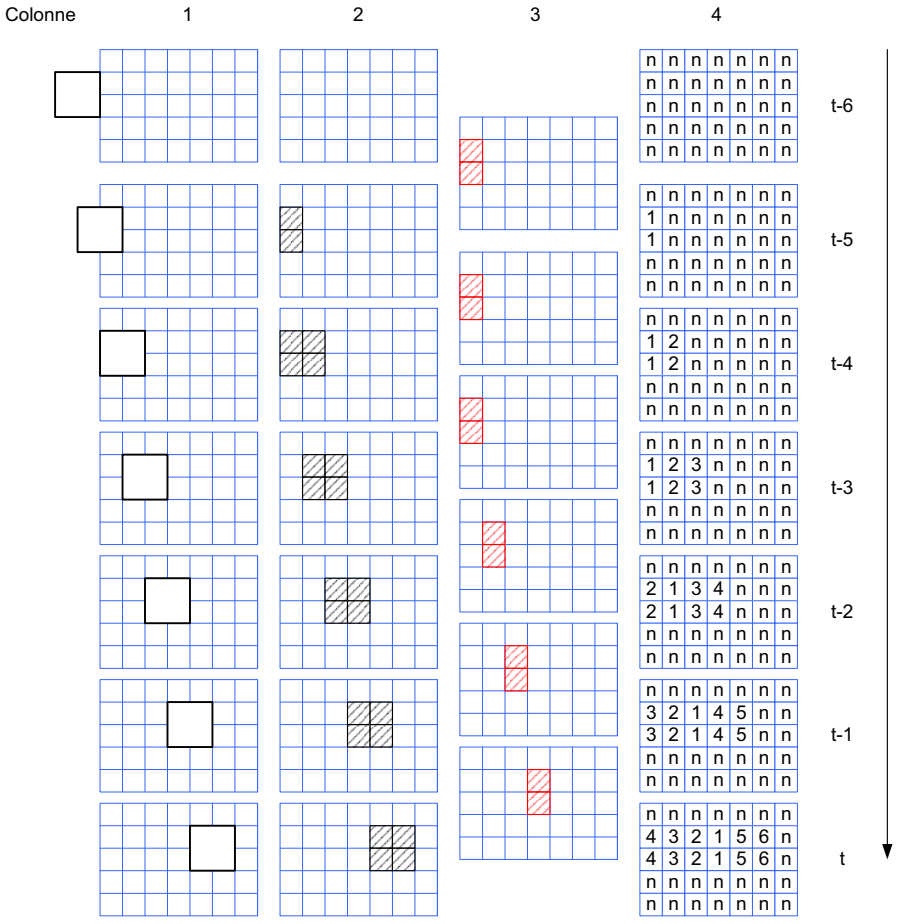


Figure 3.5 – Illustration du calcul de l'horizon temporel d'un tube spatio-temporel. Les lignes correspondent à des images successives de $t-6$ à t. De gauche à droite, la première colonne présente un objet en déplacement. La seconde colonne présente la carte des blocs *outlier* en terme de mouvement. La troisième colonne présente la carte des changements de classification (*inlier*/*outlier*) entre les blocs de deux images successives et en tenant compte du mouvement de ces blocs entre les deux images. La quatrième colonne présente pour chaque image, la taille en nombres d'images de l'horizon temporel des tubes spatio-temporels partant de chaque bloc, la valeur n correspondant à la durée maximale d'un tube soit 400ms exprimée en nombre d'images.

3.6 Filtrage temporel des distorsions spatiales dans les tubes

Le filtrage temporel des distorsions spatiales est réalisé dans l'étape 3.2 de la figure 3.3, ce qui correspond à la première branche de notre modèle. L'objectif de cette étape est d'obtenir un niveau de distorsion moyen sur la durée de la fixation. Étant donné que les grandes variations temporelles de distorsions sont les plus perceptibles, leurs contributions doivent être plus importantes que celles de variations temporelles de distorsions plus limitées. Pour ce faire, un filtrage temporel est utilisé. Il s'agit d'un filtre récursif dont les caractéristiques sont ajustées en fonction de l'importance des variations temporelles de distorsions. La contribution des variations temporelles importantes de distorsions est augmentée par rapport à la contribution des variations temporelles de distorsions plus limitées en modulant la constante de temps du filtre. Celle-ci varie en fonction de la valeur du gradient de distorsion correspondant. Le gradient de distorsion est déterminé dans l'autre branche du traitement (cf. section 3.7). Une constante de temps α_1 est utilisée si la valeur absolue de la valeur du gradient de distorsion est supérieure à un seuil μ, sinon c'est une constante de temps α_2 avec $\alpha_2 > \alpha_1$ qui est utilisée. La constante α_2 a été fixée à la durée moyenne d'une fixation et α_1 à sa moitié. Le seuil μ, déduit empiriquement, est utilisé dans la seconde branche de notre modèle.

La sortie de cette étape est une carte $\overline{VE}^{tube}_{t,k,l}$ où dans chaque bloc (k,l) à l'instant t on a le résultat du filtrage temporel des distorsions spatiales dans chaque tube spatio-temporel se terminant à l'instant t.

3.7 Évaluation temporelle des distorsions dans les tubes

La seconde branche de notre modèle commence par l'étape notée 3.3 de la figure 3.3. Elle consiste à évaluer les variations temporelles des distorsions. Dans cette première étape, les gradients temporels des distorsions spatiales dans les tubes sont calculés, afin d'évaluer les variations temporelles de distorsions les plus perceptuellement importantes au cours des fixations. Dans un tube, le gradient temporel de distorsion $\nabla VE^{tube}_{t_i,k,l}$ à l'instant t_i est calculé comme suit :

$$\nabla VE^{tube}_{t_i,k,l} = \frac{\delta VE^{tube}_{t_i,k,l}}{\delta t} \left| \begin{array}{l} \delta t = t_i - t_{i-1} \\ t_i \in Horizon\,temporel \end{array} \right. , \tag{3.3}$$

où $VE^{tube}_{t_i,k,l}$ est la valeur de distorsion à l'instant t_i.

Les faibles variations temporelles de distorsion, qui ne sont probablement pas gênantes, ne sont pas prises en compte. L'objectif de cette étape, notée 3.4 sur la figure 3.3, est de les supprimer. Dans cette étape, une opération de seuillage est effectuée sur les valeurs absolues de gradients. Si la valeur absolue de gradient temporel est inférieure à μ, la valeur de gradient est mise à zéro, elle reste inchangée sinon. Cette opération de seuillage est également utilisée pour sélectionner la constante de temps du filtrage récursif temporel de l'étape 3.2 tel que décrit précédemment.

Les caractéristiques des distorsions temporelles, comme la fréquence et l'amplitude des variations, ont un impact sur la perception de ces distorsions. L'objectif de l'étape suivante, notée 3.5 est de prendre en compte cela. Pour ce faire, un filtrage temporel des gradients de distorsion est réalisé dans chaque tube d'une image à

l'instant t. Cette opération de filtrage temporel est réalisée en comptant le nombre de changements de signe des gradients de distorsion $nS^{tube}_{t,k,l}$ se produisant à l'intérieur du tube. Le maximum (en valeur absolue) du gradient temporel de distorsion, noté $Max\nabla VE^{tube}_{t,k,l}$, est calculé et exploité. La sortie du filtrage temporel est donnée par :

$$\breve{VE}^{tube}_{t,k,l} = Max\nabla VE^{tube}_{t,k,l} \cdot fs(nS^{tube}_{t,k,l}) \,, \tag{3.4}$$

où la fonction fs est une fonction du nombre n de changements de signe :

$$fs(n) = \frac{g_s}{\sigma_s\sqrt{2\pi}} \cdot e^{-\frac{(n-\mu_s)^2}{2\sigma_s^2}} \,, \tag{3.5}$$

pour $n \geq 0$.

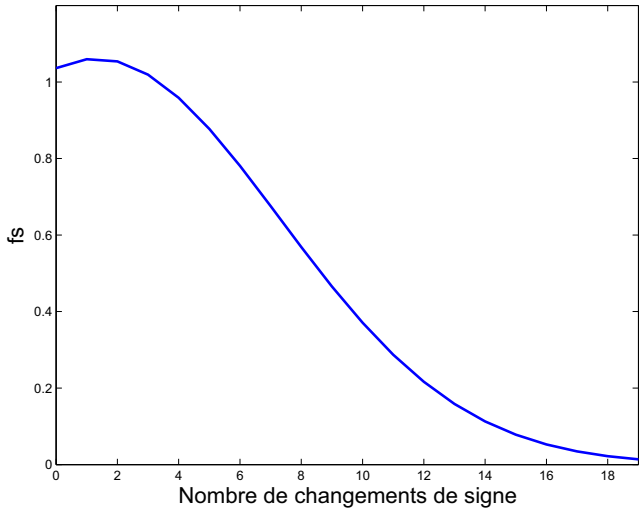


Figure 3.6 – Fonction fs. Cette fonction atteint son maximum autour d'un changement de signe des gradients de distorsion par durée de fixation.

La réponse de la fonction $fs(n)$ est donnée par la figure 3.6. La fonction $fs(n)$ donne plus d'importance aux variations temporelles de distorsions à basse-moyenne fréquence plutôt qu'à celles à basse ou moyenne ou haute fréquence. La sensibilité du système visuel humain est maximale pour des variations temporelles aux alentours de 2 à 3 cy/s, ce qui correspond à environ un changement de signe sur la durée moyenne d'une fixation. La sortie de cette étape est la carte $\breve{VE}^{tube}_{t,k,l}$ où chaque bloc (k,l) est le résultat du filtrage temporel des gradients de distorsion dans chaque tube spatio-temporel se terminant à l'instant t.

Les sorties provenant des deux branches sont ensuite combinées dans la dernière étape (notée 3.6). Celle-ci effectue le cumul au niveau fixation, où les cartes $\overline{VE}_{t,k,l}$ et $\breve{VE}_{t,k,l}$ sont fusionnées afin d'obtenir la carte finale des distorsions spatio-temporelles $\overline{\overline{VE}}_{t,k,l}$. S'il n'y a pas de variations temporelles des distorsions dans la séquence vidéo la carte finale $\overline{\overline{VE}}_{t,k,l}$ reste identique à la carte $\overline{VE}_{t,k,l}$. Cependant, en présence de variations temporelles de distorsions, la carte $\overline{VE}_{t,k,l}$ est renforcée multiplicativement par des variations temporelles de distorsion encodée dans la carte $\breve{VE}_{t,k,l}$. La carte finale est donnée par :

$$\overline{\overline{VE}}_{t,k,l} = \overline{VE}_{t,k,l} \cdot (1 + \beta \cdot \breve{VE}_{t,k,l}) \,. \tag{3.6}$$

La valeur du paramètre β a été déduite de manière empirique par des expérimentations sur des séquences de synthèse. Ces expérimentations visaient à obtenir des cartes de distorsions spatio-temporelles pertinentes sur des séquences de synthèse où des distorsions ont été introduites. Nous avons fixé β à la valeur 3. Les séquences de synthèse sont constituées d'« extraits » d'images (de forme circulaire dans l'exemple de la figure 3.7) en déplacement ou pas et ayant subissant des dégradations à différentes fréquences temporelles, le tout sur un fond gris uniforme. Dans l'exemple de la figure 3.7, on observe que sur la carte de distorsions VE (purement spatiale) les distorsions sont quasiment les mêmes sur les quatre « extraits », alors que sur la carte $\overline{\overline{VE}}$ les distorsions spatio-temporelles ne sont pas les mêmes. Sur la carte $\overline{\overline{VE}}$ on observe que les distorsions des « extraits » d'images 1 et 3 sont plus importantes que celles des « extraits » d'images 2 et 4. Ceci s'explique par le fait que les distorsions des « extraits » d'images 2 et 4 ne varient pas temporellement.

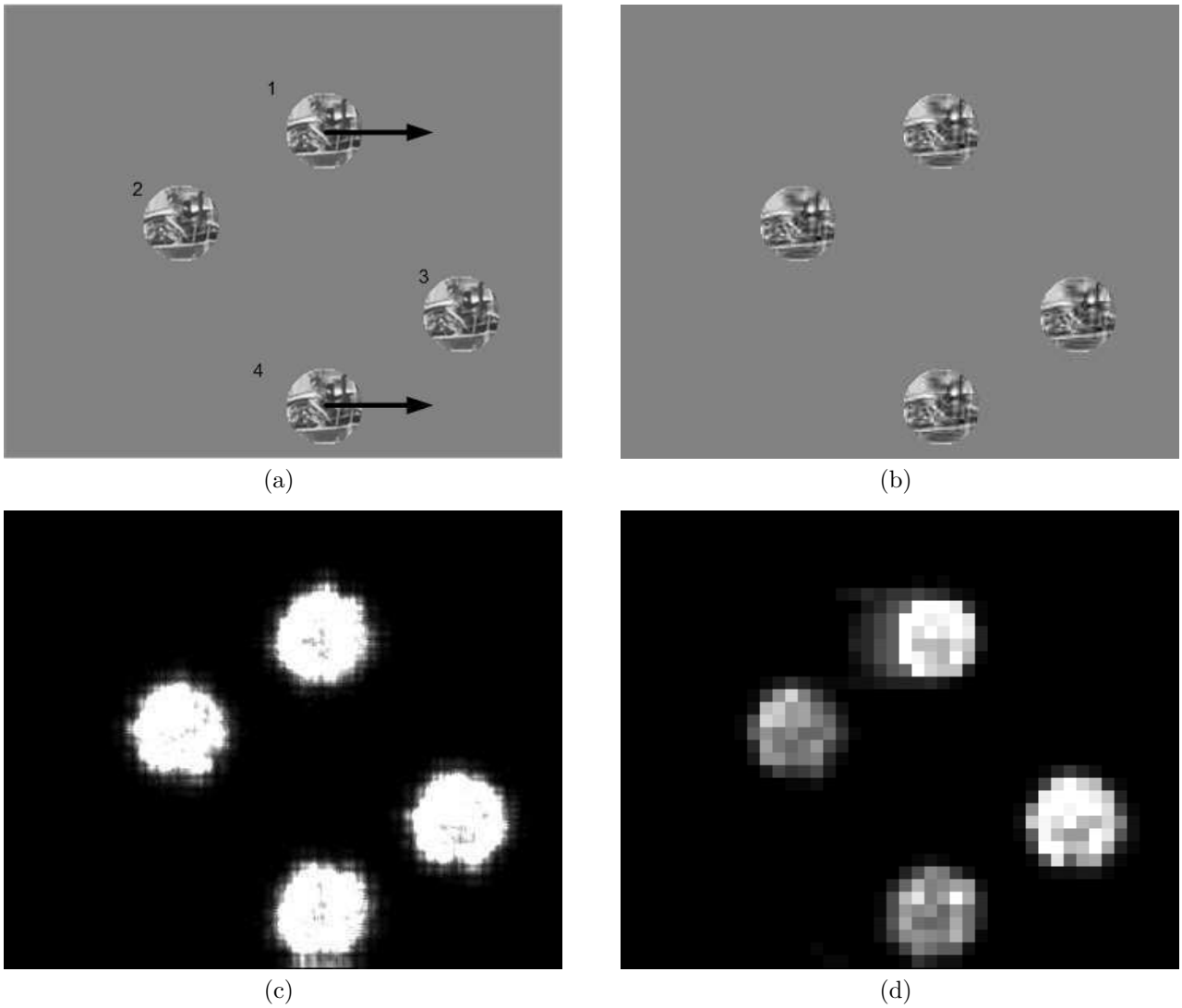
3.8 Résultats qualitatifs

La prise en compte des fluctuations temporelles de distorsions au niveau des fixations oculaires ayant été faite et modélisée, il est maintenant possible de calculer des séquences de cartes d'erreurs perceptuelles représentant les distorsions visibles entre une séquence originale et une séquence dégradée.

La figure 3.8 illustre deux cartes de distorsions pour la même image de la séquence *CrowdRun*. L'une des cartes (figure 3.8(c)) est issue d'une séquence de distorsions purement spatiale, ce qui correspond aux cartes VE (cf. figure 3.2) en entrée de notre cumul temporel court terme. L'autre carte (figure 3.8(d)) est le résultat obtenu après le cumul temporel court terme, noté $\overline{\overline{VE}}$. On observe que sur les cartes obtenues après le cumul temporel court terme, les valeurs de distorsions sont données par blocs. Chaque bloc correspond à un tube spatio-temporel calculé en remontant dans le passé de l'image considérée. La valeur du bloc correspond aux résultats du cumul temporel court terme réalisé dans le tube spatio-temporel correspondant. Le fait d'avoir une valeur de distorsion par bloc, au lieu d'une valeur par pixel, n'est pas incohérent dans le sens où le système visuel humain est plus sensible à des regroupements d'erreurs plutôt qu'à des erreurs isolées. Cependant, on peut critiquer le fait que la position des blocs soit fixe. En revanche, d'un point de vue utilisation dans des algorithmes fondés sur des blocs, cela n'est pas du tout gênant.

La figure 3.9 donne l'évolution temporelle sur l'ensemble de la séquence de la distorsion moyenne évaluée dans deux blocs (le bloc 1 et le bloc 2), avec (figure 3.9(b)) et sans (figure 3.9(a)) le cumul temporel court-terme. On peut observer l'effet de lissage temporel obtenu après le cumul temporel court terme. Le cumul temporel court terme a supprimé localement les fluctuations temporelles de distorsions auxquelles nous sommes le moins sensibles, tout en tenant compte du niveau moyen des distorsions.

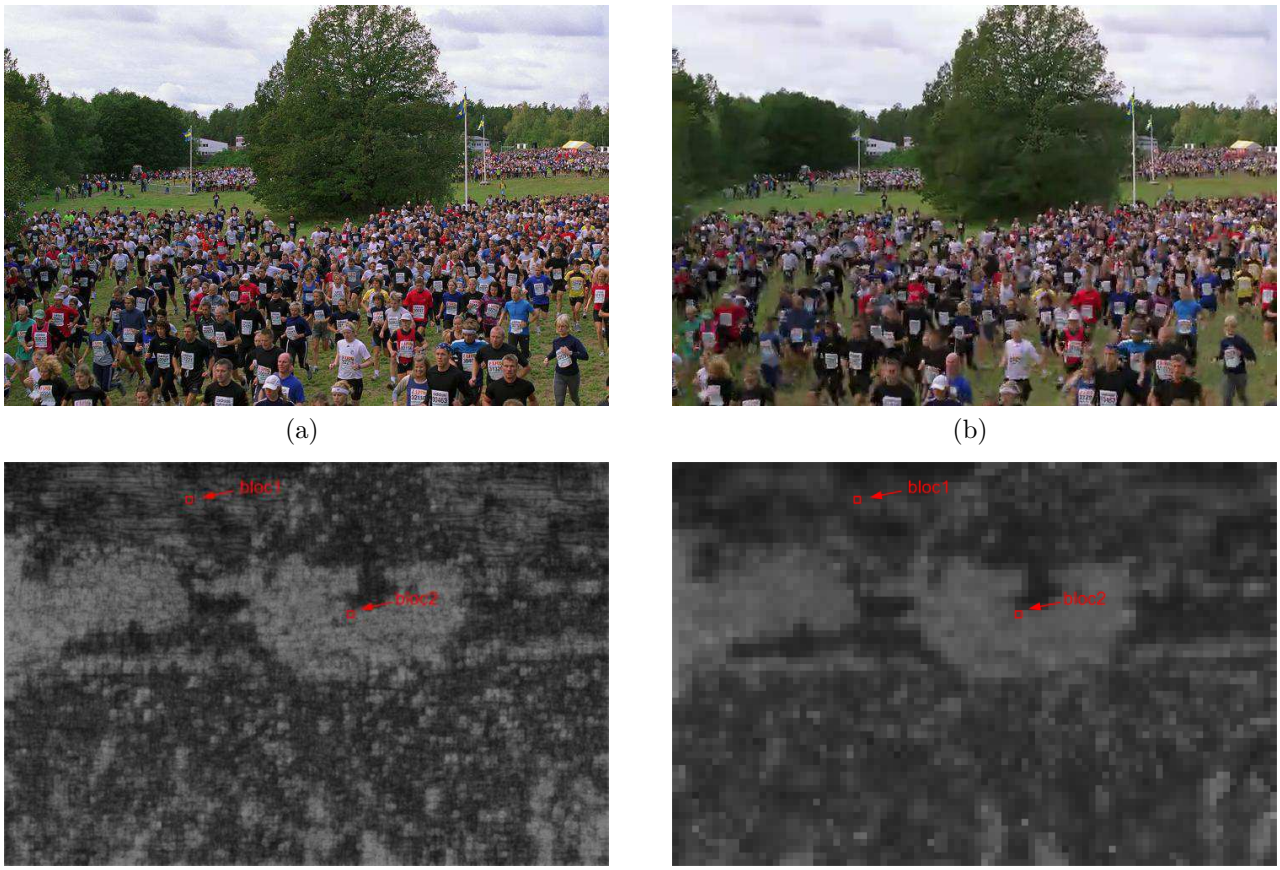
Dans la seconde partie de ce mémoire (chapitre 6), nous utiliserons cette méthode de cumul temporel au niveau des fixations dans une métrique de qualité vidéo et nous évaluerons quantitativement son apport.



(a) (b)

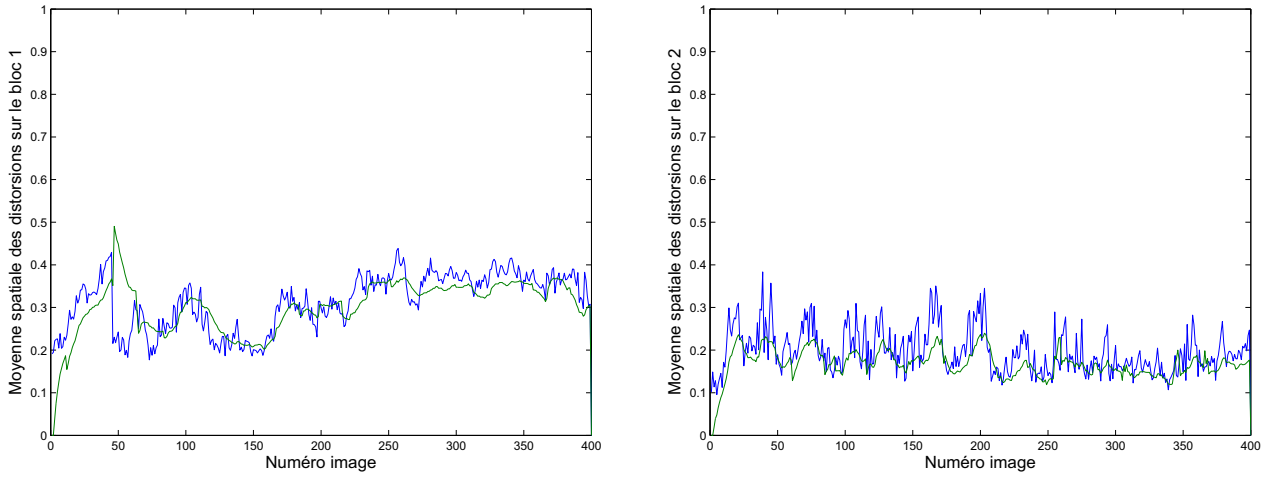
(c) (d)

Figure 3.7 – Exemple d'une séquence de synthèse : (a) version originale (avec annotations), (b) version dégradée, (c) et (d) cartes de distorsions VE (purement spatiale) et $\overline{\overline{VE}}$ (spatio-temporelle) respectivement. Dans cette séquence deux « extraits » d'images sont en mouvement de translation vers la droite (notés 1 et 4) et deux sont fixes (notés 2 et 3). Deux « extraits » d'images (1 et 3) sont dégradées à une fréquence temporelle et les deux autres (2 et 4) ne varient pas temporellement.



(a) (b) (c) (d)

Figure 3.8 – Illustration sur l'image 45 de la séquence *CrowdRun* : (a) version originale, (b) version dégradée, (c) et (d) cartes de distorsions VE (purement spatiale) et $\overline{\overline{VE}}$ (spatio-temporelle) respectivement.



(a) (b)

Figure 3.9 – Évolution temporelle de la distorsion moyenne des cartes de distorsions purement spatiale VE (en bleu) et spatio-temporelle $\overline{\overline{VE}}$ (en vert) dans le bloc 1 (a), et le bloc 2 (b) de la séquence *CrowdRun* (cf. figure 3.8).

3.9 Conclusion

L'objectif de ce chapitre concernait la conception de séquences de distorsions visuelles prenant en compte la dimension temporelle. Après avoir présenté les différentes approches de la littérature, nous avons proposé une méthode d'évaluation locale des distorsions temporelles aboutissant à la construction de séquences de distorsions visuelles spatio-temporelles. Cette méthode repose sur une modélisation fovéale du système visuel humain. Dans cette modélisation, on considère qu'une séquence vidéo est explorée par le système visuel humain au travers de la zone fovéale de la rétine grâce aux mécanismes de l'attention visuelle. Les mécanismes de l'attention visuelle produisent une succession de fixations et de mouvements de poursuite, qui vont entraîner la « fragmentation » de la vidéo en un ensemble de segments spatio-temporels qui sont évalués par l'observateur. La méthode proposée tente de reproduire ces mécanismes au travers d'un cumul temporel court terme des distorsions spatiales. Les segments spatio-temporels sont simulés par des tubes spatio-temporels dans lesquels sont cumulés les fluctuations des distorsions spatiales. Les distorsions spatiales sont évaluées par le modèle WQA proposé et présenté dans le chapitre 2.

La perception des distorsions temporelles est liée à deux facteurs : le niveau moyen de distorsions et les évolutions temporelles de distorsions. Ces deux facteurs sont pris en compte dans le cumul temporel court terme. Dans les tubes spatio-temporels, les variations temporelles des distorsions sont filtrées afin d'obtenir le niveau moyen de distorsions perceptible sur le segment spatio-temporel, les faibles variations temporelles de distorsions sont éliminées. Seules les variations temporelles de distorsions les plus importantes perceptuellement sont prises en compte. L'amplitude des variations et leur fréquence déterminent cette importante perceptuelle.

Conclusion

La première partie de ce mémoire était consacrée à l'évaluation locale des distorsions. Comme nous l'avons évoqué précédemment, l'évaluation locale des distorsions est un des besoins des concepteurs de systèmes de traitement d'images ou de vidéos. Pour répondre à ce besoin, nous avons présenté des méthodes d'évaluation locale des distorsions avec référence complète tant pour les images fixes que pour les vidéos. Les méthodes proposées reposent sur une modélisation du système visuel humain.

Concernant les images fixes, nos travaux ont consisté d'une part à simplifier la complexité de calcul d'une modélisation existante du système visuel humain en proposant une décomposition en sous-bande basée sur la transformée en ondelettes, d'autre part à proposer une meilleure modélisation des effets de masquage par la prise en compte du masquage semi-local en plus du masquage de contraste.

Concernant les vidéos, nos travaux ont consisté à proposer une nouvelle approche d'évaluation locale des distorsions temporelles. Cette approche repose sur un cumul temporel court terme des distorsions spatiales. Ce cumul temporel court terme est une modélisation fovéale du système visuel humain simulant l'évaluation des distorsions d'une vidéo réalisée au travers des mécanismes de sélection de l'attention visuelle.

L'évaluation de la pertinence des cartes de distorsions visuelles pour les images, ou des séquences de distorsions visuelles pour les vidéos n'a pu être que qualitative à cause de l'absence d'une vérité terrain qui aurait permis de réaliser cette évaluation. La construction d'une telle vérité à partir de tests subjectifs serait bénéfique pour toute la communauté travaillant sur le sujet. Cependant cette tâche est loin d'être simple et elle nécessiterait un véritable effort de recherche. Dans l'état actuel des connaissances, les tests subjectifs ne permettent que d'obtenir une note globale de qualité pour une image ou une vidéo, ou une série de notes pour une vidéo. Les performances des méthodes objectives d'évaluation sont donc évaluées à partir de ces notes. C'est d'ailleurs l'objet de la seconde partie de ce mémoire. Nous nous intéresserons à un autre besoin des concepteurs de systèmes de traitement d'images ou de vidéos. Il ne s'agit plus d'une évaluation locale des distorsions, mais plutôt d'une évaluation globale automatique de la qualité d'une image ou d'une vidéo.

Deuxième partie

Critères objectifs de qualité visuelle d'images et de vidéos

Introduction

La seconde partie de ce mémoire est consacrée à l'évaluation de la qualité d'images et de vidéos.

Après s'être intéressé dans la première partie de ce mémoire à l'évaluation locale des distorsions dans les images et les vidéos, cette seconde partie est dédiée à la construction du jugement de la qualité globale des images ou vidéos. Comme nous l'avons évoqué dans l'introduction générale, il est possible de considérer quatre niveaux d'analyse dans la construction du jugement de qualité. Alors que la première partie de ce mémoire se situait plutôt aux niveaux de la visibilité et de la perception des dégradations, la seconde partie se situe plutôt aux niveaux de la gêne qu'elles procurent et de la qualité visuelle qu'elles génèrent. La difficulté est de passer des niveaux de visibilité et de perception à un niveau de gêne puis à un niveau de qualité.

Dans cette partie, nous tentons de répondre à un besoin des concepteurs de systèmes de traitement d'images ou de vidéos en proposant des méthodes objectives d'évaluation de la qualité visuelle. Le produit de ces méthodes est une note globale de qualité tant pour l'image ou pour la vidéo considérée. Idéalement, cette note doit correspondre au jugement que donnerait un observateur humain standard, c'est-à-dire, celle que donnerait, en moyenne, un panel de taille suffisante d'observateurs.

Comme dans la partie précédente, les méthodes proposées seront des méthodes d'évaluation avec référence complète, ce qui sous-entend d'une part que la version à évaluer est comparée à une version de référence qui doit être disponible et d'autre part qu'il n'est pas nécessaire d'avoir des connaissances a priori sur les types de distorsions rencontrées.

Cette partie se décompose en trois chapitres.

Dans le chapitre 4 nous présentons un état de l'art d'une part sur les méthodes d'évaluation subjective de qualité et sur les techniques permettant d'évaluer les performances des métriques de qualité, d'autre part sur les métriques de qualité par elles-mêmes.

Le chapitre 5 est dédié à l'évaluation objective de la qualité d'images. Nous y proposerons des métriques de qualité fondées sur les travaux présentés dans le chapitre 2. Ces métriques sont évaluées quantitativement à partir des résultats issus d'un ensemble de tests subjectifs de qualité.

Le chapitre 6 est consacré à l'évaluation objective de la qualité de vidéos. Nous y proposerons une méthode innovante reposant sur les travaux présentés dans le chapitre 3. Les performances de cette métrique sont évaluées par comparaison avec les données issues de tests subjectifs de qualité.

Chapitre 4

État de l'art sur l'évaluation subjective et objective de la qualité visuelle d'images et de vidéos

4.1 Introduction

L'objet de ce chapitre est l'évaluation subjective et objective de la qualité d'images et de vidéos. Il s'agit dans les deux cas de donner une note représentant le niveau de qualité visuelle d'une image ou d'une vidéo. Contrairement à l'évaluation locale des distorsions, qui a fait l'objet de la première partie de ce mémoire, en évaluation de qualité il est possible de constituer une vérité terrain. Cette vérité est déduite de tests expérimentaux appelés tests subjectifs de qualité. Ces tests consistent à présenter des images ou des vidéos à des observateurs selon un protocole particulier, afin qu'ils évaluent leur qualité visuelle. L'organisation d'un test subjectif n'est pas triviale et la pertinence des résultats d'un test subjectif dépend de la façon dont celui-ci a été mené. Il convient de prêter une attention particulière aux éléments parasites et aux sources de biais, aussi bien dans l'élaboration et le déroulement du test que dans l'exploitation des résultats.

La première partie de ce chapitre sera consacrée aux tests subjectifs. Nous y décrirons les informations importantes et nécessaires à l'élaboration de tests subjectifs de qualité, les différents protocoles existants et la façon d'exploiter leurs résultats dans le but d'évaluer les performances de métriques de qualité. Dans la seconde partie de ce chapitre nous ferons une synthèse structurée sur les métriques objectives de qualité d'images et de vidéos. Nous décrirons des approches de type purement signal, des approches structurelles et des approches reposant sur une modélisation du système visuel humain. Dans la première partie de ce mémoire nous nous sommes intéressés à évaluer localement les distorsions perçues, ici ce qui nous intéresse ce sont les mécanismes de construction d'un jugement de qualité à partir des distorsions perçues.

4.2 Tests subjectifs d'évaluation de qualité

Cette section présente les conditions de déroulement des tests subjectifs d'évaluation de qualité qui permettent d'associer des notes de qualité à des images ou des vidéos (dégradées ou non). Ces notes de qualité représentent sous une forme très synthétique la perception que des observateurs ont de la qualité de ces images ou de ces vidéos. Ces tests permettent de constituer la vérité terrain qu'un critère objectif de qualité doit essayer d'approcher au mieux.

4.2.1 Tests subjectifs : maîtriser l'environnement

Les tests subjectifs nécessitent une normalisation des conditions de tests. Cette normalisation facilite l'appréciation des résultats et minimise l'influence de paramètres perturbateurs. L'I.T.U. (*International Telecommunication Union*) propose plusieurs recommandations dont par exemple la recommandation BT.500-10 [ITU-R Rec. BT.500-10 00]. Cette recommandation contient un certain nombre de règles pour normaliser l'environnement de test. Ces règles initialement prévues pour des tests de qualité d'images de télévision peuvent être utilisées plus généralement pour l'évaluation subjective de qualité d'images ou de vidéos. Dans cette section, trois éléments définissant la structure d'un test sont présentés : l'espace de visualisation, les observateurs et le déroulement d'une séance.

4.2.1.1 L'espace de visualisation

Les éléments les plus importants de l'espace de visualisation à maîtriser sont la distance d'observation, la luminosité ambiante et les caractéristiques de l'écran. La distance de visualisation a une influence directe sur la perception ; de cette distance dépend la répartition des fréquences spatiales de l'image ou de la vidéo projetée sur la rétine. Le contrôle de luminosité ambiante est important car il n'y a qu'une faible partie du champ visuel qui est excité par l'image ou la vidéo de test, le reste est excité par l'environnement. Celui-ci influence la perception en modifiant l'adaptation en luminance du système visuel. De plus, il est souhaitable d'adapter la luminosité ambiante afin de limiter la fatigue visuelle.

4.2.1.2 Les observateurs

La composition du panel d'observateurs est un point critique, car elle influence les résultats des tests. Parmi les critères influençant les résultats on peut citer : l'âge et le sexe des observateurs, la spécialisation professionnelle et les défauts de vision. Idéalement, le panel devrait être statistiquement représentatif de la population selon ces critères, excepté pour les défauts de vision. En effet, les observateurs ne peuvent faire partie du panel qu'à condition qu'ils n'aient pas de défaut de vision, ou qu'alors leurs défauts optiques soient corrigés (lunettes, lentilles, etc.). L'I.T.U. recommande que le panel d'observateurs soit constitué d'au moins 15 observateurs non initiés, c'est-à-dire qui ne sont pas confrontés dans leur activité professionnelle aux problématiques d'évaluation de qualité.

4.2.1.3 Les séances de test

L'organisation d'une séance de tests peut varier en fonction du protocole de tests utilisé. Cependant, une base commune existe pour les différents protocoles. Une séance de test est composée d'un certain nombre de présentations, chaque présentation correspondant à l'évaluation d'une image ou d'une vidéo potentiellement dégradée. Les différentes présentations d'une séance doivent être effectuées dans un ordre pseudo-aléatoire. De plus, les séances ne doivent pas dépasser trente minutes afin d'éviter la déconcentration et la fatigue visuelle. Une séance doit aussi comprendre des explications et quelques présentations préliminaires. Les explications doivent porter sur le protocole utilisé et sur l'échelle de notations. Les présentations préliminaires doivent permettre aux observateurs de stabiliser leur jugement en leur montrant des cas représentatifs de la gamme de qualité sur laquelle porte le test. Bien évidemment, les notations lors de ces séances préliminaires ne sont pas prises en compte dans les résultats finaux.

4.2.2 Tests subjectifs : les sources de biais

Les évaluations subjectives sont souvent considérées dans la littérature comme des références parfaites. Pourtant, c'est loin d'être le cas car il existe nombre de facteurs ayant une influence sur les résultats. La prise en compte de l'existence de ces interactions pendant l'élaboration des tests subjectifs peut permettre de limiter leur impact. De même que leur prise en compte à la suite de tests subjectifs peut permettre de corriger certains de leurs effets, ou au moins de mettre en perspective les résultats obtenus. Ces facteurs peuvent être rassemblés en trois types d'effets : l'effet contextuel, les styles cognitifs et les facteurs psychologiques.

4.2.2.1 L'effet contextuel

Cet effet exprime la dépendance de la réponse de l'observateur à un stimulus donné en fonction des stimuli précédents. Un premier type d'effet contextuel, dit de « dynamique », est observé lorsque seulement une portion de l'échelle de notation est utilisée par les observateurs. Un moyen de corriger cet effet est de proposer aux observateurs des conditions d'ancrage (*anchor conditions*) [Corriveau 99] correspondant aux dégradations extrêmes qu'ils vont rencontrer pendant le test. Si ce problème de dynamique est observé non pas sur l'ensemble des observateurs, mais de façon différente en fonction des observateurs, il est possible d'effectuer une correction des notes avant d'en calculer la moyenne. La correction à appliquer est la transformée en *Z-score* que nous décrivons section 4.2.4.4. Un autre effet contextuel est mis en évidence lorsque la note subjective varie en fonction de l'intensité des dégradations dans la présentation précédente. Par exemple, une image modérément dégradée qui serait notée de façon plus sévère après une présentation contenant une image faiblement dégradée qu'après une présentation contenant une image fortement dégradée. Pour limiter cet effet, on choisit généralement l'ordre des présentation de façon aléatoire, même si cette méthode a été critiquée dans [Corriveau 99]. Un dernier effet contextuel, plus difficile à maîtriser, est la multidimensionnalité de la qualité. Les dégradations que les observateurs doivent juger peuvent être de différentes natures. Il est possible alors que les observateurs utilisent dans ce cas une échelle de dégradation interne propre à chaque type de défaut. Les effets de blocs par exemple,

engendrent une gêne importante systématique alors que le flou est parfois perçu comme une dégradation plus « naturelle ».

4.2.2.2 Les styles cognitifs

Les styles cognitifs correspondent au fait que les observateurs ne perçoivent pas les stimuli de manière identique, indépendamment de la physiologie du système visuel humain [Sallio 77]. Certains observateurs sont portés directement vers des détails très localisés alors que d'autres ont une approche plus synthétique. Les styles cognitifs font partie des principaux facteurs qui interviennent dans une évaluation subjective. C'est un fait indépendant de la stabilité des observateurs et qui relève de l'individu et de ses caractéristiques cognitives propres. L'expérience ou la sensibilisation des observateurs à l'évaluation de qualité a une influence sur le style cognitif, et donc sur le jugement. Par exemple, un expert aura tendance à regarder certains détails pour juger une vidéo, alors qu'un observateur naïf risque d'avoir une approche plus globale. L'aspect culturel peut aussi avoir une influence sur le style cognitif. On peut citer une étude de Nisbett *et al.* [Nisbett 05] qui consistait à enregistrer les mouvements oculaires d'étudiant chinois et américains à qui étaient présentées des photos présentant une région fortement saillante au premier plan et un arrière plan complexe. Cette étude montre que les étudiants asiatiques portaient plus leur attention sur le fond que les étudiants américains, ces derniers ayant essentiellement concentré leur attention sur les objets saillants.

Une méthode pour limiter une partie du biais introduit par les styles cognitifs est d'insister sur les explications qui doivent être claires et rigoureuses. Il faut bien expliquer l'objectif du test, son déroulement et la tâche que l'on demande à l'observateur.

4.2.2.3 Les facteurs psychologiques

Comme les styles cognitifs, les facteurs psychologiques sont liés aux observateurs. La disposition psychologique de chaque observateur influence de manière importante les résultats des tests subjectifs d'évaluation de qualité. L'initiation, la motivation et l'attention sont trois facteurs psychologiques importants pouvant influencer ces résultats [Sallio 77]. Les effets dus à l'initiation peuvent être réduits par des présentations préliminaires et disparaissent généralement après deux ou trois séances. Les effets dus à la motivation peuvent être limités en impliquant les observateurs le plus possible à l'expérimentation : explication sérieuse du contexte d'application et de l'importance de l'expérience, association des commentaires éventuels des observateurs dans l'exposé des résultats, présentation aux observateurs des conclusions d'une campagne de tests. Les effets dus à l'attention peuvent être réduits en limitant la durée des séances de tests, mais il est également préférable que l'ordre de présentation permette d'équilibrer, séance après séance les différents effets (fatigue, adaptation, etc.).

4.2.3 Les différents protocoles de tests subjectifs

Les tests subjectifs d'évaluation de qualité consistent à présenter à des observateurs des images ou des vidéos potentiellement dégradées et dont ils doivent juger la qualité. Ces images ou ces vidéos peuvent être accompagnées de leur version originale.

Les tests subjectifs d'évaluation de qualité peuvent être divisés en trois grandes familles :

- les méthodes comparatives,
- les méthodes à simple stimulus,
- les méthodes à double stimuli.

Les conditions de déroulement de ces différentes méthodes sont normalisées par les recommandations de l'I.T.U. [ITU-R Rec. BT.500-10 00].

4.2.3.1 Méthodes comparatives

Ces méthodes consistent à présenter deux images ou deux vidéos aux observateurs qui doivent caractériser la relation entre les deux présentations. Pour ce faire, les observateurs peuvent avoir à disposition deux types d'échelle d'évaluation : une échelle d'évaluation par catégorie ou une échelle d'évaluation continue. Les deux types d'échelle indiquent la présence de différences perceptibles, et parfois le degré de différences perceptibles, comme celles indiquées sur la figure 4.1. Les échelles par catégorie limitent le choix des observateurs à un ensemble de catégories définies sémantiquement, alors que les échelles continues offrent davantage de souplesse en permettant aux observateurs de choisir tout point d'une droite tracée entre les différents qualificatifs.

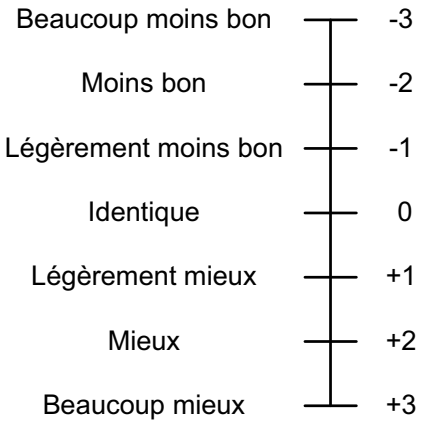


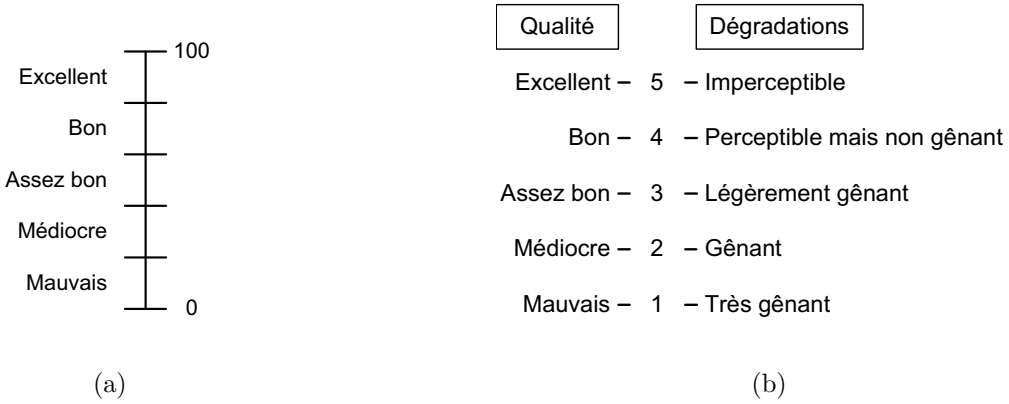
Figure 4.1 – Échelle comparative de l'I.T.U.

4.2.3.2 Méthodes à simple stimulus

Ces méthodes consistent à présenter à l'observateur une seule image, ou une seule vidéo à partir de laquelle il doit juger et noter la qualité globale. Elles sont nommées SSCQS (*Single Stimulus Continuous Quality Scale*) si elles utilisent une échelle continue de notation de la qualité, ou SSIS (*Single Stimulus Impairment Scale*) si l'échelle de qualité n'est constituée que de quelques catégories. Le sigle ACR (*Absolute Category Rating*) est aussi utilisé dans la littérature pour désigner les méthodes à simple stimulus de type SSIS.

En pratique, les échelles continues de qualité sont en réalité discrétisées mais conservent un nombre de catégories supérieures à celui dont l'observateur est conscient. Souvent, une échelle à cent valeurs est utilisée.

Les différents types d'échelle sont présentés sur la figure 4.2(a,b). Pour les échelles à cinq catégories, l'I.T.U. recommande l'utilisation d'une échelle à cinq notes (de qualité ou de dégradations). Cependant, si la dynamique des dégradations est importante, des échelles à six ou sept notes, voire même à onze notes, peuvent permettre une évaluation plus juste des dégradations importantes [CCIR 94].



(a) (b)

Figure 4.2 – Exemple d'échelle de notation : (a) continue permettant d'évaluer la qualité d'une image ou d'une vidéo, (b) par catégories de l'I.T.U. permettant d'évaluer la qualité (à gauche), ou les dégradations (à droite) d'une image ou d'une vidéo.

Dans le cadre de l'évaluation subjective de la qualité de vidéos, d'autres méthodes ont été imaginées. Contrairement aux méthodes SSCQS ou SSIS ne permettant d'obtenir qu'une note globale de la vidéo à évaluer, la méthode appelée SSCQE (*Single Stimulus Continuous Quality Evaluation*) [Alpert 97] permet d'obtenir une évaluation continue (2 notes par seconde) de la qualité de la vidéo. Dans cette méthode, les observateurs notent la qualité de la vidéo présentée au moyen d'un dispositif coulissant qu'ils déplacent dans un sens, ou dans l'autre, sur une échelle de notation continue, en fonction de leur perception momentanée de la qualité de la vidéo. La durée des vidéos présentées avec cette méthode est généralement plus importante qu'avec les méthodes SSCQS ou SSIS (de l'ordre de dix secondes en SSCQS ou SSIS, contre plusieurs minutes en SSCQE).

4.2.3.3 Méthodes à double stimuli

A la différence des méthodes à simple stimulus, ces méthodes consistent à présenter à l'observateur deux images ou deux vidéos : une version de référence et une version à évaluer qui est potentiellement dégradée. Comme pour les méthodes à simple stimulus SSCQS et SSIS, les méthodes à double stimuli sont nommées DSCQS (*Double Stimuli Continuous Quality Scale*) si elles reposent sur une échelle continue de notation de qualité, et DSIS (*Double Stimuli Impairment Scale*) si cette échelle de qualité ne contient que quelques catégories. Le sigle DCR (*Degradations Category Rating*) est aussi utilisé dans la littérature pour désigner les méthodes à double stimuli de type DSIS.

Dans la méthode DSCQS les images ou les vidéos sont présentées par paire. Chaque paire contient une référence (non dégradée) et une version à juger. Il n'y a pas d'a priori sur l'ordre des présentations et chaque

paire peut être présentée plusieurs fois avant que l'observateur note la qualité. L'observateur doit noter la qualité des deux présentations. La figure 4.3 illustre le déroulement de l'évaluation d'une paire avec la méthode DSCQS.

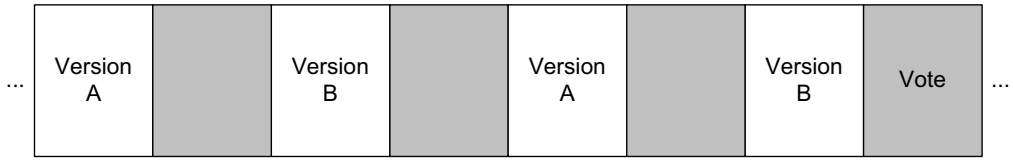


Figure 4.3 – Illustration de la méthode DSCQS. Les versions A et B sont présentées deux fois. La référence peut être indifféremment la version A ou la version B.

La méthode DSIS consiste, elle, à présenter aux observateurs les images ou les vidéos dans un ordre particulier : d'abord la référence (clairement identifiée), puis seulement après la version dont la qualité est à évaluer. A la suite de quoi l'observateur doit noter la qualité de la deuxième version par rapport à la première version (la référence). Chaque séance obéit à plusieurs règles (par exemple sa durée est limitée à une quarantaine de présentations). Une présentation comprend quatre phases :

- T1 : affichage de la référence,
- T2 : temps mort de séparation,
- T3 : affichage de la version à évaluer,
- T4 : temps de vote.

La figure 4.4 illustre le déroulement d'une présentation avec la méthode DSIS.

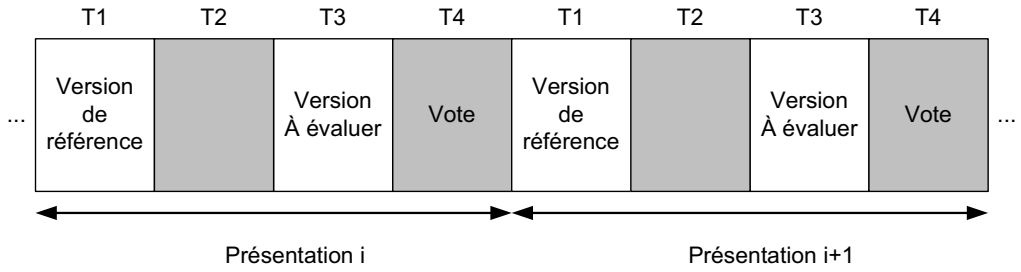


Figure 4.4 – Illustration de la méthode DSIS sur deux présentations consécutives.

L'I.T.U. recommande d'utiliser une échelle à cinq notes (de qualité ou de dégradations). Certains auteurs utilisent des échelles différentes comme le laboratoire LIVE [1] qui effectue des tests avec des notes entre un et cent.

Dans le cadre de l'évaluation subjective de la qualité de vidéos, il existe aussi une méthode d'évaluation continue de la qualité à double stimuli. Cette méthode est appelée DSCQE (*Double Stimulus Continuous Quality Evaluation*) [Alpert 97]. Le fonctionnement est le même que pour la méthode SSCQE, à la différence que la vidéo originale est présentée en même temps que la vidéo à évaluer. Dans cette méthode, les observateurs notent la

1. `http://live.ece.utexas.edu/research/quality`

qualité de la vidéo présentée au moyen d'un curseur qu'ils déplacent dans un sens ou dans l'autre selon les variations de qualité qu'ils observent. De même que pour les méthodes à simple stimulus, la durée des vidéos présentées avec cette méthode est généralement plus importante qu'avec les méthodes DSCQS ou DSIS.

4.2.3.4 Méthodes à stimuli multiple

Ces méthodes consistent à présenter à l'observateur plusieurs images ou vidéos à évaluer en laissant à l'observateur une grande liberté sur l'ordre de présentation des différentes versions. Dans la méthode SAMVIQ (*Subjective Assessment Methodology for Video Quality*), l'observateur doit évaluer une référence dite cachée car non identifiée explicitement par l'observateur et plusieurs versions dégradées. Cette méthode est très différente des autres méthodes présentées en particulier dans la façon dont les vidéos sont présentées à l'observateur. Il est donné beaucoup plus de liberté à l'observateur qui peut revoir plusieurs fois chaque version et qui peut corriger des notations. L'observateur peut comparer les versions dégradées entre elles, ainsi que par rapport à la référence explicite. Dans cette méthode une échelle de notation continue est utilisée (cf. figure 4.2(a)). L'intérêt de cette méthode est de rendre plus cohérentes les diverses notations et donc de réduire les erreurs de notation. Cependant, en contrepartie de la liberté donnée à l'observateur, la durée des tests n'est plus réellement maîtrisée. La durée de présentation de chaque vidéo est comparable avec les autres méthodes d'évaluation non continue (de l'ordre de dix secondes).

4.2.3.5 Récapitulatif des protocoles

Les différents protocoles de tests subjectifs sont récapitulés dans le tableau 4.1.

	Évaluation globale (images ou vidéos)		Évaluation continue (vidéos)
	Échelle par catégorie	Échelle continue	
Méthodes à simple stimulus	SSIS/ACR	SSCQS	SSCQE
Méthodes à double stimuli	DSIS/DCR	DSCQS	DSCQE
Méthodes à multiple stimuli	×	SAMVIQ	–
Méthodes comparatives	×	×	–

Table 4.1 – Récapitulatif des différents protocoles de tests subjectifs (Symbole × : existe mais pas de nom connu ; symbole – : n'existe pas).

4.2.4 Traitement des résultats obtenus lors de tests

A cause de la multiplicité des facteurs influents dans les tests et de la nature même, variable, des jugements subjectifs, les notes de qualité collectées auprès des observateurs peuvent être biaisées. Les observateurs peuvent avoir jugé plus sévèrement des images qui étaient moins dégradées que d'autres images qu'ils ont notées moins sévèrement. Les observateurs n'ont pas non plus forcément utilisé la même dynamique sur l'échelle de notation. Les notes doivent donc être « filtrées » avant de pouvoir être utilisées.

4.2.4.1 Note moyenne de qualité (MOS)

Suite à des tests subjectifs d'évaluation de qualité, la première étape consiste à déterminer la note de qualité (MOS : *Mean Opinion Score*) de chaque image ou vidéo présentée. Pour cela, le MOS est estimé par la moyenne des notes fournies par un panel d'observateurs jugeant indépendamment la qualité :

$$MOS_{jk} = \frac{1}{N_{obs}} \sum_{i=1}^{N_{obs}} X_{ijk} \tag{4.1}$$

avec :

- N_{obs} : nombre d'observateurs ayant noté les image ou les vidéos,
- X_{ijk} : note fournie par l'observateur i ayant noté l'image ou la vidéo issue de l'image ou vidéo originale j et ayant subi la version k de la dégradation.

4.2.4.2 Intervalle de confiance à 95%

Pour évaluer la fiabilité des résultats obtenus, à chaque note moyenne (MOS_{jk}) est associée un intervalle de confiance. Nous prendrons l'intervalle de confiance à 95% classiquement utilisé, selon lequel 95% des réponses se trouvent dans l'intervalle :

$$[MOS_{jk} - e_{jk}, MOS_{jk} + e_{jk}] \tag{4.2}$$

avec (en considérant que MOS_{jk} suit une loi gaussienne) :

$$e_{jk} = 1.96 \times \frac{\sqrt{\frac{1}{N_{obs}-1} \sum_{i=1}^{N_{obs}} (X_{ijk} - MOS_{jk})^2}}{\sqrt{N_{obs}}} \tag{4.3}$$

4.2.4.3 Rejet des observateurs

Lors d'un test, il peut arriver qu'un observateur fournisse des résultats non cohérents en donnant de bien meilleures notes aux images assez dégradées qu'aux images très peu dégradées ou en donnant des notes très différentes à des images de qualités comparables. Il est donc souhaitable de détecter ces erreurs afin d'exclure ces notes incohérentes des résultats. L'I.T.U recommande la procédure suivante :

- rejet des notations différentes d'au moins deux catégories (sur une échelle à cinq catégories) pour la même image,
- élimination des observateurs incohérents.

La cohérence des observateurs est mesurée comme suit :

- on calcule MOS_{jk} et son écart-type σ_{jk},
- on vérifie que la distribution est normale en calculant le coefficient d'aplatissement de la loi des variables aléatoires (qui est le rapport du moment du quatrième ordre sur le carré du moment du deuxième ordre, appelé aussi coefficient de Kurtosis) :

$$\beta_{\mathbf{2}jk} = \frac{\frac{1}{N_{obs}} * \sum_{i=1}^{N_{obs}} (X_{ijk} - MOS_{jk})^4}{(\frac{1}{N_{obs}} * \sum_{i=1}^{N_{obs}} (X_{ijk} - MOS_{jk})^2)^2} \tag{4.4}$$

  Si $\beta_{\mathbf{2}jk}$ est compris entre 2 et 4, la distribution est considérée comme normale.

– il faut alors déterminer P_i et Q_i, de la manière suivante :

$$\begin{array}{lll} \text{si} & X_{ijk} \geq MOS_{jk} + 2 * \sigma_{jk} & \text{(pour une distribution normale)} \\ \text{ou} & X_{ijk} \geq MOS_{jk} + \sqrt{20} * \sigma_{jk} & \text{(pour une distribution non normale)} \\ \text{alors} & P_i = P_i + 1 & \end{array} \tag{4.5}$$

$$\begin{array}{lll} \text{si} & X_{ijk} \leq MOS_{jk} - 2 * \sigma_{jk} & \text{(pour une distribution normale)} \\ \text{ou} & X_{ijk} \leq MOS_{jk} - \sqrt{20} * \sigma_{jk} & \text{(pour une distribution non normale)} \\ \text{alors} & Q_i = Q_i + 1 & \end{array} \tag{4.6}$$

Si $\frac{P_i+Q_i}{N_{images}} > 0.05$ et $\frac{P_i-Q_i}{P_i+Q_i} < 0.3$ alors l'observateur est éliminé des résultats. Le premier rapport reflète les écarts importants par rapport à la moyenne et le deuxième rapport indique si ces écarts se produisent toujours dans le même sens (positifs si P_i est important, négatifs si Q_i est important). Pour qu'un observateur soit éliminé des résultats, il faut qu'il ait un jugement trop variable et que cette variabilité s'exprime aussi bien positivement que négativement. Ainsi, un observateur qui sur-évalue ou sous-évalue constamment la qualité par rapport aux autres observateurs verra ses notes conservées et la dynamique de ses notes sera normalisée.

4.2.4.4 Normalisation de la dynamique utilisée : transformation en Z-scores et transformation inverse

Comme indiqué dans la section 4.2.2.1, les observateurs peuvent n'utiliser qu'une partie de l'échelle de notation disponible, suivant qu'ils soient initiés ou pas. Pour pallier ce problème, une transformation en Z-scores peut s'avérer utile. Cette méthode peut servir à corriger les notes de chaque observateur pour les ramener à une même dynamique comparable entre observateur.

Un Z-score donne une information sur la différence normalisée d'une note par rapport à la moyenne d'un observateur. Le Z-score est calculé de la manière suivante :

$$Z_{ijk} = \frac{X_{ijk} - \overline{X_i}}{\sigma_i} \tag{4.7}$$

avec :

$$\overline{X_i} = \frac{1}{N_{\text{deg}}} \frac{1}{N_{\text{im}}} \sum_{j=1}^{N_{\text{deg}}} \sum_{k=1}^{N_{\text{im}}} X_{ijk} \tag{4.8}$$

$$\sigma_i^2 = \frac{1}{N_{\text{deg}} - 1} \frac{1}{N_{\text{im}} - 1} \sum_{j=1}^{N_{\text{deg}}} \sum_{k=1}^{N_{\text{im}}} (X_{ijk} - \overline{X_i})^2 \tag{4.9}$$

où :

– N_{deg} est le nombre de dégradations par image ou vidéo originale,

– N_{im} est le nombre d'images ou vidéos originales,

– X_{ijk} est la note donnée par l'observateur i, à une image ou vidéo (j, k) dégradée ou non.

La transformation inverse en Z-scores permet de revenir à des notes de qualité comprises dans l'échelle de notation. Elle consiste à transformer linéairement les Z-scores de manière à retrouver la dynamique qu'avaient les notes de qualité, tous observateurs confondus, avant la transformation.

4.2.5 Conclusion

Dans cette section nous avons décrit différents aspects à prendre en compte dans l'élaboration et la conduite de tests subjectifs de qualité. Le problème est complexe et les sources de biais sont nombreuses. Nous avons vu que l'influence de nombreux effets parasites peut être réduite d'une part grâce aux recommandations existantes sur l'environnement des tests, d'autres part grâce à plusieurs techniques de traitement et d'analyse des résultats des tests subjectifs.

Les principaux protocoles de tests d'évaluation subjective de la qualité ont également été présentés. Le choix parmi tous ces protocoles n'est pas évident, même si certains protocoles sont plus faciles à écarter que d'autres. Dans notre cas, la durée des tests est un critère de choix, ainsi les méthodes comparatives et les méthodes à multiple stimuli sont écartées pour cette raison. Concernant les méthodes à évaluation continue, elles ne sont pas adaptées à l'obtention d'une note globale. Le choix parmi les autres méthodes est moins clair.

4.3 Indicateurs de performance de critères objectifs de qualité visuelle

Afin d'évaluer la pertinence de métriques de qualité, il est nécessaire de confronter les notes prédites avec le jugement des observateurs. Pour cela il est nécessaire de constituer une base de test d'images ou de vidéos. La qualité des images ou des vidéos de cette base de test est ensuite évaluée subjectivement au moyen de tests subjectifs d'évaluation de qualité et objectivement par les métriques de qualité dont on veut évaluer les performances. Pour évaluer les performances d'une métrique de qualité, on dispose donc de deux séries de donnés :

- les MOS, issus des tests subjectifs et correspondant au jugement d'un observateur moyen,
- les notes objectives, issues de la métrique objective à évaluer, notées $MOSp$.

Les couples $(MOS, MOSp)$, obtenus pour chaque image ou vidéo de la base de test peuvent être représentés sous forme de graphe comme celui présenté figure 4.5.

Si ces graphes donnent une indication qualitative des performances d'une métrique de qualité, des indicateurs numériques sont plus pratiques à utiliser. Le groupe de travail international de standardisation des méthodes d'évaluation de qualité VQEG (*Video Quality Experts Group*[2]) utilise et recommande plusieurs indicateurs pour évaluer les performances d'une métrique de qualité [VQEG 00]. Ces indicateurs expriment la précision (coefficient de corrélation linéaire), la monotonicité (coefficient de corrélation de rang), et la cohérence (*outlier ratio*) des notes objectives $MOSp$ par rapport aux mesures subjectives. Compte tenu de la façon dont les humains construisent un jugement catégoriel sous forme d'une note, le groupe de travail VQEG recommande d'utiliser une transformation non linéaire qui permet de passer de mesures objectives de qualité Q_{obj} à des notes prédites de qualité (MOSp) pour les comparer avec les MOS. Cette transformation non linéaire est de type fonction

2. `http://www.vqeg.org/`

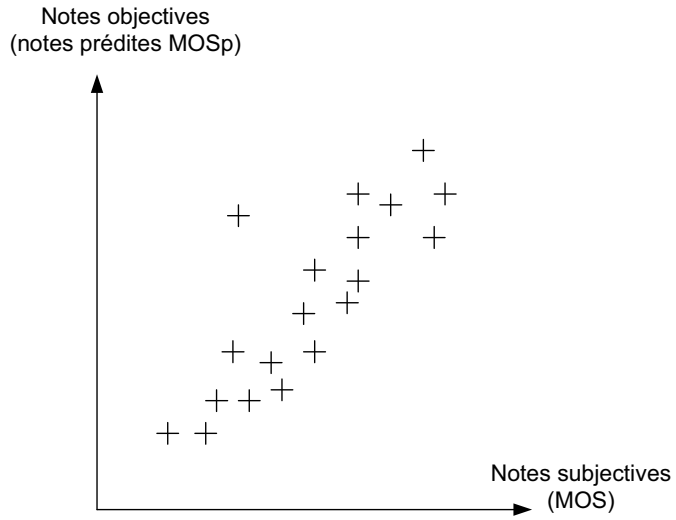


Figure 4.5 – Représentation graphique des couples $(MOS, MOSp)$.

psychométrique :

$$\text{MOSp} = \frac{b_1}{1 + e^{-b_2 \cdot (Q_{obj} - b_3)}}\,, \tag{4.10}$$

où b_1, b_2 et b_3 sont les trois paramètres de la fonction psychométrique. Le calcul des indicateurs de performance est effectué sur l'ensemble des couples $(MOS, MOSp)$. Cette transformation permet de transposer dans la même dynamique (celle des MOS), la dynamique propre à chaque métrique de qualité Q_{obj}. Elle permet en outre d'effectuer une correction non-linaire de la dynamique des métriques. Cette transformation permet donc la comparaison de différentes métriques de qualité en rendant pertinente l'utilisation des indicateurs de performance que nous allons décrire dans les sections suivantes.

4.3.1 Coefficient de corrélation linéaire : indicateur de précision

Le coefficient de corrélation linéaire (*Pearson linear correlation coefficient*), noté CC exprime la dépendance linéaire entre les mesures objectives MOSp et les notes subjectives MOS. Il est donné par la relation suivante :

$$CC = \frac{\sum_{j=1}^{N_{\text{deg}}} \sum_{k=1}^{N_{\text{im}}} (MOS_{jk} - \overline{MOS})(MOSp_{jk} - \overline{MOSp})}{\sqrt{\sum_{j=1}^{N_{\text{deg}}} \sum_{k=1}^{N_{\text{im}}} (MOS_{jk} - \overline{MOS})^2} \cdot \sqrt{\sum_{j=1}^{N_{\text{deg}}} \sum_{k=1}^{N_{\text{im}}} (MOSp_{jk} - \overline{MOSp})^2}}\,, \tag{4.11}$$

avec :

- MOS_{jk} et $MOSp_{jk}$ étant respectivement les MOS et MOSp pour l'image ou la vidéo dégradée issue de l'image ou vidéo originale k ayant subi la dégradation j,
- N_{im} et N_{deg} étant respectivement le nombres d'images ou de vidéos originales et le nombre de versions dégradées,
- $\overline{MOS}$ et $\overline{MOSp}$ étant respectivement le MOS moyen et le MOSp moyen.

La valeur du coefficient de corrélation linéaire est comprise entre -1 et 1. Plus sa valeur est proche de -1 ou de 1, plus la dépendance linéaire entre les deux séries de nombres est forte.

4.3.2 Coefficient de corrélation de rang : indicateur de monotonicité

Le coefficient de corrélation de rang, noté SROCC (*Spearman rank order correlation coefficient*), est une mesure de la monotonicité, c'est-à-dire qu'il caractérise le degré avec lequel les mesures objectives MOSp et les notes subjectives MOS évoluent dans le même sens. Ces deux séries de nombres évoluent dans le même sens si la fonction $f(MOS) = MOSp$ (ou $f(MOSp) = MOS$) est monotone. Pour calculer le SROCC, les MOS et MOSp de toutes les images ou vidéos sont ordonnés (par ordre croissant ou décroissant) afin de déterminer le rang de chacun, puis le SROCC est donné par la relation suivante :

$$SROCC = 1 - \frac{6 \sum_{j=1}^{N_{\text{deg}}} \sum_{k=1}^{N_{\text{im}}} d_{jk}^2}{N_{tot}(N_{\text{tot}}^2 - 1)}\,, \tag{4.12}$$

avec :

- d_{jk}^2 : différence de classement, ou de rang, entre le MOS et MOSp de l'image ou de la vidéo jk (image ou vidéo originale k ayant subi la dégradation j),
- N_{im}, N_{deg} et N_{tot} représentent respectivement le nombre d'images ou de vidéos originales, le nombre de dégradations, le nombre total d'images ou vidéos évaluées ($N_{\text{tot}} = N_{\text{im}} \cdot N_{\text{deg}}$).

Un coefficient de corrélation de rang proche de 1 est recherché car alors cela signifie que la métrique de qualité classe les images ou les vidéos, de la moins bonne à la meilleure qualité, selon le même ordre que les observateurs (SROCC = -1 indiquerait un classement dans l'ordre inverse).

4.3.3 Outlier ratio : indicateur de cohérence

Cet indicateur permet de mesurer l'aptitude de la métrique de qualité à prédire une note de qualité qui ne soit pas trop éloignée du MOS. Cet indicateur, appelé *Outlier ratio* et noté OR, exprime le nombre d'images ou de vidéos mal notées par rapport au nombre d'images ou de vidéos testées. Il est déterminé par le rapport suivant :

$$OR = \frac{\text{Nombre de notes aberrantes}}{\text{Nombre total de notes}}\,, \tag{4.13}$$

où une note est déclarée « aberrante » si elle est en dehors de l'intervalle (intervalle de confiance à 95% environ) :

$$[MOS_{jk} - 2SE_{jk}, MOS_{jk} + 2SE_{jk}]\,, \tag{4.14}$$

où SE_{jk} désigne l'erreur type de la moyenne MOS_{jk} des notes subjectives pour l'image ou la vidéo originale k ayant subi la dégradation j. L'erreur type SE_{jk} (*standard error*) est calculé à partir de l'écart type σ_{jk} selon la relation :

$$SE_{jk} = \frac{\sigma_{jk}}{\sqrt{N_{obs}}}\,, \tag{4.15}$$

où N_{obs} correspond au nombre d'observateurs ayant noté l'image ou la vidéo jk. La proportion la plus faible possible de notes « aberrantes » est souhaitée.

4.3.4 Erreur de prédiction de la qualité

Dans la littérature l'erreur quadratique moyenne, ou plutôt sa racine carrée RMSE (*Root mean square error*), est utilisée comme indicateur complémentaire. Le RMSE permet de mesurer la distance $L2$ entre les mesures objectives MOSp et les notes subjectives MOS. Elle est donnée par la relation suivante :

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N_{\text{deg}}} \sum_{k=1}^{N_{\text{im}}} (MOS_{jk} - MOSp_{jk})^2}{N_{\text{im}} \cdot N_{\text{deg}}}}, \tag{4.16}$$

avec :

- MOS_{jk} et $MOSp_{jk}$ représentent respectivement les MOS et MOSp pour l'image ou la vidéo originale k ayant subi la dégradation j,
- N_{im} et N_{deg} représentent respectivement le nombres d'images ou de vidéos originales, le nombre de dégradations.

Plus RMSE est faible, plus la dispersion des MOSp autour de leur MOS respectif est faible. Cet indicateur est souvent utilisé pour comparer plusieurs métriques de qualité entre elles. La métrique de qualité ayant la plus faible RMSE est recherchée.

4.3.5 Tests de significativité ou tests statistiques de différence significative

En comparant les performances respectives de plusieurs métriques de qualité sur la même base de test d'images ou de vidéos, il est possible de positionner les métriques les unes par rapport aux autres. Cependant, il ne suffit pas que l'indicateur de performance d'une métrique soit supérieur à celui d'une autre pour que cette métrique lui soit significativement supérieure. Pour vérifier ce résultat il est nécessaire d'effectuer des tests statistiques afin d'établir si la différence entre les performances des métriques est significative ou pas. Dans cette section plusieurs tests statistiques classiquement utilisés dans la littérature sont présentés.

4.3.5.1 Test de significativité sur les coefficients de corrélation

Il existe une méthode statistique pour comparer deux coefficients de corrélation (*Pearson correlation*). Cette méthode est d'ailleurs utilisée par le groupe de travail VQEG dans le rapport [VQEG 03]. La première étape de ce test consiste à appliquer la transformation de Fisher (*Fisher z' transformation*) sur les coefficients de corrélation. Cette transformation est justifiée par le fait que la distribution des coefficients de corrélation de Pearson ne suit pas une loi normale. Cette transformation est donnée par :

$$z' = \frac{1}{2}(ln(1 + CC) - ln(1 - CC)), \tag{4.17}$$

où CC est le coefficient de corrélation de Pearson.

Il s'agit ensuite de construire l'intervalle de confiance entre deux coefficients de corrélation CC_1 et CC_2. Ces coefficients sont donc transformés en z'_1 et z'_2, et l'intervalle de confiance est alors défini par la relation :

$$z'_1 - z'_2 \pm z\sigma_{z'_1 - z'_2}. \tag{4.18}$$

La valeur de z est définie à partir d'une table (*z table*) et de la précision que l'on veut donner à l'intervalle de confiance. Par exemple, pour un intervalle de confiance à 95%, la valeur de z sera 1.96. La valeur de l'erreur standard $\sigma_{z'_1-z'_2}$ est donnée par la relation suivante :

$$\sigma_{z'_1-z'_2} = \sqrt{\frac{1}{N_1-3}.\frac{1}{N_2-3}} \tag{4.19}$$

où N_1 et N_2 correspondent au nombre de paires de valeurs ayant servi à calculer les coefficients CC_1 et CC_2 respectivement. Il reste ensuite à calculer la valeur des bornes de l'intervalle de confiance puis à leur appliquer la transformation inverse de *Fisher z'* afin de vérifier si la différence $CC_1 - CC_2$ est comprise, ou pas dans cet intervalle. Si la différence $CC_1 - CC_2$ n'est pas comprise dans cet intervalle, alors la différence entre les deux coefficients de corrélation est considérée comme significative.

4.3.5.2 Test de significativité sur les valeurs de différence résiduelle entre les MOS et les MOSp

Pour tester si deux métriques sont significativement différentes, il est possible d'effectuer un F-test sur les valeurs de différence résiduelle entre les MOS et les MOSp, cette méthode est utilisée par exemple dans [Sheikh 06b, Chandler 07].

L'hypothèse nulle du F-test est que la variance des valeurs de différence résiduelle $MOS - MOSp_1$ d'une métrique est égale à la variance des valeurs de différence résiduelle $MOS - MOSp_2$ d'une autre métrique. Le F-test nous donne alors la probabilité que l'hypothèse nulle ne soit pas rejetée. Il suffit ensuite de comparer cette valeur avec la valeur de confiance du test de significativité. Par exemple, pour une confiance à 95%, il faut que la probabilité que l'hypothèse nulle ne soit pas rejetée, soit inférieure à 0.05. L'utilisation de ce test suppose que la distribution des variances suive une loi normale. Cette supposition doit être testée comme partie intégrante du test. Cependant, dans les cas où la base de test n'est pas de taille assez importante, il peut être difficile de vérifier cette hypothèse.

4.4 Métriques de qualité visuelle avec référence complète

L'évaluation objective de la qualité d'images ou de vidéos consiste à produire une mesure de qualité la plus proche possible de celle que fournirait en moyenne un panel d'observateurs lors de tests d'évaluations de la qualité visuelle. Les métriques de qualité peuvent permettre par exemple de comparer différentes méthodes de codage ou de compression d'information d'images ou de vidéos afin de déterminer leur performance en terme de qualité visuelle.

Comme nous l'avons déjà évoqué en introduction générale, on trouve dans la littérature de nombreuses métriques de qualité. Ces métriques peuvent être réparties selon les trois catégories suivantes :

- les métriques avec référence complète, pour lesquelles la version originale et la version à évaluer doivent être disponibles,
- les métriques avec référence réduite, pour lesquelles la version à évaluer et une description (réduite) de la version originale doivent être disponibles,

– les métriques sans référence, pour lesquelles seule la version à évaluer est nécessaire.

Notre objectif étant la création de métriques de qualité avec référence complète, nous ne présenterons que la littérature les concernant. Parmi cette catégorie de métriques de qualité nous distinguerons, les métriques reposant sur la création d'une carte d'erreurs des autres approches. Comme nous l'avons évoqué dans la première partie, la création d'une carte d'erreurs est souvent la première étape de bon nombre de métriques de qualité. Cette étape ayant déjà fait l'objet de la première partie, le lecteur pourra s'y reporter pour plus de détails. Nous nous intéresserons plutôt à la seconde étape qui consiste à cumuler les erreurs en une note de qualité.

4.4.1 Approches reposant sur le calcul de cartes de distorsions : cumul spatial des distorsions

Dans les approches reposant sur le calcul de cartes de distorsions, une étape de cumul des distorsions est nécessaire pour construire la note de qualité. Dans le cas des métriques pour images fixes, le cumul des distorsions est un cumul purement spatial. La figure 4.6 représente la structure générale d'une métrique de qualité d'images reposant sur le calcul de cartes d'erreurs. L'étape de cumul des erreurs y est encadrée en rouge.

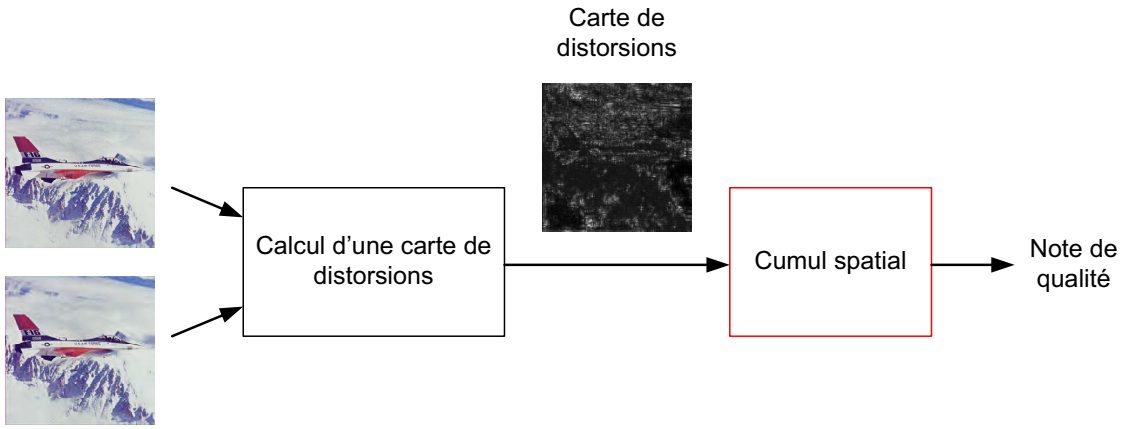


Figure 4.6 – Structure générale d'une métrique de qualité d'images reposant sur le calcul de cartes de distorsions.

4.4.1.1 Approches purement de type signal

L'approche de type signal la plus utilisée en image est le PSNR (*Peak Signal to Noise Ratio*). Il est tout particulièrement utilisé en compression afin de quantifier les performances des codeurs en mesurant la qualité de reconstruction d'une image par rapport à la version originale. Le PSNR est défini par :

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{d^2}{EQM}\right), \tag{4.20}$$

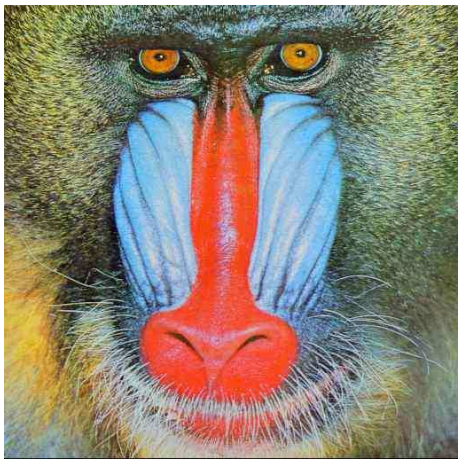
où d est l'amplitude maximale (crête) du signal. Dans le cas standard d'une image où les composantes d'un pixel sont codées sur 8 bits, $d = 255$.

EQM est l'erreur quadratique moyenne (*MSE, Mean square error*), définie par :

$$\mathrm{EQM} = \frac{1}{M \cdot N} \sum_{m=1}^{M} \sum_{n=1}^{N} \|I_o(m,n) - I_d(m,n)\|^2, \tag{4.21}$$

où I_o et I_d sont respectivement l'image originale et l'image dégradée, M et N les dimensions des images. Dans ce cas, le cumul spatial des erreurs est une moyenne.

Les approches purement de type signal ne modélisent pas le système visuel humain, elles font l'hypothèse que la qualité visuelle décroît quand la distorsion du signal augmente. En fait, la qualité visuelle peut rester inchangée si les distorsions engendrées restent sous le seuil différentiel de visibilité. La qualité ne dépend pas seulement des distorsions mais aussi de nombreux autres paramètres comme le contenu de l'image, ou encore de la localisation des dégradations. Les mesures de qualité fournies par les critères classiques du traitement du signal ne sont pas bien corrélées avec le jugement humain. La figure 4.7 illustre les mauvaises performances du PSNR sur deux images.



(a)

(b)

Figure 4.7 – (a) image *Mandrill* dégradée (MOS=4.62, PSNR=26.83), et (b) image *Lena* dégradée (MOS=1.12, PSNR=28.95). D'après les MOS, l'image *Mandrill* et moins dégradée que l'image *Lena*, alors que le PSNR indique l'inverse.

L'image *Lena* est notée moins sévèrement par le PSNR que l'image *Mandrill*, alors qu'il apparaît clairement, en regardant ces deux images, que l'image *Mandrill* est de meilleure qualité. Cette observation qualitative est confirmée par les MOS issus de tests subjectifs de qualité. Une explication vient de ce que l'image *Mandrill* contient beaucoup de détails fins ayant un potentiel de masquage important et donc un niveau plus élevé du seuil différentiel de visibilité. Les erreurs sont souvent masquées ou faiblement perçues.

4.4.1.2 Les approches structurelles

Dans les approches structurelles d'évaluation d'images, dont la construction des cartes d'erreurs a été exposée section 2.2.2, le cumul est également assez basique. Dans la version purement spatiale [Wang 04a], la note de qualité, notée MSSIM est la valeur moyenne des similarités locales :

$$\text{MSSIM} = \frac{1}{M \cdot N} \sum_{m=1}^{M} \sum_{n=1}^{N} SSIM(I_o(m,n), I_d(m,n)) \,, \tag{4.22}$$

où I_o et I_d représentent respectivement l'image originale et l'image dégradée, M et N les dimensions des images. La même fonction de cumul spatial est utilisée dans la version multi-échelle [Wang 03].

4.4.1.3 Les approches basées sur une modélisation du système visuel humain

Dans les approches modélisant le système visuel humain, le cumul des erreurs est classiquement réalisé au moyen d'une sommation de Minkowski [Winkler 00, Le Callet 01]. Si l'on considère une métrique de qualité d'images dont la carte des distorsions est $E(m,n)$, alors la note globale de distorsions D est calculée comme suit :

$$D = (\frac{1}{M \cdot N} \sum_{m=1}^{M} \sum_{n=1}^{N} (E(m,n))^{\beta})^{1/\beta}, \tag{4.23}$$

Dans ces métriques, la modélisation du système visuel porte sur la conception des cartes de distorsions. La façon de construire le cumul spatial n'est pas liée explicitement à une modélisation du système visuel humain. Cependant, la sommation de Minkowski permet de donner plus d'importance aux erreurs les plus importantes dans l'élaboration de la note finale (avec un exposant supérieur à 1), ce qui semble plus proche du jugement humain que de prendre une simple moyenne des erreurs. Plus l'exposant β est grand, plus les distorsions de forte amplitude sont renforcées. Si on fait tendre β vers l'infini, l'expression revient à prendre la valeur maximale des distorsions. Pour $\beta = 1$, l'expression revient à prendre la moyenne des distorsions.

4.4.2 Approches reposant sur le calcul de séquences temporelles de cartes de distorsions : cumul spatial et temporel des distorsions

Les approches reposant sur le calcul d'une séquence temporelle de cartes de distorsions nécessitent aussi une étape de cumul des distorsions pour construire la note de qualité. Dans le cas des métriques pour vidéos, les distorsions doivent être cumulées à la fois spatialement et temporellement. La figure 4.8 représente la structure générale d'une métrique de qualité reposant sur le calcul d'une séquence de cartes de distorsions. L'étape de cumul des distorsions y est encadrée en rouge.

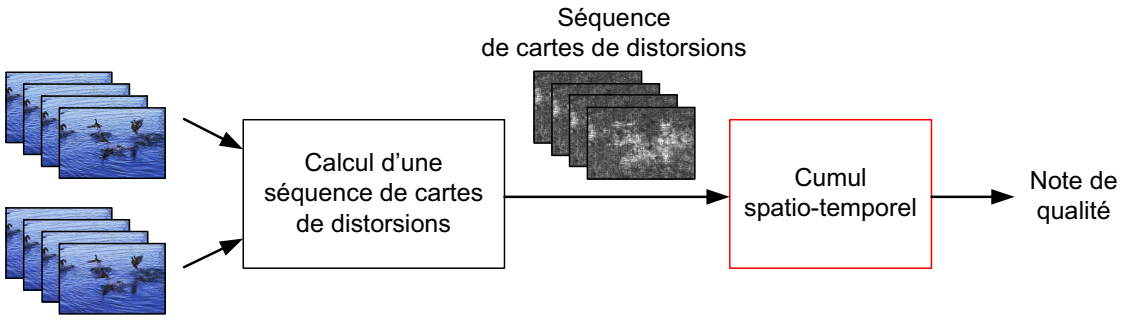


Figure 4.8 – Structure générale d'une métrique de qualité de vidéos reposant sur le calcul de séquences de cartes de distorsions.

4.4.2.1 Approches purement de type signal

L'approche de type signal la plus utilisée en vidéo numérique est aussi le PSNR (cf. relation (4.20)). Dans la version du PSNR appliquée à la vidéo, l'EQM est définie pour 2 séquences d'images par :

$$\text{EQM} = \frac{1}{M \cdot N \cdot T} \sum_{t=1}^{T} \sum_{m=1}^{M} \sum_{n=1}^{N} \|I_o(m,n,t) - I_d(m,n,t)\|^2 \,, \tag{4.24}$$

où I_o et I_d sont respectivement les images originales et les images dégradées, M et N sont les dimensions des images et T est le nombre d'images de la vidéo.

Comme nous l'avons vu dans la section 4.4.1.1, les approches de type signal ne modélisent pas le système visuel humain, ce qui pose des problèmes dans l'évaluation spatiale des distorsions. Dans le cas des métriques pour vidéos un autre problème vient s'ajouter : le cumul temporel de ces approches mathématiques ne prend pas en compte l'impact perceptuel des variations temporelles des distorsions.

4.4.2.2 Les approches structurelles

Dans les approches structurelles d'évaluation de séquences d'images, dont la construction des cartes d'erreurs a été exposée section 3.2.2, les cumuls sont assez simples. Dans la version étendue à la vidéo [Wang 04b], le cumul des erreurs est décomposé en deux étapes. La première étape est un cumul spatial des cartes d'erreurs structurelles réalisé au moyen d'une moyenne pondérée, et permettant d'obtenir une note Q_i par image i :

$$Q_i = \frac{\sum_{j=1}^{R_s} w_{ij} \text{MSSIM}_{ij}}{\sum_{j=1}^{R_s} w_{ij}} \,, \tag{4.25}$$

où w_{ij} est une pondération dépendante de la luminance explicitée relation (3.1), et R_s correspondant à l'échantillonnage spatial utilisé pour le calcul des valeurs de SSIM. Contrairement à la version spatiale où les valeurs de SSIM sont calculées sur toutes les fenêtres possibles, dans la version étendue à la vidéo les auteurs ont observé expérimentalement qu'en sélectionnant convenablement les fenêtres leur nombre pouvait être considérablement diminué. Cette sélection a pour effet de réduire le temps de calcul tout en maintenant un niveau correct de robustesse de la mesure.

La seconde étape est le cumul des notes de chaque image réalisé également au moyen d'une moyenne pondérée, et permettant d'obtenir une note globale Q :

$$Q = \frac{\sum_{j=1}^{F} W_i Q_i}{\sum_{i=1}^{F} W_i} \,, \tag{4.26}$$

où F est le nombre d'images de la séquence, et W_i est une pondération des notes par image dépendante du mouvement inter-image et de la luminance :

$$W_i = \begin{cases} \sum_{j=1}^{R_S} w_{ij} & M_i \leq 0.8 \\ ((1.2 - M_i)/0.4). \sum_{j=1}^{R_S} w_{ij} & 0.8 < M_i \leq 1.2 \\ 0 & M_i > 1.2 \end{cases} \tag{4.27}$$

où M_i représente la quantité de mouvement dans l'image i. Le paramètre M_i est calculé comme suit :

$$M_i = \frac{(\sum_{j=1}^{R_S} m_{ij})/R_S}{K_M}, \tag{4.28}$$

où m_{ij} représente la norme du vecteur de mouvement de la fenêtre d'échantillonnage j de l'image i, et K_M est un facteur de normalisation de la quantité de mouvement par image. Les inconvénients de cette méthode ont été introduits section 3.2.2, on peut leur rajouter que le cumul temporel des distorsions ne tient pas réellement compte de leurs variations temporelles mais seulement du mouvement dans la vidéo d'origine.

4.4.2.3 Les approches basées sur une modélisation du système visuel humain

Comme nous l'avons déjà évoqué dans la section 4.4.1.3 pour l'évaluation des images fixes, les approches modélisant le système visuel humain utilisent classiquement une sommation de Minkowski pour effectuer le cumul des distorsions. Dans ce cas, les distorsions doivent être cumulées dans les dimensions spatiale et temporelle :

$$D = (\frac{1}{M \cdot N \cdot T} \sum_{t=1}^{T} \sum_{m=1}^{M} \sum_{n=1}^{N} (E(m,n,t))^{\beta})^{1/\beta}, \tag{4.29}$$

où $E(m,n,t)$ est la carte des distorsions perceptuelles pour l'image t. M et N sont les dimensions des images et T est le nombre d'images de la vidéo. L'exposant β étant parfois différent entre la dimension spatiale et la dimension temporelle. Comme dans le cas du cumul spatial des distorsions, plus l'exposant est grand et supérieur à 1, plus les distorsions de forte amplitude sont renforcées.

On peut citer par exemple Winkler, qui dans son PDM [Winkler 99], utilise une sommation de Minkowski pour le cumul des différentes sous-bandes, pour le cumul spatial et pour le cumul temporel. De même, Van den Branden Lambrecht utilise une sommation de Minkowski dans l'étape de cumul des métriques de qualité de vidéos proposées dans [van den Branden Lambrecht 96c] et [van den Branden Lambrecht 96a]. Watson aussi utilise une sommation de Minkowski pour cumuler les résultats des différentes dimensions du modèle sur lequel repose sa métrique DVQ [Watson 01]. La sommation de Minkowski est une méthode de cumul bien connue qui permet de donner plus d'importance aux erreurs les plus importantes dans l'élaboration de la note finale (avec un exposant supérieur à 1). Cependant, cette fonction de cumul ne permet pas non plus de prendre en compte les variations temporelles de qualité.

4.4.3 Autres approches

Dans la littérature, on trouve d'autres approches qui ne sont pas basées sur la création de cartes d'erreurs. Nous ne nous étendrons pas sur ces approches qui sont en marge de notre sujet d'étude. Ces approches se focalisent en général sur la mesure et la combinaison de caractéristiques des images ou des vidéos, ayant un sens en terme d'évaluation de la qualité. On citera cependant l'approche de la *National Telecommunications and Information Administration (NTIA)* pour ses bonnes performances dans les tests du groupe de travail VQEG [VQEG 03]. NTIA a développé une métrique de qualité pour de la vidéo (*Video Quality Metric : VQM*) [Pinson 04] adopté par l'ANSI (*American National Standards Institute*) comme une norme nationale

aux États-Unis [ANSI T1.801.03 03] et comme norme internationale au travers des recommandations de l'ITU [IUT-T Rec. J.144 04, IUT-R Rec. BT.1683 04]. Les recherches de NTIA se sont concentrées sur le développement de paramètres indépendants des technologies et modélisant la façon dont les observateurs humains perçoivent la qualité vidéo. Ces paramètres sont ensuite combinés linéairement pour obtenir une note de qualité. Le modèle général contient sept paramètres indépendants. Quatre paramètres sont basés sur les caractéristiques spatiales extraites des gradients de la composante de luminance Y. Deux paramètres sont basés sur les caractéristiques extraites du vecteur (C_B, C_R) formé par les deux composantes de chrominance. Un paramètre est basé sur le produit de caractéristiques qui mesurent le contraste et le mouvement, et qui sont toutes les deux extraites de la composante de luminance Y. Ce dernier paramètre porte sur le fait que la perception spatiale des distorsions peut être influencée par la quantité de mouvement.

Les paramètres sont calculés en divisant la vidéo en régions spatio-temporelles. Cette division varie d'un paramètre à l'autre. Pour obtenir la valeur de chaque paramètre, des fonctions de cumul (appelées *collapsing function*) spatial et temporel sont utilisées. Ces fonctions sont différentes d'un paramètre à l'autre. Pour les fonctions de cumul spatial, les auteurs utilisent : la moyenne des valeurs, l'écart type des valeurs, la moyenne de n% des valeurs ordonnées. Pour le cumul temporel, ils utilisent : la moyenne des valeurs ou sélectionnant le niveau du n^{ieme} centile supérieur ou inférieur. A défaut d'être toujours justifiées perceptuellement, les fonctions de cumul utilisées sont plus élaborées que des valeurs moyennes ou des sommations de Minkowski.

4.4.4 Conclusion

Nous venons de voir que la construction du jugement de qualité à partir des distorsions est très souvent réalisée à partir de fonctions de cumul assez simple. La sommation de Minkowski, avec des valeurs d'exposant différentes, est d'ailleurs très souvent utilisée dans les métriques objectives de qualité d'images fixes, ainsi que dans les métriques de vidéos.

Pour les images on peut s'en contenter comme une première approximation, même si une modélisation du système visuel humain plus poussée peut être envisagée comme nous le verrons dans la troisième partie de ce mémoire avec les mécanismes de sélection de l'attention visuelle. Cette première approximation nous permettra, dans le chapitre 5 d'évaluer quantitativement les performances de métriques de qualité reposant sur les modélisations du système visuel humain proposées dans le chapitre 2.

Par contre pour les vidéos, il nous semble important de tenir compte des variations temporelles des distorsions, et se limiter à une sommation de Minkowski nous semble trop simpliste. Contrairement à la construction de séquences temporelles de distorsions (cf. chapitre 3), il ne s'agit plus des variations de distorsions au niveau des fixations (cumul court terme), mais il s'agit cette fois des variations de distorsions au niveau de la séquence entière (cumul long terme). L'élaboration d'un cumul temporel long terme sera d'ailleurs l'objet du chapitre 6.

4.5 Conclusion

Ce chapitre était consacré à l'état de l'art sur l'évaluation subjective et objective de la qualité d'images et de vidéos. Concernant l'évaluation subjective de la qualité, nous avons insisté sur l'importance de maîtriser l'environnement pendant des tests subjectifs afin de limiter l'influence des facteurs perturbateurs comme par exemple les conditions d'observation. Les sources de biais, comme les styles cognitifs ou les facteurs psychologiques, ont été soulignés et des moyens de réduire leurs effets ont été proposés. Nous avons ensuite présenté les protocoles existants en les regroupant en méthodes comparatives, en méthodes à simple stimulus, en méthodes à double stimuli et en méthode à stimuli multiple. Ensuite, des techniques permettant d'évaluer les performances de métriques de qualité à partir des résultats de tests subjectifs d'évaluation de la qualité ont été présentées.

Concernant l'évaluation objective de la qualité d'images ou de vidéos, nous avons passé en revue différentes approches dites « à référence complète » de la littérature. Ces approches évaluent la qualité d'images ou de vidéos lorsque la version originale et la version dégradée sont toutes les deux disponibles. Nous nous sommes intéressés plus particulièrement aux approches dont la première étape consiste à calculer des cartes d'erreurs et dont la seconde étape réside dans le cumul de ces erreurs. La première étape étant l'objet de la première partie de ce mémoire, nous n'avons décrit ici que la seconde étape.

Dans les chapitres suivants nous allons ajouter l'étape de cumul des distorsions aux modélisations proposées dans la première partie de ce mémoire. Pour l'évaluation de la qualité des images, une méthode de cumul simple nous permettra d'évaluer quantitativement les performances de ces modélisations. Pour les vidéos, nous proposerons une nouvelle méthode de cumul temporel long terme qui viendra compléter le cumul temporel court terme proposé dans le chapitre 3. Les performances de cette métrique de qualité seront aussi évaluées quantitativement.

Chapitre 5

Critères objectifs de qualité visuelle d'images

5.1 Introduction

Comme nous l'avons évoqué précédemment, l'évaluation objective de la qualité d'images fait partie des besoins de l'industrie du traitement d'images ou de vidéos. D'où la nécessité de développer des méthodes objectives permettant d'évaluer la qualité visuelle de façon automatique. Idéalement, ces méthodes doivent permettre la construction d'une note de qualité visuelle correspondant au jugement que donnerait un observateur standard. L'objet de ce chapitre est de présenter des critères objectifs de qualité visuelle d'images avec référence complète et d'évaluer leurs performances. Les critères objectifs de qualité présentés sont fondés sur les évaluations locales des distorsions proposées dans le chapitre 2 et sur un cumul spatial de ces distorsions locales. Les métriques ainsi obtenues seront évaluées quantitativement à partir des résultats de tests subjectifs de qualité ainsi qu'à partir des techniques présentées dans le chapitre précédent.

La première partie de ce chapitre est dédiée à la fonction de cumul spatial utilisée. La seconde partie est consacrée à l'évaluation des modèles proposés. Cette partie commence par la description des tests subjectifs dont sont issues les notes subjectives nécessaires à l'évaluation des performances, pour se terminer par la présentation des résultats et de leurs commentaires.

5.2 Cumul spatial

L'objectif du cumul spatial est d'obtenir une note objective représentant la qualité visuelle d'une image à partir des distorsions perçues dans cette image. Cette note est obtenue en combinant les distorsions visuelles spatiales. En première approximation, nous avons choisi d'utiliser un cumul spatial classique mais perceptuellement plausible, à savoir une sommation de Minkowski. Avec une valeur d'exposant adaptée ($\beta > 1$), et contrairement à une simple moyenne des distorsions, cette fonction de cumul permet d'accentuer la contribution des distorsions les plus importantes à la construction de la note finale. Ce comportement est plausible avec la construction d'un jugement de qualité d'un observateur car ce sont les distorsions les plus importantes auxquelles celui-ci est le

plus sensible. Cependant, ceci ne représente qu'un aspect limité de la construction du jugement de qualité. Comme nous l'avons déjà évoqué, nous explorerons dans la troisième partie de ce mémoire d'autres aspects liés au système visuel pouvant influencer la construction du jugement de qualité. La sommation de Minkowski est aussi utilisée dans les modèles du système visuel humain sur lesquels sont basés les méthodes d'évaluation locale des distorsions proposées dans le chapitre 2. Elle est utilisée pour effectuer les cumuls fréquentiels des différentes sous-bandes : cumul selon les orientations et cumul selon les fréquences radiales.

Dans les métriques proposées, la note de qualité Q est calculée à partir de la carte de distorsions visuelles spatiales $VE_{m,n}$ (cf. section 2.8) :

$$Q = \left(\frac{1}{M \cdot N} \sum_{m=1}^{M} \sum_{n=1}^{N} \left(VE_{m,n} \right)^{\beta_s} \right)^{\frac{1}{\beta_s}}, \tag{5.1}$$

où M et N sont respectivement la hauteur et la largeur des images, β_s est l'exposant de Minkowski.

Ce cumul spatial est appliqué sur une carte de distorsions visuelles spatiales $VE_{m,n}$. La méthode de cumul reste la même pour les deux modèles proposés : le modèle fondé sur l'analyse de Fourier (appelé FQA) et le modèle fondé sur une décomposition en ondelettes (appelé WQA).

5.3 Expérimentations

5.3.1 Bases d'évaluation subjective

Les notes subjectives d'évaluation de qualité d'images utilisées dans cette section sont issues de différents tests subjectifs réalisés sur deux bases d'images.

5.3.1.1 *Base IVC*

L'une des bases, que nous appellerons *base IVC*, est issue des travaux de Le Callet [Le Callet 01]. Elle a été élaborée à partir de dix images de scènes naturelles. Ces images, illustrées figure 5.1, sont de taille 512×512 pixels et ont été sélectionnées parmi celles communément utilisées par les concepteurs de méthodes de compression d'images. Pour chacune d'elles, des versions plus ou moins dégradées ont été générées en utilisant trois systèmes dégradants :

- une compression JPEG, celle-ci est basée sur une quantification et un codage opérant sur des blocs transformés par DCT ;
- une compression JPEG2000, qui utilise la transformée en ondelettes séparables ;
- un système de dégradation d'images introduisant du flou par filtrage linéaire 2D (de type gaussien).

Ces trois systèmes dégradants permettent d'obtenir des distorsions de natures différentes et couvrant une grande partie des types de dégradations rencontrées. L'utilisation de plusieurs taux de compression et de plusieurs niveaux de flou conduisent à un total de 120 images dégradées.

Les évaluations subjectives de qualité ont été menées à une distance de visualisation de quatre fois la hauteur de l'image affichée sur un écran CRT et dans des conditions normalisées [ITU-R Rec. BT.500-10 00]. Le protocole

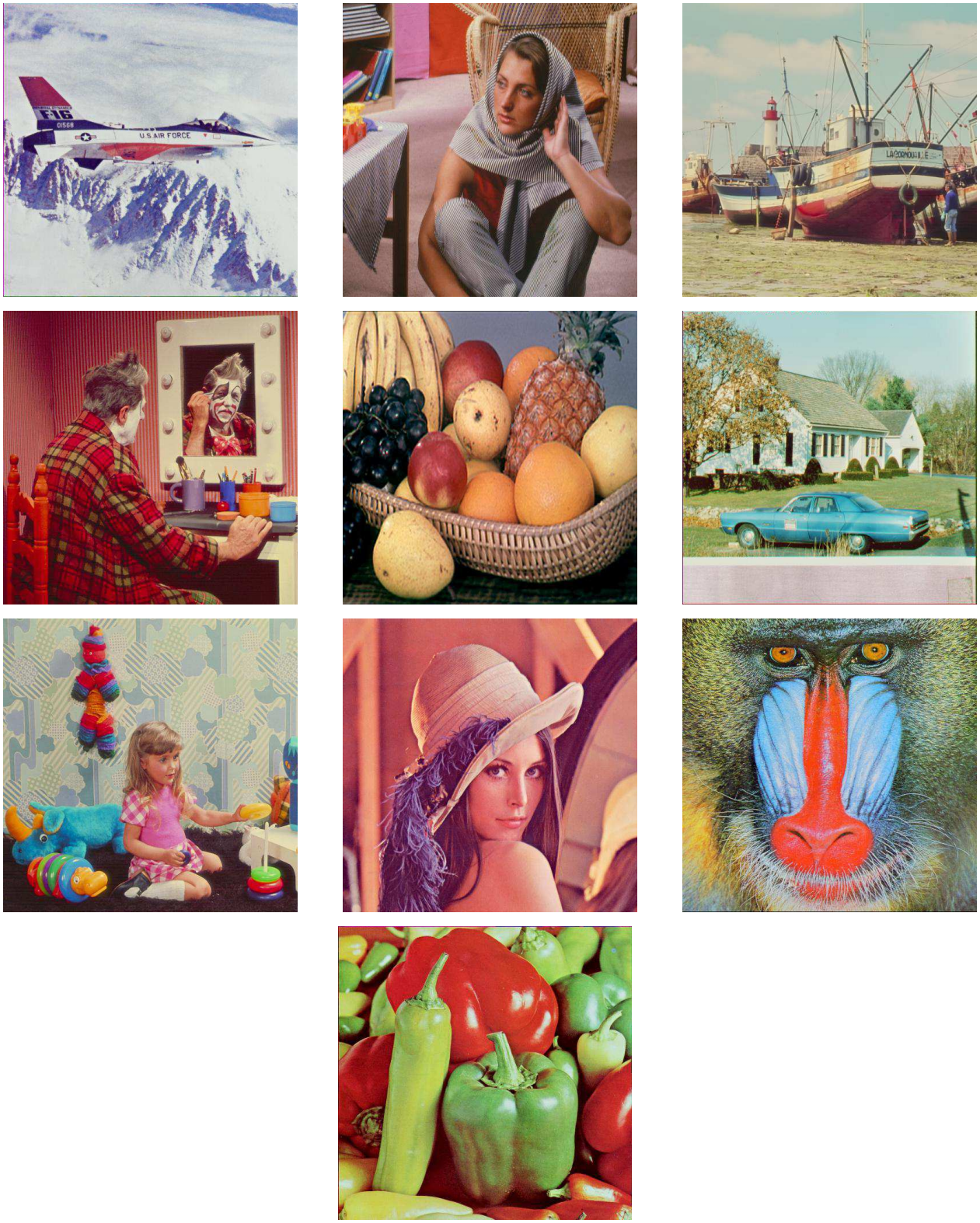


Figure 5.1 – Images originales de la base de test *IVC*.

de test DSIS (*Double Stimulus Impairment Scale*) a été utilisé avec une échelle de dégradations à cinq catégories (cf. chapitre 5).

Les images ont été notées par vingt observateurs ayant une acuité visuelle normale (avec ou sans correction optique) selon le test de Monoyer et une perception normale des couleurs selon le test d'Ichihara. Les observateurs n'étaient pas des experts en traitement d'images et ne connaissaient pas l'expérimentation (la nature des dégradations).

Les données de qualité visuelle ont été traitées selon les méthodes décrites dans la section 4.2.4, afin d'éliminer les notes *suspectes*. Avec le protocole de test DSIS, les images originales ne sont pas évaluées, il n'est donc possible de déduire de ces expérimentions que le MOS de chaque image dégradée.

5.3.1.2 *Base Toyama*

La seconde base de données d'évaluations subjectives, que nous appellerons *base Toyama*, est issue des travaux de Sazzad *et al.* [Sazzad 07] de l'Université de Toyama [1] au Japon. Elle a été élaborée à partir de quatorze images de scènes naturelles. Ces images, illustrées figure 5.2, sont de taille 768×512 pixels. Pour chacune d'elles des versions plus ou moins dégradées ont été générées en utilisant deux systèmes dégradants :

- une compression JPEG ;
- une compression JPEG2000.

Ces deux systèmes dégradants permettent d'obtenir des distorsions de natures différentes, et couvrant une grande partie des types de dégradations rencontrées en compression d'images. L'utilisation de plusieurs taux de compression permet d'obtenir un total de 168 images dégradées.

Les évaluations subjectives ont été menées à une distance de visualisation de quatre fois la hauteur de l'image affichée sur l'écran et également dans des conditions normalisées [ITU-R Rec. BT.500-10 00]. Deux campagnes d'évaluations subjectives différentes ont été menées sur cette base. Pour les deux campagnes de test, le protocole de test ACR (*Absolute Category Rating*) a été utilisé avec une échelle de dégradations à cinq catégories. Les notes subjectives provenant de ces deux campagnes de tests seront référencées sous les noms de *base Toyama1* et *base Toyama2*. La campagne de tests *base Toyama1* a été réalisé à l'Université de Toyama [Sazzad 07], alors que la campagne de tests *base Toyama2* a été réalisé à l'Université de Nantes [2] [Tourancheau 08]. Pour la *base Toyama1*, les images étaient présentées sur un écran CRT, alors que pour la base *base Toyama2* les images étaient présentées sur un écran LCD.

Les images ont été notées par des observateurs ayant une acuité visuelle normale (sans ou avec correction optique) selon le test de Monoyer et une perception des couleurs normale selon le test d'Ichihara. Les observateurs n'étaient pas des experts en traitement d'image et ne connaissaient pas l'expérimentation. Pour la *base Toyama1*, les observateurs étaient japonais et au nombre de seize, alors que pour la *base Toyama2* les observateurs étaient français et au nombre de vingt sept.

1. *Graduate School of Engineering, University of Toyama*
2. Laboratoire IRCCyN, Université de Nantes

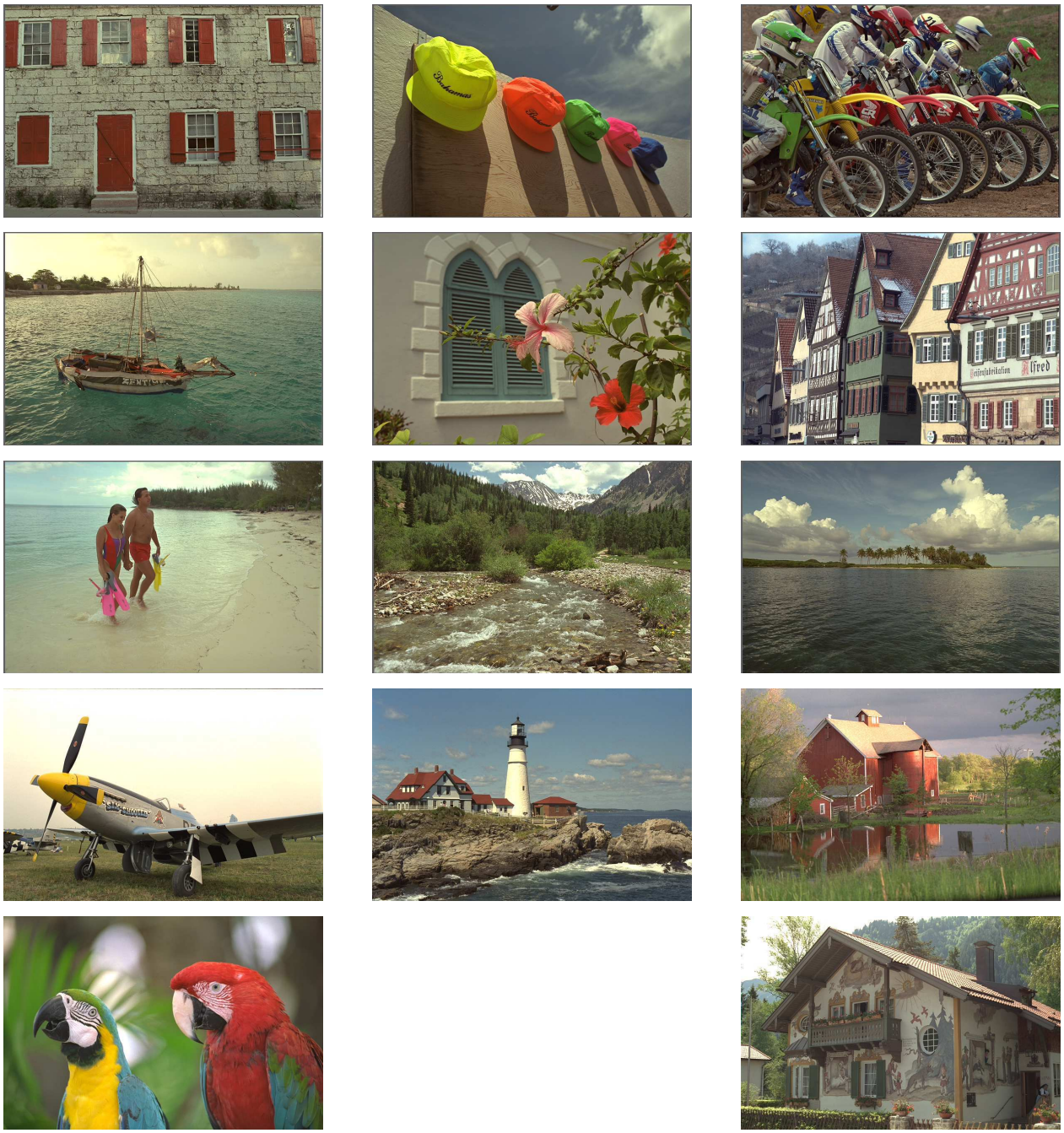


Figure 5.2 – Images originales de la base de test *Toyama*.

Les notes subjectives ont été traitées selon les méthodes décrites dans la section 4.2.4, afin d'éliminer les notes *suspectes*. Avec le protocole de test ACR, les images originales sont elles aussi notées en référence cachée, il est donc possible de déduire de ces expérimentions à la fois des MOS et des DMOS (*Differential MOS*). Les DMOS sont obtenus en calculant, pour chaque image dégradée, la différence entre le MOS de la version originale (non dégradée) et le MOS de l'image dégradée considérée.

Nous disposons donc de deux bases d'images et de notes subjectives issues de trois campagnes de test. Les caractéristiques de ces campagnes de test sont résumées dans le tableau 5.1.

Campagne de test	*IVC*	*Toyama1*	*Toyama2*
Format	512×512	768×512	768×512
#Contenus/#Images dégradées	10/120	14/168	14/168
Protocole	DSIS	ACR	ACR
Conditions d'observation	4H (ITU-R BT 500.10)		
Écran	CRT	CRT	LCD
Population	Français (20)	Japonais (16)	Français (27)

Table 5.1 – Description des expérimentations subjectives.

5.3.2 Évaluation des performances

Dans cette section, nous allons évaluer les performances des modèles proposés (FQA et WQA). Nous allons également évaluer l'impact du masquage sur les performances en particulier l'impact du masquage semi-local (sLM). Nous rappelons que le masquage semi-local diffère du masquage de contraste pur de part le support spatial pris en compte. Les deux effets de masquage expliquent la modification de la sensibilité due à la présence de fort contraste, mais tandis que le masquage de contraste est appliqué localement quasiment point à point, le masquage semi-local considère explicitement une semi-localité autour d'un point, c'est-à-dire son proche voisinage (cf. section 1.3.5.1).

Les performances de quatre métriques de qualité sont évaluées et comparées (cf. tableau 5.2). Les deux premières sont deux versions du modèle FQA : sans puis avec masquage semi-local, notées respectivement FQA_1 et FQA_2. Les deux dernières sont deux versions du modèle WQA : sans et avec masquage semi-local, notées respectivement WQA_1 et WQA_2.

Le modèle de masquage de contraste pur est celui de Daly de base et le modèle de masquage semi-local est le modèle de Daly que nous avons amélioré (cf. chapitre 2).

Métriques	Décomposition en sous-bandes	Masquage de contraste	Masquage semi-Local
FQA_1	FFT (DCP)	$\surd$	
FQA_2 (sLM)	FFT (DCP)	$\surd$	$\surd$
WQA_1	DWT	$\surd$	
WQA_2 (sLM)	DWT	$\surd$	$\surd$

Table 5.2 – Les quatres métriques de qualité comparées.

Comme décrit section 4.2.4, les mesures issues des métriques objectives de qualité sont transformées en MOS

prédits (MOSp), au moyen d'une fonction psychométrique (cf. équation 4.10), avant d'être comparées. Afin de permettre aux lecteurs de se faire une opinion sur les bases de test, les performances de trois métriques bien connues de la littérature sont données en sus. La première est le PSNR, la seconde est la métrique « structurelle » SSIM et la troisième est la version multi-résolution de la SSIM (MS-SSIM). Dans les sections suivantes, nous évaluerons les performances des métriques sur la *base IVC*, puis sur la *base Toyama* (*Toyama1* et *Toyama2*).

Nous avons étayé la comparaison des performances par des tests statistiques de significativité (cf. section 4.3.5). Comme dans [Sheikh 06b], le test statistique utilisé est un F-test sur les résidus MOS-MOSp. L'hypothèse de nullité des résidus indique que la variance des résidus d'une métrique et la variance des résidus de l'autre métrique sont identiques.

5.3.2.1 Résultats sur la base IVC

Nous allons commencer par comparer les performances des métriques sur la *base IVC*. Les résultats sur l'ensemble de la *base IVC* sont présentés tableau 5.3. Les résultats des tests statistiques de significativité sont présentés dans le tableau 5.4.

Métriques	CC	SROCC	RMSE
FQA_1	0.897	0.895	0.549
FQA_2 sLM	0.941	0.938	0.422
WQA_1	0.892	0.896	0.562
WQA_2 sLM	0.923	0.921	0.48
PSNR	*0.768*	*0.77*	*0.795*
SSIM	*0.832*	*0.844*	*0.691*
MS-SSIM	*0.917*	*0.922*	*0.504*

Table 5.3 – Résultats sur la *base IVC* entière.

MOSp \ MOSp	PSNR	SSIM	MS-SSIM	FQA_1	FQA_2 sLM	WQA_1	WQA_2 sLM
PSNR	*1.0*	0.13573	0.00000 ($p < 0.05$)	0.00008 ($p < 0.05$)	0.00000 ($p < 0.05$)	0.00024 ($p < 0.05$)	0.00000 ($p < 0.05$)
SSIM	0.13573	*1.0*	0.00065 ($p < 0.05$)	0.01335 ($p < 0.05$)	0.00000 ($p < 0.05$)	0.02771 ($p < 0.05$)	0.00009 ($p < 0.05$)
MS-SSIM	0.00000 ($p < 0.05$)	0.00065 ($p < 0.05$)	*1.0*	0.34125	0.06180 ($p < 0.10$)	0.21978	0.60670
FQA_1	0.00008 ($p < 0.05$)	0.01335 ($p < 0.05$)	0.34125	*1.0*	0.00497 ($p < 0.05$)	0.78248	0.14293
FQA_2 sLM	0.00000 ($p < 0.05$)	0.00000 ($p < 0.05$)	0.06180 ($p < 0.10$)	0.00497 ($p < 0.05$)	*1.0*	0.00207 ($p < 0.05$)	0.17519
WQA_1	0.00024 ($p < 0.05$)	0.02771 ($p < 0.05$)	0.21978	0.78248	0.00207 ($p < 0.05$)	*1.0*	0.08193 ($p < 0.10$)
WQA_2 sLM	0.00000 ($p < 0.05$)	0.00009 ($p < 0.05$)	0.60670	0.14293	0.17519	0.08193 ($p < 0.10$)	*1.0*

Table 5.4 – Tests statistiques sur les résidus entre les MOS et les MOSp. Chaque valeur donne pour le couple de métriques (ligne,colonne) la probabilité que l'hypothèse nulle d'égalité des variances soit rejetée. Si la valeur est inférieure à 0.05, les deux métriques sont significativement différentes avec une confiance de 95%. Si la valeur est inférieure à 0.10, les deux métriques sont significativement différentes avec une confiance de 90%.

Les quatre modèles multi-canaux (FQA et WQA) présentent des performances significativement meilleures (confiance à 95%) que la SSIM et le PSNR, en termes de coefficient de corrélation linéaire (CC), de coefficient de corrélation de rang (SROCC) et de racine carrée d'erreur quadratique moyenne (RMSE). Dans le cas du critère de qualité SSIM, le ΔCC entre les modèles multi-canaux et la SSIM varie de +0.06 à +0.109. Un critère de qualité de type structurel comme SSIM semble donc moins performant pour prédire la qualité des images qu'une approche simulant le système visuel humain. Par contre la MS-SSIM présente des performances proches de celles des modèles FQA et WQA. La tendance semble privilégier les quatre modèles multi-canaux par rapport à la MS-SSIM, cependant il n'y a que le modèle FQA_2 qui lui soit significativement supérieur (confiance à 90%).

Le modèle FQA_1 (sans masquage semi-local) surpasse la version du modèle WQA_1 (sans masquage semi-local), en termes de RMSE et de CC. Le ΔCC entre ces deux métriques est de +0.05.

De la même manière le modèle FQA_2 obtient de meilleures performances que le modèle WQA_2, en termes de CC, SROCC et RMSE. Le ΔCC entre ces deux métriques est de +0.018.

Ces observations montrent que les modèles exploitant une décomposition fréquentielle (domaine de Fourier) produisent des performances très légèrement supérieures à celles des modèles basés ondelettes, même si ces différences ne sont pas significatives (confiance à 95%). L'explication réside sans doute dans une meilleure simulation du comportement multi-canal du système visuel humain par la décomposition en canaux perceptuels (DCP) que par la décomposition basée ondelettes. Cependant, les performances des modèles basés sur une décomposition en ondelettes restent tout à fait intéressantes, en particulier en ce qui concerne le modèle avec masquage semi-local dont le CC avec les notes subjectives atteint 0.923.

Les résultats différenciés selon la nature des images dégradées sont présentés tableaux 5.5 et 5.6. Les résultats des tests statistiques de significativité sont présentés dans le tableau 5.7. Les résultats sur le sous-ensemble d'images compressées JPEG2000 présentés dans le tableau 5.5, conduisent à la même conclusion. Toutefois, les résultats sur les sous-ensembles d'images JPEG, JPEG+JPEG200 et Flou, présentés dans les tableaux 5.5 et 5.6 sont différents. Pour les sous-ensembles JPEG et JPEG+JPEG2000, le modèle FQA_1 surpasse le modèle WQA_1, mais le modèle FQA_2 est surpassé par le modèle WQA_2. Sur le sous-ensemble d'images Flou, le modèle WQA_1 surpasse le modèle FQA_1 ce qui est tout à fait surprenant. Globalement on observe la supériorité des modèles FQA, même si les performances varient avec le type de distorsions. Toutefois, ces tendances ne sont pas confirmées par les tests de significativité.

	JPEG			JPEG2000		
Métriques	CC	SROCC	RMSE	CC	SROCC	RMSE
FQA_1	0.857	0.862	0.599	0.936	0.947	0.457
FQA_2 (sLM)	0.938	0.939	0.403	0.947	0.952	0.414
WQA_1	0.851	0.854	0.611	0.906	0.916	0.549
WQA_2 (sLM)	0.96	0.965	0.327	0.94	0.946	0.439

Table 5.5 – Résultats sur les sous-ensembles JPEG et JPEG2000 de la *base IVC*.

L'utilisation du masquage semi-local augmente systématiquement les performances du modèle en termes de

	JPEG+JPEG2000			Flou		
Métriques	CC	SROCC	RMSE	CC	SROCC	RMSE
FQA_1	0.899	0.904	0.544	0.88	0.837	0.542
FQA_2 (sLM)	0.938	0.941	0.431	0.949	0.932	0.36
WQA_1	0.877	0.886	0.598	0.97	0.943	0.277
WQA_2 (sLM)	0.943	0.942	0.415	0.912	0.893	0.47

Table 5.6 – Résultats sur les sous-ensembles JPEG+JPEG2000 et Flou de la *base IVC*

MOSp / MOSp	FQA_1	FQA_2 sLM	WQA_1	WQA_2 sLM
FQA_1	*0000* *0000*	1010 1011	0001 0001	1010 1010
FQA_2 sLM	1010 1011	*0000* *0000*	1010 1110	0000 0000
WQA_1	0001 0001	1010 1110	*0000* *0000*	1011 1011
WQA_2 sLM	1010 1010	0000 0000	1011 1011	*0000* *0000*

Table 5.7 – Tests statistiques sur les résidus entre les MOS et les MOSp par type de dégradations. Pour chaque couple de métriques (ligne,colonne) deux séquences de caractères ('0' ou '1') indiquent si l'hypothèse nulle d'égalité des variances est rejetée pour les différents sous-ensembles d'images. Le caractère '1' indique que la différence est significative, tandis que le caractère '0' indique que la différence n'est pas significative. La première séquence de caractères indique les résultats avec une confiance de 95%, tandis que la seconde séquence indique les résultats avec une confiance de 90%. Les séquences de caractères représentent, dans l'ordre, les résultats pour les sous-ensembles d'images : JPEG, JPEG2000, JPEG+JPEG2000 et Flou.

CC, SROCC et RMSE dans les deux configurations (FQA_1 vs FQA_2, WQA_1 vs WQA_2). Cette observation est valable sur l'ensemble de la *base IVC*, ainsi que sur les sous-ensembles JPEG, JPEG2000 et JPEG+JPEG2000. Sur la base entière les ΔCC entre l'utilisation ou non d'un modèle de masquage semi-local sont respectivement $+0,044$ et $+0,031$ pour FQA, WQA. La même tendance est observée en termes de SROCC et de RMSE. Les tests de significativité montrent que cette différence est significative avec une confiance de 95% pour FQA, et avec une confiance de 90% pour WQA. Les résultats sont plus modérés sur les dégradations de type flou, où l'amélioration est sensible avec le modèle FQA, mais c'est l'inverse avec le modèle WQA. Une explication possible réside dans la façon dont l'activité semi-locale est prise en compte. Les dégradations de type flou conduisent à une réduction significative de l'activité semi-locale entre l'image de référence et l'image dégradée. Comme l'erreur est normalisée par le maximum des élévations de seuil calculées sur l'image de référence et sur l'image dégradée (cf. l'équation (2.29)), les effets de masquage semi-local peuvent être surestimés dans ce cas. Ces observations montrent l'impact positif du masquage semi-local, et prouvent que la prise en compte des effets de masquage ne doit pas se limiter au masquage de contraste.

Les figures 5.3 et 5.4 représentent graphiquement les couples $(MOS, MOSp)$. La figure 5.3 permet d'analyser l'impact du masquage semi-local sur l'évaluation des différents types de distorsions. L'amélioration de la prédiction due au masquage semi-local n'est pas spécifique à un type particulier de dégradations. Quel que soit le type de distorsions, le masquage semi-local apporte une amélioration significative. La figure 5.4 permet d'analyser

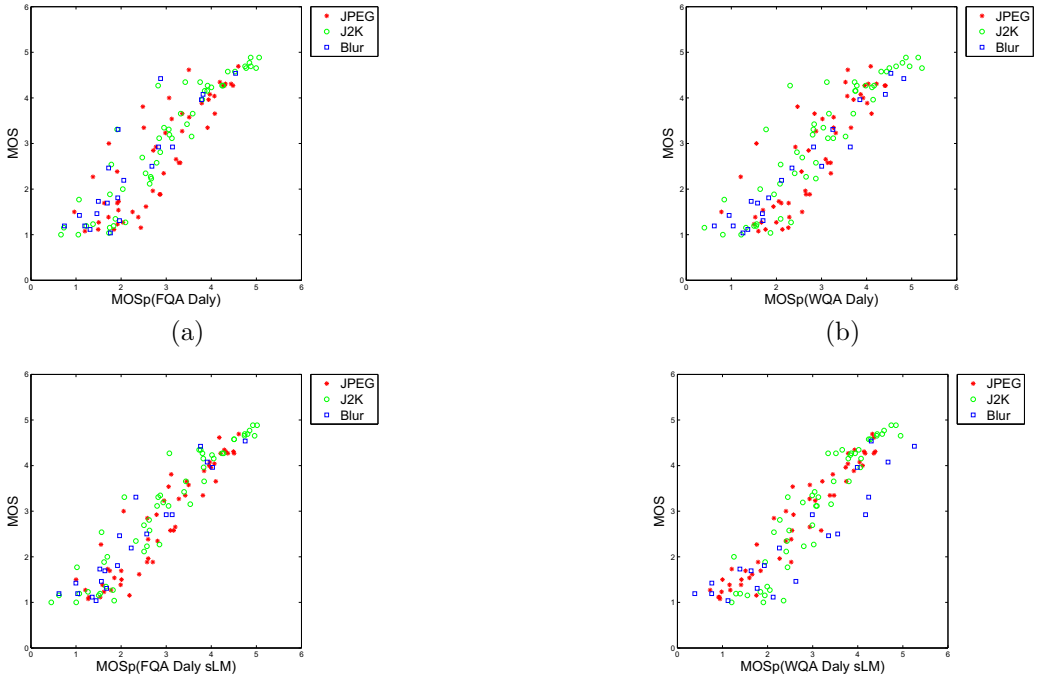


Figure 5.3 – Nuage de points des couples $(MOS, MOSp)$ par type de distorsions : (a), (c) pour FQA_1 et FQA_2 respectivement ; (b), (d) pour WQA_1 et WQA_2 respectivement.

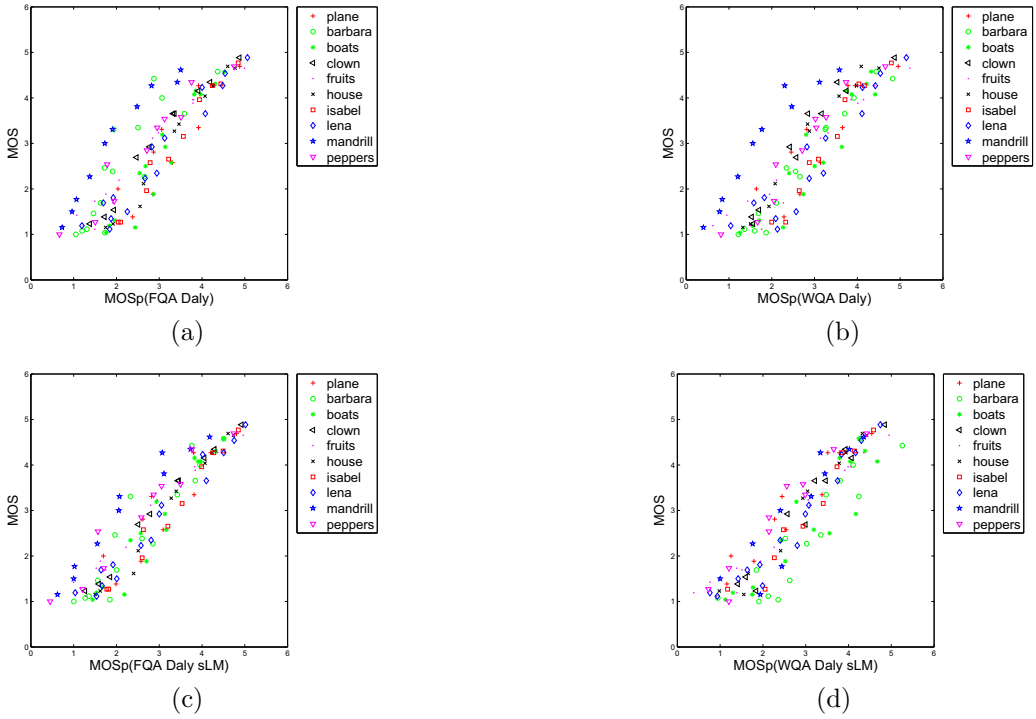


Figure 5.4 – Nuage de points des couples $(MOS, MOSp)$ par image de référence : (a), (c) pour FQA_1 et FQA_2 respectivement ; (b), (d) pour WQA_1 et WQA_2 respectivement.

l'impact du masquage semi-local sur l'évaluation des différents contenus (versions dégradées de la même image de référence). Cette figure montre que l'amélioration due au masquage semi-local est plus spécifique aux images noté ☆, qui proviennent de la même image de référence : *Mandrill* (cf. figure 2.22(a)). Le contenu de ces images est particulier en raison de leurs grandes zones texturées (les fourrures et des moustaches). La qualité de ces images est sous-estimée par les modèles sans masquage semi-local, ou, en d'autres termes, les distorsions sont surestimées. L'utilisation du masquage semi-local améliore sensiblement la qualité de l'évaluation de ces images. Cette observation tend à montrer que les distorsions surestimées sont situées dans les zones ayant une activité spatiale importante. Cela confirme l'intérêt de l'utilisation du masquage semi-local dans ce type de zones.

5.3.2.2 Résultats sur la base Toyama

Afin de consolider les résultats de la métrique WQA (WQA_1 et WQA_2), ses performances ont été évaluées sur les données subjectives des bases *Toyama1* et *Toyama2*. Les résultats sont présentées dans le tableau 5.8. Les performances sont évaluées séparément sur les MOS et sur les DMOS, ces derniers étant disponibles grâce au protocole de test utilisé (ACR). Les résultats des tests statistiques de significativité sont présentés dans le tableau 5.9.

		Toyama2 (ACR)			*Toyama1 (ACR)*		
		CC	SROCC	RMSE	CC	SROCC	RMSE
MOS	WQA_1	0.851	0.855	0.571	0.837	0.844	0.71
	WQA_2 (sLM)	0.937	0.941	0.38	0.919	0.923	0.514
	PSNR	0.699	0.685	0.777	0.685	0.678	0.943
	SSIM	0.823	0.826	0.618	0.814	0.82	0.754
	MS-SSIM	0.91	0.925	0.455	0.898	0.912	0.576
DMOS	WQA_1	0.874	0.874	0.535	0.85	0.85	0.68
	WQA_2 (sLM)	0.943	0.942	0.367	0.932	0.93	0.468
	PSNR	0.73	0.717	0.752	0.691	0.683	0.931
	SSIM	0.833	0.838	0.61	0.805	0.81	0.766
	MS-SSIM	0.911	0.921	0.458	0.9	0.902	0.564

Table 5.8 – Résultats sur les bases entières *Toyama1* et *Toyama2* (MOS et DMOS).

Pour information et pour permettre aux lecteurs de faire leurs propres opinions sur la base d'images *Toyama*, les performances des critères PSNR, SSIM et MS-SSIM sont également données.

Comme sur la *base IVC*, les deux versions du modèle WQA (sans puis avec masquage semi-local) surpassent les critères PSNR et SSIM en termes de CC, SROCC et RMSE dans les différents cas considérés. Les différences entre les deux versions du modèle WQA et le PSNR sont significatives (confiance à 95%), de même que la différence entre le modèle WQA_2 et SSIM. Par contre la différence entre WQA_1 et SSIM n'est significative (confiance à 90%) que pour *Toyama2*(DMOS).

Concernant la MS-SSIM, les résultats montrent qu'elle surpasse le modèle WQA_1 (confiance à 95%) en termes de CC, SROCC et RMSE dans les différents cas considérés. Cependant, elle est surpassée par le modèle WQA_2 en termes de CC, SROCC et RMSE dans les différents cas considérés. La différence entre la MS-SSIM et WQA_2 est significative (confiance à 95%) dans tous les cas sauf dans celui de l'évaluation subjective *Toyama1*(MOS).

	PSNR	SSIM	MS-SSIM	WQA_1	WQA_2 (sLM)
PSNR	*0000*	1111	1111	1111	1111
	0000	1111	1111	1111	1111
SSIM	1111	*0000*	1111	0000	1111
	1111	*0000*	1111	0001	1111
MS-SSIM	1111	1111	*0000*	1111	0111
	1111	1111	*0000*	1111	0111
WQA_1	1111	0000	1111	*0000*	1111
	1111	0001	1111	*0000*	1111
WQA_2 (sLM)	1111	1111	0111	1111	*0000*
	1111	1111	0111	1111	*0000*

Table 5.9 – Tests statistiques sur les résidus entre les notes prédites et notes subjectives pour la base Toyama. Pour chaque couple de métriques (ligne,colonne) deux séquences de caractères ('0' ou '1') indiquent si l'hypothèse nulle d'égalité des variances est rejetée. Les séquences de caractères représentent, dans l'ordre, les résultats pour les évaluations subjectives : *Toyama1*(MOS), *Toyama1*(DMOS), *Toyama2*(MOS) et *Toyama2*(DMOS). Le caractère '1' indique que la différence est significative, tandis que le caractère '0' indique que la différence n'est pas significative. La première séquence de caractères indique les résultats avec une confiance de 95%, tandis que la seconde séquence indique les résultats avec une confiance de 90%.

Quelles que soient les données subjectives (*Toyama1* ou *Toyama2*), l'utilisation du masquage semi-local améliore les performances du modèle WQA_1 en termes de CC, SROCC et RMSE. La différence entre le modèle WQA_1 et le modèle WQA_2 est significative (confiance à 95%) dans les différents cas considérés. Cette observation se fait aussi bien avec les MOS et qu'avec les DMOS. Sur la *base IVC*, le ΔCC entre l'utilisation ou non d'un modèle de masquage semi-local était de $+0,031$. Sur la base *Toyama2*, les ΔCC concernant les MOS et les DMOS sont respectivement de $+0,086$ et $+0,069$. Sur la base *Toyama1*, les ΔCC concernant les MOS et les DMOS sont respectivement $+0,082$ et $+0,082$. La même tendance est observée en termes de SROCC et RMSE. Ces observations montrent et confirment l'impact positif du masquage semi-local. Cette amélioration est retrouvée sur plusieurs tests subjectifs dont les conditions de test ne sont pas identiques. Elle reste valable :

- que l'écran soit un CRT ou un LCD,
- que les observateurs soient français ou japonais,
- que le protocole de test soit DSIS ou ACR,
- qu'on considère le MOS ou le DMOS.

Ces facteurs modifient les résultats des tests subjectifs, ainsi que les performances des métriques testées. Cependant, ces facteurs ne suppriment pas l'amélioration des performances entre les métriques qui n'utilisent pas le masquage semi-local et les métriques qui l'utilisent. Cette observation montre que l'intérêt du masquage semi-local n'est pas limité à des conditions de test particulières, mais bien qu'il est général.

Dans la suite de ce mémoire, et sans autre précision de notre part, nous ferons référence à WQA comme étant la version de la métrique incluant les effets de masquage semi-local, c'est-à-dire la version précédemment notée WQA_2. En effet, cette version est la plus performante en terme de capacité de prédiction. Les cartes de distorsions calculées à partir de ce modèle sont par ailleurs les plus pertinentes.

5.4 Conclusion

L'objectif de ce chapitre était la présentation et l'évaluation de critères objectifs de qualité visuelle d'images avec référence complète. Les critères proposés permettent de construire une note de qualité en cumulant spatialement les distorsions locales évaluées à partir de modèles du système visuel humain. Ces critères reposent sur une modélisation des caractéristiques principales du système visuel humain : le comportement multi-canal, la sensibilité au contraste et les effets de masquage.

Tout d'abord, il a été montré que simuler le comportement multi-canal du système visuel humain avec une décomposition en canaux perceptuels dans le domaine de Fourier, ou avec une décomposition en sous-bandes dans le domaine des ondelettes conduit à des performances proches même si on observe une légère tendance en faveur des modèles reposant sur une décomposition en canaux perceptuels dans le domaine de Fourier. Les modèles utilisant une décomposition en ondelettes obtenant de bonnes performances, la transformée en ondelettes peut donc être considérée comme une bonne alternative pour réduire la complexité de calcul.

Deuxièmement, on a observé l'impact positif sur les performances de l'utilisation d'un modèle de masquage semi-local. En effet, le masquage semi-local permet d'améliorer significativement les performances des critères proposés. Le masquage semi-local est complémentaire au masquage de contraste. L'intégration de ce type de masque dans les mesures de qualité améliore les performances en terme de prédiction de la qualité, ainsi que la pertinence des cartes d'erreurs perceptuelles, surtout pour des dégradations de type compression d'images.

Dans le chapitre suivant, nous nous intéresserons à la construction du jugement de qualité à partir de distorsions spatio-temporelles.

Chapitre 6

Critères objectifs de qualité visuelle de vidéos

6.1 Introduction

L'objet de ce chapitre est d'étendre aux images animées la problématique du chapitre précédent qui était l'évaluation objective de la qualité visuelle des images. Le développement de méthodes objectives automatiques d'évaluation de la qualité de vidéos répond à un besoin fort de l'industrie de l'image et de la vidéo. De même que dans le cas des images fixes, ces méthodes doivent permettre la construction d'une note de qualité visuelle correspondant au jugement que donnerait un observateur humain standard. La dimension temporelle intrinsèque des vidéos a pour conséquence d'ajouter une dimension temporelle aux distorsions. Les distorsions spatio-temporelles ont été évalué localement aux niveaux des fixations et des mouvements de poursuites dans le chapitre 3. Cependant, la dimension temporelle doit aussi être considérée dans la construction du jugement de qualité visuelle d'une vidéo. La problématique devient alors celle du cumul spatio-temporel des distorsions dans le but d'obtenir une note objective de qualité visuelle. La construction du jugement de qualité d'un observateur présente plusieurs caractéristiques intéressantes à prendre en compte comme l'existence d'un comportement asymétrique (*quick to criticize, slow to forgive*) et celle d'un effet de saturation perceptuelle.

Dans ce chapitre nous présentons et nous évaluons un critère objectif de qualité visuelle de vidéos avec référence complète. Le critère proposé se décompose en deux phases. La première phase, qui était l'objet du chapitre 3, consiste à évaluer localement les distorsions visuelles par un cumul temporel court terme. La seconde phase consiste à construire une note de qualité à partir de l'évaluation locale des distorsions visuelles réalisée précédemment, en cumulant spatio-temporellement les distorsions locales se produisant au cours de la séquence. En fait, les distorsions vont d'abord être cumulées spatialement puis temporellement. Ce cumul temporel est qualifié de cumul temporel long terme et l'échelle de temps est celle de la séquence. L'évaluation des performances de la métrique proposée a nécessité la conception et la réalisation de tests subjectifs d'évaluation de la qualité de vidéos.

La première partie de ce chapitre est consacrée à la description de la fonction de cumul temporel long terme.

La seconde partie est dédiée à l'évaluation des performances de la métrique proposée. Nous décrivons d'abord les tests subjectifs que nous avons menés et dont sont issues les notes subjectives nécessaires à l'évaluation des performances. Puis, les résultats de l'évaluation des performances sont présentés et commentés.

6.2 Cumul spatial et cumul temporel

Dans la première partie de ce mémoire, nous avons conçu et décrit une méthode de construction des distorsions visuelles localisées pour lesquelles les localités sont en fait des tubes spatio-temporels. Il s'agit maintenant de construire un jugement global de qualité à partir de ces distorsions localisées spatio-temporellement. Mais que doit prendre en compte ce cumul spatio-temporel des distorsions afin de reproduire le jugement humain ?

Les variations temporelles de distorsions sur une séquence vidéo jouent un rôle important sur l'appréciation de la qualité visuelle globale. Le niveau moyen de distorsions sur l'ensemble de la séquence n'est pas suffisant pour évaluer sa qualité. Le jugement de qualité dépend non seulement du niveau moyen de distorsions sur toute la séquence mais aussi des variations temporelles de distorsions visuelles sur la durée de la séquence. Concernant ce dernier point une particularité du processus d'évaluation d'un observateur peut être caractérisée par la phrase suivante : *quick to criticize, slow to forgive*, ce que l'on peut traduire par « prompt à critiquer et lent à pardonner ». Ceci illustre le comportement temporellement asymétrique des observateurs humains dans la construction de leur jugement de qualité.

L'existence d'un mécanisme long terme dans le cumul temporel des distorsions a été introduit par les travaux de Masry et Hemami [Masry 04]. Ce mécanisme y est modélisé par un traitement récursif opéré sur les notes de qualité par image, celles-ci ayant été préalablement lissées par un mécanisme court terme. Ce cumul temporel long terme comprend un comportement asymétrique et un effet de saturation perceptuelle. Le comportement asymétrique donne davantage d'importance à la diminution de la qualité qu'à son amélioration. Au-delà d'un certain niveau de distorsions, l'augmentation de celles-ci n'a plus autant d'impact sur le jugement, c'est ce qui est appelé l'effet de saturation perceptuelle. Ces deux propriétés seront reprises dans notre modélisation.

Une question importante concerne l'ordre dans lequel doit être réalisé le cumul spatio-temporel des distorsions. Doit-on les cumuler d'abord spatialement puis temporellement, ou bien doit-on les cumuler temporellement puis spatialement ? Si l'on tente de rapprocher ces deux options du processus d'évaluation de la qualité par un observateur, la première option peut être interprétée de la façon suivante : l'observateur construit son jugement à partir d'une évaluation globale des images successives de la séquence vidéo. La seconde option, quant à elle, peut être interprétée par : l'observateur construit son jugement de qualité à partir d'une évaluation sur toute la séquence de chaque « zone » des images de la séquence, pour ensuite en déduire une note globale de qualité. La première option semble plus probable, elle permet d'ailleurs de faire le lien avec les méthodes d'évaluation continue de la qualité de vidéos (cf. section 4.2.3) comme dans les travaux de Masry et Hemami [Masry 04]. C'est cette option que nous avons retenue dans nos travaux.

Dans notre approche, l'évaluation spatio-temporelle des distorsions sur l'ensemble d'une vidéo se décompose en deux étapes indiquées figure 6.1. Les erreurs perceptuelles des cartes spatio-temporelles sont d'abord cumu-

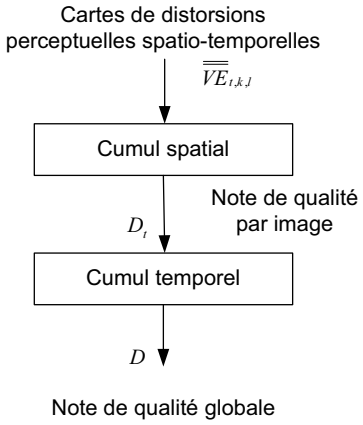


Figure 6.1 – Schéma du cumul long terme.

lées spatialement, puis temporellement. Nous rappelons que les cartes de distorsions spatio-temporelles sont le résultat du cumul temporel court terme (cf. chapitre 3) ce qui implique que les distorsions visuelles à l'instant t_i ne dépendent pas uniquement de l'image I_i, mais dépendent aussi des images la précédent. Le cumul temporel long terme est la dernière étape de la construction de la note globale de qualité d'une séquence vidéo. Il permet d'élaborer un jugement global de qualité (note) en tenant compte des distorsions apparaissant tout au long d'une vidéo.

6.2.1 Cumul spatial

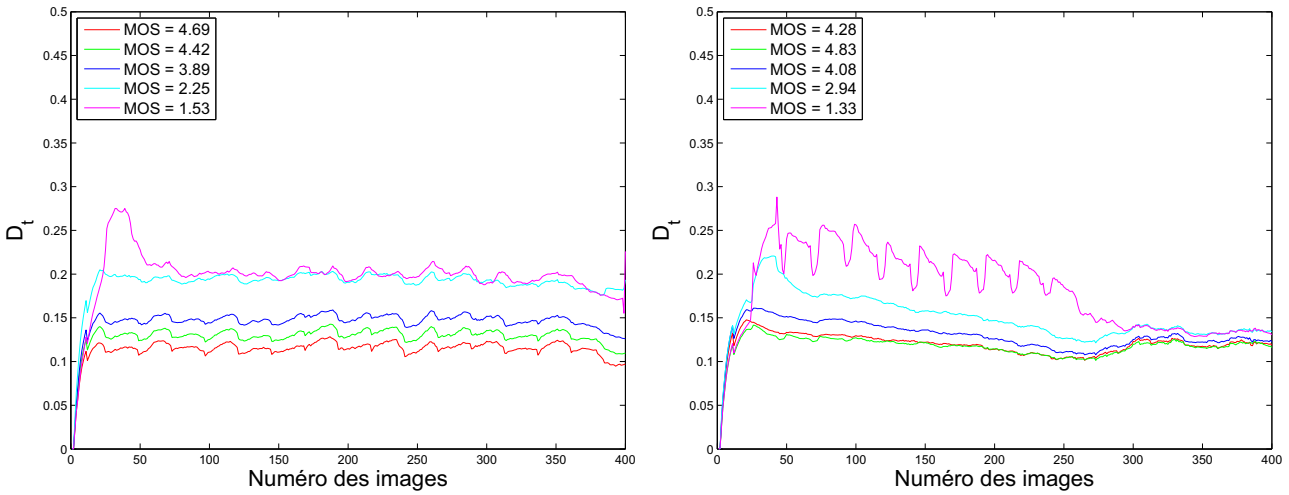
Le but de cette étape est d'obtenir une note objective représentant le niveau de distorsions perceptuelles à chaque instant (pour chaque image) de la séquence. Une note D_t représentant le niveau de distorsions perceptuelles par image est calculée à partir de la carte de distorsions perceptuelles spatio-temporelles $\overline{\overline{VE}}_{t,k,l}$ de chaque image au moyen d'une sommation Minkowski :

$$D_t = \left(\frac{1}{K \cdot L} \sum_{k=1}^{K} \sum_{l=1}^{L} \left(\overline{\overline{VE}}_{t,k,l} \right)^{\beta_s} \right)^{\frac{1}{\beta_s}}, \tag{6.1}$$

où K et L sont respectivement la hauteur et la largeur des cartes spatio-temporelles de distorsions et β_s est l'exposant de Minkowski ($\beta_s = 2$). La fonction de cumul spatial (sommation Minkowski) utilisée dans cette étape est similaire à celle utilisée dans les métriques de qualité pour images fixes du chapitre précédent. Cependant, dans le cas présent les distorsions perceptuelles sont spatio-temporelles et non pas purement spatiales.

A l'issue de ce cumul les séquences de distorsions spatio-temporelles sont projetées sur l'axe temporel, sous la forme de notes intermédiaires. La figure 6.2 illustre le résultat du cumul spatial pour deux séquences vidéos et cinq niveaux de dégradation.

Une note intermédiaire est calculée à la même fréquence temporelle que celle d'affichage des images de la séquence. Une note intermédiaire ne correspond pas à un cumul spatial des distorsions spatiales d'une image de



(a) (b)

Figure 6.2 – Évolution temporelle des notes intermédiaires D_t, pour deux séquences de la base de test (cf. section 6.3.1) et cinq niveaux de dégradation : (a) *ParkRun*, (b) *MobCal*. L'axe horizontal représente le numéro d'images, et axe vertical représente l'échelle de distorsions (de 0 pour la meilleure qualité à 0.5 pour la pire qualité). Pour chaque niveau de dégradations les MOS sont donnés.

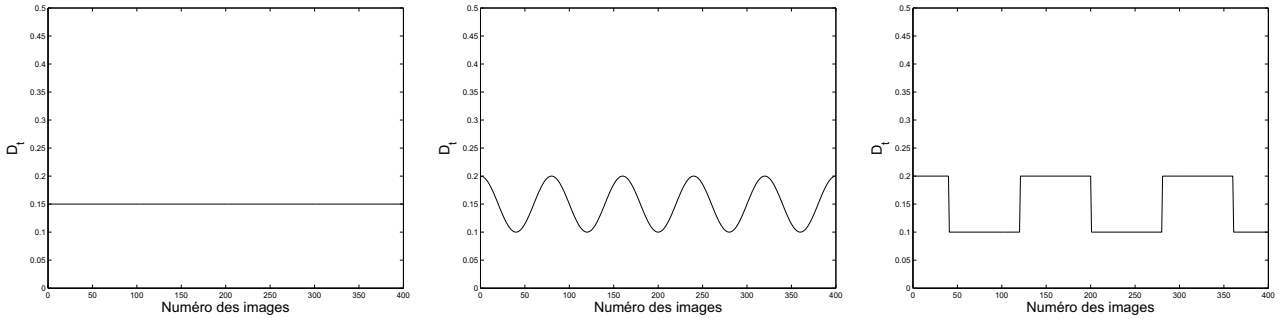
la séquence, mais elle correspond à un cumul spatial des distorsions spatio-temporelles calculées dans les tubes spatio-temporels débouchant sur cette image.

6.2.2 Cumul temporel

Le niveau de distorsions par image D_t évolue tout au long de la séquence. La note objective globale du niveau de distorsions perçues dans une séquence, appelée D, dépend à la fois du niveau moyen de distorsions dans la séquence et des variations temporelles de D_t tout au long de la séquence. En fait le niveau de distorsions réellement perçues est accru par les variations temporelles de distorsions. La figure 6.3 illustre ce phénomène. Dans cette figure, les 4 exemples proposés ont temporellement le même niveau moyen de distorsion, par contre l'exemple (a) est jugé bien moins gênant que les autres exemples. En effet, l'exemple (a) présente un niveau de distorsion constant temporellement, alors que les exemples (b) et (c) présentent des évolutions temporelles du niveau de distorsions. Les exemples (b) et (c) sont considérés comme les plus gênants.

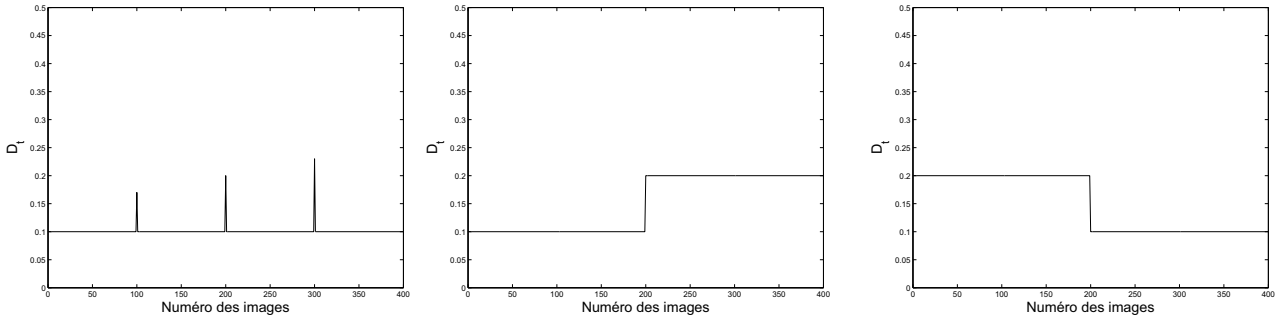
Dans un contexte d'évaluation de qualité, les observateurs sont moins sensibles à de nouveaux changements de qualité au-delà de certains seuils, soit vers une meilleure qualité, soit vers une qualité plus mauvaise [Tan 98]. C'est ce que nous avons appelé l'effet de saturation perceptuelle. La figure 6.4(a) présente trois « pics » de distorsions à trois instants différents. Si l'on considère que le premier pic est juste au niveau de la saturation perceptuelle, les deux autres pics vont provoquer la même gêne que le premier.

Le comportement asymétrique est le fait que les humains sont plus à même de se rappeler les expériences désagréables plutôt que les expériences agréables (*quick to criticize, slow to forgive*) [Tan 98]. Les figures 6.4(b) et (c) présentent deux variations de distorsions. Dans le premier cas la variation est l'augmentation du niveau



(a) (b) (c)

Figure 6.3 – Exemples d'évolutions temporelles des notes intermédiaires D_t. Les différents exemples ont temporellement le même niveau moyen de distorsion. L'axe horizontal représente le numéro d'images, et l'axe vertical représente l'échelle de distorsions (de 0 pour la meilleure qualité à 0.5 pour la pire qualité).



(a) (b) (c)

Figure 6.4 – Exemples d'évolutions temporelles des notes intermédiaires D_t : (a) illustration de la saturation perceptuelle, (b,c) illustration du comportement asymétrique. L'axe horizontal représente le numéro d'images, et l'axe vertical représente l'échelle de distorsions (de 0 pour la meilleure qualité à 0.5 pour la pire qualité).

de distorsion, alors que dans le second cas c'est une diminution. Les deux variations ont la même amplitude, cependant la gêne occasionnée dans le premier cas sera supérieure à celle occasionnée dans le second cas.

Le cumul temporel proposé intègre ces deux propriétés importantes :

– un effet de saturation perceptuelle,

– un comportement asymétrique.

Le niveau global de distorsions D d'une vidéo est calculé à partir des notes de distorsions par image D_t, comme la somme de la moyenne temporelle des distorsions $\overline{D}$, et d'un terme Δ_D représentant la variation temporelle des distorsions sur la séquence. Afin de limiter l'influence des variations trop élevées de distorsions, D est calculé avec un effet de saturation comme suit :

$$D = \begin{cases} \bar{D} + \Delta_D & \text{pour } \Delta_D < \lambda_1 \cdot \overline{D} \\ \bar{D} + \lambda_1 \cdot \overline{D} & \text{pour } \Delta_D \geq \lambda_1 \cdot \overline{D} \end{cases} \quad \text{avec } \lambda_1 > 0. \tag{6.2}$$

Le niveau global de distorsions D augmente linéairement avec les variations temporelles jusqu'à un seuil de saturation proportionnel à $\overline{D}$. De ce fait, les exemples de la figure 6.3, qui ont la même moyenne temporelle des distorsions $\overline{D}$, n'auront pas le même niveau de distorsions D

Dans la construction du jugement de qualité, nous faisons l'hypothèse que, parmi toutes les variations de distorsions se produisant au cours d'une séquence, ce sont les variations les plus importantes qui contribuent le plus à la gêne visuelle et donc à la qualité. Dans notre modèle, c'est le terme Δ_D qui permet de favoriser les variations de distorsions les plus importantes. Il est calculé comme suit :

$$\Delta_D = \lambda_2 \cdot avg_{n\ \%}(abs(\nabla' D_t))\,, \tag{6.3}$$

où $\nabla' D_t$ est l'ensemble des valeurs de gradient temporel des distorsions par image D_t après la transformation asymétrique, $abs(X)$ est la valeur absolue de X, $avg_{n\ \%}(X)$ est la moyenne des valeurs de X au-dessus du nième percentile de X. La valeur de n est fixée à 95%. Cette fonction permet de ne prendre en compte que les variations de distorsions les plus importantes. L'utilisation du nième percentile permet de ne pas fixer de seuil pour la sélection des variations à prendre en compte. Ceci permet de s'adapter à la dynamique de l'ensemble des valeurs de gradient temporel des distorsions par image D_t pour chaque vidéo évaluée.

Le comportement asymétrique du jugement humain est simulé par une transformation asymétrique des valeurs de gradient, celle-ci est calculée comme suit :

$$\nabla' D_t = \left\{ \begin{array}{ll} \lambda_3 \cdot \nabla D_t & \text{for } \nabla D_t < 0 \\ \nabla D_t & \text{for } \nabla D_t \geq 0 \end{array} \right| \lambda_3 \leq 1, \tag{6.4}$$

où la valeur de λ_3 contrôle le comportement asymétrique. Si $\lambda_3 < 1$, plus de poids est donné aux augmentations de distorsions qu'aux diminutions. Dans les cas (b) et (c) de la figure 6.4, la contribution du gradient de la variation du cas (c) serait diminuée par rapport à celle du cas (b).

Ces différents traitements permettent d'obtenir le niveau global de distorsions perceptuelles D. Finalement, la note de qualité globale VQA est construite à partir du niveau de distorsions perceptuelles D en utilisant une

fonction psychométrique, tel que recommandé par le groupe de travail VQEG [VQEG 00] :

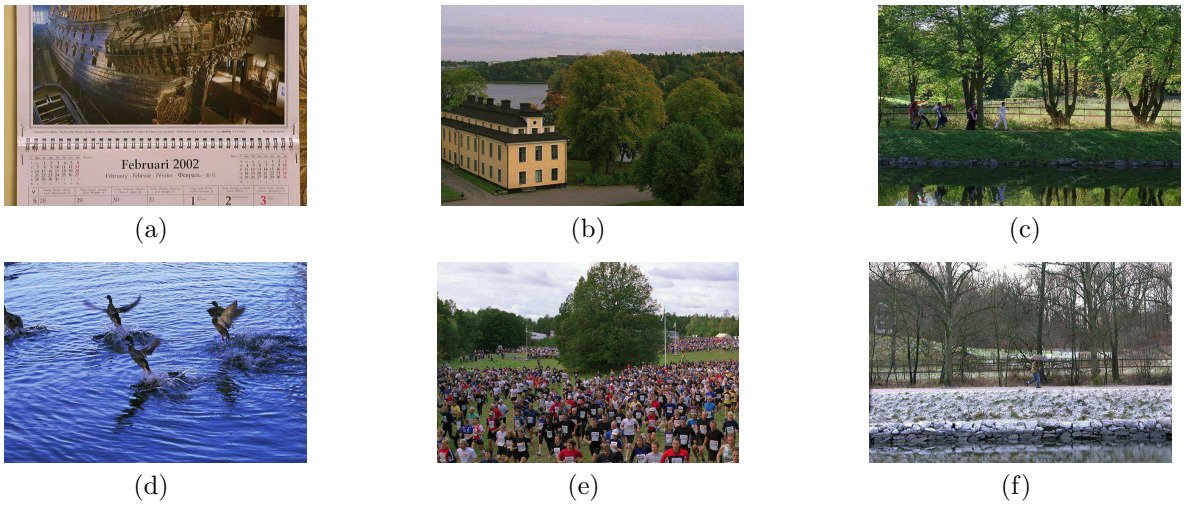
$$\mathrm{VQA} = \frac{b_1}{1 + e^{-b_2 \cdot (D - b_3)}} , \tag{6.5}$$

où b_1, b_2 et b_3 sont les trois paramètres de la fonction psychométrique.

6.3 Expérimentations

6.3.1 Base d'évaluation subjective

De manière à disposer de données subjectives d'évaluation de la qualité de vidéos, nous avons mené de nouveaux tests sur une nouvelle base de vidéos. La base de données d'évaluations subjectives comprend 60 vidéos de 8 secondes. Elle a été construite à partir de 10 séquences de référence considérées d'une qualité irréprochable (sans dégradation). Les séquences de référence sont des séquences de scènes naturelles de contenu divers comme l'illustre la figure 6.5. La résolution spatiale des séquences est 720x480 avec une fréquence temporelle de 50 Hz pour un mode de balayage progressif. Toutes les séquences vidéo de référence ont été dégradées par un système de compression H.264/AVC avec cinq débits différents. Il en résulte cinquante séquences vidéo dégradées. Les cinq débits ont été choisis afin de générer, pour chaque séquence de référence, des dégradations couvrant toute la gamme des dégradations de l'échelle à cinq catégories utilisées au cours des tests subjectifs : de « imperceptible » à « très gênant ».



(a) (b) (c)

(d) (e) (f)

Figure 6.5 – Exemples de séquences vidéo extraits de la base de test : (a) la séquence *MobCal*, (b) la séquence *InToTree*, (c) la séquence *ParkJoy*, (d) la séquence *DucksTakeOff*, (e) la séquence *CrowdRun*, et (f) la séquence *ParkRun*.

Les dégradations introduites par l'encodage ne sont pas stationnaires ni temporellement ni spatialement, et dépendent des contenus des séquences vidéo.

Les évaluations subjectives ont été menées à une distance de quatre fois la hauteur de l'image affichée sur un écran CRT et dans des conditions normalisées de visualisation [ITU-R Rec. BT.500-10 00]. Le protocole de

test DSIS (*Double Stimulus Impairment Scale*) a été utilisé avec une échelle de dégradations à cinq catégories.

Les séquences ont été notées par trente six observateurs ayant une acuité visuelle normale (sans ou avec correction optique) selon le test de Monoyer. Leur perception des couleurs était normale selon le test d'Ichihara. Les observateurs n'étaient pas des experts en traitement d'images et ne connaissaient pas l'expérimentation.

Les données de qualité visuelle ont été traitées selon les méthodes décrites dans la section 4.2.4, afin d'éliminer les notes aberrantes.

6.3.2 Évaluation des performances

Les performances de plusieurs métriques de qualité ont été évaluées par comparaison avec les notes (MOS) issues des tests subjectifs réalisés. Les métriques objectives de qualité testées ont été :

- la métrique de qualité vidéo proposée, appelée VQA (version achromatique).
- le PSNR (version achromatique). Le PSNR sur l'ensemble de la vidéo étant la moyenne temporelle des valeurs de PSNR par image.
- la VSSIM développée par Wang *et al.* [Wang 04b]. Nous avons utilisé tous les paramètres décrits dans [Wang 04b], à l'exception du facteur de normalisation K_M du mouvement inter-image qui a été adapté à la fréquence d'affichage des séquences de notre base de test (cf. section 4.4.2.2).
- le VQM développé par NTIA [Pinson 04]. Parmi les différents modèles de VQM, nous avons choisi d'utiliser le *modèle général* qui est considéré comme le plus précis. Le *modèle général* est connu sous le nom de métrique H dans le plan de test (Phase II) du groupe de travail VQEG [VQEG 03].

Afin d'évaluer les différentes étapes de la métrique VQA, trois notes supplémentaires de distorsions perceptuelles sont calculées en plus de la note de qualité finale. Elles sont issues de trois versions plus ou moins simplifiées (VQA_1, VQA_2 et VQA_3) de la métrique VQA.

La première note intermédiaire est une note purement spatiale du niveau de distorsions perceptuelles, appelée VQA_1. Elle est calculée à partir des cartes de distorsions purement spatiales de la métrique pour image fixe WQA [1], comme suit :

$$VQA_1 = \frac{1}{T}\sum_{t=1}^{T} d_t \,, \tag{6.6}$$

où T est le nombre total d'images et d_t est une note moyenne de distorsions visuelles par image t, calculée comme suit :

$$d_t = \left(\frac{1}{K \cdot L} \sum_{k=1}^{K} \sum_{l=1}^{L} \left(VE_{t,k,l} \right)^{\beta_s} \right)^{\frac{1}{\beta_s}} , \tag{6.7}$$

où $VE_{t,k,l}$ sont les cartes de distorsions spatiales (calculées avec WQA), K et L sont la hauteur et la largeur des cartes de distorsions spatiales respectivement, et β_s est l'exposant de Minkowski.

Dans la deuxième note intermédiaire de distorsions perceptuelles, appelée VQA_2, le cumul temporel court terme (niveau fixation) est désactivé, ce qui signifie que les distorsions perceptuelles sont calculées à partir du cumul temporel long terme (cf. équation 6.2) où D_t est remplacé par d_t. D_t est la note par image de distorsions

1. Il s'agit de la version de cette métrique incluant les effets de masquage semi-local (cf. chapitre 5)

spatio-temporelles (avec le cumul temporel au niveau fixation), tandis que d_t est la note par image des distorsions purement spatiales (sans le cumul temporel au niveau fixation).

Dans la troisième note intermédiaire de distorsions perceptuelles, appelée VQA_3, le cumul temporel court terme (niveau fixation) est bien présent, par contre le cumul temporel long terme est remplacé par une simple moyenne temporelle :

$$VQA_3 = \frac{1}{T}\sum_{t=1}^{T} D_t \,, \tag{6.8}$$

où T est le nombre total d'images et D_t est une note par image t, le cumul temporel au niveau fixation étant activé.

Les cumuls temporels des différentes versions sont récapitulées dans le tableau 6.1. La comparaison entre les

Versions	Cumul temporel court terme	Cumul temporel long terme
VQA	Oui (D_t)	Oui
VQA_1	Non (d_t)	Non (moyenne temporelle)
VQA_2	Non (d_t)	Oui
VQA_3	Oui (D_t)	Non (moyenne temporelle)

Table 6.1 – Récapitulatif des différentes versions.

métriques VQA_2 et VQA permet d'évaluer l'amélioration due à l'évaluation des distorsions spatio-temporelles au niveau fixation, autrement dit au cumul temporel court terme. La comparaison entre les métriques VQA_1 et VQA d'une part, les métriques VQA_3 et VQA d'autre part, permet d'évaluer l'amélioration due au cumul temporel long terme et à son interaction avec le cumul temporel court terme. L'étude de l'impact du cumul temporel court terme sur les performances permet de valider quantitativement les travaux proposés dans le chapitre 3 pour la conception de séquences de distorsions visuelles de vidéos.

Comme nous l'avons mentionné précédemment pour les critères de qualité d'images, avant d'évaluer les métriques de qualité pour vidéos, une fonction psychométrique (cf. relation (4.10)) est utilisée pour transformer les différentes notes objectives en MOS prédit, noté MOSp, tel que recommandé par le groupe de travail VQEG [VQEG 00].

Les résultats, présentés dans le tableau 6.2, sont indiqués pour les différentes métriques (VSSIM, VQM et VQA) ainsi que pour les trois notes de qualité intermédiaires (VQA_1, VQA_2 et VQA_3) de VQA. Les résultats du PSNR sont fournis également à titre d'information, pour permettre aux lecteurs de faire leurs propres opinions sur la base de test. Les figures 6.6 et 6.7 montrent les nuages de points représentant les couples (MOS,MOSp) sur la base de test, produits d'une part par le PSNR, la VSSIM, le VQM, la VQA, d'autre part par les trois versions simplifiées (VQA_1, VQA_2 et VQA_3) de VQA. Comme pour les critères de qualité d'images fixes, les performances sont évaluées par trois indicateurs : le coefficient de corrélation linéaire (CC), le coefficient de corrélation de rang (SROCC) et la racine carrée d'erreur quadratique moyenne (RMSE).

Les résultats des tests statistiques de significativité (cf. section 4.3.5) sont présentés dans le tableau 6.3. Comme dans [Sheikh 06b], le test statistique est un F-test sur les résidus MOS-MOSp. L'hypothèse nulle est l'égalité entre la variance des résidus d'une métrique et la variance des résidus d'une autre métrique.

Métriques (MosP)	CC	SROCC	RMSE
PSNR	0.516	0.523	0.982
VQM	0.854	0.898	0.597
VSSIM	0.738	0.758	0.773
VQA	0.892	0.903	0.519
VQA_1	0.831	0.872	0.638
VQA_2	0.834	0.863	0.633
VQA_3	0.84	0.878	0.621

Table 6.2 – Performances des métriques de qualité visuelle sur toute la base de test. Comparaison en termes de CC, SROCC et RMSE.

	MOSp(PSNR)	MOSp(VSSIM)	MOSp(VQM)	MOSp(VQA)
MOSp(PSNR)	*1.0*	0.09690 ($p < 0.10$)	0.00066 ($p < 0.05$)	0.00002 ($p < 0.05$)
MOSp(VSSIM)	0.09690 ($p < 0.10$)	*1.0*	0.07259 ($p < 0.10$)	0.00610 ($p < 0.05$)
MOSp(VQM)	0.00066 ($p < 0.05$)	0.07259 ($p < 0.10$)	*1.0*	0.33157
MOSp(VQA)	0.00002 ($p < 0.05$)	0.00610 ($p < 0.05$)	0.33157	*1.0*

Table 6.3 – Tests statistiques sur les résidus entre les MOS et les MOSp. Chaque valeur donne pour le couple de métriques (ligne,colonne) la probabilité que l'hypothèse nulle d'égalité des variances soit rejetée. Si la valeur est inférieure à 0.05 les deux métriques sont significativement différentes avec une confiance de 95%. Si la valeur est inférieure à 0.10 les deux métriques sont significativement différentes avec une confiance de 90%.

Le PSNR ne conduit pas à une bonne prédiction de la qualité. En effet, le CC du PSNR avec les notes subjectives est seulement de 0.516. Ce résultat donne une idée sur le niveau de difficultés à prédire la qualité des séquences vidéo de la base de test. D'après les tableaux 6.2 et 6.3, le PSNR a des performances significativement inférieures à celles de VQM et VQA avec une confiance de 95%, et à celles de VSSIM avec une confiance de 90%.

Le critère proposé (VQA) produit de bons résultats comparé aux autres approches. VQA est statistiquement équivalent à VQM sur la base de test utilisée. Cependant, avec une confiance de 95%, VQA est statistiquement meilleure que VSSIM alors que VQM ne l'est pas. VQM est statistiquement meilleure que VSSIM avec une confiance inférieure (90%). Il est important de mentionner que les paramètres de la méthode proposée (VQA) ont été choisis empiriquement, sans aucune optimisation sur les vidéos de la base de test ($\lambda_1 = 1$, $\lambda_2 = 10$, $\lambda_3 = 0,25$, et $n = 95$).

La taille des échantillons est un point important pour réaliser des tests statistiques. La base de test utilisée ne contient que cinquante vidéos. L'interprétation des tests statistiques est donc à mettre en perspective avec la taille des échantillons. Par exemple, notre base de test n'est sans doute pas assez importante pour permettre de distinguer statistiquement VQA et VQM. Néanmoins la tendance est favorable à VQA qui obtient un CC de 0.892 alors que pour VQM, le CC vaut 0.854.

6.3.2.1 Influence du contenu des séquences sur les performances

Les figures 6.6 et 6.7 montrent que les performances des métriques varient en fonction du contenu des séquences vidéo. Cependant, le contenu des séquences ne perturbe pas les différentes métriques de la même

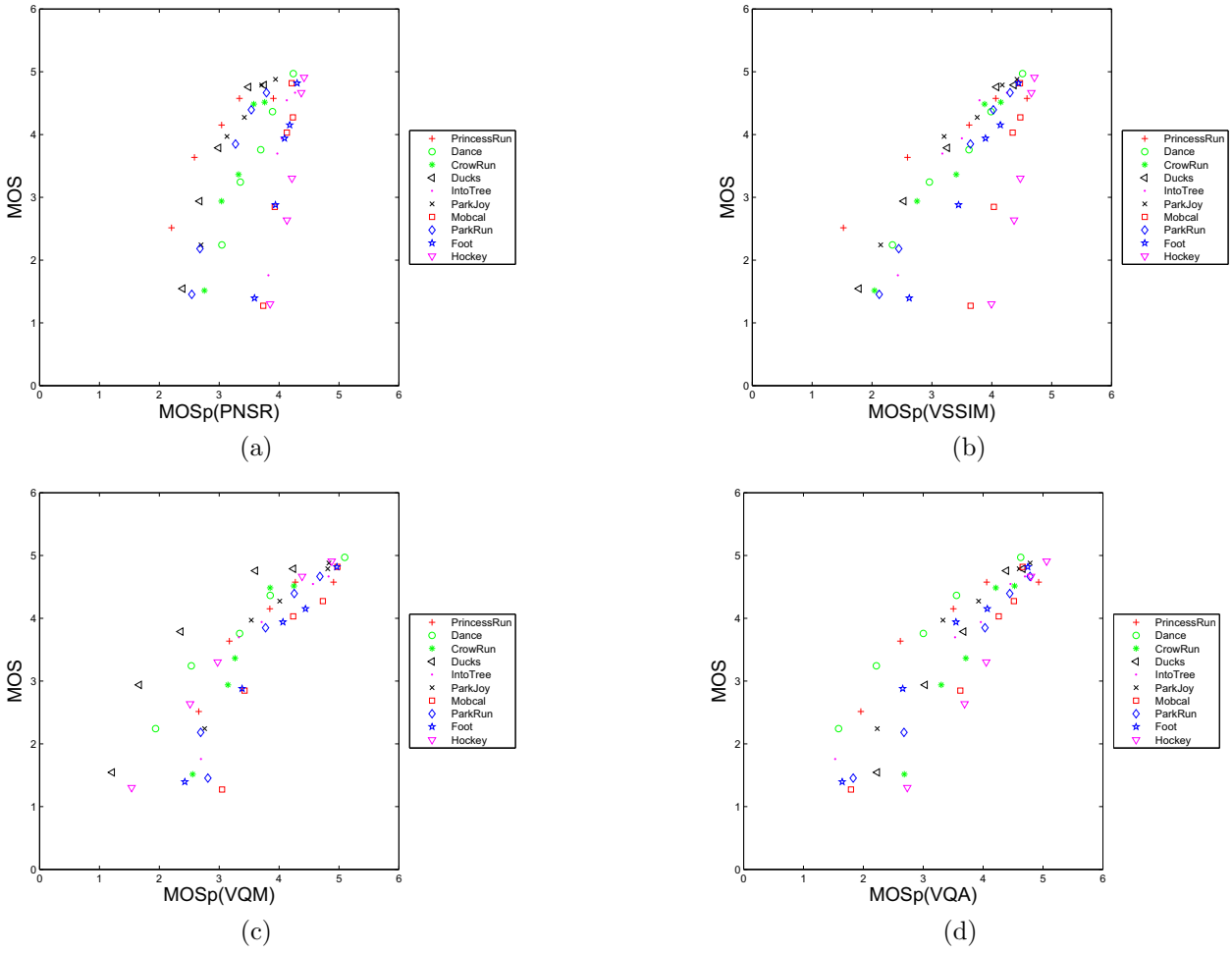


Figure 6.6 – Nuage de points des couples $(MOS, MOSp)$ par vidéo de référence. Chaque point représente une séquence vidéo. Le même symbole est utilisé pour toutes les vidéos dégradées issues de la même vidéo de référence : (a) PSNR, (b) VSSIM, (c) VQM et (d) VQA.

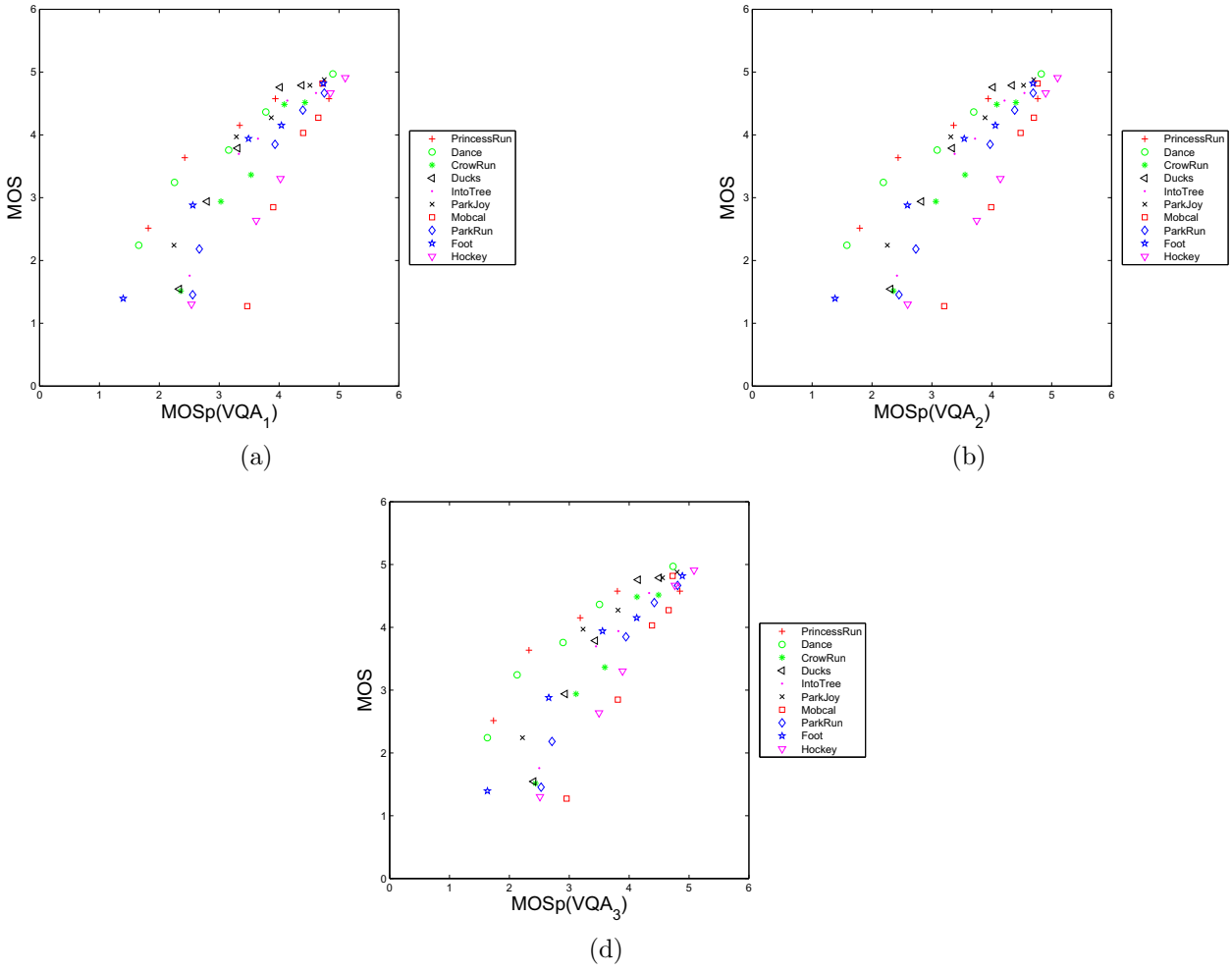
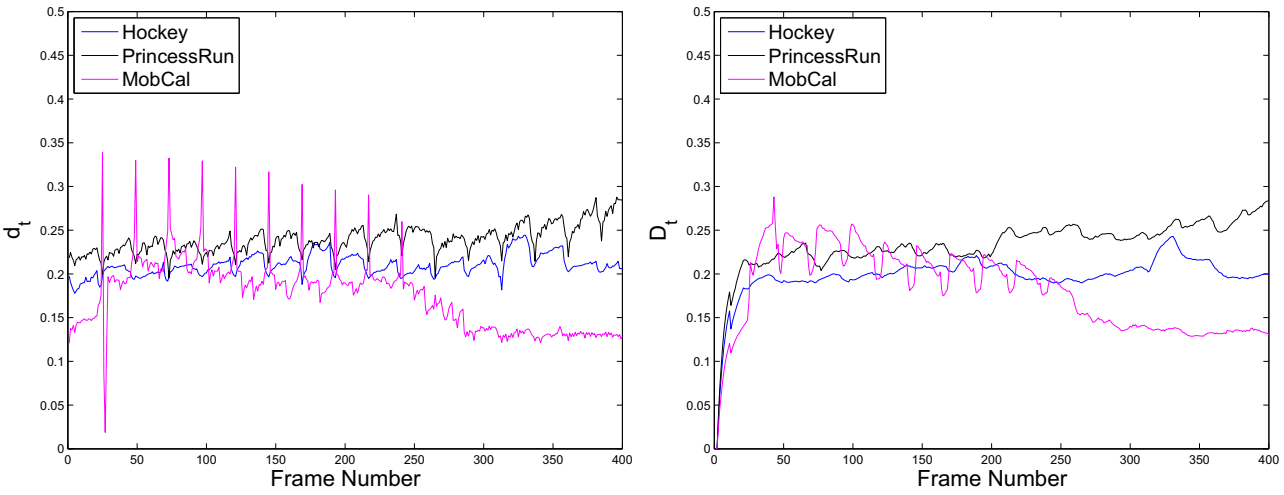


Figure 6.7 – Nuage de points des couples $(MOS, MOSp)$ par vidéo de référence. Chaque point représente une séquence vidéo. Le même symbole est utilisé pour toutes les vidéos dégradées issues de la même vidéo de référence : (a) VQA_1, (b) VQA_2 et (c) VQA_3.

manière. Par exemple, VQM sous-estime la qualité de la séquence *Ducks*, alors que la VQA ne la sous-estime pas. VQA sous-estime la qualité des séquences *PrincessRun* et *Dance*, et surestime la qualité de la séquence de *Hockey*. Une explication possible réside dans le fait que les distorsions spatiales sont respectivement surestimées et sous-estimées. La figure 6.8 montre que les notes de distorsions par image (d_t et D_t) de la séquence *Hockey* sont plus faibles que les notes de distorsions par image de la séquence *PrincessRun*, alors que les MOS de la séquence *Hockey* sont inférieurs aux MOS de la séquence de *PrincessRun*. Dans ces séquences, les variations temporelles des distorsions n'expliquent pas les erreurs de prédiction de la qualité. Cela montre que, dans la métrique proposée, une bonne évaluation des distorsions spatiales est nécessaire. L'évaluation des distorsions temporelles est donc dépendante des performances de la première étape de la métrique.



(a) (b)

Figure 6.8 – Évolution temporelle des notes de distorsions par image d_t (a), et D_t (b), pour trois séquences dégradées de la base de test : *Hockey* (MOS=1.4), *PrincessRun* (MOS=2.6) and *MobCal* (MOS=1.3). L'axe horizontal représente le numéro d'images, et l'axe vertical représente l'échelle de distorsions (de 0 pour la meilleure qualité à 0.5 pour la pire qualité).

6.3.2.2 Influence des cumuls temporels de VQA sur les performances

La comparaison entre les résultats de VQA_1, VQA_2, VQA_3 et VQA montre une contribution positive des différentes étapes de la métrique proposée. On constate une amélioration de la prédiction de la qualité entre la version purement spatiale (VQA_1) et la version spatio-temporelle (VQA). Par exemple, le ΔCC entre ces deux configurations est de $+0,061$. Comme on pouvait s'y attendre, cela montre que les distorsions temporelles jouent un rôle important dans l'évaluation de la qualité vidéo. L'amélioration de la prédiction de la qualité entre VQA_2, et VQA montre l'importance de l'évaluation spatio-temporelle des distorsions au niveau fixation (cumul temporel court terme). Cette étape paraît fondamentale avant le cumul temporel long terme. Une explication possible vient de l'effet de lissage des variations temporelles de distorsions introduit par le cumul temporel court terme. Cet effet permet une meilleure analyse des conséquences long terme des variations temporelles

des distorsions, en éliminant les variations temporelles de distorsions parasites (d'un point de vue perceptuel). Cet effet de lissage est illustré figure 6.8, en comparant les variations temporelles des notes de distorsions d_t (figure 6.8(a)) et D_t (figure 6.8(b)). Le cumul temporel au niveau fixation temporelle ne permet pas seulement d'améliorer la prédiction de la métrique mais il améliore également la pertinence des cartes de distorsions. La comparaison entre VQA_3 et VQA montre encore le bénéfice de l'association du cumul temporel court terme et du cumul temporel long terme. En effet, si l'on remplace le cumul temporel long terme par une simple moyenne temporelle on observe une chute des performances. Par exemple, le ΔCC entre VQA et VQA_3 est de $-0,052$.

6.3.2.3 Influence des paramètres du cumul temporel long terme sur les performances

Les résultats, présentés dans le tableau 6.4, sont indiqués pour VQA et pour différentes valeurs des paramètres λ_3 et n.

λ_3	$n$ième percentile	CC	SROCC	RMSE
0	0	0.85	0.874	0.605
0	80	0.879	0.892	0.547
0	85	0.885	0.893	0.535
0	90	0.892	0.901	0.518
0	95	0.895	0.912	0.512
0.25	0	0.851	0.874	0.601
0.25	80	0.88	0.892	0.545
0.25	85	0.885	0.893	0.533
0.25	90	0.892	0.901	0.518
0.25	95	0.895	0.912	0.511
0.5	0	0.853	0.875	0.599
0.5	80	0.877	0.89	0.551
0.5	85	0.883	0.895	0.539
0.5	90	0.89	0.901	0.522
0.5	95	0.894	0.912	0.513
0.75	0	0.854	0.878	0.597
0.75	80	0.872	0.89	0.561
0.75	85	0.876	0.893	0.552
0.75	90	0.883	0.896	0.538
0.75	95	0.892	0.91	0.519
1	0	0.854	0.877	0.596
1	80	0.867	0.883	0.571
1	85	0.87	0.886	0.565
1	90	0.875	0.89	0.554
1	95	0.887	0.908	0.53

Table 6.4 – Comparaison des performances de VQA pour différentes valeurs des paramètres λ_3 et n, en termes de CC, SROCC et RMSE. Les paramètres λ_1 et λ_2 sont choisis pour optimiser les performances. Les résultats concernent toute la base de test.

Dans cette expérience, les valeurs des paramètres λ_1 et λ_2 sont sélectionnées pour maximiser les performances. Le paramètre λ_3 modifie le comportement asymétrique du cumul temporel long terme. La modification de la prédiction de la qualité en fonction de λ_3 montre que le cumul temporel long terme avec un comportement symétrique ($\lambda_3 = 1$), obtient de moins bons résultats qu'un cumul temporel long terme avec un comporte-

ment asymétrique. Il est intéressant de noter que, pour atteindre la meilleure performance, le comportement asymétrique doit donner, au moins, deux fois plus de poids à l'augmentation du niveau de distorsions qu'à sa diminution. En outre, le choix empirique de la valeur de λ_3 ($\lambda_3 = 0.25$), semble être une bonne option.

Le paramètre n modifie le poids accordé aux gradients temporels maximums des valeurs de distorsions par image. Les plus mauvais résultats sont obtenus lorsque tous les gradients temporels des valeurs de distorsion par image sont pris en compte ($n = 0$). La modification de la prédiction de la qualité en fonction de n montre que le cumul temporel long terme tire avantage de l'utilisation des gradients temporels maximums des valeurs de distorsions par image. Même si les meilleures performances sont obtenues avec $n = 95$, les résultats sont robustes à une variation de n autour de cette valeur. Il est intéressant de noter que $n = 95$ signifie que les plus importantes variations de distorsions se produisant 5% du temps sont les plus significatives en terme de performance de la prédiction de la qualité. Cela renforce le fait que les variations de distorsions d'amplitude importante doivent être prises en considération dans l'élaboration de la note de qualité d'une séquence vidéo.

λ_3	$n$ième percentile	CC	SROCC	RMSE
0	0	0.831	0.872	0.638
0	80	0.831	0.872	0.638
0	85	0.831	0.872	0.638
0	90	0.831	0.872	0.638
0	95	0.832	0.869	0.636
0.25	0	0.831	0.872	0.638
0.25	80	0.831	0.872	0.638
0.25	85	0.831	0.868	0.638
0.25	90	0.832	0.867	0.636
0.25	95	0.834	0.863	0.633
0.5	0	0.831	0.872	0.638
0.5	80	0.831	0.868	0.638
0.5	85	0.832	0.866	0.636
0.5	90	0.833	0.87	0.634
0.5	95	0.839	0.866	0.624
0.75	0	0.831	0.872	0.638
0.75	80	0.832	0.868	0.636
0.75	85	0.833	0.867	0.635
0.75	90	0.834	0.869	0.633
0.75	95	0.846	0.869	0.611
1	0	0.831	0.872	0.638
1	80	0.832	0.867	0.636
1	85	0.833	0.87	0.634
1	90	0.835	0.869	0.632
1	95	0.85	0.865	0.605

Table 6.5 – Comparaison des performances de VQA_2 pour différentes valeurs des paramètres λ_3 et n, en termes de CC, SROCC et RMSE. Les paramètres λ_1 et λ_2 sont choisis pour optimiser les performances. Les résultats concernent toute la base de test.

Les résultats sont également indiqués pour VQA_2 (désactivation du cumul temporel court terme niveau fixation) et présentés dans le tableau 6.5 pour différentes valeurs des paramètres λ_3 et n. Dans cette expérience,

les valeurs des paramètres λ_1 et λ_2 sont choisies pour maximiser les performances. Les résultats montrent que le cumul temporel long terme n'améliore pas les performances lorsque le cumul temporel court terme est désactivé. Cette observation est valable quelles que soient les valeurs des paramètres λ_1, λ_2, λ_3 et n. Par conséquent, la nature fondamentale de l'étape de cumul court terme est renforcée par ces résultats.

6.4 Conclusion

Les objectifs de ce chapitre était la conception, le développement et l'évaluation d'un critère objectif de qualité visuelle de vidéos avec référence complète. La métrique que nous avons proposée est basée sur l'étude des variations temporelles des distorsions spatiales. Les variations temporelles de distorsions sont évaluées à deux niveaux : au niveau des fixations oculaires et sur l'ensemble de la séquence. Ces deux niveaux sont assimilés respectivement à un cumul temporel court terme et à un cumul temporel long terme. Le cumul temporel court terme était l'objet du chapitre 3. Le cumul temporel long terme a été présenté dans ce chapitre. Il comprend un comportement asymétrique permettant de donner un poids différent aux augmentations et aux diminutions des distorsions, et un effet de saturation perceptuelle permettant de limiter le poids des distorsions au-delà d'un certain niveau. Des tests subjectifs d'évaluation de qualité ont été menés afin d'évaluer les performances de cette métrique. La métrique objective de qualité proposée a été comparée aux notes subjectives, ainsi qu'à d'autres métriques issues de la littérature.

Tout d'abord, les résultats montrent les bonnes performances de la métrique proposée par rapport à des métriques de la littérature. Elle obtient des résultats significativement meilleurs que le PSNR et la VSSIM sur la base de tests utilisée. Par exemple, VQA obtient un CC de 0.892 alors que la VSSIM n'obtient qu'un CC de 0.738. Elle semble aussi supérieure à VQM, même si pour cette métrique la taille de la base de test n'est pas assez importante pour les différencier d'un point de vue statistique.

Les résultats obtenus par les trois versions simplifiées de notre critère VQA montrent aussi l'importance des différentes étapes de la métrique proposée. En particulier, la présence d'un cumul temporel court terme est fondamentale pour le cumul temporel long terme. En effet, sans le cumul temporel court terme les performances de la métrique diminuent. Par ailleurs, un aspect intéressant de la métrique proposée réside dans le fait que les cartes de distorsions spatiales peuvent être considérées comme une « entrée ». Nous avons utilisé ici la métrique WQA, issue de nos travaux, mais on peut imaginer la remplacer par une autre méthode produisant des cartes de distorsions visuelles spatiales encore plus réalistes.

Conclusion

La seconde partie de ce mémoire était consacrée à l'évaluation de la qualité d'images et de vidéos, ou autrement dit, à la construction d'un jugement global de qualité. Comme nous l'avons évoqué précédemment, l'évaluation de la qualité est un des besoins de l'industrie de l'image et de la vidéo. Pour répondre à ce besoin, nous avons présenté des métriques de qualité avec référence complète pour les images fixes, ainsi que pour les vidéos. Nous avons aussi présenté des méthodes d'évaluation subjective de la qualité. Ces méthodes permettent de construire une vérité terrain à partir de laquelle il est possible d'évaluer quantitativement les performances de métriques de qualité.

Concernant les images fixes, nos travaux ont consisté à concevoir des métriques de qualité reposant sur les modèles du système visuel humain proposé dans le chapitre 2 et utilisant une méthode de cumul spatial de la littérature. Ces métriques ont été évaluées à partir de plusieurs tests subjectifs. Les résultats les plus importants montrent d'une part l'importance de prendre en compte le masquage semi-local dans la modélisation des effets de masquage et d'autre part qu'il est possible de simuler le comportement multi-canal du système visuel à partir d'une transformée en ondelettes, sans pour autant que les performances n'en pâtissent.

Concernant les vidéos, nos travaux ont consisté à proposer une nouvelle approche d'évaluation de la qualité. Cette approche repose sur un cumul temporel long terme ainsi que sur le cumul temporel court terme proposé précédemment (cf. chapitre 3). Le cumul temporel long terme intègre un comportement asymétrique sur les variations instantanées de distorsions et un effet de saturation perceptuelle. Afin d'évaluer les performances de notre approche et de les comparer avec celles de métriques de la littérature, des tests subjectifs d'évaluation de la qualité de vidéo ont été menés. Les résultats montrent les bonnes performances de l'approche proposée ainsi que la nécessité des deux cumuls temporels utilisés. Par exemple, le PSNR, la VSSIM et VQA obtiennent respectivement un CC de 0.892, 0.516 et 0.738.

Cette partie nous a permis de construire un jugement de qualité d'images et de vidéos. Pour cela nous avons pris en compte toutes les distorsions visuelles. Cela suppose de considérer qu'un observateur était capable de toutes les voir et que toutes contribuaient à son jugement de qualité. La réalité est pourtant différente. En effet, les mécanismes de l'attention visuelle opèrent une sélection de l'information visuelle disponible. Il est donc naturel de se demander quelle est l'influence de l'attention visuelle dans la construction du jugement de qualité et si l'on peut s'en servir pour améliorer l'évaluation de la qualité. C'est ce que nous allons examiner dans la troisième et dernière partie de ce mémoire.

Troisième partie

Attention visuelle et construction du jugement de qualité visuelle

Introduction

Nous avons consacré la première partie de ce mémoire à évaluer localement la perception des distorsions dans les images et dans les vidéos. Ensuite, la seconde partie de ce mémoire fut consacrée à la construction d'une note objective de qualité visuelle à partir des distorsions perceptuelles locales évaluées dans la première partie. Pour y parvenir, nous nous sommes appuyés sur des modélisations de certaines parties du système visuel humain. Ces modélisations faisaient l'hypothèse que nous avions affaire à un « super observateur » capable de saisir la totalité de l'image ou de la vidéo à évaluer en une seule fois. Autrement dit, que toutes les zones spatiales pour les images ou toutes les zones spatio-temporelles pour les vidéos étaient regardées.

Les modélisations proposées dans les parties précédentes peuvent être qualifiées de fovéales, car elles modélisent la sensibilité de cette partie de la vision humaine. En réalité, le « super observateur » n'existe pas et un observateur humain n'est pas capable de traiter toute l'information visuelle disponible avec la sensibilité de la zone fovéale. En effet, l'énorme quantité d'informations visuelles disponibles dans notre environnement dépasse les capacités de traitement du système visuel humain dont les ressources sensorielles et mécaniques sont limitées. Pour faire face à ces limitations, le système visuel humain possède la faculté de sélectionner l'information pertinente localisée spatialement dans son champ visuel. On parle alors d'attention visuelle. L'attention visuelle permet de déplacer la zone fovéale de l'oeil successivement sur différentes parties d'une image ou d'une vidéo. On peut donc s'interroger sur le rôle de l'attention visuelle dans un contexte d'évaluation de la qualité visuelle.

Les modélisations proposées dans les parties précédentes, sont des modélisations « du pire des cas », dans le sens où toutes les distorsions de l'image ou de la vidéo sont prises en compte pour construire la note objective de qualité. Un observateur humain construit son jugement de qualité à partir de l'information qu'il a perçue. L'attention visuelle jouant un rôle dans la sélection de cette information, elle devrait donc aussi jouer un rôle dans la construction du jugement de qualité.

La première idée qui vient à l'esprit est de ne prendre en compte que les zones de l'image ou de la vidéo qui ont été effectivement regardées pour construire une note objective de qualité. Mais comment déterminer les zones qui ont été perçues ? Il existe bien des modèles d'attention visuelle, mais quelle attention visuelle doit-on modéliser ? En effet, les notes subjectives de qualité que l'on utilise pour évaluer les performances des métriques de qualité sont collectées au travers de tests subjectifs de qualité. Est-ce que dans ce contexte d'évaluation de qualité, l'attention visuelle est la même que dans un contexte d'exploration libre ? Même en considérant que nous connaissons les zones que les observateurs ont perçues, comment en tenir compte dans une métrique de qualité ? Est-ce vraiment aussi simple que de donner plus de poids aux zones les plus regardées ?

Le rôle de l'attention visuelle dans l'évaluation subjective et objective de la qualité d'images ou de vidéos soulève nombre de questions auxquelles nous allons tenter d'apporter des réponses dans cette troisième partie. Celle-ci se décompose en trois chapitres. Dans le premier chapitre nous allons décrire l'attention visuelle et la littérature la reliant à l'évaluation de la qualité visuelle. Les deux chapitres suivants sont consacrés à l'étude de l'attention visuelle dans un contexte d'évaluation subjective et objective de la qualité visuelle. Ces deux

chapitres sont dédiés pour l'un à l'évaluation de la qualité d'images et pour l'autre à l'évaluation de la qualité de vidéos.

Chapitre 7

État de l'art sur l'attention visuelle

7.1 Introduction

L'objectif de ce chapitre est de présenter un état de l'art sur les mécanismes de l'attention visuelle et de faire un lien avec la littérature sur l'évaluation de la qualité visuelle d'images ou de vidéos. Ainsi, la première partie est consacrée à l'attention visuelle proprement dite. Nous verrons les différents mécanismes de l'attention visuelle ainsi que les mouvements oculaires permettant de les traduire concrètement. Dans une seconde partie, nous nous intéressons à la littérature traitant conjointement de l'attention visuelle et de l'évaluation de la qualité visuelle. Ce chapitre nous servira de base pour comprendre et interpréter les résultats des chapitres suivants.

7.2 Les mouvements oculaires et l'attention visuelle

7.2.1 Les mouvements oculaires

Comme nous l'avons introduit dans la première partie de ce mémoire, le système visuel humain est intrinsèquement limité. Pour pallier cette limitation, celui-ci utilise les mouvements oculaires pour mobiliser ses ressources de traitement de l'information visuelle. Ces mouvements oculaires prennent la forme de mouvements de poursuites, de convergences, de saccades ou encore de fixations. Les deux mouvements oculaires principaux, associés à la focalisation dite *overt*, sont les fixations et les saccades. Contrairement à ce que laisse penser leur nom, les fixations sont considérées comme un type de mouvement. Ces deux types de mouvements sont décrits dans les paragraphes suivants.

7.2.1.1 Les fixations

Une phase de fixations correspond à une phase pendant laquelle le regard est stationnaire. Cette phase se produit lorsque l'oeil fixe un objet de l'environnement visuel d'un observateur. Le terme fixation vient de l'apparente position stationnaire de l'oeil durant cette phase. Cependant, les fixations sont considérées comme des mouvements oculaires à cause des mouvements résiduels de l'oeil pendant cette phase. Ces légers mouvements permettent de modifier continuellement la zone examinée par la fovéa afin que cette dernière ne soit pas exposée à un signal constant. Si l'oeil était réellement stationnaire, c'est-à-dire en vision complètement stabilisée, la

perception visuelle disparaîtrait progressivement à cause d'adaptation rapide des cellules photoréceptrices et en particulier des cônes. Comme expliqué en annexe A, l'acuité visuelle est maximale au centre de la fovéa, là où sont concentrés les cônes. Les phases de fixations sont généralement séparées par des saccades.

7.2.1.2 Les saccades

Les saccades sont des mouvements oculaires très rapides (vitesse entre 100 et 700 degrés par seconde) [Salvucci 99]. Ce type de mouvement permet de déplacer le regard d'un endroit à un autre afin de les inspecter avec la partie la plus performante (en terme de résolution spatiale) de la rétine : la fovéa. Les saccades sont souvent considérées comme un mécanisme favorisant la sélection des informations visuelles pertinentes de notre champ visuel. L'exploration de notre environnement visuel se fait donc par une série de sauts permettant le déplacement rapide de nos ressources sensorielles d'un point à un autre. Le passage d'un point à un autre ne se fait pas forcément par le plus court chemin, c'est-à-dire la ligne droite. La trajectoire peut en effet être incurvée. De plus, plusieurs saccades peuvent être nécessaires pour atteindre une cible spécifique, tout dépend de la précision de la première la saccade. Durant ces mouvements très rapides, le pouvoir d'analyse du système visuel est très faible et pratiquement aucune information visuelle n'est traitée.

7.2.1.3 Les autres types de mouvement

Les autres mouvements oculaires, d'importance secondaire, sont brièvement décrits ci-dessous :

- les mouvements de poursuite : ce type de mouvement oculaire permet de maintenir notre regard sur un objet en mouvement. Son rôle est important, car il permet de stabiliser sur la rétine l'objet en mouvement. Ainsi, l'image de cet objet peut être examinée par la fovéa avec un fort pouvoir de résolution. La vitesse angulaire maximale de poursuite est d'environ $30°/s$;
- les mouvements de vergence (convergence et divergence) sont des mouvements pour lesquels les axes visuels des deux yeux se déplacent dans des directions horizontales opposées. Ces mouvements permettent d'assurer la fusion binoculaire, c'est-à-dire le processus permettant d'avoir la sensation de percevoir une image unique à partir des deux images rétiniennes (droite et gauche). Ce type de mouvement est utile pour acquérir des informations visuelles d'un objet situé dans un plan focal différent de celui de notre regard (plus proche ou plus éloigné). Par exemple, lorsque l'objet fixé par un observateur se rapproche de lui, ses yeux ajustent leur position en rapprochant leur axe visuel (mouvement de convergence) pour maintenir la fusion binoculaire alors que, lorsque l'objet s'éloigne de lui, les axes visuels de ses yeux ont plutôt tendance à s'écarter (mouvement de divergence). Ces mouvements sont négligeables lorsque l'on regarde un écran car toutes les informations visuelles appartiennent à un même plan.

7.2.2 L'attention visuelle sélective

7.2.2.1 Définition

L'attention visuelle désigne le mécanisme de sélection des informations visuelles spatio-temporelles pertinentes du monde visible. Notre environnement visuel produisant une quantité d'informations visuelles supérieure à la capacité de traitement de notre système visuel, celui-ci s'est adapté en mettant en place des mécanismes ou des stratégies bien particulières pour sélectionner les informations à effectivement traiter. En d'autres termes, l'attention visuelle nous permet d'utiliser de façon optimisée nos ressources biologiques ; ainsi, seule une petite partie des informations incidentes est transmise aux aires supérieures de notre cerveau [Ballard 91]. En 1993 R. Milanese [Milanese 93], puis plus tard en 1995 J. K. Tsotsos [Tsotsos 95] décrivent le mécanisme d'attention visuelle comme étant des répétitions de phases de sélection (détection et localisation) et de focalisation (mouvement oculaire ou focalisation interne).

7.2.2.2 Les mécanismes de sélection dits passifs

Les mécanismes de sélection dits passifs de l'information sont liés aux caractéristiques intrinsèques du système visuel humain (cf. Partie I), abstraction faite des mouvements oculaires. Les principaux mécanismes passifs de sélection de l'information visuelle sont rappelés ci-dessous :

- le premier mécanisme et le plus évident concerne la transduction photoélectrique (transformation de la lumière en signal interprétable par le cerveau). Cette transformation ne concerne qu'une bande étroite du spectre global de la lumière incidente, appelée la lumière visible ;
- l'information est échantillonnée par les cellules photosensibles de façon non uniforme : la restitution de la résolution spatiale de l'information est plus importante dans la zone fovéale que sur le reste de la rétine ;
- les cellules visuelles présentent une sensibilité aux fréquences spatiales ; en d'autres termes, nous ne sommes pas en mesure d'apprécier tous les détails de notre environnement visuel avec le même degré de précision ;
- les cellules rétiniennes et corticales suppriment la redondance d'informations ; elles répondent uniquement aux contrastes.

7.2.2.3 Les mécanismes de sélection dit actifs

Une illustration souvent utilisée pour décrire l'attention visuelle est la métaphore du faisceau lumineux (*spotlight of attention*) [Neisser 67] où l'attention est comparée à un faisceau lumineux illuminant les zones de notre champ visuel qui sont inspectées. La focalisation d'attention, c'est-à-dire l'inspection d'une zone particulière, peut se faire de deux façons : une focalisation dite *overt* ou une focalisation dite *covert*. Le premier type de focalisation se manifeste directement par un mouvement oculaire. Le deuxième, quant à lui, ne met pas directement en jeu un mouvement oculaire. Cette focalisation utilise la vision périphérique, c'est-à-dire la vision en bordure du champ visuel (zone parafovéale). Cette forme d'attention est particulièrement bien mise en évidence chez les malentendants [Bavelier 00, Muir 03]. En effet, des expériences oculométriques, utilisant des séquences d'images présentant une personne traduisant un discours en langage de signes, ont montré que l'attention fovéale des

malentendants se portait essentiellement sur le visage de la traductrice. En dépit du fait qu'ils ne fixaient pas directement les mains de la traductrice, ils étaient tout à fait capables de retranscrire le discours.

La théorie binaire de l'attention visuelle, dont J.Braun et D. Sagi [Braun 90] sont à l'origine, distingue deux mécanismes :

- un mécanisme exogène (pré-attentif) [Posner 80] ou plus communément appelé *Bottom-Up* sélectionnant les informations visuelles selon leur saillance. C'est un mécanisme involontaire, et relativement éphémère, guidé par les caractéristiques des différentes zones de notre champ visuel (déplacement oculaire vers les zones capturant notre attention). Ce mécanisme de sélection se fait donc sans aucune connaissance a priori.
- le second mécanisme est dit endogène (attentif) [Posner 80] ou *Top-Down*. Notre attention et le déplacement oculaire s'effectuent sous un contrôle volontaire et cognitif. En d'autres termes, ce mécanisme est guidé par la tâche à accomplir. Par conséquent, le déploiement de l'attention visuelle (le trajet oculaire), sera lui aussi dépendant de la tâche.

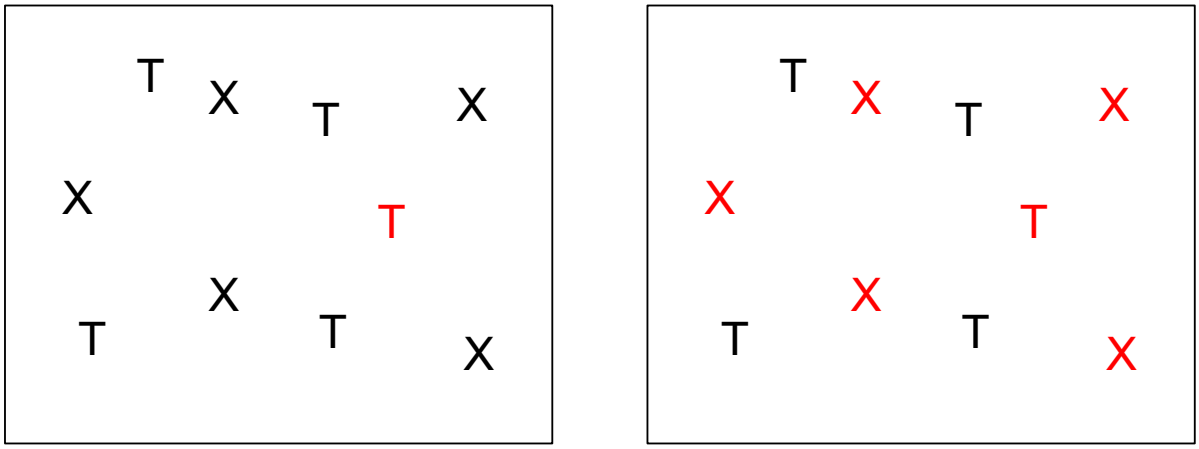
La description de ces mécanismes peut être complétée par quelques mots sur la théorie de l'intégration de caractéristiques (*Feature Integration Theory*, abrégé FIT) de A. Treisman et G. Gelade [Treisman 80]. Ces travaux reposent sur des expériences de recherche visuelle. Le principe de ces expériences consiste à mesurer le temps de réaction nécessaire pour discriminer un objet cible enfoui parmi d'autres objets communément appelés distracteurs. Les objets peuvent être simples, c'est-à-dire constitués d'une seule dimension visuelle (la couleur, l'orientation, la forme, etc.) ou composés de plusieurs dimensions (objet coloré orienté par exemple). Les expériences effectuées révèlent deux comportements distincts :

- si la cible diffère des distracteurs d'au moins une caractéristique visuelle, cas disjonctif (exemple de la figure 7.1(a)), alors le temps de réaction nécessaire pour résoudre la recherche visuelle est constant et cela quel que soit le nombre de distracteurs. Bien souvent, on considère que la cible saute aux yeux (dans la littérature scientifique, le verbe anglais *to pop-out* est très souvent utilisé) ;
- par contre, si la cible est une combinaison de caractéristiques (exemple de la figure 7.1(b)), le temps de réaction augmente linéairement avec le nombre de distracteurs. Dans ce cas, appelé cas conjonctif, la recherche de la cible est séquentielle puisque tous les objets sont scrutés afin de déterminer la cible.

Ainsi, le cas disjonctif est à rapprocher du mécanisme *Bottom-up* qui, finalement, permet de traiter les caractéristiques visuelles d'une scène rapidement et d'une façon massivement parallèle. Le cas conjonctif, quant à lui, est à rapprocher du mécanisme *Top-Down* qui est un mécanisme lent et traitant les informations visuelles de façon séquentielle ou série. On parle également de la dichotomie attentif/pré-attentif, mentionnée dans les travaux de A. Treisman et G. Gelade [Treisman 80] et de J.M. Wolfe [Wolfe 04], qui supposent un premier traitement automatique sur l'ensemble du champ visuel suivi d'un traitement localisé déployé par l'observateur.

7.2.2.4 Le mécanisme inhibiteur de l'attention visuelle

En marge de l'attention volontaire *Top-Down* ou involontaire *Bottom-Up*, un autre mécanisme intéressant, appelé inhibition de retour, en abrégé IOR (*Inhibition Of Return*), est à considérer. L'inhibition de retour



(a) (b)

Figure 7.1 – Exemples d'expériences de recherche visuelle : (a) une seule lettre T rouge et les autres lettres noires : cas disjonctif (traitement parallèle) ; (b) mélange de lettres T et X rouges et noires : cas conjonctif (traitement série).

consiste à inhiber une zone inspectée afin d'éviter le retour continuel de l'attention visuelle sur cette même zone. Grâce à ce mécanisme, les différentes régions du champ visuel sont explorées séquentiellement. Par exemple, dans le cadre d'une recherche visuelle de type conjonctive, l'inhibition de retour est primordiale puisqu'elle évite à l'observateur de continuellement re-tester les mêmes objets [Klein 99]. D'après les études de M. Posner et Y. Cohen [Posner 84], l'inhibition de retour n'a lieu que lorsque la durée d'inspection d'une zone est supérieure à 300 ms. Cela correspond à une phase de fixations dont la durée est supérieure à la durée moyenne des fixations mesurées dans nos expérimentations en exploration libre d'images (cf. figure 8.3.1).

7.2.2.5 Les caractéristiques visuelles attirant le regard

Nous venons de voir le caractère sélectif de l'attention visuelle humaine, ce qui signifie que le système visuel humain répond de façon privilégiée à certains types de signaux provenant des objets et des évènements de notre environnement. Parmi ces signaux on peut citer le cas typique de l'apparition inattendue d'un objet dans une scène [Yantis 96]. De façon plus générale, l'attention visuelle réagit à ce qui est appelé des singularités locales [Treisman 80]. La figure 7.1(a) donne un exemple courant de singularité locale, basée ici sur la couleur rouge d'une lettre qui « saute aux yeux » en comparaison de la couleur noire des autres lettres. Par ailleurs, la sémantique joue aussi un rôle important dans le déploiement de l'attention visuelle, et l'incohérence d'objets avec le contexte de la scène attire notre attention [Henderson 99], on parle d'objets saillants sémantiquement. Enfin, différentes études [Mannan 97, Reinagel 99] ont cherché à estimer les caractéristiques visuelles attirant notre regard, à partir de points de fixation réels. Ces études montrent d'une part que les régions fixées présentent un contraste (de luminance, de couleur, de texture [Parkhurst 04], de mouvement, etc.) plus important que les autres régions, d'autre part que les régions fixées différent de leur voisinage. Ces conclusions révèlent que le système visuel tend à maximiser l'information à transmettre au cerveau en minimisant la redondance spatio-temporelle de celle-ci.

7.3 Attention visuelle et évaluation de qualité

7.3.1 Attention visuelle et évaluation subjective de la qualité

L'attention visuelle a été peu étudiée dans un contexte d'évaluation subjective de la qualité d'images ou de vidéos. L'une des rares études sur le sujet est celle de Vuori et *al.* [Vuori 04, Vuori 06]. Les auteurs ont réalisé conjointement des tests oculométriques et des tests subjectifs d'évaluation de la qualité d'images afin de découvrir s'il était possible de prédire la qualité subjective d'images en se basant sur les mouvements oculaires. Deux types de dégradations ont été introduits dans les images évaluées à savoir : une modification du contraste et du flou. Trois tâches étaient demandées successivement aux observateurs : évaluer la qualité globale, évaluer la qualité des couleurs et une tâche particulière pour chaque image comme par exemple : compter le nombre de bâtiments dans l'image X, ou évaluer l'état émotionnel de la personne présente sur l'image Y. Seule la dimension temporelle de l'attention visuelle a été étudiée, et cela au travers de la durée des saccades. Ces expérimentions montrent deux résultats intéressants :

- la tâche demandée aux observateurs a une influence sur la durée des saccades. Les saccades sont significativement plus longues dans les tâches d'évaluation de qualité que dans la troisième tâche ;
- la qualité des images présentées a une influence significative sur la durée des saccades. La durée des saccades augmente avec la diminution de la qualité subjective des images.

Vuori et *al.* avancent donc l'hypothèse que la qualité subjective des images a une influence sur le déploiement de l'attention visuelle.

Plus récemment, les travaux de Vu et *al.* [Vu 08] se sont intéressés à l'influence des distorsions sur le déploiement spatial de l'attention visuelle. Leur étude repose sur la réalisation de tests oculométriques. Ces tests ont été menés sur des images fixes en exploration libre ainsi qu'en évaluation de qualité. Ces travaux étant postérieurs à ceux que nous présentons dans la troisième partie de ce mémoire, ils ne sont pas inclus dans l'état de l'art. Cependant, nous reviendrons sur ces travaux dans la section 8.3.

A notre connaissance, il n'existe pas d'études dans le cas de la vidéo.

7.3.2 Attention visuelle et évaluation objective de la qualité

Si les mécanismes de l'attention visuelle ont été peu étudiés dans un contexte d'évaluation subjective de la qualité d'images ou de vidéos, plusieurs auteurs l'ont utilisée dans le but d'améliorer les performances de métriques objectives de qualité d'images. L'idée sous-jacente étant que les zones les plus saillantes d'une image doivent avoir un poids plus important dans l'évaluation de qualité que les autres zones. La prise en compte de l'attention visuelle dans un critère objectif de qualité se résume généralement par une pondération spatiale des distorsions calculées par le critère en fonction de cartes d'importance, ou de cartes de régions d'intérêt (*Region of Interest : ROI*). Ces cartes sont calculées à partir de modèles d'attention visuelle plus au moins élaborés.

On peut citer les travaux d'Osberger et *al.* [Osberger 98], qui utilisent une carte d'importance IM (*Importance Map*) pour pondérer les cartes d'erreurs perceptuelles calculées dans une métrique de qualité d'images avec

référence. Ces cartes d'importante sont calculées à partir de 5 caractéristiques des différentes régions de l'image de référence obtenues par segmentation : le contraste, la taille, la forme, la position, et une classification premier plan/arrière plan. La pondération de la carte d'erreurs perceptuelles $PDM(x, y)$ par la carte d'importance $IM(x, y)$ est réalisée selon la relation suivante :

$$IPDM(x, y) = PDM(x, y) \cdot IM(x, y)^{\gamma}, \tag{7.1}$$

où $IPDM(x, y)$ représente la carte d'erreurs perceptuelles pondérée. $IM(x, y)$ variant entre 0 et 1, et $\gamma = 1$. Les auteurs obtiennent une amélioration de la qualité de la prédiction lorsque la carte d'importance est utilisée. Les tests subjectifs ont été réalisés par 18 observateurs. La base de test utilisée contient 44 images dont 32 sont dégradées par du codage par ondelettes et du codage JPEG, les 12 autres images sont encodées à haut débit sur les régions d'intérêt et à bas débit ailleurs hors ces régions.

Dans [Barland 06], les auteurs proposent une métrique de qualité sans référence basée sur la combinaison de mesures locales de flou et de *ringing*. Une carte d'importance est calculée par un modèle de saillance multi-résolution basé sur les contrastes achromatique et couleur. Une pondération linéaire des mesures locales de flou et de *ringing* par la carte d'importance est aussi proposée. Les pondérations des deux mesures locales par la carte d'importance sont similaires. Par exemple, la pondération de la mesure de flou BM est réalisée par :

$$BM = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} IM(i,j) \cdot A'_{Edge}(i,j) \cdot I_A^2(i,j)}{\sum_{i=1}^{M}\sum_{j=1}^{N} IM(i,j) \cdot A_{Edge}(i,j) \cdot I_A^2(i,j)} \cdot \frac{\frac{N(A'_{Edge})}{M \cdot N}}{\frac{N(A_{Edge})}{M \cdot N}}, \tag{7.2}$$

où $IM(i, j)$ représente la carte d'importance, $I_A(i, j)$ la composante achromatique, $A_{Edge}(i, j)$ et $A'_{Edge}(i, j)$ sont des images binaires représentant respectivement la détection des contours de l'image et son complémentaire. $N(A_{Edge})$ et $N(A'_{Edge})$ représentent respectivement le nombre de pixels non nuls de $A_{Edge}(i, j)$ et $A'_{Edge}(i, j)$. Les auteurs montrent une légère amélioration des performances lorsque la carte d'importance est utilisée. Les performances sont évaluées sur un sous-ensemble des images dégradées par un codage JPEG2000 de la base LIVE [1].

Dans un contexte vidéo, Lu et *al.* dans [Lu 04, Lu 05] calculent des cartes d'importance de la qualité perçue *Perceptual Quality Significance Map : PQSM*) à partir d'un modèle d'attention visuelle *Bottom-Up* et *Top-Down*. La modélisation du mécanisme *Bottom-Up* est réalisée à partir du mouvement, du contraste de couleur et du contraste de texture. La modélisation du mécanisme *Top-Down* est réalisée à partir de la détection de visage et de teinte chair. Les auteurs utilisent ces cartes pour moduler des modèles de JND (*Just Noticeable Difference*), lesquels sont ensuite utilisés pour piloter l'insertion de bruit dans des vidéos. Des tests de préférences montrent que les séquences préférées sont celles dont l'insertion de bruit a été pilotée par les modèles de JND modulés par les cartes d'importance de la qualité perçue (PQSM). Ce résultat encourage l'utilisation de l'attention visuelle dans un contexte d'évaluation de qualité. Par contre leurs résultats ne permettent pas de mettre en évidence la contribution des cartes d'importance de la qualité perçue seules pour améliorer des métriques de qualité. En effet, dans leurs expérimentations sur des métriques existantes, les auteurs comparent les critères PSNR

1. `http://live.ece.utexas.edu/research/quality`

et SSIM (moyenne spatio-temporelle) avec des versions pondérées par les modèles de JND modulés par les cartes d'importance de la qualité perçue. Il n'est donc pas possible d'évaluer uniquement l'impact des cartes d'importance de la qualité perçue.

Il est intéressant de noter que dans la littérature présentée les pondérations proposées sont généralement des pondérations linéaires. Les résultats ont tendance à montrer un impact positif de l'utilisation de ce qui est appelé « carte d'importance ». Cependant, l'interprétation des résultats de ces travaux est délicate, car deux problématiques sont toujours cumulées. La première pose la question de la validité des modèles d'attention visuelle utilisés. La seconde problématique est celle de l'utilisation des cartes de saillance, ou des cartes d'importance, dans l'évaluation objective de la qualité. Comment combiner l'information de saillance avec les distorsions afin de construire une note objective de qualité ?

Les travaux récents de Larson *al.* [Larson 08] abordent la seconde problématique en utilisant l'information de saillance issue de tests oculométriques pour pondérer des métriques de qualité existantes. Ces travaux étant postérieurs à ceux que nous présentons dans la troisième partie de ce mémoire, ils ne sont pas inclus dans l'état de l'art. Cependant, nous reviendrons sur ces travaux dans la section 8.4.

7.4 Conclusion

La vocation de ce chapitre était d'une part de présenter des connaissances générales sur l'attention visuelle, d'autre part d'introduire les liens possibles entre évaluation de qualité et attention visuelle.

La première partie de ce chapitre était consacrée à l'attention visuelle. Nous y avons décrit les principaux types de mouvements oculaires permettant le déploiement de l'attention visuelle : les fixations, les mouvements de poursuite et les saccades. Nous avons également présenté les principaux mécanismes de l'attention visuelle dont les mécanismes de sélection actifs dit *Bottom-up* et *Top-down*.

La seconde partie de ce chapitre était dédiée à la littérature étudiant simultanément l'attention visuelle et l'évaluation subjective ou objective de la qualité visuelle. La littérature sur le sujet est assez restreinte et beaucoup de questions restent en suspens. En ce qui concerne l'évaluation subjective de la qualité, il semble que la tâche d'évaluation de qualité, ainsi que les distorsions présentent dans les images, aient une influence sur le déploiement de l'attention visuelle des observateurs, en particulier concernant la durée des saccades. Cependant les résultats expérimentaux restent assez limités pour les images fixes et semblent inexistants pour les vidéos. En ce qui concerne l'évaluation objective de la qualité, les auteurs cherchent à utiliser des modèles d'attention visuelle pour améliorer les performances de métriques de qualité en pondérant linéairement les distorsions par des cartes d'importance. Les résultats semblent encourageants, cependant ces travaux cumulent deux problématiques : d'une part la pertinence des modèles d'attention visuelle et d'autre part l'utilisation de l'attention visuelle pour améliorer les performances de ces métriques. Dans les deux prochains chapitres nous allons tenter d'apporter des éléments de réponses à une seule des deux problématiques.

Chapitre 8

Attention visuelle et construction du jugement de qualité d'images

8.1 Introduction

L'objet de ce chapitre est l'étude de l'attention visuelle en évaluation de qualité d'images. Deux aspects distincts sont abordés dans ce chapitre. Le premier aspect, concernant plutôt l'évaluation subjective, s'intéresse à la stratégie visuelle déployée par les observateurs durant une campagne de tests subjectifs d'évaluation de qualité. Il s'agit d'étudier l'impact d'une tâche d'évaluation de qualité sur l'attention visuelle par rapport à une situation d'exploration libre. Le second aspect, concernant plutôt l'évaluation objective, s'intéresse à l'utilisation de l'information de saillance pour améliorer des métriques de qualité. Contrairement à la littérature sur le sujet, l'information de saillance utilisée dans cette étude n'est pas issue d'un modèle d'attention visuelle, mais provient de tests réalisés sur des observateurs. L'utilisation de l'information de saillance issue de tests oculométriques permet de séparer les deux problématiques suivantes : la pertinence des modèles d'attention visuelle et l'utilisation de l'attention visuelle pour améliorer les performances de métriques de qualité.

Afin de collecter les données nécessaires à l'étude de ces deux aspects, nous avons effectué des tests oculométriques sur un ensemble d'observateurs. Un test oculométrique consiste à enregistrer les mouvements oculaires des observateurs. Ces tests ont été réalisés d'une part dans une situation d'exploration libre et d'autre part, durant une campagne d'évaluation subjective de la qualité d'images.

La première partie de ce chapitre est consacrée à la description des tests expérimentaux. La seconde partie est dédiée à l'étude de l'impact de la tâche d'évaluation de qualité sur l'attention visuelle. Finalement, la troisième partie de ce chapitre est focalisée sur l'utilisation de la saillance en évaluation objective de la qualité.

8.2 Expérimentations oculométriques et tests subjectifs de qualité

Comme introduit précédemment, les tests oculométriques ont été réalisés dans deux situations différentes, correspondant chacune à une tâche particulière :

– une tâche d'exploration libre (*Free viewing task* ou *Free-task*), où les observateurs ont pour indication de

regarder les images le plus naturellement possible.

– une tâche d'évaluation de qualité (*Quality-task*), où les observateurs ont pour indication d'évaluer la qualité des images qui leur sont présentées.

Les images de la base de test *IVC* (décrite section 5.3.1.1) ont été utilisées à la fois pour les tests oculométriques en exploration libre et à la fois pour les tests oculométriques en tâche d'évaluation de qualité.

Dans cette section nous allons décrire d'abord le dispositif oculométrique, puis les différents tests oculométriques, et enfin la construction de la saillance spatiale.

8.2.1 Dispositif oculométrique : l'oculomètre

Comme nous l'avons déjà évoqué, le dispositif oculométrique, appelé aussi oculomètre, est un appareil permettant d'enregistrer les mouvements oculaires d'un observateur humain. Le dispositif[1] utilisé dans nos expérimentations est présenté figure 8.1.

Figure 8.1 – Dispositif oculométrique.

L'avantage de ce dispositif est qu'il ne perturbe pas la vision de l'observateur, par contre son inconvénient est son relatif inconfort. Ce dispositif illumine l'oeil avec une source de lumière infrarouge (donc invisible), et capture une image infrarouge de l'oeil au moyen d'une caméra, elle aussi, infrarouge. Cette image infrarouge de l'oeil est ensuite utilisée pour calculer la direction du regard. Cette image est analysée pour y détecter la position de la pupille, et la position de deux des quatre images dites de « Purkinje » (ou de « Purkinje-Sanson »). Les deux images de Purkinje exploitées ici, sont les reflets de la source infrarouge sur différentes parties de l'oeil :

– la première image de Purkinje (P1) est le reflet de la source infrarouge sur la cornée,

– la quatrième image de Purkinje (P4) est le reflet de la source infrarouge sur le cristallin.

La direction du regard est ensuite déterminée à partir des positions relatives de la pupille et des deux images (ou reflets) de Purkinje.

Afin de permettre au dispositif de fonctionner, il est nécessaire d'immobiliser au maximum la tête de l'observateur. L'observateur doit donc poser son menton sur un support rigide horizontal réglable en hauteur et son

1. L'oculomètre de la société *Cambridge Research System*

front contre une sangle transversale. Le dispositif est composé d'une structure verticale sur laquelle est fixée (cf. figure 8.1) :

- La caméra, le miroir et la source infrarouge,
- Le support de menton,
- Le support frontal.

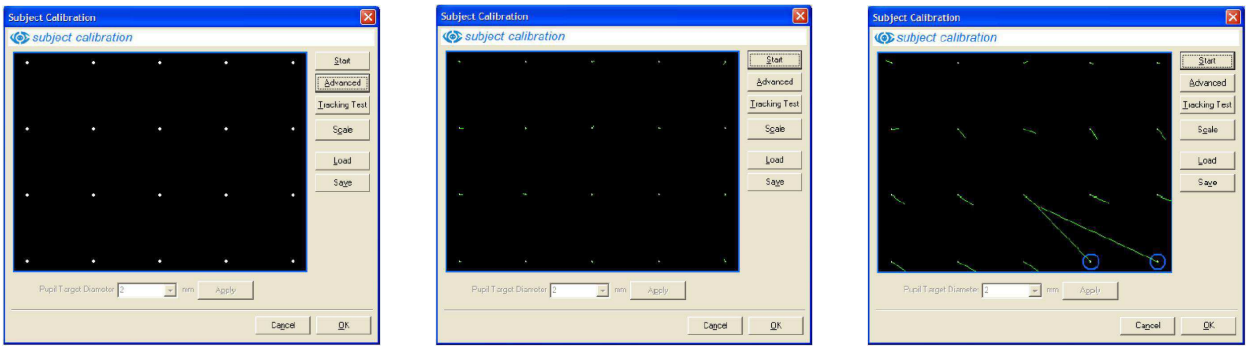
Les caractéristiques techniques de l'oculomètre sont données dans le tableau 8.1

Caractéristiques techniques	
Technique de mesure	pupille et images de Purkinje
Fréquence d'échantillonnage	50Hz
Résolution	0.1°
Précision	$0.25 - 0.5°$
Excursion horizontale, verticale	$\pm 40°$, $\pm 20°$
Mouvement de tête autorisé	$\pm$10mm

Table 8.1 – Caractéristiques techniques de l'oculomètre.

L'utilisation du dispositif oculométrique nécessite une phase de calibrage. Durant cette phase, il est demandé à l'observateur de fixer une série de points s'affichant séquentiellement sur l'écran. La correspondance entre la position des points affichés sur l'écran d'une part, et la position du regard correspondant mesuré par le dispositif d'autre part, est calculée et donnée comme résultat à cette phase :

- si la correspondance est bonne (cf. figure 8.2(b)) le test oculométrique peut commencer,
- si la correspondance n'est pas bonne (cf. figure 8.2(c)), le calibrage est à refaire. Les principales raisons d'un échec de la phase de calibrage sont un mauvais réglage de l'oculomètre (position de l'observateur par rapport à la caméra, réglage de la caméra, etc.), ou la morphologie de l'oeil de l'observateur comme par exemple une paupière trop basse venant recouvrir une partie de la pupille. Lorsque le problème est lié à l'observateur, il n'est pas toujours possible de trouver une solution à l'échec du calibrage et dans certains cas l'observateur doit être éliminé.



(a) (b) (c)

Figure 8.2 – Phase de calibrage du dispositif oculométrique : (a) les points séquentiellement affichés sur l'écran, (b) succès de la phase de calibrage, (c) échec de la phase de calibrage.

8.2.2 Exploration libre

Durant cette partie des tests, vingt images sont présentées aux observateurs. Dix images sont les images originales de la base *IVC*, et dix images sont des versions fortement dégradées des dix images originales. Chaque image est présentée pendant huit secondes, pendant lesquelles l'observateur est libre d'explorer l'image à sa convenance. Entre deux présentations, une image uniforme de gris moyen est affichée pendant 3 secondes.

Les vingt images sont regroupées et présentées en trois listes : deux listes de huit images et une liste de quatre images. L'affichage d'une liste d'images commence systématiquement par une phase de calibrage. Le déroulement de l'affichage d'une liste d'images est illustré figure 8.3.

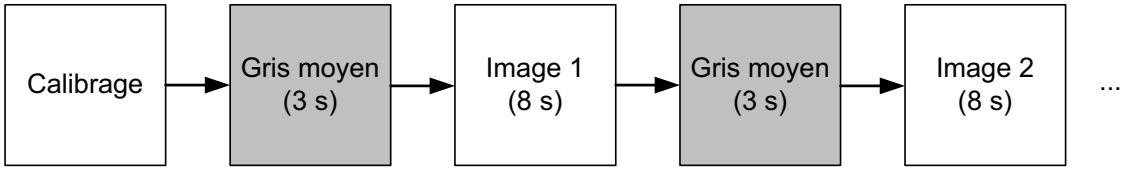


Figure 8.3 – Déroulement de l'affichage d'une liste d'images en expérimentation libre.

8.2.3 Tâche d'évaluation de qualité

Durant cette partie des tests, une campagne de tests subjectifs d'évaluation de qualité est menée de bout en bout. Les résultats de cette campagne sont d'ailleurs utilisés dans la section 5.3. Les 120 images dégradées de la base *IVC* sont évaluées en utilisant le protocole DSIS (*Double Stimulus Impairment Scale*), et chaque présentation permet d'évaluer la qualité d'une image. Le déroulement d'une présentation est illustré figure 8.4.

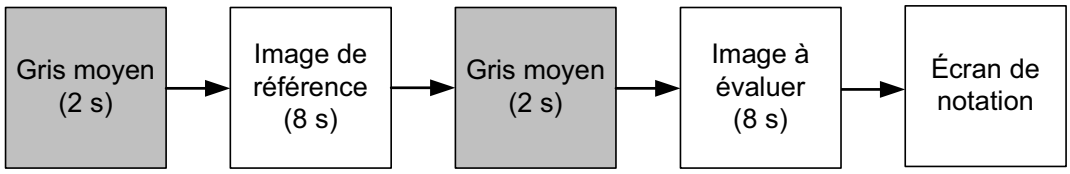


Figure 8.4 – Déroulement d'une présentation en tâche d'évaluation de qualité (protocole DSIS).

Une présentation consiste à afficher successivement une image de référence, et une version à évaluer de cette image. Les images de référence correspondent aux images originales de la base. Chaque image est présentée pendant huit secondes, une image uniforme de gris moyen étant présentée pendant deux secondes avant chacune d'entre elle. Une fois le couple d'images présenté, un écran de notation est affiché (cf. figure 8.5). Cet écran de notation permet à l'observateur d'utiliser le dispositif oculométrique pour enregistrer sa note, autrement dit de « voter avec ses yeux ».

L'intérêt de ce système de notation est d'éviter une phase de calibrage à chaque notation. En effet, si l'observateur devait utiliser un autre système (papier, clavier, voix, etc.) pour enregistrer sa note, le risque qu'il

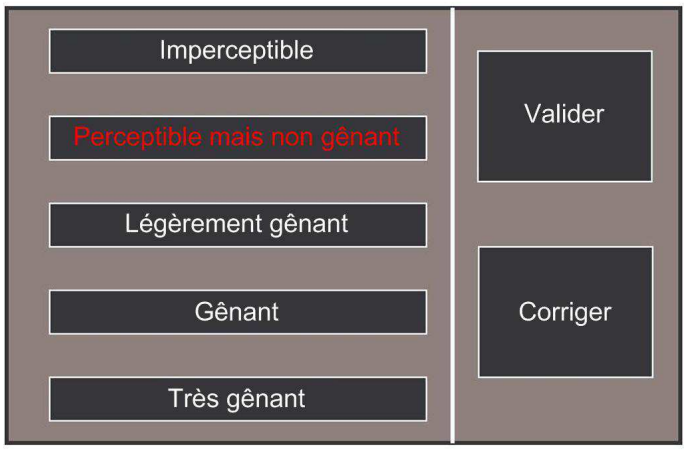


Figure 8.5 – Écran de notation présenté à la fin de chaque présentation et reprenant l'échelle de dégradations à cinq niveaux.

bouge la tête serait trop important et une phase de calibrage serait alors nécessaire pour garantir la qualité des mesures oculométriques suivantes.

Les 120 images sont regroupées et présentées en 30 listes de quatre présentations chacune. L'affichage d'une liste de présentations commence systématiquement par une phase de calibrage. Le déroulement de l'affichage d'une liste d'images est illustré figure 8.6.

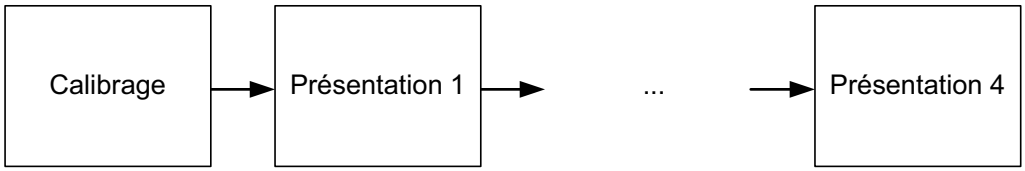


Figure 8.6 – Déroulement de l'affichage d'une liste de présentations en tâche d'évaluation de qualité.

8.2.4 Déroulement de l'ensemble des tests oculométriques

Les deux parties des tests oculométriques (exploration libre + tâche d'évaluation de qualité) sont décomposées en trois séances de manière à minimiser la fatigue visuelle et la lassitude des observateurs. Les séances durent entre 20 et 40 minutes et n'ont pas lieu la même journée. La répartition des différentes listes d'images ou de présentations dans les différentes séances est la suivante :

- première séance : les trois listes d'images (exploration libre), et sept listes de présentations (tâche d'évaluation de qualité) ;
- seconde séance : quinze listes de présentations (tâche d'évaluation de qualité) ;
- troisième séance : huit listes de présentations (tâche d'évaluation de qualité).

8.2.5 Construction d'une saillance spatiale

A l'issue des tests oculométriques, nous disposons d'un enregistrement oculométrique, pour chaque observateur et pour chaque image visualisée. Chaque enregistrement contient les positions successives, dans le référentiel de l'image, du point d'intersection entre la direction du regard de l'observateur et le plan dans lequel est affichée l'image. Ces positions successives sont enregistrées toutes les vingt millisecondes. A partir de toutes ces données, une carte de fixation (ou carte de saillance) est construite pour chaque observateur et pour chaque image visualisée. Ces cartes encodent le degré de saillance de chaque site (x, y) des images. Ces cartes sont souvent comparées à des cartes de relief [Wooding 02], constituées de pics et de vallées, où les pics représentent les régions d'intérêt de l'observateur.

La première étape de construction des cartes de saillance consiste à parcourir chaque enregistrement oculométrique afin d'identifier les périodes de fixations ainsi que les périodes de saccades. En effet, les données relatives aux saccades seront supprimées. L'algorithme suivant est utilisé :

```
-----------------------------------------------------------------------------------------
Algorithme d'identification des fixations
-----------------------------------------------------------------------------------------
Pour chaque échantillon :
1 : Calculer la vitesse point à point de chaque échantillon.

2 : Étiqueter chaque échantillon dont la vitesse point à point est inférieure
à un seuil (25 deg/s) comme fixation et les autres comme saccade.

3 : Regrouper les échantillons étiquetés en fixation consécutifs en groupe de fixations
et supprimer des échantillons étiquetés en saccade.

4 : Supprimer les groupes de fixations dont la durée est inférieure à 100 millisecondes.

5 : Associer à chaque groupe d'échantillon en fixation, une fixation dont les coordonnées
correspondent au barycentre des coordonnées de tous les échantillons du groupe.
-----------------------------------------------------------------------------------------
```

Cet algorithme est inspiré des travaux de Salvucci et Goldberg [Salvucci 00]. Les fixations sont identifiées grâce, d'une part à la vitesse angulaire du regard ($< 25°/s$), d'autre part à une durée minimale (100ms).

Une fois les mouvements de saccades supprimés, les cartes de fixation peuvent être calculées. Une carte de fixation $CS^{(k)}$ pour un observateur k peut être calculée de plusieurs façons : soit l'intérêt des zones de l'image (les pics de la carte de fixation) dépend du nombre de fixations, soit il dépend du nombre de fixations et de la durée de celles-ci.

Une carte de fixation ne dépendant que du nombre de fixations est calculée selon la relation :

$$CS^{(k)}(x, y) = \sum_{j=1}^{M} \Delta(x - x_j, y - y_j), \tag{8.1}$$

où M est le nombre total de fixations et Δ est le symbole de Kronecker.

Une carte de fixation dépendant du nombre de fixations et de leur durée est calculée selon la relation :

$$CS^{(k)}(x,y) = \sum_{j=1}^{M} \Delta(x - x_j, y - y_j) \cdot d(x_j, y_j), \tag{8.2}$$

où M est le nombre total de fixations, Δ est le symbole de Kronecker et d est la durée de la fixation.

Afin de déterminer le comportement d'un observateur moyen, les cartes de fixation de tous les observateurs sont combinées en une carte de fixation moyenne CS :

$$CS(x,y) = \frac{1}{N} \sum_{k=1}^{N} CS^{(k)}(x,y), \tag{8.3}$$

où N est le nombre d'observateurs. La carte de fixation moyenne représente les zones les plus attractives de l'image lorsqu'un nombre important d'observateurs est considéré.

Malgré son intérêt, la carte de fixation moyenne ne représente pas vraiment la réalité. Tout d'abord, l'oeil ne fixe pas un point sur une image, mais plutôt une zone ayant une taille visuelle proche de celle de la fovéa. De plus, les cartes de saillance sont obtenues à partir d'expérimentations faisant intervenir un appareillage à la précision limitée. En effet, si on se réfère de nouveau à ses caractéristiques (cf. tableau 8.1), la précision du dispositif oculométrique est comprise entre 0.25° et 0.5° de champ visuel. A partir de ces deux considérations, les densités de saillance DS sont obtenues par convolution des cartes de fixation moyenne CS avec un filtre gaussien bi-dimensionnel :

$$DS(x,y) = CS(x,y) * g_\sigma(x,y), \tag{8.4}$$

où g_σ est la gaussienne bi-dimensionnelle donnée par :

$$g_\sigma(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-\mu_x)^2 - (y-\mu_y)^2}{2\sigma^2}}, \tag{8.5}$$

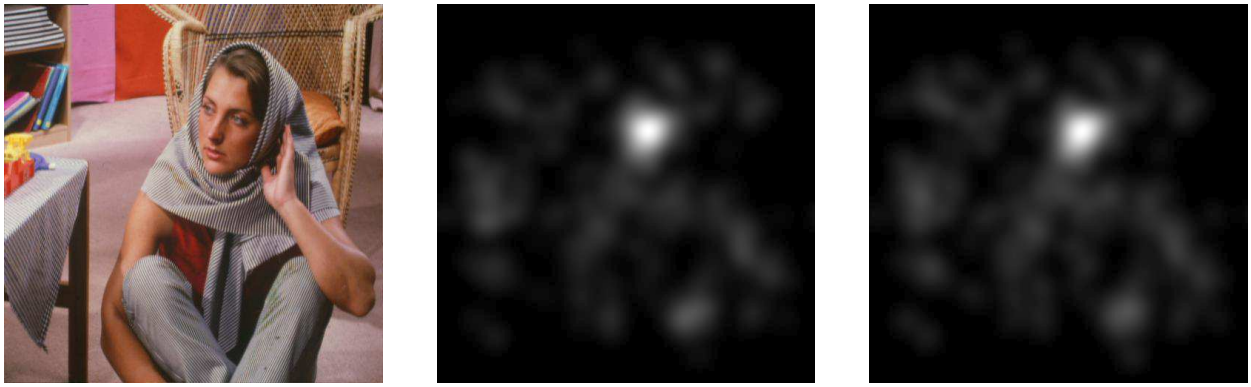
L'écart type σ vaut 0.5°. La figure 8.7 donne deux exemples de cartes de densités de saillance moyenne pour l'image *Barbara*. On observe sur cet exemple que la zone d'intérêt est clairement la tête de jeune femme.

Sauf précision de notre part, ce que nous appellerons par la suite « carte de saillance », fera référence en fait à une densité de saillance.

8.3 Impact de la tâche d'évaluation de la qualité sur l'attention visuelle

L'étude de l'attention visuelle peut être un moyen d'améliorer l'évaluation de la qualité d'image. Par exemple, un artefact qui apparaît sur une région d'intérêt est beaucoup plus gênant qu'une dégradation apparaissant sur une zone de moindre intérêt. Dans cette section nous nous posons la question suivante : quelles sont les différences en termes de stratégies visuelles entre la visualisation d'une image en exploration libre et la visualisation d'une image en tâche d'évaluation de qualité ?

Pour tenter de répondre à cette question, nous avons analysé les données oculométriques collectées lors des tests décrits en section 8.2. Nous rappelons que les images originales présentées en exploration libre correspondent



(a) (b) (c)

Figure 8.7 – Exemples de cartes de densités de saillance moyenne pour l'image *Barbara* (a) : (b) est une carte de densités de saillance moyenne calculée à partir du nombre et des durées de fixation, et (c) est une carte de densités de saillance moyenne calculée à partir du nombre de fixation. Plus la valeur tend vers le blanc, et plus la saillance est élevée.

aux images de références présentées en tâche d'évaluation de qualité. Afin d'éviter les ambiguïtés, nous utiliserons par la suite le terme « référence » pour désigner les images originales à la fois en exploration libre et à la fois en tâche d'évaluation de qualité.

8.3.1 Tâche et durée des fixations

La durée moyenne de fixation est calculée lorsque les cas suivants sont considérés :

- l'image de référence est visualisée par les observateurs en exploration libre.
- l'image de référence est visualisée par les observateurs en tâche d'évaluation de qualité. Cette image est présentée juste avant la version dégradée à évaluer.
- une version dégradée (à évaluer) de la même image est visualisée par les observateurs en tâche d'évaluation de qualité.

A partir des données oculométriques, la durée moyenne de fixation est calculée pour chaque observateur et pour chaque image. La durée moyenne de fixation par image présentée est obtenue en calculant la moyenne des durées moyennes de fixation de chaque observateur pour cette image.

La figure 8.8 donne la durée moyenne de fixation dans les trois cas mentionnés ci-dessus.

Cette analyse indique que les durées moyennes de fixation sont similaires lorsqu'on considère l'image de référence en exploration libre et l'image dégradée en tâche d'évaluation de qualité. Dans ce cas, le comportement oculomoteur n'est pas perturbé par la tâche. Il est important de souligner que ce résultat ne signifie pas que les observateurs regardent les mêmes zones. Cela signifie seulement que l'un des paramètres de la stratégie visuelle est inchangé.

Par contre, si l'on considère le cas de l'image de référence visualisée en tâche d'évaluation de qualité, la durée des fixations est beaucoup plus longue que dans les cas précédents. Dans ce cas, le comportement oculomoteur est clairement modifié. Une explication possible réside dans le fait que, dans ce cas, les observateurs s'efforcent

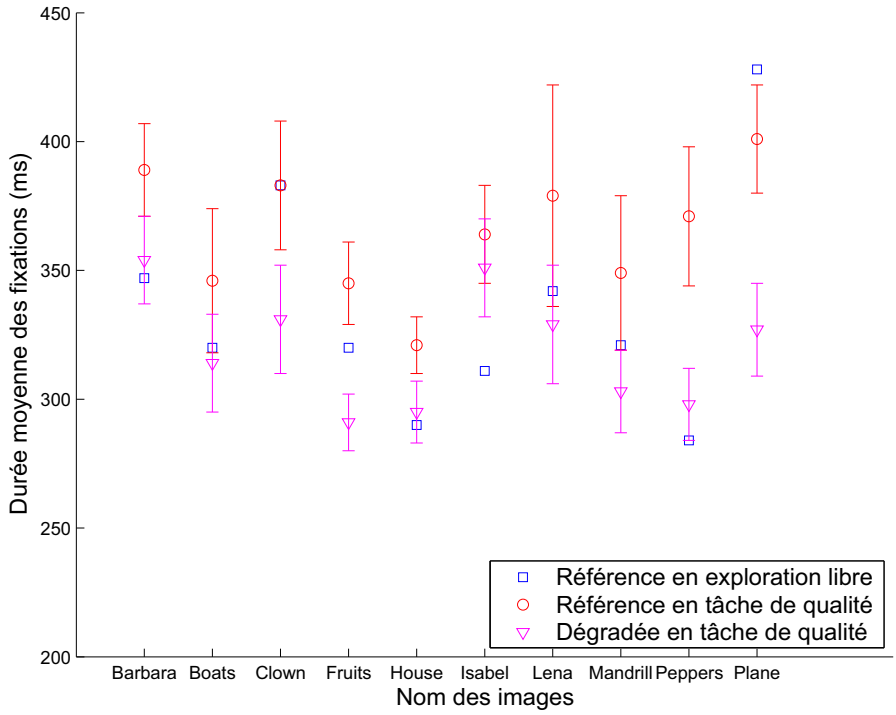


Figure 8.8 – Durée moyenne des fixations par image en exploration libre et en tâche d'évaluation de qualité. L'intervalle de confiance à 95% est donné pour chaque cas.

de bien mémoriser certaines parties de l'image en préparation à l'image dégradée qui suit et qu'ils vont devoir évaluer. La mémorisation spatiale semble ici importante pour réaliser la tâche demandée.

8.3.2 Tâche et cartes de saillance

Afin de tester la correspondance entre les différentes cartes saillances, deux indicateurs sont utilisés : la divergence de Kullback-Leibler et le coefficient de corrélation. Le premier indicateur évalue le degré de dissimilarité qui existe potentiellement entre deux fonctions de densité de probabilité. La divergence de Kullback-Leibler, notée KL, est donnée par la relation suivante :

$$KL(p|h) = \sum_x p(x) Log(\frac{p(x)}{h(x)}) \tag{8.6}$$

où p et h sont les fonctions de densité de probabilité. Lorsque les deux densités de probabilité sont strictement égales, la valeur du KL est zéro. Le second indicateur est le coefficient de corrélation linéaire. Il est noté CC et il mesure la force de la relation linéaire existant entre deux variables. Sa valeur évolue entre -1 et $+1$. Lorsque la valeur du CC est proche de ± 1, la relation linéaire entre les deux variables est presque parfaite. Le coefficient de corrélation CC est donnée par :

$$CC(p,h) = \frac{cov(p,h)}{\sigma_p \sigma_h} \tag{8.7}$$

où p et h représentent les cartes de saillance, $cov(p,h)$ est la valeur de covariance entre p et h. σ_p et σ_h représentent respectivement l'écart-type des cartes de saillance p et h.

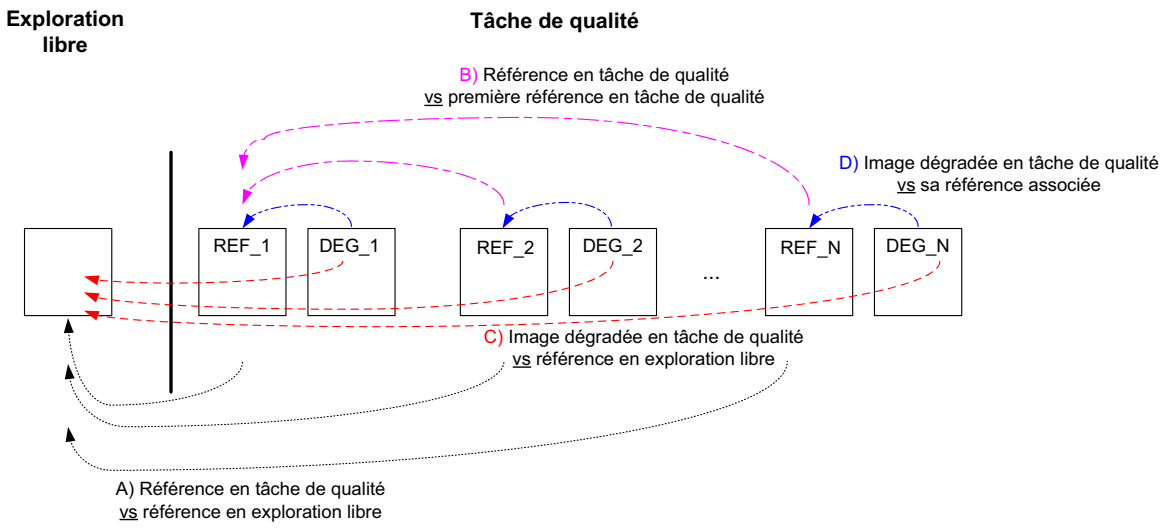


Figure 8.9 – Illustration des différentes comparaisons effectuées entre les cartes de saillance.

La figure 8.9 illustre les différentes comparaisons que nous avons effectuées :

- test A), *référence en tâche de qualité versus référence en exploration libre* : dans ce premier essai, nous mettons l'accent sur l'influence de la tâche sur le comportement oculomoteur. Est-ce que les observateurs regardent les mêmes régions ?
- test B), *référence en tâche de qualité versus première référence en tâche de qualité* : l'objectif ici est de montrer (ou pas) que les observateurs adaptent leur stratégie visuelle pour inspecter l'image de référence dans une tâche d'évaluation de qualité. Tentent-ils d'apprendre quelque chose afin d'affiner leur jugement de qualité ?
- test C), *image dégradée en tâche de qualité versus référence en exploration libre* : il est bien connu que la tâche agit sur le déploiement de l'attention visuelle. Mais nous ne savons pas dans quelle mesure une tâche de qualité modifie l'attention visuelle. Cette question est abordée ici en comparant les cartes de saillance mesurées en exploration libre et en tâche de qualité. En outre, les dégradations modifient-elles les cartes de saillance ?
- test D), *image dégradée en tâche de qualité versus sa référence associée en tâche de qualité* : lors de l'utilisation du protocole DSIS, la stratégie visuelle est-elle la même pour la référence et pour l'image dégradée ?

Les résultats des deux premiers tests sont illustrés sur les figures 8.10 et 8.11, alors que ceux des deux derniers sont illustrés figures 8.13 et 8.14.

8.3.2.1 Influence de la tâche sur l'exploration des images de référence

Comme prévu, le degré de dissimilarité entre les cartes de saillance issues des images de référence est important lorsque deux tâches différentes sont considérées (cf. figures 8.10 et 8.11).

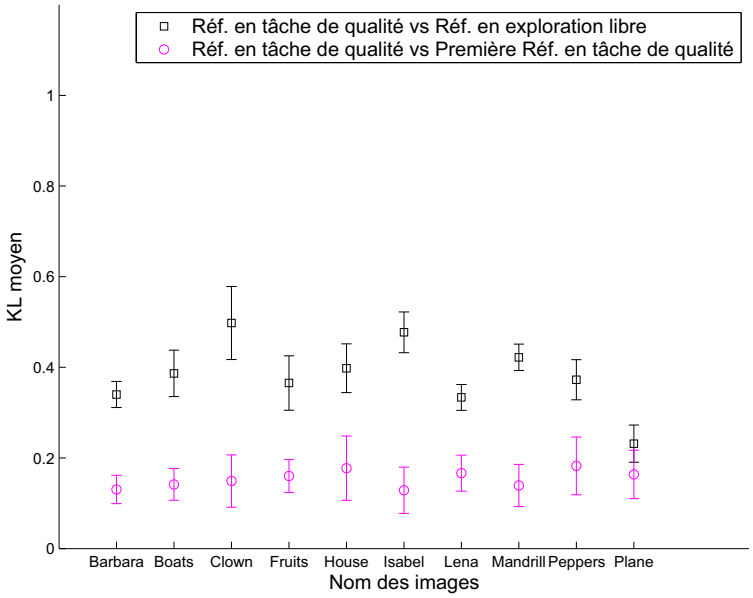


Figure 8.10 – Moyennes des divergences de Kullback-Leibler calculées pour chaque image d'origine. Comme le montre la figure 8.9, les valeurs de KL sont calculées, d'une part entre la carte de saillance de l'image de référence en tâche de qualité et la carte de saillance d'image de référence en exploration libre (test A), d'autre part entre la carte de saillance de l'image de référence en tâche de qualité et la première carte de saillance provenant de la première présentation de l'image de référence en tâche de qualité (test B). Pour chaque valeur l'intervalle de confiance à 95% est donné.

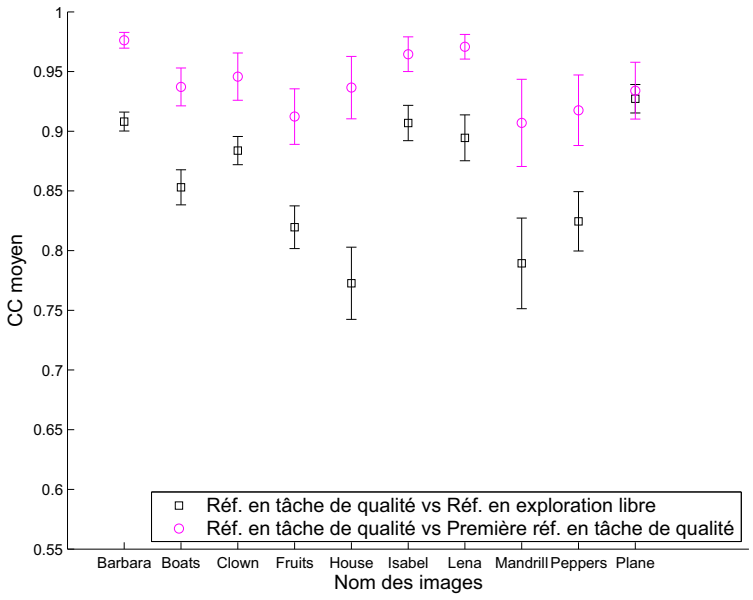
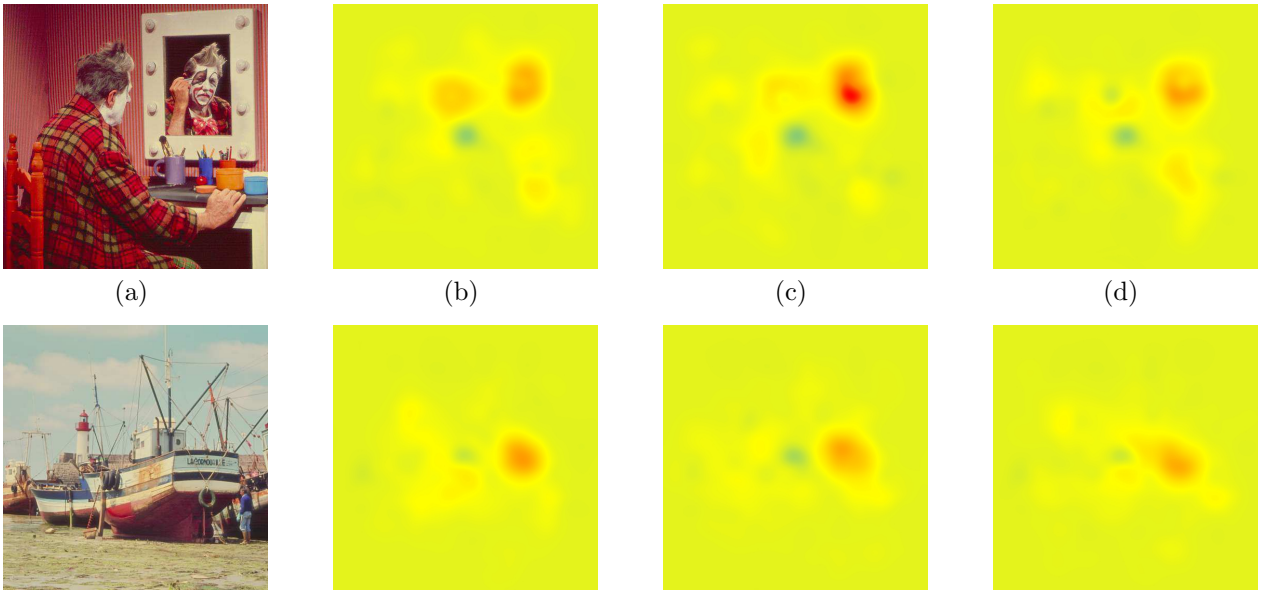


Figure 8.11 – Moyennes des Coefficients de Corrélation calculées pour chaque image d'origine. Comme le montre la figure 8.9, les valeurs de CC sont calculées, d'une part entre la carte de saillance de l'image de référence en tâche de qualité et la carte de saillance d'image de référence en exploration libre (test A), d'autre part entre la carte de saillance de l'image de référence en tâche de qualité et la première carte de saillance provenant de la première présentation de l'image de référence en tâche de qualité (test B). Pour chaque valeur l'intervalle de confiance à 95% est donné.

Les valeurs de KL sont comprises dans l'intervalle $[0.3, 0.5]$. La même tendance est observée pour les valeurs de CC qui sont dans la gamme de $[0.77, 0.92]$. La tâche a donc un impact sur l'exploration des images de référence.

Pour aller plus loin dans cette analyse, la différence entre les deux cartes de saillance issues des images de référence visualisées en exploration libre et en tâche de qualité est calculée. Plusieurs exemples sont donnés dans la figure 8.12. Les zones jaunes sont considérées avec la même intensité à la fois en exploration libre et en tâche de qualité. Les zones rouges correspondent aux zones qui sont davantage inspectées en tâche de qualité.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 8.12 – Première ligne : (a) image de référence *Clown*, (b) différence entre la carte de saillance en exploration libre et la carte de saillance de la première présentation de l'image de référence en tâche de qualité, (c) différence entre la carte de saillance en exploration libre et la carte de saillance de la troisième présentation de l'image de référence en tâche de qualité ; (d) différence entre la carte de saillance en exploration libre et la carte de saillance de la cinquième présentation de l'image de référence en tâche de qualité. Deuxième ligne : (e) image de référence *Boats*, (f) différence entre la carte de saillance en exploration libre et la carte de saillance de la première présentation de l'image de référence en tâche de qualité, (g) différence entre la carte de saillance en exploration libre et la carte de saillance de la troisième présentation de l'image de référence en tâche de qualité ; (h) différence entre la carte de saillance en exploration libre et la carte de saillance de la cinquième présentation de l'image de référence en tâche de qualité. Les zones rouges sont plus regardées en tâche de qualité qu'en exploration libre. Les zones en bleu sont moins regardées en tâche de qualité qu'en exploration libre.

Pour l'image *Clown*, les différences les plus importantes concernent le visage du clown dans le miroir, sa tête et sa main. Les observateurs prennent plus de temps pour inspecter ces zones en tâche d'évaluation de qualité qu'en exploration libre. Ceci est cohérent avec le précédent résultat de la figure 8.8 qui laissait penser que les observateurs essayaient de mémoriser certaines zones. Concernant l'image *Boats*, la différence des cartes de saillance indique principalement que la zone se situant autour du nom du bateau a été davantage regardée en tâche de qualité. Cette zone est donc sans doute utilisée comme une zone de comparaison pour l'évaluation

de la qualité des versions dégradées de cette image.

8.3.2.2 Adaptation de la stratégie visuelle en tâche de qualité

Le second résultat des figures 8.10 et 8.11 concerne l'adaptation de la stratégie visuelle en tâche de qualité. En tâche de qualité, les observateurs ont vu plusieurs fois la même image de référence, la mémoire à court terme et la capacité des observateurs à apprendre comment évaluer la qualité de l'image (par exemple, pour évaluer la qualité d'une image, il est préférable de parcourir les zones uniformes plutôt que les zones texturées) peuvent probablement modifier la stratégie visuelle. Bien qu'il était raisonnable de penser que les observateurs soient de plus en plus performants, les résultats indiquent que cette hypothèse est fausse en terme de stratégie visuelle. Dans ce cas, à la fois le degré de dissimilarité et l'intervalle de confiance sont faibles. Les valeurs de KL et de CC sont respectivement comprises dans les intervalles $[0.12, 0.18]$ et $[0.9, 0.97]$.

8.3.2.3 Exploration des images dégradées : influence de la tâche et des dégradations

Les figures 8.13 et 8.14 permettent d'aborder deux aspects :

- Quelles sont les différences de stratégie visuelle entre l'exploration libre et la tâche de qualité lorsque les images dégradées sont considérées ?
- Les dégradations ont-elles la capacité d'attirer ou de modifier le déploiement de l'attention visuelle ?

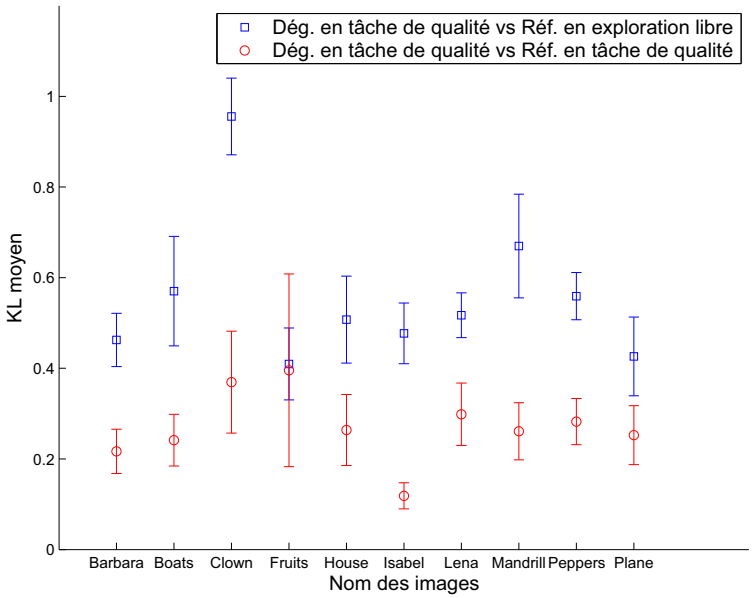


Figure 8.13 – Moyennes des divergences de Kullback-Leibler calculées pour chaque image d'origine, quelle que soit la dégradation. Comme le montre la figure 8.9, les valeurs de KL sont calculées, d'une part entre la carte de saillance de l'image dégradée en tâche de qualité et la carte de saillance de l'image de référence en exploration libre (test C), d'autre part entre la carte de saillance de l'image dégradée en tâche de qualité et la carte de saillance provenant de son image de référence associées en tâche de qualité (test D). Pour chaque valeur l'intervalle de confiance à 95% est donné.

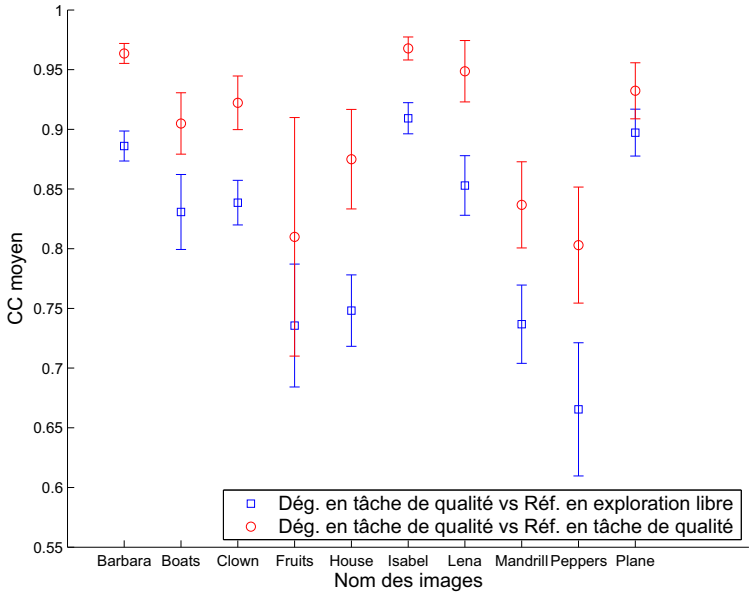


Figure 8.14 – Moyennes des Coefficients de Corrélation calculées pour chaque image d'origine, quelle que soit la dégradation. Comme le montre la figure 8.9, les valeurs de CC sont calculées, d'une part entre la carte de saillance de l'image dégradée en tâche de qualité et la carte de saillance de l'image de référence en exploration libre (test C), d'autre part, entre la carte de saillance de l'image dégradée en tâche de qualité et la carte de saillance provenant de son image de référence associées en tâche de qualité (test D). Pour chaque valeur l'intervalle de confiance à 95% est donné.

Concernant le premier point, les résultats (test C) indiquent qu'il existe une différence importante entre les stratégies visuelles qui sont déployées entre les images de référence en exploration libre et les versions dégradées en tâche de qualité. Ces résultats confirment l'influence de la tâche observée sur les stratégies visuelles déployées entre les images de référence en exploration libre et les images de référence en tâche de qualité (les résultats des figures 8.10 et 8.11 sont retrouvés : test A). Les valeurs de KL et de CC sont respectivement dans la gamme $[0.42, 0.95]$ et $[0.66, 0.9]$.

En outre, les intervalles de confiance sont bien plus importants que ceux des figures 8.10 et 8.11. Cela semble indiquer que le type de dégradation (flou, JPEG, JPEG2000) et le niveau de dégradation ont une influence sur le déploiement de l'attention visuelle. La figure 8.15 présente trois cartes de différences entre les cartes de saillance obtenues sur des images de référence en exploration libre et les cartes de saillance obtenues sur des images dégradées en tâche de qualité. Il est à noter qu'il n'y a effectivement pas de similitude frappante entre ces cartes. Par exemple, pour l'image *Boats* les différences sont fortement concentrées sur le nom de bateau pour les dégradations de type flou (f), alors qu'elles sont plus dispersées pour les dégradations de type JPEG2000 (g) et JPEG (h).

Ces résultats sont cohérents avec les résultats d'une étude récente réalisée par Vu et *al.* [Vu 08]. Dans cette étude, les auteurs se sont intéressés à l'influence des distorsions sur le déploiement spatial de l'attention visuelle. Leur étude repose aussi sur la réalisation de tests oculométriques. Ces tests ont été menés sur des images fixes en exploration libre ainsi qu'en évaluation de qualité. Cinq observateurs ont participé aux tests, et un

seul ne connaissait pas les motifs de l'expérimentation. Les tests en exploration libre ont été réalisés sur 29 images originales de la base LIVE [2]. Les tests en évaluation de qualité ont été réalisés sur un sous-ensemble d'images dégradées de la base LIVE. Les 100 images utilisées sont issues de 10 images originales, de 5 types de dégradations (flou, bruit blanc, JPEG, JPEG2000 et perte de paquets) et de deux niveaux de dégradations pour chaque type de dégradations : un niveau faible proche du seuil de visibilité et un niveau élevé où les dégradations étaient clairement visibles. Le déploiement de l'attention visuelle est étudié en comparant qualitativement les cartes obtenues pour chaque image présentée en moyennant les positions du regard enregistrées pour les 5 observateurs. Les expérimentations réalisées dans cette étude montrent principalement deux résultats :

- En évaluation de la qualité, lorsque les distorsions sont de type flou ou bruit blanc, les observateurs ont tendance à regarder les mêmes régions qu'en exploration libre. L'explication avancée par les auteurs est que dans ce cas la répartition des distorsions est uniforme sur les images, ce qui ne donne pas aux observateurs de nouvelles régions à fixer.
- Dans le cas de distorsions de type JPEG et JPEG2000 et aussi de type « perte de paquets » où les distorsions sont plus localisées, les observateurs ont tendance à regarder les distorsions lorsqu'elles sont clairement visibles, modifiant ainsi leur comportement par rapport à l'exploration libre. Les auteurs avancent donc l'hypothèse que les distorsions localisées spatialement peuvent attirer le regard des observateurs lorsque leur niveau est suffisamment élevé.

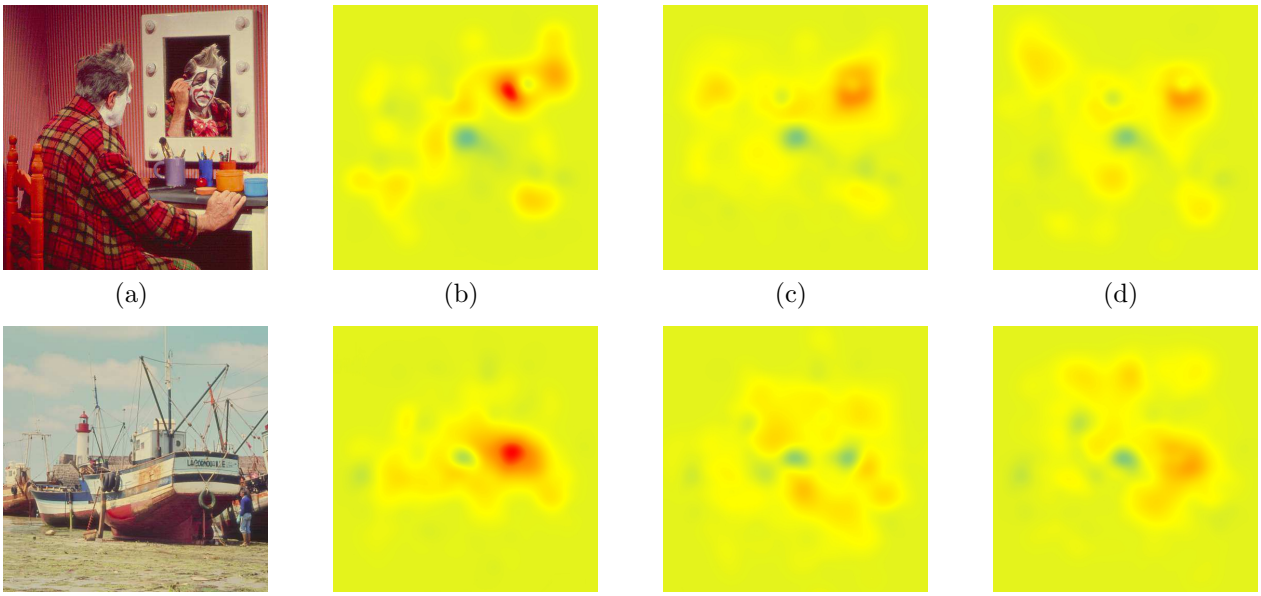
Ces résultats sont intéressants bien qu'il soit regrettable que la représentation du déploiement de l'attention visuelle à partir des mesures oculométriques ne soit pas très élaborée (pas de création de cartes de saillance) et que les comparaisons entre les différentes images présentées ne soient que qualitatives. Le nombre d'observateurs mériterait aussi d'être augmenté afin d'améliorer la fiabilité des résultats.

L'influence des dégradations sur le déploiement de l'attention visuelle entre exploration libre des images de référence et évaluation de la qualité des images dégradées (test C) doit être nuancée par les résultats de la comparaison des stratégies visuelles déployées en évaluation de qualité entre les images de référence et les images dégradées (test D). En effet, les valeurs de KL et de CC sont comprises respectivement dans les intervalles $[0.11, 0.28]$ et $[0.8, 0.96]$, ce qui conduit à la conclusion qu'il existe peu de différences entre les cartes de saillance obtenues en évaluation de qualité à partir des images de référence et des versions dégradées. Cette différence est en tout cas moins importante qu'entre les cartes de saillance obtenues à partir des images de référence en exploration libre et des versions dégradées en évaluation de qualité. Sur les images dégradées, il semble que la tâche ait une influence plus importante que les dégradations sur le déploiement de l'attention visuelle.

8.3.3 Discussion

Comme prévu, la tâche d'évaluation de qualité a un effet significatif sur les mouvements oculaires. Le premier résultat montre une augmentation de la durée moyenne de fixation sur l'image de référence en tâche d'évaluation de qualité, par rapport à tous les autres cas. Cela peut signifier que lors de l'évaluation de qualité avec le

2. `http://live.ece.utexas.edu/research/quality`



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 8.15 – Première ligne : (a) image de référence *Clown*, (b) différence entre la carte de saillance en exploration libre et la carte de saillance d'une version dégradée (JPEG2000) en tâche de qualité, (c) et (d) différences entre la carte de saillance en exploration libre et la carte de saillance de deux versions dégradée (JPEG) en tâche de qualité. Deuxième ligne : (e) image de référence *Boats*, (f) différence entre la carte de saillance en exploration libre et la carte de saillance d'une version dégradée (flou) en tâche de qualité, (g) différence entre la carte de saillance en exploration libre et la carte de saillance d'une version dégradée (JPEG2000) en tâche de qualité ; (h) différence entre la carte de saillance en exploration libre et la carte de saillance d'une version dégradée (JPEG) en tâche de qualité. Les zones rouges sont plus regardées en tâche de qualité qu'en exploration libre. Les zones en bleu sont moins regardées en tâche de qualité qu'en exploration libre.

protocole DSIS, les observateurs essayent de mémoriser certaines parties de l'image de référence en préparation de l'évaluation de l'image qui suit et qu'ils vont devoir évaluer.

Le deuxième résultat important concerne la variation de la stratégie visuelle tout au long d'une campagne de tests subjectifs de qualité utilisant le protocole DSIS. Nous montrons que les observateurs ne sont pas plus compétitifs à la fin des tests qu'au début. En d'autres termes, il n'y a pas d'adaptation visuelle ou d'apprentissage de la tâche d'évaluation de qualité (du point de vue de l'attention visuelle) lié au contenu. Ce résultat conforte l'utilisation du protocole DSIS dans lequel la présentation systématique de l'image de référence avant chaque version à évaluer aurait pu être la source d'un biais.

Enfin, la comparaison des intervalles de confiance des valeurs de KL et de CC montre qu'ils sont moins importants pour le test A (référence en tâche de qualité versus référence en exploration libre) que pour le test C (image dégradée en tâche de qualité versus image de référence en exploration libre). On observe que les dégradations ont une influence sur la stratégie visuelle comme l'avait noté aussi T. Vuori [Vuori 06] et, plus récemment, Vu et *al.* [Vu 08].

On peut donc déduire de ces résultats que dans un contexte d'évaluation de la qualité d'images avec référence, plusieurs stratégies visuelles différentes se distinguent. L'utilisation de l'attention visuelle dans des métriques de qualité avec référence se trouve confrontée à une première question. Si la stratégie visuelle n'est pas la même sur les images de référence et sur les images dégradées, quelle stratégie doit-on utiliser ? Doit-on utiliser l'attention visuelle déployée sur l'image de référence ? Ou bien celle déployée sur l'image dégradée ? Ou bien les deux ? Utiliser l'attention visuelle déployée sur l'image dégradée semble la solution la plus cohérente dans le sens où l'observateur construit son jugement à partir des dégradations perçues lors de son exploration de l'image dégradée. Par conséquent, l'utilisation de l'attention visuelle déployée uniquement sur l'image de référence ne semble pas cohérente car, même si l'observateur juge l'image dégradée en comparaison de son exploration de l'image de référence, c'est dans l'exploration de l'image dégradée qu'il trouve les dégradations à la base de son jugement. L'utilisation conjointe des deux stratégies est une seconde solution cohérente, mais certainement plus complexe, c'est pourquoi dans notre étude nous allons déjà explorer la solution consistant à utiliser l'attention visuelle déployée sur l'image dégradée.

Faisons maintenant l'hypothèse que nous ayons compris comment l'attention visuelle pouvait nous permettre d'améliorer les performances de métriques de qualité d'images. La prochaine interrogation serait de savoir comment prédire cette attention visuelle. En effet, si dans notre étude nous avons accès à la vérité terrain, cette information n'est évidemment pas disponible dans la pratique. Il existe dans la littérature des modèles d'attention visuelle dont le but est généralement d'évaluer l'attention visuelle déployée en exploration libre. Or, nous venons de voir que la tâche de qualité avait une influence sur le déploiement de l'attention visuelle. Il faut donc envisager que les modèles d'attention visuelle existant ne soient pas adaptés à cette application. Pour cela, il faudrait d'une part que ces modèles soient capables de prendre en compte l'influence de la tâche de qualité, d'autre part qu'ils soient capables de simuler les différences de stratégies déployées entre les images de référence et les images dégradées.

8.4 Impact de l'attention visuelle sur les performances de métriques de qualité

Comme nous l'avons déjà mentionné, l'attention visuelle est pressentie comme un moyen d'améliorer l'évaluation de la qualité d'images. Par exemple, un artefact qui apparaît sur une région d'intérêt serait plus gênant qu'une dégradation apparaissant sur une zone de moindre intérêt. Ainsi, une idée de base pour améliorer les métriques de qualité, par le biais de l'information de saillance, est de donner plus d'importance aux dégradations situées sur les régions de forte saillance au détriment des dégradations situées sur les régions de moindre intérêt. Beaucoup de mesures de qualité avec référence complète sont mises en oeuvre en deux étapes. Dans la première phase, les distorsions des images sont évaluées localement et permettent la création d'une carte de distorsion. Dans la deuxième étape, une fonction de cumul spatial des distorsions est utilisée pour combiner les valeurs de distorsion en une note globale de qualité.

Dans la littérature, certains auteurs tentent d'utiliser l'information liée à l'attention visuelle pour améliorer la prédiction de métriques de qualité [Osberger 98, Barland 06]. Néanmoins, l'interprétation de ces études est compliquée par le fait que deux problèmes étroitement liés ne sont pas étudiés séparément. Le premier problème est la détermination de la saillance avec des modèles d'attention visuelle, tandis que le second problème est l'utilisation de l'information de saillance dans des fonctions de cumul spatial des distorsions. L'objectif de cette section est d'étudier uniquement le second problème.

Cette étude est rendue possible grâce aux tests oculométriques réalisés pendant la campagne d'évaluation de qualité. Ces tests sont décrits dans la section 8.2. Nous disposons d'une part des MOS, d'autre part de l'information de saillance correspondant aux zones de l'image où les observateurs ont regardé pour construire leur jugement de qualité. Nous allons examiner différentes fonctions de cumul spatial basées sur l'information de saillance. Nous nous efforcerons de répondre à la question suivante : est-ce que le recours à l'information de saillance dans une fonction de cumul spatial des distorsions permet d'améliorer la prédiction d'une métrique de qualité d'images ?

8.4.1 Métriques de qualité basées saillance

Dans nos expérimentations, plusieurs métriques de qualité simples basées sur la saillance ont été testées. Ces métriques adoptent une implantation en deux étapes. Pour chaque métrique, une carte de distorsion est d'abord évaluée à partir de l'image de référence et de l'image dégradée. Ensuite, une note de qualité est calculée à partir de la carte des distorsions en utilisant une fonction de cumul spatial des distorsions exploitant la saillance visuelle. L'information de saillance est constituée par les cartes de saillance réelles issues des tests oculométriques détaillés section 8.2.

8.4.1.1 Cartes des distorsions spatiales

Trois méthodes sont utilisées pour calculer les cartes de distorsion. La première méthode est une simple différence absolue calculée entre l'image de référence et l'image dégradée. La seconde méthode est le critère

SSIM [Wang 04a] (cf. section 2.2.2) calculé entre l'image de référence et l'image dégradée. La troisième est le modèle basé sur une décomposition en ondelettes utilisé dans la métrique WQA (cf. chapitre 5).

8.4.1.2 Fonction de cumul spatial basée saillance

L'idée est d'utiliser l'information locale de saillance afin de pondérer la valeur locale de distorsion. La forme générale d'une telle fonction de pondération spatiale est donnée par :

$$Q = \frac{\sum_{x=1}^{W} \sum_{y=1}^{H} w_i(x,y) \cdot q(x,y)}{\sum_{x=1}^{W} \sum_{y=1}^{H} w_i(x,y)}, \tag{8.8}$$

où Q est la note de qualité objective, W et H sont respectivement la largeur et la hauteur de l'image, $w_i(x,y)$ est la pondération attribuée au site (x,y), i définissant la façon de concevoir la pondération, $q(x,y)$ est la valeur de la distorsion au site (x,y).

Quatre fonctions différentes w_i, utilisant l'information locale de saillance, sont comparées. Ces fonctions sont données par :

$$\left| \begin{array}{l} w_1(x,y) = SM_n(x,y) \\ w_2(x,y) = 1 + SM_n(x,y) \\ w_3(x,y) = SM(x,y) \\ w_4(x,y) = 1 + SM(x,y) \end{array} \right. \tag{8.9}$$

où $SM(x,y) \in [0; S_{max}]$ est la carte de saillance non normalisée, et $SM_n(x,y) \in [0,1]$ est la carte de saillance normalisée. Les cartes de saillance sont calculées de deux façons différentes, comme cela a été expliqué dans la section 8.2.5 :

- en fonction du nombre de fixations (FN),
- en fonction du nombre de fixations et de la durée des fixations (FD).

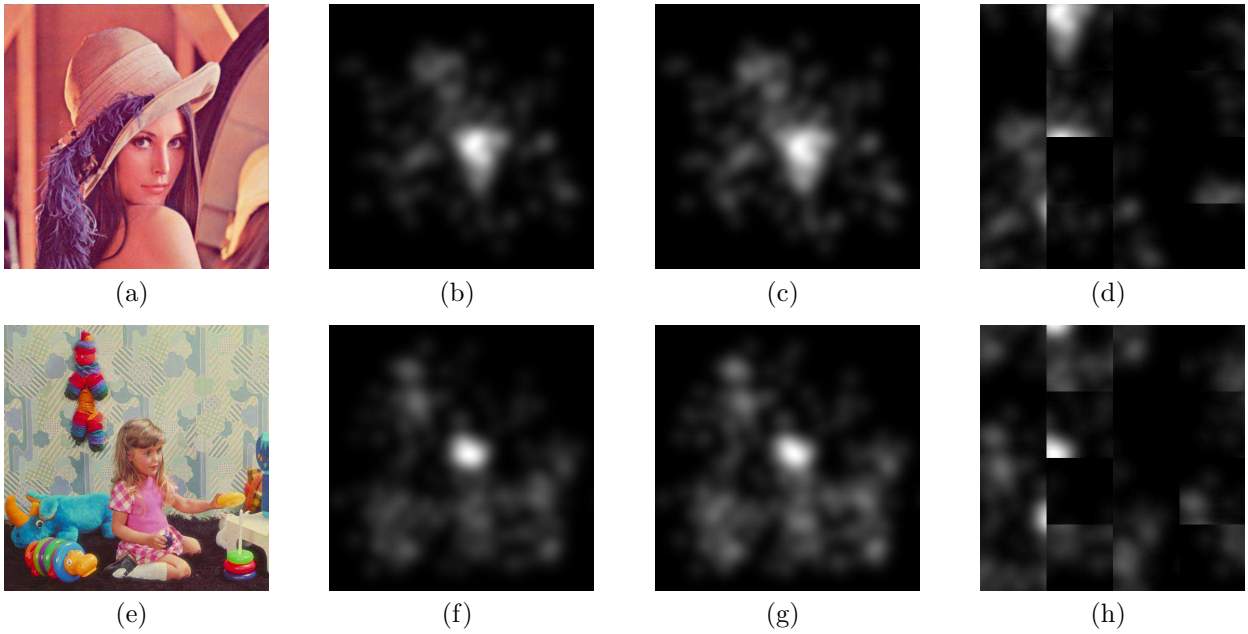
8.4.2 Analyse quantitative

Les mesures objectives de qualité d'images testées regroupent deux types de carte de distorsion, quatre fonctions de cumul spatial, deux types de carte de saillance. Une version *permutée* des cartes de saillance est également testée (cf. figure 8.16d). Dans cette version *permutée*, la carte de saillance est divisée en 16 blocs et chaque bloc est remplacé par un autre. Par conséquent, l'information locale de saillance est perdue tout en conservant la dynamique des cartes, la proportion de zones couvertes, ainsi qu'une certaine cohérente spatiale par blocs. Cette version *permutée* est utilisée afin d'évaluer l'influence de la localisation spatiale de la saillance des cartes utilisées, en modifiant le moins possible les autres caractéristiques de ces cartes.

Avant d'évaluer les mesures objectives de qualité d'images, une fonction psychométrique est utilisée pour transformer le niveau de distorsion Q en MOS prédit (MOSp), tel que recommandé par le groupe de travail VQEG [VQEG 00] :

$$\text{MOSp} = \frac{b_1}{1 + e^{-b_2 \cdot (Q - b_3)}}, \tag{8.10}$$

où b_1, b_2 et b_3 sont les trois paramètres de la fonction psychométrique.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 8.16 – (a) images de référence, (b) cartes de saillance moyennes basées sur la durée des fixations (FD), (c) cartes de saillance moyennes basées sur le nombre de fixations (FN), et (d) versions *permutées* de (b).

Pour évaluer l'impact de l'information de saillance, les différentes métriques basées saillance sont comparées aussi aux approches classiques ($w_i = 1$ dans l'équation (8.8)).

8.4.2.1 Évaluation à partir de la stratégie visuelle et du jugement de qualité d'un observateur moyen

L'évaluation objective de la qualité visuelle consistant à prédire le jugement de qualité d'un observateur standard ou moyen, nous nous intéressons ici à son comportement. L'information de saillance utilisée correspond donc aux cartes de saillance moyennées sur l'ensemble des observateurs. Les métriques de qualité sont évaluées en comparant les MOS et les MOSp sur l'ensemble de la base d'images et en utilisant deux indicateurs : le coefficient de corrélation linéaire (CC) et l'erreur quadratique moyenne (RMSE). Les résultats sont présentés dans les tableaux 8.2, 8.3 et 8.4.

Dans le cas des cartes de distorsion fondées sur la différence absolue (cf. tableau 8.2), une amélioration de la prédiction est observée avec la fonction de cumul w_3, quel que soit le type de saillance utilisé (FN ou FD). Les CC sont respectivement de 0.83 et 0.82 en utilisant les cartes de saillance FD et FN, contre 0.74 sans l'utilisation de l'information de saillance. Une amélioration de la prédiction est également observée avec la w_4 fonction du poids, mais seulement avec les cartes de saillance de type FN. Le CC est de 0.825 avec l'information de saillance, contre 0.74 sans l'information de saillance. Aucune autre amélioration de prédiction n'est observée, mais des détériorations de la prédiction sont observées. Par exemple, pour la fonction de cumul w_1, les CC sont respectivement de 0.51 et 0.5 pour les cartes de saillance FD et FN, contre 0.74 sans utilisation de la saillance. Les mêmes observations sont faites avec RMSE.

Cumul		FD (saillance)		FN (saillance)	
Saillance	w_i	CC	RMSE	CC	RMSE
Aucune	***1***	***0.742***	***0.814***	***0.742***	***0.814***
Réelle	w_1	0.510	1.044	0.504	1.049
	w_2	0.733	0.826	0.731	0.829
	w_3	0.830	0.678	0.821	0.692
	w_4	0.825	0.686	0.754	0.797
Permutée	w_1	0.387	1.142	0.388	1.140
	w_2	0.725	0.836	0.722	0.840
	w_3	0.764	0.815	0.750	0.835
	w_4	0.778	0.787	0.744	0.813

Table 8.2 – Comparaison des performances des métriques de qualité, lorsque que la différence absolue est utilisée pour calculer la carte de distorsion.

Avec les cartes de saillance « permutée », aucune véritable amélioration de la prédiction n'est observée en termes de CC et de RMSE, et même parfois des détériorations de la prédiction sont observées. Le même constat est fait avec les fonctions de cumul w_3 et w_4, ce qui signifie que les améliorations de la prédiction constatées avec la saillance réelle et ces deux fonctions de cumul ne sont pas dues au hasard. On peut noter que les performances obtenues sans l'information de saillance étant faibles, il est sans doute plus facile de les améliorer que dans les autres cas étudiés. Ceci peut expliquer en partie la nette amélioration des performances obtenue dans ce cas.

Dans le cas des cartes de distorsions issues de SSIM (cf. tableau 8.3), aucune véritable amélioration de la prédiction n'est observée, ni en terme de RMSE, ni en terme de CC, et cela quels que soient la fonction de cumul et le type de saillance utilisés. Les observations sont les mêmes en utilisant les cartes de saillance *permutée*. Ces observations semblent montrer que la façon de prendre en compte l'information saillance n'est pas efficace.

Cumul		FD (saillance)		FN (saillance)	
Saillance	w_i	CC	RMSE	CC	RMSE
Aucune	***1***	***0.827***	***0.686***	***0.827***	***0.686***
Réelle	w_1	0.820	0.696	0.821	0.695
	w_2	0.827	0.686	0.827	0.686
	w_3	0.820	0.696	0.821	0.695
	w_4	0.825	0.688	0.828	0.685
Permutée	w_1	0.811	0.713	0.818	0.701
	w_2	0.827	0.684	0.828	0.684
	w_3	0.811	0.713	0.818	0.701
	w_4	0.826	0.685	0.828	0.684

Table 8.3 – Comparaison des performances des métriques de qualité, lorsque que la SSIM est utilisée pour calculer la carte de distorsion.

Dans le cas des cartes de distorsions calculées selon le même modèle que dans la métrique WQA (cf. tableau 8.4), aucune véritable amélioration de la prédiction n'est observée, ni en terme de RMSE, ni en terme de CC, et cela quels que soient la fonction de cumul et le type de saillance utilisés. Comme dans le cas des cartes de distorsions issues de SSIM, l'utilisation les cartes de saillance *permutée* conduit aux mêmes observations.

En conséquence, l'impact positif de l'information de saillance sur la prédiction n'est pas aussi clairement établi

Cumul		FD (saillance)		FN (saillance)	
Saillance	w_i	CC	RMSE	CC	RMSE
Aucune	***1***	***0.885***	***0.580***	***0.885***	***0.580***
Réelle	w_1	0.842	0.671	0.847	0.660
	w_2	0.883	0.583	0.884	0.582
	w_3	0.842	0.671	0.847	0.660
	w_4	0.866	0.621	0.884	0.583
Permutée	w_1	0.860	0.634	0.861	0.633
	w_2	0.885	0.580	0.884	0.581
	w_3	0.860	0.634	0.861	0.633
	w_4	0.878	0.595	0.885	0.580

Table 8.4 – Comparaison des performances des métriques de qualité, lorsque que le modèle basé ondelettes de la métrique WQA est utilisé pour calculer la carte de distorsion.

que prévu. La prédiction n'est pas améliorée, même s'il existe des exceptions lorsque l'on utilise la fonction de cumul w_3 et les cartes de distorsions reposant sur la différence absolue. Les quatre fonctions de cumul utilisées favorisent les distorsions présentes sur les zones de saillance au détriment des autres. Les fonctions de cumul w_1 et w_3 sont plus pénalisantes pour les distorsions présentent dans les zones n'attirant pas l'attention, que les fonctions de cumul w_2 et w_4. Les fonctions de cumul et la fabrication de la saillance moyenne peuvent être suspectées pour expliquer la non-amélioration de la prédiction. Cette dernière est examinée dans la section suivante.

8.4.2.2 Évaluation à partir de la stratégie visuelle et du jugement de qualité d'observateurs particuliers

Afin d'écarter les causes éventuelles de la non-amélioration de la prédiction dues à l'étude du comportement de l'observateur moyen, les comportements de huit observateurs particuliers sont étudiés. Leur carte de saillance et leur note de qualité individuelle sont utilisées pour évaluer les mesures objectives de qualité. Pour chaque observateur étudié, les métriques de qualité sont évaluées en comparant les notes individuelles et les MOSp (où plutôt devrait-on-dire ici les notes prédites) sur l'ensemble de la base d'images et en utilisant deux indicateurs : le coefficient de corrélation linéaire (CC) et l'erreur quadratique moyenne (RMSE). Les résultats sont présentés pour un observateur dans le tableau 8.5 et pour les huit observateurs considérés en annexe C.

Dans le cas des cartes de distorsion basées sur la différence absolue, aucune véritable amélioration de la prédiction n'est observée quels que soient la fonction de cumul, le type de saillance utilisé et l'observateur considéré. Les moyennes des ΔCC sont respectivement $-0,08$ et $-0,07$ pour la saillance de type FN et FD. Les résultats sont très variables d'un observateur à l'autre.

Dans le cas des cartes de distorsion issues de SSIM, les observations sont les mêmes. Il n'y a pas d'amélioration de la prédiction, quels que soient la fonction de cumul, le type de saillance utilisé et l'observateur considéré. Les moyennes des ΔCC sont respectivement $-0,01$ et $-0,02$ pour la saillance de type FN et FD. Les résultats sont aussi très variables d'un observateur à l'autre.

On n'observe pas d'amélioration de la prédiction, même si, pour un observateur, on utilise la saillance réelle

Carte de distorsions	Cumul		FD (saillance)		FN (saillance)	
	Saillance	w_i	CC	RMSE	CC	RMSE
DiffAbs	***Aucune***	***1***	***0.646***	***0.924***	***0.646***	***0.924***
	Réelle	w_1	0.560	1.002	0.571	0.994
		w_2	0.649	0.920	0.649	0.921
		w_3	0.599	0.969	0.644	0.926
		w_4	0.609	0.960	0.660	0.909
	Permutée	w_1	0.469	1.069	0.456	1.077
		w_2	0.645	0.925	0.643	0.926
		w_3	0.530	1.026	0.550	1.010
		w_4	0.546	1.014	0.645	0.925
SSIM	***Aucune***	***1***	***0.741***	***0.814***	***0.741***	***0.814***
	Réelle	w_1	0.711	0.851	0.712	0.850
		w_2	0.741	0.813	0.741	0.813
		w_3	0.711	0.851	0.712	0.850
		w_4	0.732	0.824	0.741	0.813
	Permutée	w_1	0.702	0.863	0.700	0.865
		w_2	0.741	0.813	0.741	0.813
		w_3	0.702	0.863	0.700	0.865
		w_4	0.733	0.824	0.741	0.814

Table 8.5 – Observateur n°1 : Comparaison des performances des métriques de qualité, lorsque que la différence absolue et la SSIM sont utilisées pour calculer les cartes de distorsion.

correspondant aux zones de l'image qu'il a fixé pour construire son jugement de qualité. En conséquence, la construction de la saillance moyenne n'explique pas la non-amélioration de la prédiction lorsque l'observateur moyen est considéré.

8.4.3 Discussion

Quatre fonctions de cumul basées saillance ont été testées dans le but d'améliorer l'évaluation objective de la qualité d'image. L'attention visuelle réelle, enregistrée au travers des mouvements oculaires des observateurs lors d'une campagne d'évaluation de la qualité, est utilisée. Les résultats globaux montrent que l'amélioration de la prédiction n'est pas clairement établie. Certes l'amélioration de la prédiction sur certains cas particuliers montrent que l'attention visuelle peut être intéressante, mais la non-amélioration générale laisse penser que la façon de prendre en compte l'attention visuelle ne peut se limiter à une simple pondération spatiale.

Les résultats des travaux récents de Larson *al.* [Larson 08] mènent à des conclusions similaires. Dans [Larson 08], les auteurs pondèrent des métriques existantes (PSNR, SSIM [Wang 04a], VIF [Sheikh 06a], VSNR [Chandler 07] et WSNR) par de l'information de saillance issue de tests expérimentaux. Les tests oculométriques sont les mêmes que dans [Vu 08] et ont été décrits section 8.3.2.3. Les données oculométriques sont utilisées pour segmenter les images en trois types de régions, cela en fonction du nombre de fixations par pixel :

- les régions de non-intérêt (*non-ROI*, jamais fixées),
- les régions d'intérêt secondaire (*2nd-ROI*, nombre de fixations par pixel inférieur à la moyenne),
- les régions d'intérêt primaire (*1st-ROI*, nombre de fixations par pixel supérieur à la moyenne).

La pondération d'une métrique est réalisée en calculant une note pour chaque type de régions puis en combinant linéairement les 3 notes obtenues :

$$E_{tot} = \alpha_{1st-ROI} \cdot E_{1st-ROI} + \alpha_{2nd-ROI} \cdot E_{2nd-ROI} + \alpha_{non-ROI} \cdot E_{non-ROI} \,, \tag{8.11}$$

où E_{tot} est la note finale, $E_{1st-ROI}$, $E_{2nd-ROI}$ et $E_{non-ROI}$ sont les notes pour les trois types de régions, et $\alpha_{1st-ROI}$, $\alpha_{2nd-ROI}$ et $\alpha_{non-ROI}$ sont les poids accordés aux notes de chaque type de régions. Les résultats montrent une tendance à l'amélioration des performances des métriques lorsque la pondération est utilisée, cependant les résultats sont variables d'une métrique à l'autre et aucune amélioration n'est trouvée comme vraiment significative.

A la vue de ces résultats, une pondération des cartes de distorsions par la saillance, qui se limite à donner plus d'importance aux distorsions appartenant aux zones de forte saillance, ne semble pas être la bonne approche. La construction du jugement de qualité à partir des distorsions spatiales semble plus complexe. Au cours de son processus d'évaluation, un observateur va explorer l'image à évaluer, et au cours de cette exploration il va rencontrer des distorsions plus ou moins importantes. Toutes les distorsions qu'il aura rencontrées au cours son exploration de l'image vont plus ou moins contribuer à la construction de son jugement de qualité. Son exploration de l'image ne va pas être uniforme et l'observateur passera plus ou moins de temps à explorer les différentes zones de l'image. Il est possible qu'il passe beaucoup de temps à chercher des distorsions dans des zones où celles-ci ne sont pas évidentes. De même, il est probable qu'il ne passe pas beaucoup de temps sur les zones où les dégradations sont évidentes préférant s'intéresser à des zones pas encore explorées.

Un observateur peut donc passer moins de temps sur une dégradation évidente que sur une dégradation plus discrète. Dans le premier cas, la saillance est faible, mais la contribution à la note de qualité est élevée. La saillance est faible car l'observateur n'est resté que peu de temps sur cette distorsion, et la contribution à la note de qualité est élevée car la distorsion est importante par conséquent la gêne occasionnée doit l'être aussi. Dans le second cas, la saillance est élevée et la contribution au jugement de qualité est plus faible. La saillance est élevée car l'observateur est resté plus longtemps sur cette distorsion, et la contribution à la note de qualité est faible car la distorsion est faible par conséquent la gêne occasionnée doit l'être aussi. Il semble que l'information de saillance et l'intensité des dégradations doivent être considérées conjointement dans la fonction de cumul spatial. Des fonctions de pondération plus complexe doivent donc être élaborées.

Ces résultats ne sont pas cohérents avec tous les travaux antérieurs [Osberger 98, Barland 06]. L'amélioration de la prédiction observée dans ces travaux pourrait être expliquée par une amélioration de la cohérence spatiale des erreurs plutôt que par l'information de saillance elle-même. Comme nous l'avons évoqué dans le chapitre 7, l'information de saillance utilisée dans ces travaux pose problème. En effet, la saillance est déterminée à partir de modèle d'attention visuelle dont la validité n'a pas été démontrée. Comme nous l'avons montré dans la première partie de ce chapitre, les stratégies visuelles déployées sur une image en exploration libre et sur la même image en tâche d'évaluation de qualité ne sont pas les mêmes. Or, les modèles d'attention visuelle ne prennent pas en compte la tâche de qualité. Ils déterminent plutôt la stratégie visuelle déployée dans le cas d'une exploration

libre.

8.5 Conclusion

Ce chapitre était consacré à l'étude de l'attention visuelle en évaluation de qualité d'images. Cette étude a reposé sur des tests oculométriques réalisés d'une part dans une situation d'exploration libre, d'autre part, durant une campagne d'évaluation subjective de la qualité d'images.

Le premier aspect étudié concernait l'évaluation subjective. A partir des données oculométriques, nous avons montré que la tâche d'évaluation de qualité avait un impact sur la stratégie visuelle. Nous avons montré aussi que du point de vue de l'attention visuelle il n'y avait pas d'adaptation visuelle ou d'apprentissage de la tâche d'évaluation de qualité lié au contenu entre le début et la fin de la campagne de tests subjectifs d'évaluation de qualité que nous avons menée. Ce résultat permet de conforter d'une manière générale l'utilisation du protocole DSIS dans une campagne de tests subjectifs d'évaluation de qualité. Nous avons également observé que les dégradations avaient une influence sur la stratégie visuelle, même si cette influence semble moins importante que celle liée à la tâche.

Le second aspect étudié concernait l'évaluation objective, et voir comment l'attention visuelle participait à la construction du jugement de qualité. En utilisant l'information de saillance réelle, nous avons montré que l'utilisation de fonctions de pondération donnant simplement plus d'importance aux zones de forte saillance, ne permettait pas d'améliorer de façon générale la prédiction de métriques de qualité. Ces résultats ne confirment pas des études de l'art antérieur [Osberger 98, Barland 06] qui montraient une amélioration des performances. Cependant et contrairement à nos travaux, les cartes d'importance utilisées dans ces études n'étaient pas issues de tests expérimentaux, mais étaient issues de modèles. Ces études mixaient, sans donc les séparer, deux problématiques : la pertinence des modèles d'attention visuelle et l'utilisation de l'attention visuelle pour améliorer les performances de métriques.

Davantage de recherches doivent être menées pour mieux comprendre les mécanismes de l'attention visuelle dans une tâche d'évaluation de la qualité d'images. Il semble que l'information de saillance et l'intensité des dégradations doivent être considérées de façon conjointe.

Chapitre 9

Attention visuelle et construction du jugement de qualité de vidéos

9.1 Introduction

L'objet de ce chapitre est l'étude de l'attention visuelle en évaluation de qualité de vidéos. Deux aspects distincts sont abordés dans ce chapitre. Le premier aspect, concernant plutôt l'évaluation subjective, s'intéresse à la stratégie visuelle déployée par les observateurs durant une campagne de tests subjectifs d'évaluation de qualité. Il s'agit d'étudier l'impact d'une tâche d'évaluation de qualité sur l'attention visuelle par rapport à une situation d'exploration libre. Le second aspect, concernant plutôt l'évaluation objective, s'intéresse à l'utilisation de l'information de saillance pour améliorer des métriques objectives de qualité. Dans le chapitre précédent nous avons réalisé une étude similaire mais concernant la présentation d'images fixes. Dans le cas des images, le contenu présenté n'évolue pas au cours du temps, alors que dans le cas des vidéos celui-ci évolue temporellement. Dans le cas de la vidéo, la dimension temporelle du déploiement de l'attention visuelle, c'est-à-dire l'ordre dans lequel les différentes zones sont explorées, est bien plus importante que dans le cas des images fixes. Pour les images fixes, l'attention visuelle était étudiée à l'aide de cartes de saillance, lesquelles étaient construites en cumulant temporellement toutes les fixations. Pour les vidéos cette façon de faire n'est plus valable, et il devient nécessaire de construire non plus des cartes de saillance, mais des séquences de saillance afin de prendre en compte la dimension temporelle du déploiement de l'attention visuelle.

L'étude de l'attention visuelle dans le cas des vidéos soulève les mêmes questions que dans les cas des images fixes. Les stratégies visuelles sont-elles les mêmes dans les deux situations ? Les stratégies sont-elles les mêmes sur les vidéos de référence et sur les vidéos dégradées ? Pour tenter de répondre à ces questions, des tests oculométriques ont été menés dans les deux situations : en exploration libre et durant une campagne de tests subjectifs d'évaluation de qualité de vidéos.

La première partie de ce chapitre est consacrée à la description des tests expérimentaux. La seconde partie est dédiée à l'étude de l'impact de la tâche d'évaluation de qualité sur l'attention visuelle. Finalement, la troisième partie de ce chapitre est consacrée à l'utilisation de la saillance en évaluation objective de la qualité.

9.2 Expérimentations oculométriques et tests subjectifs de qualité

Comme introduit précédemment, les tests oculométriques ont été réalisés dans deux situations différentes, correspondant chacune à une tâche particulière :

- une tâche d'exploration libre (*Free viewing task* ou *Free-task*), où les observateurs ont pour seule indication de regarder les vidéos le plus naturellement possible.
- une tâche d'évaluation de qualité (*Quality-task*), où les observateurs ont pour indication d'évaluer la qualité visuelle des vidéos qui leur sont présentées.

Les vidéos de la base de test décrite section 6.3.1 ont été utilisées pour les tests oculométriques en exploration libre et aussi pour les tests oculométriques en tâche d'évaluation de qualité.

9.2.1 Exploration libre

Durant cette partie des tests, vingt vidéos sont présentées aux observateurs. Dix vidéos sont les vidéos originales de la base de test et les dix autres sont des versions fortement dégradées des dix vidéos originales. Chaque vidéo présentée dure huit secondes, pendant lesquelles l'observateur est libre d'explorer la vidéo à sa convenance. Chaque vidéo est précédée d'une séquence de transition. Celle-ci consiste à afficher une image de gris uniforme pendant quatre secondes. Pendant les deux premières secondes un point noir clignotant apparaît successivement à deux positions différentes et pseudo-aléatoires, puis pendant les deux dernières secondes seule l'image de gris uniforme est affichée. Cette séquence illustrée figure 9.1 permet d'« initialiser » l'attention visuelle des observateurs avant chaque nouvelle vidéo. Dans le protocole de test, il est demandé aux observateurs de fixer le point noir clignotant à chaque fois qu'il apparaît. En fixant le point noir l'observateur n'est plus focalisé sur la dernière zone fixée de la vidéo précédente. Les deux dernières secondes ne comportant qu'un écran gris permettent à l'observateur de reposer un peu son système visuel avant la présentation suivante. Par ailleurs, cette séquence permet de vérifier la synchronisation entre l'affichage des vidéos et l'enregistrement des mouvements oculaires par l'oculomètre.

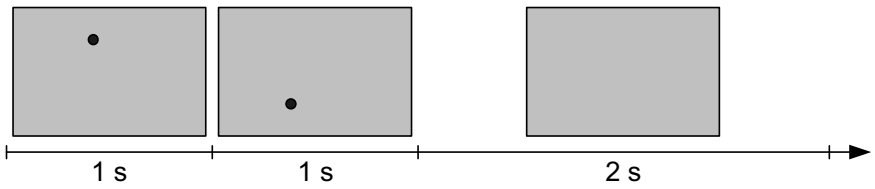


Figure 9.1 – Illustration de la séquence de transition affichée avant chaque vidéo de tests : affichage d'un écran gris uniforme pendant 4 secondes où un point noir clignotant apparaît successivement à deux endroits différents durant les 2 premières.

Les vingt vidéos sont regroupées et présentées en deux listes de dix vidéos. L'affichage d'une liste de vidéos commence systématiquement par une phase de calibrage du dispositif oculométrique. Le calibrage du dispositif est une étape très importante qui conditionne le bon déroulement des tests oculométriques. Nous ne reviendrons

pas sur cette étape qui a déjà été détaillée section 8.2.1 dans le cadre des tests oculométriques sur images. Le déroulement de l'affichage d'une liste de vidéos est illustré figure 9.2.

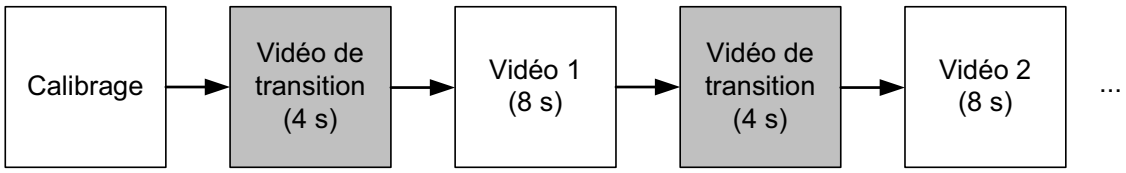


Figure 9.2 – Déroulement de l'affichage d'une liste de vidéos en expérimentation libre.

9.2.2 Tâche d'évaluation de qualité

Durant cette partie des tests, une campagne de tests subjectifs d'évaluation de qualité est menée de bout en bout. Les résultats de cette campagne sont d'ailleurs utilisés dans la section 6.3. Comme dans les expérimentations sur les images du chapitre précédent, le protocole DSIS (*Double Stimulus Impairment Scale*) est utilisé pour évaluer la qualité des vidéos. Le déroulement d'une présentation est illustré figure 9.3. Chaque vidéo est présentée pendant huit secondes et une séquence de transition est présentée pendant quatre secondes avant chacune d'entre elles. La séquence de transition est du même type que celle décrite dans la section 9.2.1. Une fois le couple de vidéos présenté, l'écran de notation est affiché comme dans les tests oculométriques sur les images (cf. figure 8.5).

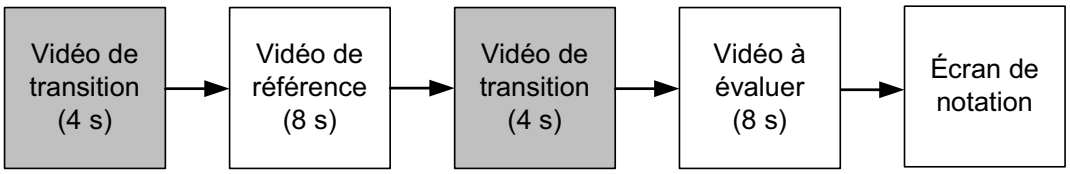


Figure 9.3 – Déroulement d'une présentation en tâche d'évaluation de qualité (protocole DSIS).

Les cinquante vidéos sont regroupées et présentées en dix listes de cinq présentations chacune. L'affichage d'une liste de présentations commence systématiquement par une phase de calibrage. Le déroulement de l'affichage d'une liste de vidéos est illustré figure 9.4.

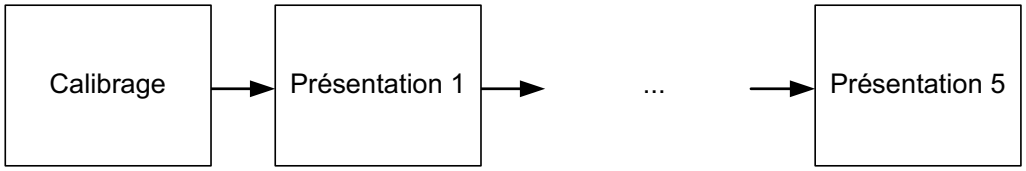


Figure 9.4 – Déroulement de l'affichage d'une liste de présentations en tâche d'évaluation de qualité.

9.2.3 Déroulement de l'ensemble des tests oculométriques

Les deux parties des tests oculométriques (exploration libre et tâche d'évaluation de qualité) sont décomposées en deux séances de manière à minimiser la fatigue visuelle et la lassitude des observateurs. Les séances durent entre 30 et 35 minutes et n'ont pas lieu la même journée. La répartition des différentes listes de vidéos ou de présentations entre les deux séances est la suivante :

- première séance : les deux listes de vidéos (exploration libre), une liste d'entraînement (tâche d'évaluation de qualité) et quatre listes de présentations (tâche d'évaluation de qualité) ;
- seconde séance : six listes de présentations (tâche d'évaluation de qualité) ;

Afin d'éviter l'attraction du centre de l'écran, l'affichage des vidéos n'est pas centré sur celui-ci. Chaque vidéo est affichée à une position pseudo-aléatoire sur l'écran. Afin de réduire les sources de biais, une liste de vidéos d'entraînement est présentée aux observateurs au début de la campagne de tests subjectifs d'évaluation de qualité. Ces vidéos d'entraînement permettent aux observateurs, d'une part, de se familiariser avec l'outil de notation, d'autre part de se faire une idée sur l'importance des dégradations qu'ils vont devoir évaluer.

9.2.4 Construction d'une saillance spatio-temporelle

A l'issue des tests oculométriques sur les vidéos, et comme dans le cas des tests sur les images, nous disposons d'un enregistrement oculométrique, pour chaque observateur et pour chaque vidéo visualisée. Chaque enregistrement contient les positions successives, dans le référentiel des images de la vidéo, du point d'intersection entre la direction du regard de l'observateur et le plan dans lequel est affichée la vidéo. Ces positions successives sont enregistrées toutes les vingt millisecondes. Ces données vont nous permettre de construire une représentation spatio-temporelle de la saillance. Comme nous l'avons évoqué en introduction il n'est pas possible d'utiliser des cartes de saillance pour étudier la saillance sur des vidéos. Les images sont des objets à deux dimensions (2D) qui ne varient pas au cours de temps, il n'est donc pas nécessaire de prendre en compte la dimension temporelle pour déterminer le degré de saillance de chaque site (x, y). Par contre, les vidéos sont des objets ayant une dimension temporelle (2D+t), il est donc nécessaire de considérer celle-ci pour déterminer le degré de saillance de chaque site (t, x, y).

La première conséquence de la prise en compte de la dimension temporelle est donc le remplacement des cartes de saillance par la création de séquences de saillance. A partir des données oculométriques collectées, une séquence de saillance est construite pour chaque observateur et pour chaque vidéo visualisée. Par analogie avec les cartes de saillance (cf. section 8.2.5), les séquences de saillance encodent le degré de saillance de chaque site (t, x, y) des vidéos.

La seconde conséquence de la prise en compte de la dimension temporelle de l'attention visuelle se situe au niveau des mouvements oculaires possibles. Dans le cas des images, les deux principaux type de mouvements oculaires sont les fixations et les saccades. Par contre, dans le cas des vidéos un troisième type de mouvements est à considérer : les mouvements de poursuite. En effet, les mouvements de poursuite étant les mouvements oculaires permettant de suivre un objet en mouvement, ils ne peuvent se produire que dans le cas des vidéos.

De la même manière que les fixations étaient symptomatiques de la saillance des objets fixes, les mouvements de poursuite sont symptomatiques de la saillance des objets en mouvement.

9.2.4.1 Identification des mouvements oculaires

La première étape de construction des séquences de saillance consiste à parcourir chaque enregistrement oculométrique afin d'identifier les périodes de fixations, les mouvements de poursuite ainsi que les périodes de saccade. Comme pour la construction des cartes de saillance, les données relatives aux saccades seront supprimées. L'algorithme suivant est utilisé :

```
-----------------------------------------------------------------------------------------
Algorithme d'identification des fixations et des poursuites
-----------------------------------------------------------------------------------------
Pour chaque échantillon :
1 : Calculer la vitesse point à point de chaque échantillon.

2 : Étiqueter chaque échantillon dont la vitesse point à point est inférieure
à un seuil (25 deg/s) comme fixation ou poursuite, et les autres comme saccade.

3 : Regrouper les échantillons étiquetés en fixation consécutifs en groupe de fixations
ou poursuites et supprimer des échantillons étiquetés en saccade.

4 : Supprimer les groupes de fixations ou poursuites
dont la durée est inférieure à 100 millisecondes.
-----------------------------------------------------------------------------------------
```

Comme l'algorithme utilisé pour les images, cet algorithme est aussi inspiré des travaux de Salvucci et Goldberg [Salvucci 00]. Les fixations et les poursuites sont identifiées grâce, d'une part à la vitesse angulaire du regard (<25°/s), d'autre part par comparaison à une durée minimale (100ms). Cependant l'utilisation de cet algorithme en vidéo nécessite de prendre des précautions. Ces précautions concernent la différentiation des mouvements de poursuite par rapport aux mouvements de saccade. Le seuil utilisé est de 25°/s, cependant la capacité de poursuite de l'oeil est plus importante. Afin de déterminer si l'utilisation de ce seuil est problématique dans notre étude, nous avons analysé les mouvements des objets dans nos séquences de test. Les résultats de cette analyse sont donnés dans le tableau 9.1. Dans toutes les séquences sauf deux, on observe que les objets les plus rapides ont une vitesse inférieure à 25°/s dans les conditions d'observation de nos expérimentations. L'une des séquences dont la vitesse d'un objet dépasse le seuil est la séquence *DucksTakeOff*. Les objets concernés sont les pointes des ailes des canards lorsqu'ils s'envolent. L'amplitude du mouvement est très faible, et cette situation ne représente pas une situation de poursuite. L'autre séquence pouvant poser problème est la séquence *ParkJoy*. L'objet concerné est un arbre qui passe au premier plan en moins de 800ms. Ce n'est pas un objet d'intérêt a priori et les chances d'avoir une poursuite sur cet objet sont faibles. Dans ces conditions, nous avons donc considéré que nous pouvions utiliser l'algorithme proposé.

Par ailleurs, cet algorithme est différent de celui utilisé pour les images, car il faut considérer les mouvements de poursuite, en plus des fixations. Les groupes de fixations (ou poursuites) ne sont plus représentés par le

Séquences	Descriptions	Vitesse des objets les plus rapides	
CrowdRun	Foule qui court pendant un marathon	Coureurs au premier plan	6-8°/s
Dance	Personnes déguisées qui dansent	Le drapeau (ponctuel)	24-25°/s
DucksTakeOff	Canards sur l'eau qui s'envolent	Canards qui s'envolent Ailes qui battent (amplitude courte)	12-13°/s 30-40°/s
Foot	Football	Joueurs de foot ponctuellement	10-11°/s
Hockey	Hockey sur glace	Joueurs de hockey ponctuellement	4-5°/s
InToTree	Zoom sur une maison et un arbre	La maison	10-12°/s
MobCal	Calendrier et jouets en mouvement	Le petit train à la fin	6-7°/s
ParkJoy	Personnes qui courent sur chemin boisé	Arbre au premier plan (<800ms)	30-33°/s
ParkRun	Personne qui court sur chemin en hiver	Le décor en avant plan	4°/s
PrincessRun	Femme qui court sur chemin boisé	Arbre au premier plan	13-14°/s

Table 9.1 – Description des vidéos, et vitesse de déplacement des objets les plus rapides. Les vitesses sont calculées à partir des vecteurs de mouvements, et sont données en °/s relativement aux conditions d'observation des tests expérimentaux (quatre fois la hauteur des images affichées sur l'écran, cf. section 6.3.1).

barycentre des échantillons les composant, car sinon un mouvement de poursuite serait représenté par une position fixe située au barycentre de sa trajectoire. Chaque échantillon qui n'est pas étiqueté en saccade garde donc ses coordonnées, ce qui permet de représenter correctement les mouvements de poursuite.

9.2.4.2 Construction des séquences de saillance

Une fois les mouvements de saccade supprimés, les séquences contenant les fixations et les poursuites peuvent être calculées. Une séquence de fixations et de poursuites CS^k pour un observateur k peut être calculée de plusieurs façons : soit l'intérêt des zones de la vidéo dépend seulement de la présence d'une fixation (ou d'une poursuite), soit il dépend de la présence d'une fixation (ou d'une poursuite) et de la durée de celle-ci.

Une séquence de fixations et de poursuites ne dépendant que de la présence d'une fixation (ou d'une poursuite) est calculée selon la relation :

$$CS^{(k)}(t,x,y) = \Delta(t-t_j, x-x_j, y-y_j), \forall j \in [1;M], \tag{9.1}$$

où M est le nombre total d'échantillons j de fixations et de poursuites, et Δ est le symbole de Kronecker.

Une séquence de fixations et de poursuites dépendant de la présence d'une fixation (ou d'une poursuite) et de leur durée est calculée selon la relation :

$$CS^{(k)}(t,x,y) = \Delta(t-t_j, x-x_j, y-y_j) \cdot d(t_j, x_j, y_j), \forall j \in [1;M], \tag{9.2}$$

où M est le nombre total d'échantillons j de fixations et de poursuites, Δ est le symbole de Kronecker et d est la durée de la fixation ou la période de poursuite.

Afin de déterminer le comportement d'un observateur moyen, les séquences de fixations et de poursuites de tous les observateurs sont combinées en une séquence de fixations et de poursuites moyenne CS :

$$CS(t,x,y) = \frac{1}{N}\sum_{k=1}^{N} CS^{(k)}(t,x,y), \tag{9.3}$$

où N est le nombre d'observateurs. La séquence de fixations et de poursuites moyenne représente les zones les plus attractives de la vidéo lorsqu'un nombre important d'observateurs est considéré.

Tout comme les cartes de fixation moyenne pour les images, la séquence de fixations et de poursuites moyenne ne représente pas vraiment la réalité pour les vidéos. Tout d'abord, l'oeil ne fixe pas un point sur une vidéo, mais plutôt une zone ayant une taille visuelle proche de celle de la fovéa. De plus, les séquences de fixations et de poursuites moyennes sont obtenues à partir d'expérimentations faisant intervenir un appareillage à la précision limitée. En effet, si on se réfère de nouveau à ses caractéristiques (cf. tableau 8.1), la précision du dispositif oculométrique est comprise entre 0.25° et 0.5° de champ visuel. A partir de ces deux considérations, des séquences de densité de saillance DS sont obtenues par convolution des séquences de fixations et de poursuites moyennes CS avec une fonction gaussienne bi-dimensionnelle g_σ :

$$DS(t,x,y) = CS(t,x,y) * g_\sigma(x,y), \tag{9.4}$$

avec :

$$g_\sigma(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-\mu_x)^2-(y-\mu_y)^2}{2\sigma^2}}, \tag{9.5}$$

L'écart type σ est pris égal à 0.5°.

Sauf précision de notre part, ce que nous appellerons par la suite « séquence de saillance », fera référence en fait à une séquence de densité de saillance.

9.3 Impact de la tâche d'évaluation de la qualité sur l'attention visuelle

Nous avons observé précédemment sur les images que les stratégies visuelles ne sont pas identiques suivant les situations et que la tâche d'évaluation de qualité influençait le déploiement de l'attention visuelle des observateurs. Mais qu'en est-il pour des vidéos ?

L'analyse des données oculométriques collectées durant les expérimentations décrites dans la section 9.2, devrait nous permettre d'apporter des éléments de réponse à cette question. Nous rappelons que les séquences originales présentées en exploration libre correspondent aux séquences de références présentées en tâche d'évaluation de qualité. Afin d'éviter les ambiguïtés, nous utiliserons par la suite le terme « référence » pour désigner les séquences originales à la fois en exploration libre et à la fois en tâche d'évaluation de qualité.

9.3.1 La tâche et la durée des fixations/poursuites

A partir des données oculométriques, la durée moyenne de fixation (ou poursuite) est calculée pour chaque observateur et pour chaque vidéo dans les trois situations suivantes :

- la vidéo de référence est visualisée par les observateurs en exploration libre.
- la vidéo de référence est visualisée par les observateurs en tâche d'évaluation de qualité. Cette vidéo est présentée juste avant la version dégradée à évaluer.

– une version dégradée de la même vidéo est visualisée par les observateurs en tâche d'évaluation de qualité.

La durée moyenne de fixation par vidéo présentée est obtenue en calculant la moyenne des durées moyennes de fixation de chaque observateur pour cette vidéo. La figure 9.5 donne la durée moyenne de fixation dans les trois situations mentionnées ci-dessus. On observe que les durées moyennes de fixations sont similaires entre les trois situations. La durée moyenne de fixation varie d'une situation à une autre, ainsi que d'une vidéo à une autre entre 404 ms et 545 ms.

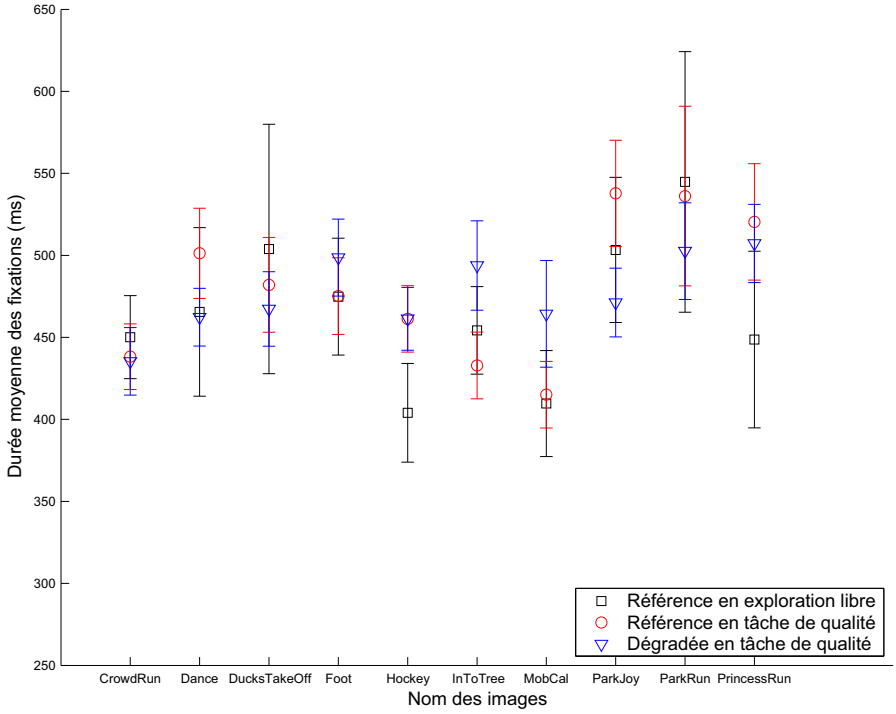


Figure 9.5 – Durée moyenne des fixations (ou poursuites) par vidéo de référence en exploration libre et en tâche d'évaluation de qualité. L'intervalle de confiance à 95% est donné pour chaque cas.

A la différence de ce que nous avions observé pour les images, cette composante de la stratégie visuelle ne semble pas être modifiée par la tâche de qualité dans le cas des vidéos. La dimension temporelle intrinsèque des vidéos joue certainement un rôle dans le rythme donné au déploiement de l'attention visuelle. Grâce au caractère intemporel des images, l'observateur peut aller et venir librement sur les différentes zones de celle-ci dans n'importe quel ordre et sans que cela ait une incidence sur le contenu qu'il va y trouver. Par contre la dimension temporelle des vidéos contraint l'observateur à adapter continuellement sa stratégie visuelle à la vidéo. Il n'est plus aussi libre dans son exploration, car le contenu des différentes zones spatiales varie temporellement. Il ne peut donc plus aller et venir sur les différentes zones spatiales sans que cela ait une incidence sur le contenu qu'il va y trouver. Dans le cas des images, nous avions émis l'hypothèse que les observateurs tentaient de mémoriser certaines zones lors des présentations des images de référence en tâche d'évaluation de qualité, ce qui avait pour effet d'allonger la durée moyenne des fixations dans ce cas. Ce comportement est moins plausible dans le cas des vidéos, car le contenu de chaque zone spatiale est susceptible de changer à tout moment. Dans

un contexte d'évaluation de qualité, il est possible que les observateurs privilégient le nombre de zones explorées, plutôt que le temps consacré à chaque zone. Cette hypothèse est plausible avec les durées moyennes de fixation observées dans les différentes situations.

9.3.2 La tâche et les séquences de saillance

A partir des données oculométriques, des séquences de saillance sont construites pour chaque vidéo présentée. Ces séquences de saillances sont comparées les unes par rapport aux autres pour différentes situations. La figure 9.6 illustre les différentes comparaisons que nous avons effectuées :

- test A), *référence en tâche de qualité versus référence en exploration libre* : dans ce premier essai, nous mettons l'accent sur l'influence de la tâche sur le comportement oculomoteur. Est-ce que les observateurs regardent les mêmes régions ?
- test B), *référence en tâche de qualité versus première référence en tâche de qualité* : l'objectif ici est de montrer (ou pas) que les observateurs adaptent leur stratégie visuelle pour inspecter la vidéo de référence dans une tâche d'évaluation de qualité.
- test C), *vidéo dégradée en tâche de qualité versus référence en exploration libre* : il est bien connu que la tâche de qualité agit sur le déploiement de l'attention visuelle. Mais nous ne savons pas dans quelle mesure une tâche modifie l'attention visuelle. Cette question est abordée ici en comparant les séquences de saillance mesurées en exploration libre et en tâche de qualité. En outre, les dégradations modifient-elles les séquences de saillance ?
- test D), *vidéo dégradée en tâche de qualité versus sa référence associée en tâche de qualité* : lors de l'utilisation du protocole DSIS, la stratégie visuelle est-elle la même pour la référence et pour la vidéo dégradée ?
- test E), *vidéo dégradée en exploration libre versus référence en exploration libre* : les dégradations ont-elles un impact sur la stratégie visuelle en exploration libre ?

Afin de comparer les différentes séquences de saillance, trois indicateurs sont utilisés : la divergence de Kullback-Leibler (KL), le coefficient de corrélation (CC), et l'aire sous les courbes ROC (AUC : *Area Under Curve*).

9.3.2.1 Indicateurs de comparaison

Les deux premiers indicateurs (KL et CC) ont déjà été présentés dans la section 8.3.2 pour la comparaison de cartes de saillance. Dans le cas des séquences de saillance, ces indicateurs sont calculés image par image avant d'être moyennés temporellement.

Le troisième indicateur est l'aire sous la courbe ROC (*Receiver Operating Characteristic*), notée AUC. Comme pour les deux premiers indicateurs, la comparaison est effectuée image par image avant d'être moyennés temporellement. La courbe ROC est un graphe qui mesure la performance d'un classificateur binaire. Nous allons utiliser cette méthode pour évaluer la ressemblance entre deux ensembles de données : deux séquences de saillance. Pour

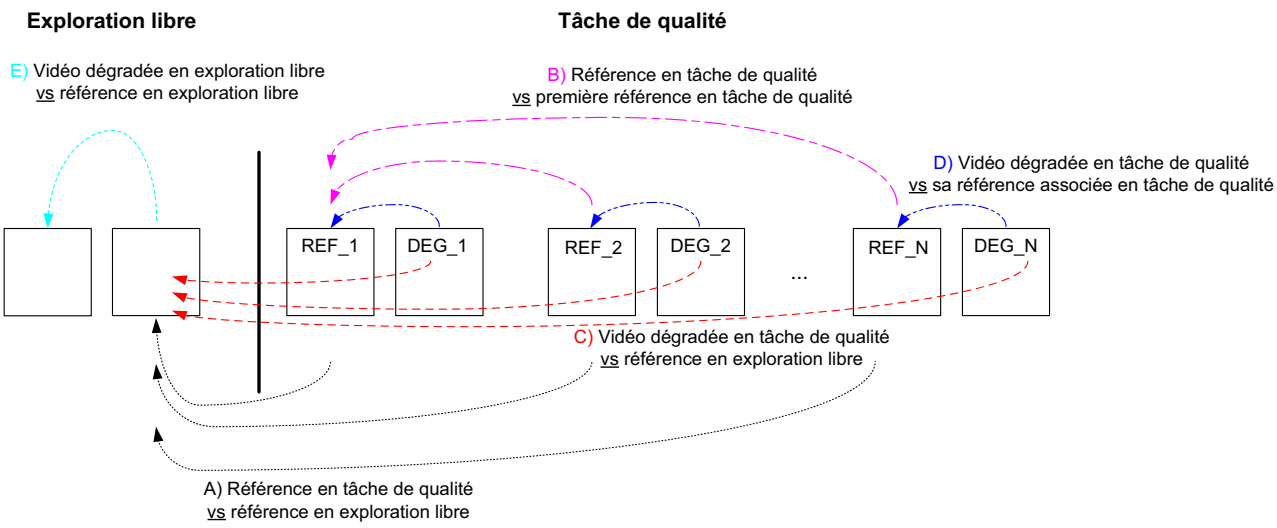


Figure 9.6 – Illustration des différentes comparaisons effectuées entre les séquences de saillance.

les cinq comparaisons effectuées (notées A, B, C, D et E sur la figure 9.6) la séquence de saillance de référence est le premier terme de chaque comparaison. Pour une comparaison donnée, les régions ayant une saillance supérieure à un certain seuil sont étiquetées comme régions fixées (ou poursuivies). La valeur du seuil est fixée à 14. Comme les séquences de saillance sont codées sur 8 bits, la valeur maximale vaut 255 et est obtenue lorsque tous les observateurs regardent au même endroit, au même moment. Étant donné que les expérimentations oculométriques impliquaient 36 observateurs, la contribution d'un observateur particulier est d'environ 7. Une région ayant une valeur de saillance supérieure à 14 correspond à une zone ayant été fixée au même instant par au moins deux observateurs.

Concernant le second terme de la comparaison, appelé « *outcome* » dans la terminologie ROC, 128 seuils sont utilisés pour binariser la séquence, allant de 0 à 254. A partir du tableau de contingence 2x2, le taux de vrais positifs (TPR : *True Positive Rate*), ainsi le taux de faux positifs (FPR : *False Positive Rate*) sont calculés pour chaque seuil. Les TPR et le FPR sont calculés pour chaque image d'une séquence donnée avant d'être moyennés temporellement. Enfin, la courbe ROC définie par des couples (FPR,TPR) est déduite.

Les propriétés de la courbe ROC sont rappelées brièvement ci-dessous :

- Si les séquences de saillance sont très « semblables », on obtiendra un point dans la partie supérieure gauche, ou de coordonnées (0,1), de l'espace ROC ;
- La ligne diagonale divise l'espace ROC entre les zones de bonne ou de mauvaise classification. Contrairement aux zones de mauvaise classification, les zones de bonne classification correspondent à une similitude entre les séquences de saillance.

L'aire sous la courbe ROC (AUC) est une mesure utile pour évaluer la similarité entre deux séquences de saillance. Une valeur d'AUC de 0.5 indique que la discrimination entre les deux ensembles est due au hasard, alors qu'une valeur d'AUC proche de 1 indique une forte similarité.

La figure 9.7 donne les courbes ROC associées à la comparaison A (cf. figure 9.6) comme exemples.

9.3.2.2 Exploration des séquences de référence : influence de la tâche ?

Les résultats des cinq comparaisons, sont présentés dans les figures 9.8, 9.9 et 9.10, pour les indicateurs KL, CC et AUC respectivement. Pour chaque indicateur, les valeurs données correspondent à des moyennes de toutes les vidéos issues de la même vidéo de référence et correspondant à une situation de comparaison (A, B, C, D ou E).

On observe que les tendances générales sont proches d'un indicateur à l'autre. La similarité entre les séquences de saillance dépend fortement du contenu de la vidéo. On observe que les vidéos présentant des objets d'intérêts clairement définis et localisés (*Foot*, *Hockey*, *ParkJoy* et *ParkRun*) sont les vidéos où les séquences de saillance sont les plus similaires d'une situation à une autre. Par exemple, l'AUC pour ces vidéos est supérieure à 0.95 quelle que soit la comparaison considérée. Par contre les vidéos ne présentant pas d'objets d'intérêts clairement définis et localisés (*CrowRun*, *MobCal*, *InToTree*) sont les vidéos où les séquences de saillance varient le plus et de façon différenciée d'une situation à une autre. La même observation peut être faite sur la dispersion inter présentation. En effet, les intervalles de confiance sont plus faibles pour les vidéos présentant une ou des zones d'intérêt localisées.

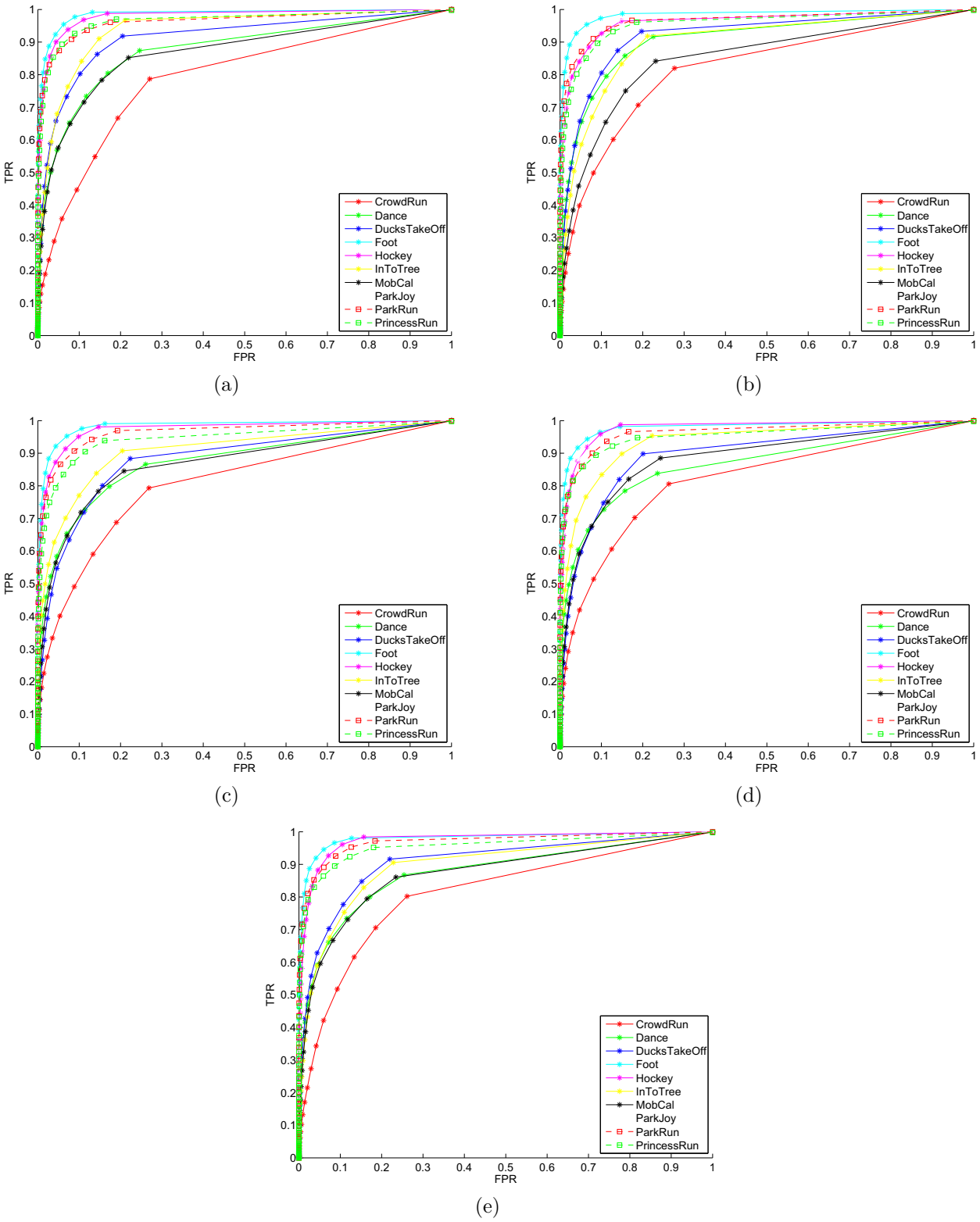
Les résultats des comparaisons A et B permettent d'étudier l'impact de la tâche de qualité sur les stratégies visuelles déployées pour regarder les vidéos de référence. Contrairement à ce que nous avions observé sur les images, dans le cas des vidéos de référence la tâche de qualité ne modifie pas clairement l'attention visuelle par rapport à une situation d'exploration libre, même si les résultats sont légèrement différents d'un indicateur à l'autre en fonction du contenu. Globalement les trois indicateurs ne permettent pas de distinguer les comparaisons A et B sauf peut être :

- pour le CC avec la vidéo *PrincessRun*,
- pour le KL avec les vidéos *Foot* et *ParkRun*,
- pour l'AUC avec les vidéos *CrownRun*, *Foot*, *Hockey*, *ParkRun* et *PrincessRun*.

Cependant l'impact de la tâche reste relativement faible. Ce résultat rejoint les observations faites sur les durées moyennes de fixation. Alors que dans le cas des images de référence en tâche d'évaluation de qualité, les observateurs avaient tendance à s'intéresser à des zones autres que celles regardées en exploration libre, dans le cas des vidéos, les observateurs ont tendance à regarder les mêmes régions des vidéos de référence qu'en exploration libre.

Contrairement à ce que nous avions observé sur les images, les résultats des comparaisons C et D sont proches, même si les résultats sont légèrement différents d'un indicateur à l'autre en fonction du contenu. Globalement les trois indicateurs ne permettent pas de distinguer les comparaisons C et D sauf :

- pour le CC avec les vidéos *Foot* et *InToTree*,
- pour le KL avec la vidéo *InToTree*,
- pour l'AUC avec la vidéo *CrownRun*.



(a) (b) (c) (d) (e)

Figure 9.7 – Courbes ROC associées à la comparaison A. Les figures (a), (b), (c), (d) et (e) représentent respectivement les graphes correspondant à la présentation de la vidéo de référence en tâche de qualité précédent la présentation des versions dégradées pour les cinq niveaux de dégradations (du niveau de dégradations le plus faible au niveau le plus élevé). Ces courbes sont utilisées pour calculer les valeurs d'AUC correspondantes.

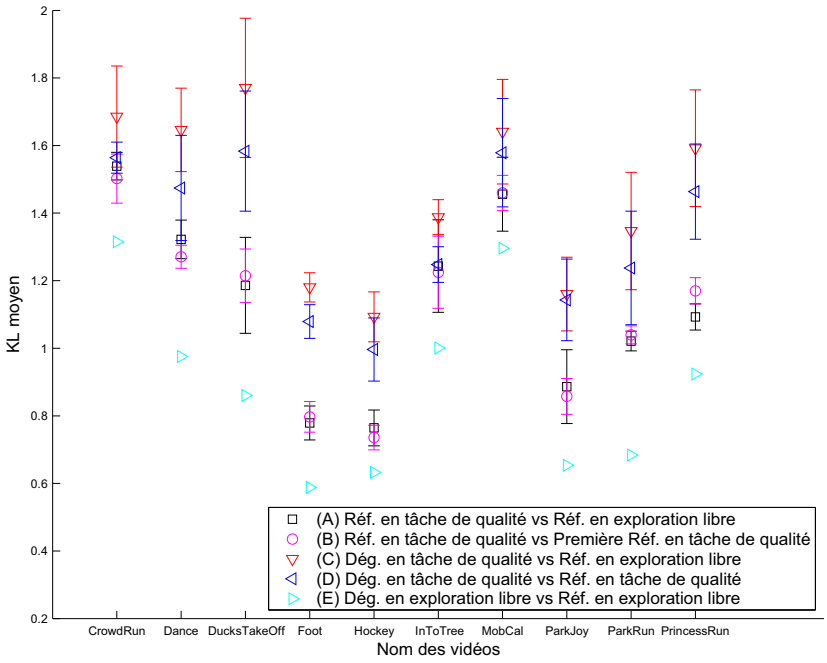


Figure 9.8 – Moyennes des divergences de Kullback-Leibler calculées pour chaque vidéo d'origine. Comme le montre la figure 8.9, les valeurs de KL sont calculées pour les cinq tests (cf. figure 9.6 : A, B, C, D et E). Pour chaque valeur l'intervalle de confiance à 95% est donné.

Ces résultats sont cohérents avec les résultats du paragraphe précédent montrant que les vidéos de référence sont plutôt explorées de la même manière quelle que soit la situation.

9.3.2.3 Influence de la tâche sur l'exploration des séquences dégradées

A l'inverse les résultats des comparaisons A et C d'une part, et B et D d'autre part, indiquent que la tâche d'évaluation de qualité a une influence sur le déploiement de l'attention visuelle en ce qui concerne les vidéos dégradées, autrement dit les vidéos à évaluer. La tendance générale montre que les différences entre les séquences de saillance des vidéos dégradées en tâche de qualité et la référence en exploration libre sont plus importantes que les différences entre les vidéos de référence en tâche de qualité et la référence en exploration libre. De même, la tendance générale montre que les différences entre les séquences de saillance des vidéos dégradées en tâche de qualité et la première référence en tâche de qualité sont plus importantes que les différences entre les vidéos de référence en tâche de qualité et la première référence en tâche de qualité. De plus, les intervalles de confiance sont plus importants dans les comparaisons avec les versions dégradées que dans les comparaisons avec les vidéos de référence, ce qui indique une plus grande dispersion inter présentation. La tâche de qualité semble donc avoir une influence sur l'attention visuelle déployée lors de l'exploration des vidéos dégradées.

Les résultats de la comparaison E indiquent qu'en exploration libre les dégradations ont très peu d'influence sur le déploiement de l'attention visuelle. De toutes les comparaisons de séquences de saillance effectuées, c'est celle pour qui les éléments comparés sont les plus similaires. Cette observation renforce l'influence de la tâche

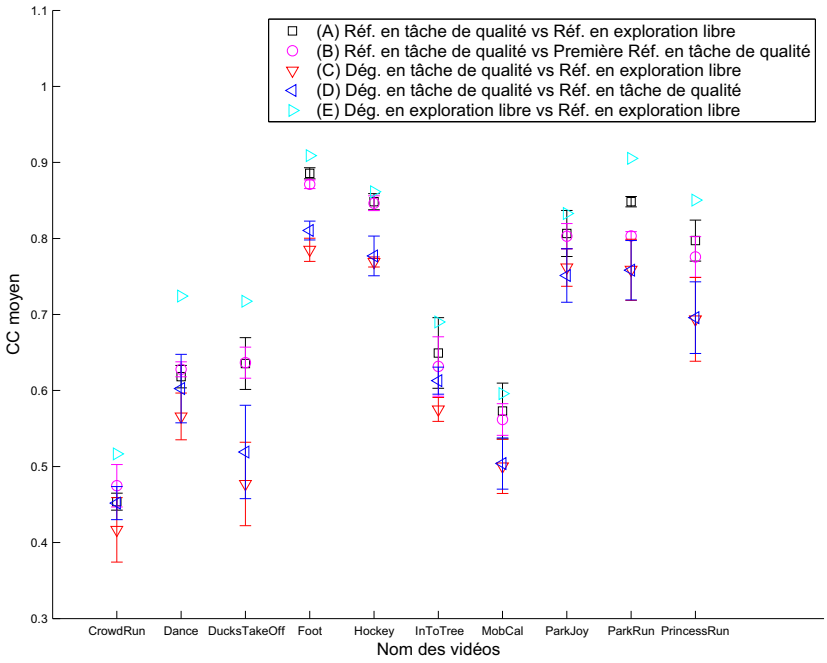


Figure 9.9 – Moyennes des Coefficients de Corrélation calculées pour chaque vidéo d'origine. Comme le montre la figure 8.9, les valeurs de CC sont calculées pour les cinq tests (cf. figure 9.6 : A, B, C, D et E). Pour chaque valeur, l'intervalle de confiance à 95% est donné.

de qualité sur les vidéos dégradées, car il apparaît clairement que les différences entre les séquences de saillance constatées dans le paragraphe précédent, ne sont pas dues aux seules dégradations, mais au fait que les vidéos dégradées se trouvent être les vidéos à évaluer.

9.3.3 Discussion

Les différentes comparaisons nous montrent deux résultats importants :

- une tâche de qualité a peu d'influence sur l'exploration des vidéos de référence,
- une tâche de qualité a de l'influence sur l'exploration des vidéos dégradées.

Le premier résultat est différent de ce que nous avions observé sur les images fixes, où la tâche semble avoir une influence aussi sur l'exploration des images de référence. Pour les images, nous avions émit l'hypothèse que lors de l'évaluation de qualité avec le protocole DSIS, les observateurs tentaient de mémoriser certaines parties de l'image de référence en préparation de l'évaluation de l'image qui suivait. Dans le cas de la vidéo ce comportement n'est pas reproduit. Comme expliqué précédemment, nous supposons que la dimension temporelle de la vidéo prive l'observateur de la liberté nécessaire à l'expression de ce comportement. La dimension temporelle des vidéos contraint l'observateur à adapter continuellement sa stratégie visuelle à la vidéo. Il n'est plus aussi libre dans son exploration, car le contenu des différentes zones spatiales varie temporellement.

On peut donc en déduire que dans le cas de la vidéo deux stratégies visuelles différentes se distinguent. Dans le cas de la vidéo, comme dans le cas des images, l'utilisation de l'attention visuelle pour évaluer la qualité se

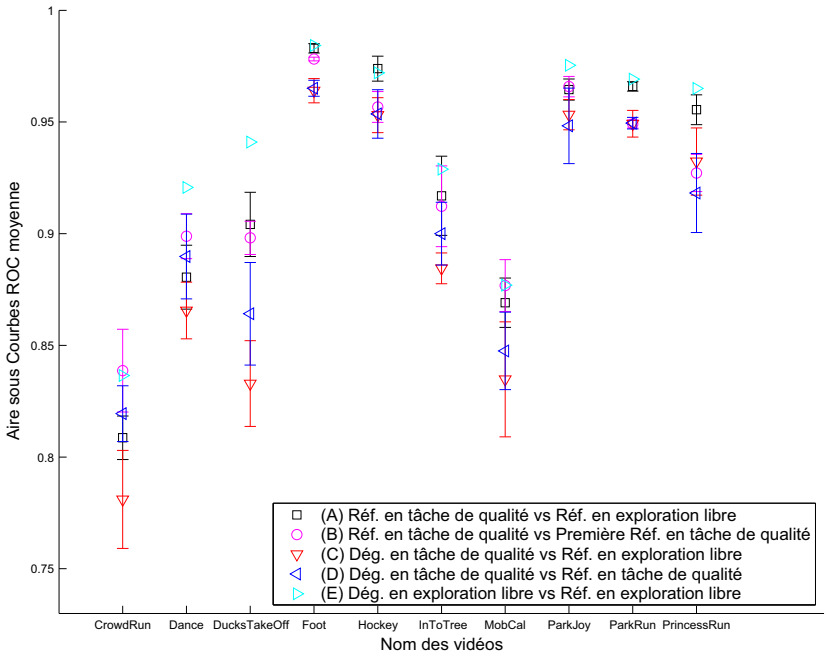


Figure 9.10 – Moyennes des Aires sous les courbes ROC (AUC) calculées pour chaque vidéo d'origine. Comme le montre la figure 8.9, les valeurs de AUC sont calculées pour les cinq tests (cf. figure 9.6 : A, B, C, D et E). Pour chaque valeur l'intervalle de confiance à 95% est donné.

trouve confrontée aux mêmes questions. Si la stratégie visuelle n'est pas la même sur les vidéos de référence et sur les vidéos dégradées, quelle stratégie doit-on utiliser ? Doit-on utiliser l'attention visuelle déployée sur la vidéo de référence ? Ou bien celle déployée sur la vidéo dégradée ? Ou bien les deux ? Comme dans le cas des images, notre étude sera dédiée à l'exploration de la solution consistant à utiliser l'attention visuelle déployée sur la vidéo dégradée.

Faisons maintenant l'hypothèse que la vérité terrain nous ait permis de comprendre comment l'attention visuelle pouvait permettre d'améliorer les performances de métriques de qualité de vidéos. L'étape suivante est de prédire l'attention visuelle. Dans cet optique, les considérations effectuées dans la section 8.3.3 sur les modèles d'attention visuelle des images restent valables pour les modèles d'attention visuelle de vidéos. Ces modèles devront d'une part être capables de prendre en compte l'influence de la tâche de qualité et d'autre part être capables de modéliser les différences de stratégies déployées entre les vidéos de référence et les vidéos dégradées.

9.4 Impact de l'attention visuelle sur les performances de métriques de qualité

L'attention visuelle est supposée pouvoir être un élément important pour améliorer l'évaluation objective de la qualité. Dans le chapitre précédent concernant les images fixes nous étions partis de l'hypothèse qu'un artefact qui apparaît sur une région d'intérêt est beaucoup plus gênant qu'une dégradation apparaissant sur

zone de moindre intérêt. Ainsi, nous avons tenté d'améliorer des métriques de qualité d'images, par le biais de l'information de saillance, en donnant plus d'importance aux dégradations situées sur les régions de forte saillance au détriment des dégradations situées sur les autres régions. Les résultats des expérimentations que nous avons menées sur les images, nous ont montré que le problème était plus complexe et qu'une simple pondération linéaire ne suffisaient pas. Mais qu'en est-il dans le cas de l'évaluation objective de la qualité de vidéos ? C'est ce que nous allons analyser dans cette section.

Cette étude est rendue possible grâce aux tests oculométriques réalisés pendant une campagne d'évaluation de qualité et décrits dans la section 9.2. Nous disposons donc d'une part des MOS, d'autre part de l'information de saillance correspondant aux zones des vidéos que les observateurs ont explorées pour construire leur jugement de qualité. Dans cette étude, différentes fonctions de cumul spatial basées sur l'information de saillance sont examinées. Nous nous efforçons de répondre à la question suivante : est-ce que le recours à l'information de saillance dans une fonction de cumul spatial des distorsions permet d'améliorer la prédiction d'une métrique de qualité de vidéos ?

9.4.1 Métriques de qualité fondées sur la saillance

Dans nos expérimentations, plusieurs métriques de qualité simples exploitant la saillance visuelle sont testées. Ces métriques sont basées sur la métrique VQA proposée dans le chapitre 6. Différentes fonctions de cumul spatial sont testées. Ces fonctions interviennent au niveau du cumul spatial par image des séquences de distorsions spatio-temporelles (le cumul temporel court terme) et en remplacement de la relation (6.1).

9.4.1.1 Fonction de cumul spatial exploitant la saillance

L'idée est d'utiliser l'information locale de saillance afin de pondérer la valeur locale des distorsions spatio-temporelles dans le même esprit que ce qui a été fait dans la littérature pour les images fixes. La forme générale d'une telle fonction de pondération spatiale adaptée à la vidéo est donnée par :

$$D_t^S = \left(\frac{\sum_{k=1}^{K} \sum_{l=1}^{L} w_i(t,x,y) \cdot \left(\overline{\overline{VE}}(t,x,y) \right)^{\beta_s}}{\sum_{k=1}^{K} \sum_{l=1}^{L} w_i(t,x,y)} \right)^{\frac{1}{\beta_s}}, \tag{9.6}$$

où D_t^S représente le niveau de distorsions par image après pondération par la saillance, K et L sont respectivement la hauteur et la largeur de l'image, $w_i(t,x,y)$ est la pondération attribuée au site (t,x,y), i définissant la façon de concevoir la fonction de pondération, et $\overline{\overline{VE}}_{t,x,y}$ représente la carte de distorsions spatio-temporelles à l'instant t, autrement dit la valeur de la distorsion au site (t,x,y). Deux valeurs de β_s ont été testées : 1 et 2. Cette fonction de pondération remplace le cumul spatial défini dans la relation (6.1).

Six fonctions différentes w_i, utilisant l'information locale de saillance, ont été testées. Ces fonctions sont

données par :

$$\left|\begin{array}{l} w_1(t,x,y) = SM_n(t,x,y) \\ w_2(t,x,y) = 1 + SM_n(t,x,y) \\ w_3(t,x,y) = SM(t,x,y) \\ w_4(t,x,y) = 1 + SM(t,x,y) \\ w_5(t,x,y) = SM_b(t,x,y) \\ w_6(t,x,y) = 1 + SM_b(t,x,y) \end{array}\right. \tag{9.7}$$

où $SM(t,x,y) \in [0; S_{max}]$ est la carte de saillance non normalisée à l'instant t, $SM_n(t,x,y) \in [0,1]$ est la carte de saillance normalisée à l'instant t, et $SM_b(t,x,y)$ est une version binarisée de la carte de saillance à l'instant t. Les cartes de saillance sont calculées en fonction du nombre de fixations. La version binarisée de la carte de saillance est réalisée avec une valeur de seuil fixée à 14, ce qui correspond aux zones fixées au même moment par au moins deux observateurs (cf. section 9.3.2.1).

Le niveau global de distorsions D d'une vidéo est calculé au moyen du cumul temporel long terme défini dans la section 6.2.2 par la relation 6.2.

9.4.2 Analyse quantitative

Les mesures objectives de qualité de vidéos testées sont donc basées sur six fonctions de cumul spatial.

Comme précédemment, une fonction psychométrique (cf. relation (6.5)) est utilisée pour transformer le niveau de distorsion D en MOS prédit (MOSp), tel que recommandé par le groupe de travail VQEG [VQEG 00] :

Pour évaluer l'impact de l'information de saillance, les différentes métriques basées saillance sont comparées aussi aux approches classiques ($w_i = 1$ dans la relation (9.6)).

Nous nous intéressons ici au comportement de l'observateur moyen, c'est pourquoi les séquences de saillance moyennées sur l'ensemble des observateurs sont utilisées. Les métriques de qualité sont évaluées en comparant les MOS et les MOSp sur l'ensemble de la base de vidéos présentée dans la section 6.3.1 et en utilisant trois indicateurs de performance : CC, SROCC et RMSE. Les résultats sont présentés dans le tableau 9.2.

Pour $\beta_s = 1$, les résultats montrent qu'il n'y a pas d'amélioration générale des performances. Les performances sont même légèrement moins bonnes pour les pondérations w_1, w_3 et w_5 que pour la version sans pondération ($w_i = 1$). Les ΔCC valent respectivement -0.014, -0.014 et -0.013 pour les pondérations w_1, w_3 et w_5. Les fonctions de cumul w_1, w_3 et w_5 sont plus pénalisantes pour les distorsions présentes dans les zones n'attirant pas l'attention, que les fonctions de cumul w_2, w_4 et w_6. En effet, les fonctions de cumul w_1, w_3 et w_5 utilisant directement la saillance, si celle-ci est nulle le poids des distorsions présentes dans les zones correspondantes l'est aussi. Par contre les fonctions de cumul w_2, w_4 et w_6 sont de la forme $(1 + Saillance)$, ce qui permet de donner un poids non nul aux distorsions présentent dans les zones de saillance nulle. Il semble donc que la pondération par la saillance ne doive pas trop pénaliser ces zones là. Cette observation était aussi valable dans le cas des images (cf. section 8.4.2.1). Pour les autres pondérations les résultats sont équivalents à la version sans pondération ($w_i = 1$).

Cumul			Indicateurs		
Saillance	w_i	β_s	CC	SROCC	RMSE
Aucune	***1***	***1***	***0.889***	***0.904***	***0.526***
Réelle	w_1	1	0.875	0.903	0.554
	w_2	1	0.889	0.904	0.525
	w_3	1	0.875	0.903	0.554
	w_4	1	0.883	0.908	0.538
	w_5	1	0.876	0.904	0.553
	w_6	1	0.89	0.906	0.524
Aucune	***1***	***2***	***0.892***	***0.9***	***0.519***
Réelle	w_1	2	0.878	0.904	0.548
	w_2	2	0.892	0.901	0.519
	w_3	2	0.878	0.904	0.548
	w_4	2	0.886	0.912	0.532
	w_5	2	0.88	0.905	0.546
	w_6	2	0.893	0.902	0.517

Table 9.2 – Comparaison des performances des métriques de qualité en fonction des différentes fonctions de pondération w_i, des valeurs β_s et pour un cumul spatial défini par la relation (9.6).

Pour $\beta_s = 2$, les résultats montrent que les tendances entre les différentes pondérations sont similaires. Cependant, on peut observer une amélioration globale des résultats par rapport à la configuration $\beta_s = 1$, ce qui conforte l'idée de donner plus de poids aux distorsions spatiales les plus importantes (cf. section 5.2).

Les résultats montrent également que la fonction de cumul w_6 produit les meilleures performances quelle que soit la valeur de β_s. Cette fonction utilisant une version binarisée de la saillance (zones saillantes, zones non saillantes), on peut remettre en cause l'intérêt d'utiliser toutes les valeurs intermédiaires de saillance pour la pondération des distorsions.

9.4.3 Discussion

Différentes fonctions de cumul spatial basées sur l'attention visuelle ont été testées dans le but d'améliorer l'évaluation objective de la qualité de vidéos. L'attention visuelle réelle, enregistrée au travers des mouvements oculaires des observateurs lors d'une campagne d'évaluation de la qualité, est utilisée. Les résultats montrent que l'amélioration de la prédiction n'est pas établie et laissent penser que la façon de prendre en compte l'attention visuelle ne peut se limiter à une simple pondération spatiale. Les résultats tendent à montrer également que la fonction de pondération ne doit pas trop pénaliser les zones de saillance nulle (ou de faible saillance) sous peine de voir les performances décroître.

Ces résultats sont concordants avec ceux obtenus sur les images et leur explication peut être similaire. De même que pour les images (cf. section 8.4.3), au cours de l'exploration d'une vidéo à évaluer Un observateur peut passer moins de temps sur une dégradation évidente que sur une dégradation plus discrète. Dans le premier cas, la saillance est faible, mais la contribution à la note de qualité est élevée. La saillance est faible car l'observateur n'est resté que peu de temps sur cette distorsion, et la contribution à la note de qualité est élevée car la distorsion est importante par conséquent la gêne occasionnée doit l'être aussi. Dans le second

cas, la saillance est élevée et la contribution au jugement de qualité est plus faible. La saillance est élevée car l'observateur est resté plus longtemps sur cette distorsion, et la contribution à la note de qualité est faible car la distorsion est faible par conséquent la gêne occasionnée doit l'être aussi. Il semble donc que l'information de saillance et l'intensité des dégradations doivent être considérées conjointement dans la fonction de cumul spatial. L'amélioration des performances de métriques de qualité reposant sur l'utilisation de l'information liée à l'attention visuelle, demande donc l'élaboration de fonctions de pondération plus complexes.

9.5 Conclusion

Ce chapitre était dédié à l'étude de l'attention visuelle en évaluation de qualité de vidéos. Cette étude fut possible grâce à la réalisation de tests oculométriques menés d'une part dans une situation d'exploration libre, d'autre part durant une campagne d'évaluation subjective de la qualité de vidéos.

Le premier aspect étudié concernait l'évaluation subjective de la qualité. A partir des données oculométriques, nous avons montré que la tâche d'évaluation de qualité avait un impact sur la stratégie visuelle déployée pour regarder les vidéos dégradées à évaluer. Par contre, nous avons montré que la tâche d'évaluation de qualité n'influençait pas clairement le déploiement de la stratégie visuelle dans le cas des vidéos de référence. L'impact de la tâche d'évaluation de qualité n'est donc pas le même suivant que l'on présente des images ou des vidéos aux observateurs. Une explication possible réside dans le caractère temporel intrinsèque des vidéos : la dimension temporelle des vidéos contraint l'observateur à adapter continuellement sa stratégie visuelle à la vidéo. Il n'est plus aussi libre dans son exploration, car le contenu des différentes zones spatiales varie temporellement. Dans une situation d'exploration libre, la part du mécanisme *bottom-up* est sans doute plus importante pour les vidéos que pour les images. Cependant, lorsqu'il s'agit de regarder les vidéos à évaluer, le mécanisme *top-down* lié à la tâche à accomplir reste prépondérant.

Le second aspect étudié concernait l'évaluation objective et l'influence de l'attention visuelle dans la construction du jugement de qualité. En utilisant l'information de saillance réelle, nous avons montré que l'utilisation de fonctions de pondération donnant simplement plus d'importance aux zones de forte saillance ne permettait pas d'améliorer de façon générale la prédiction de métriques de qualité. Ces résultats confirment ceux obtenus pour les images dans le chapitre précédent. Comme nous l'avons évoqué dans pour les images, davantage de recherches doivent être menées pour mieux comprendre les mécanismes de l'attention visuelle dans une tâche d'évaluation de la qualité de vidéos. L'information de saillance et l'intensité des dégradations semblent devoir être considérées de façon conjointe.

Conclusion

La troisième partie de ce mémoire était consacrée à l'étude du rôle de l'attention visuelle dans l'évaluation subjective et objective de la qualité d'images et de vidéos. Cette étude repose sur les données collectées lors de tests oculométriques. Ces tests ont été menés sur des images et sur des vidéos, à la fois dans une situation d'exploration libre et à la fois dans des campagnes d'évaluation subjective de la qualité visuelle. Le protocole utilisé lors des tests subjectifs d'évaluation de qualité est le protocole DSIS. Les données collectées ont permis la construction de vérités terrains représentées sous la forme de cartes de saillance pour les images et de séquences de saillance pour les vidéos.

Concernant l'évaluation subjective de la qualité, nous avons comparé les cartes et les séquences de saillance obtenues dans les différentes configurations. Les résultats les plus importants ont montré que la tâche de qualité avait une influence sur le déploiement de l'attention visuelle. Dans le cas des images, l'influence de la tâche se manifeste à la fois sur les images de référence et à la fois sur les images dégradées, et cela de façons différentes. Sur les images de référence, les observateurs semblent essayer de mémoriser certaines zones en prévision de l'image à évaluer qui suit. Par ailleurs, nous avons montré qu'il n'y avait aucun apprentissage de la tâche de qualité du point de vue de l'attention visuelle, confortant ainsi l'utilisation du protocole DSIS. Dans le cas des vidéos, l'influence de la tâche ne se manifeste pas sur les vidéos de référence mais plutôt sur les vidéos à évaluer. L'existence de ces différentes stratégies visuelles complique la modélisation de l'attention visuelle. Dans ce contexte, les modèles d'attention visuelle devraient être capables d'une part de prendre en compte l'influence de la tâche de qualité et d'autre part de simuler les différences de stratégies déployées entre les références et les dégradées. Ce qui, à notre connaissance, n'est pas le cas actuellement.

Concernant l'évaluation objective de la qualité, nous nous sommes intéressés à l'influence de l'attention visuelle dans la construction du jugement de qualité. Pour cela, nous avons utilisé l'information de saillance réelle pour pondérer les distorsions mesurées au sein de différentes métriques de qualité. Les fonctions de pondération étaient inspirées du peu de littérature sur le sujet, à la différence majeure que la saillance utilisée dans nos travaux est la saillance réelle et non pas une saillance issue d'un modèle d'attention visuelle. Nous avons montré que, dans le cas des images comme dans le cas des vidéos et contrairement à certains résultats de la littérature, une simple pondération linéaire des distorsions par l'attention visuelle ne permettait pas d'améliorer clairement les performances de méthodes d'évaluation objective de la qualité. Davantage de recherches doivent être menées dans les deux cas pour mieux comprendre les mécanismes de l'attention visuelle dans la construction du jugement de qualité. Il semble que l'information de saillance et l'intensité des dégradations doivent être considérées de façon conjointe.

Conclusion et perspectives

Plusieurs résultats importants ont été obtenus par les travaux effectués dans cette thèse. Les contributions de ce travail touchent l'évaluation locale des distorsions perceptuelles ainsi que l'évaluation globale de la qualité visuelle d'images et de vidéos. Ces contributions tentent de répondre à des besoins de l'industrie de l'image et de la vidéo.

Contributions majeures concernant l'évaluation locale des distorsions perceptuelles en images et vidéos

L'évaluation locale des distorsions est un des besoins des concepteurs de systèmes de traitement d'images ou de vidéos. Afin de répondre à ce besoin, nous avons conçu et développé des critères objectifs d'évaluation locale des distorsions avec référence complète tant pour les images fixes que pour les vidéos. Les méthodes proposées reposent sur une modélisation du système visuel humain.

Concernant les images fixes, nous avons simplifié une modélisation existante du système visuel humain en proposant une décomposition en sous-bande fondée sur la transformée en ondelettes. De plus, nous avons proposé une amélioration de la modélisation des effets de masquage par la prise en compte du masquage semi-local en plus du masquage de contraste.

Concernant les vidéos, nous avons conçu et développé une nouvelle approche d'évaluation locale des distorsions temporelles. Cette approche repose sur un cumul temporel court terme des distorsions spatiales. Ce cumul temporel court terme est une modélisation fovéale du système visuel humain simulant l'évaluation des distorsions d'une vidéo réalisée au travers des mécanismes de sélection de l'attention visuelle.

L'absence de vérité terrain ne nous a pas permis de réaliser une évaluation quantitative de la pertinence des cartes de distorsions visuelles obtenues pour les images, ou des séquences de distorsions visuelles obtenues pour les vidéos. Cependant, dans le cas des images, nous avons évalué qualitativement la pertinence de nos cartes de distorsions visuelles par la réalisation de tests comparatifs sur des observateurs. Ces tests montrent que nos cartes de distorsions visuelles ont été préférées dans 60% des cas par rapport aux cartes d'erreurs quadratiques et aux cartes de SSIM.

Contributions majeures concernant l'évaluation de la qualité visuelle d'images et de vidéos

Un autre besoin des concepteurs de systèmes de traitement d'images ou de vidéos est l'évaluation objective de la qualité d'images et de vidéos, ou autrement dit, la construction automatique d'un jugement global de qualité visuelle. Pour répondre à ce besoin, nous avons proposé des critères objectifs de qualité visuelle avec référence complète pour les images fixes et pour les vidéos.

Concernant les images fixes, nos travaux ont consisté à concevoir, développer et valider des métriques de qualité s'appuyant sur les critères objectifs développés pour évaluer localement les distorsions perceptuelles spatiales auxquelles nous avons ajouté ensuite un cumul spatial des distorsions. Ces métriques ont été évaluées à partir d'un ensemble de tests subjectifs de qualité visuelle. Les résultats les plus importants montrent d'une part que la prise en compte du masquage semi-local dans la modélisation des effets de masquage améliore significativement les performances des critères et d'autre part qu'il est possible de simuler le comportement multi-canal du système visuel à partir d'une transformée en ondelettes, avec de bonnes performances. Par exemple, dans le cas de la métrique fondée sur la transformée en ondelettes et modélisant le masquage semi-local, les coefficients de corrélation avec les notes subjectives moyennes varient entre 0.919 pour la base *Toyama1* (MOS) et 0.943 pour la base *Toyama2* (DMOS).

Concernant les vidéos, nous avons proposé une nouvelle approche d'évaluation objective de la qualité. Cette approche repose sur deux cumuls temporels : un cumul temporel long terme et un cumul temporel court terme. Le cumul temporel court terme est celui développé dans cette thèse pour l'évaluation locale des distorsions perceptuelles spatio-temporelles. Le cumul temporel long terme intègre un comportement asymétrique sur les variations instantanées de distorsions et un effet de saturation perceptuelle. Afin d'évaluer les performances de notre approche et de les comparer avec celles de métriques de la littérature, des tests subjectifs d'évaluation de la qualité de vidéo ont été menés. Les résultats montrent les bonnes performances de l'approche proposée ainsi que la nécessité des deux cumuls temporels utilisés. Par exemple, la métrique que nous avons développée (VQA) fournit un coefficient de corrélation avec les notes subjectives moyennes de 0.892 sur la base de test utilisée, contre 0.516 pour le PSNR moyenné temporellement et contre 0.738 pour le critère VSSIM.

Contributions majeures relativement à l'attention visuelle dans un contexte d'évaluation de la qualité d'images et de vidéos

Nous avons apporté les premiers éléments de réponse au rôle de l'attention visuelle en évaluation de la qualité visuelle d'images et de vidéos. Pour cela, nous avons mené des tests oculométriques à la fois dans une situation d'exploration libre et à la fois dans des campagnes d'évaluation subjective de la qualité visuelle. Les données collectées ont permis la construction de la vérité terrain nécessaire à nos travaux.

Concernant l'évaluation subjective de la qualité, nous avons comparé les cartes et les séquences de saillance obtenues dans les situations d'exploration libre et d'évaluation subjective de la qualité visuelle. Les résultats les

plus importants ont montré l'influence de la tâche de qualité sur le déploiement de l'attention visuelle ainsi que celle, dans une moindre mesure, des distorsions. L'existence de ces différentes stratégies visuelles complique la modélisation de l'attention visuelle. Dans ce contexte, les modèles d'attention visuelle devraient être capables d'une part de prendre en compte l'influence de la tâche de qualité et d'autre part de simuler les différences de stratégies déployées entre les références et les dégradées.

Concernant l'évaluation objective de la qualité, nous nous sommes intéressés à l'influence de l'attention visuelle dans la construction du jugement de qualité. Pour cela, nous avons utilisé l'information de saillance réelle pour pondérer les distorsions mesurées au sein de différentes métriques de qualité. Les fonctions de pondération étaient inspirées du peu de littérature sur le sujet, à la différence majeure que l'attention visuelle utilisée dans nos travaux était la véritable attention visuelle produite naturellement par les observateurs et non pas celle issue d'un modèle. Nous avons montré que, dans le cas des images comme dans celui des vidéos et contrairement à certains résultats de la littérature, l'utilisation de l'attention visuelle ne peut se limiter à une simple pondération linéaire des distorsions.

Perspectives

Les perspectives sont multiples et relatives aux différents sujets abordés dans cette thèse.

Concernant l'évaluation locale des distorsions. Nous avons montré que la prise en compte du masquage semi-local présentait un grand intérêt. Cependant, sa modélisation mériterait d'être améliorée car l'utilisation de l'entropie n'est sans doute pas la mesure semi-locale idéale. La définition d'une mesure semi-locale permettant de mieux qualifier l'effet de masquage semi-local est un axe de recherche intéressant. Par ailleurs, un problème récurant en évaluation locale des distorsions est l'absence de vérité terrain. La conception et la réalisation d'expérimentations permettant d'en constituer une seraient bénéfiques pour toute la communauté travaillant sur le sujet.

Concernant l'évaluation objective de la qualité visuelle globale d'images animées, d'autres façons de réaliser le cumul temporel long terme peuvent être envisagées. On pourrait remettre en question la position du cumul spatial par images le précédent en imaginant un cumul spatio-temporel reprenant le comportement asymétrique et l'effet de saturation perceptuelle du cumul temporel long terme actuel.

Concernant l'utilisation de l'attention visuelle en évaluation objective de la qualité, les recherches ne font que commencer. Ces recherches doivent passer par la mise au point de fonctions de pondération plus élaborées que de simples pondérations linéaires des distorsions. Ces fonctions doivent préalablement être validées avec l'attention visuelle réelle avant que celle-ci ne soit remplacée par le produit d'un modèle d'attention visuelle. Un axe de recherche peut être de concevoir une fonction considérant de façon conjointe l'attention visuelle et l'intensité des dégradations. Par ailleurs, des travaux peuvent être envisagés sur la modélisation de l'attention visuelle dans ce contexte afin que soit intégrer dans les modèles les aspects cognitifs nécessaires.

Annexes

Annexe A

Biologie du système visuel humain

A.1 L'oeil : organe de la vision

Du point de vue fonctionnel, l'oeil peut être comparé à un appareil photo et la rétine à une pellicule photographique. En effet, le rôle de l'appareil photo est de concentrer sur le film une image nette ni trop sombre ni trop lumineuse. On y parvient grâce à la bague de mise au point qui met l'objet au foyer, et au diaphragme qui s'ouvre et se ferme pour laisser passer juste la bonne quantité de lumière pour la sensibilité du film. Notre oeil fait exactement la même chose, à tout moment de la journée. La mise au point est assurée par la cornée et le cristallin, alors que l'iris s'occupe d'ajuster la luminosité optimale pour notre rétine. Celle-ci, avec ses nombreuses couches de neurones, est toutefois beaucoup plus complexe et sensible qu'une pellicule photographique.

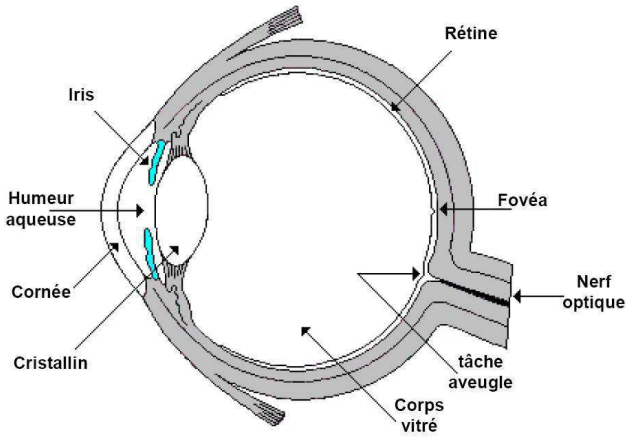


FIGURE A.1 – Coupe transversale de l'oeil.

La première membrane traversée par la lumière est la conjonctive. Il s'agit d'une fine membrane transparente qui couvre le devant de l'oeil et se replie pour tapisser l'intérieur des paupières. Avant d'atteindre les différentes régions de la rétine, la lumière traverse ensuite la cornée qui forme la surface externe transparente et légèrement bombée au centre de l'oeil (cf. Figure A.1). Comme la cornée ne possède pas de vaisseaux sanguins, elle prend ses nutriments dans le milieu qui est situé derrière, l'humeur aqueuse, ainsi que dans celui qui est situé devant,

les larmes répandues par le clignement des paupières.

La lumière traverse ensuite le cristallin, véritable lentille qui baigne entre l'humeur aqueuse et l'humeur vitrée qui remplit l'intérieur de l'oeil.

La pupille est le terme employé pour désigner l'orifice qui permet à la lumière d'entrer dans l'oeil et d'atteindre la rétine. Le diamètre de la pupille est contrôlé par l'iris, un muscle circulaire dont la pigmentation donne la couleur à l'oeil, et dont la contraction lui permet de s'adapter continuellement aux différentes conditions d'éclairage. Ainsi, la nuit, on aura de grandes pupilles noires parce que notre iris est ouvert au maximum pour laisser entrer le peu de lumière disponible. C'est ce qu'on appelle le réflexe pupillaire.

Le fond de l'oeil est pour sa part tapissé par la rétine qui capte les rayons lumineux. Le nerf optique, formé par les axones des cellules ganglionnaires de la rétine, quitte ensuite l'oeil par l'arrière pour rejoindre le premier relais visuel dans le cerveau. Cette zone de la rétine où l'information lumineuse est perdue est appelée disque optique ou tâche aveugle.

La sclérotique, ou blanc de l'oeil, est en continuité de la cornée. Elle forme la paroi dure du globe oculaire et dans laquelle sont insérées trois paires de muscles. Ce sont ces muscles oculaires qui permettent les mouvements du globe oculaire dans les orbites du crâne.

Située entre la sclérotique et la rétine, la choroïde est une couche richement vascularisée qui assure la nutrition de l'iris et de la rétine. Elle contient une couche de cellules pigmentées qui absorbent la lumière et qui font que l'intérieur de notre oeil, visible à travers la pupille, paraît noir.

Différentes parties de l'oeil participent à la focalisation de l'image sur la rétine. L'humeur aqueuse et l'humeur vitrée jouent un rôle fondamental dans la focalisation de l'image sur la rétine grâce au phénomène de réfraction. La courbure de la cornée accentue aussi la réfraction des rayons lumineux virtuellement parallèles provenant d'objets très éloignés. Le cristallin contribue également, mais dans une moindre mesure, à réfracter les rayons lumineux venant de loin pour qu'ils convergent en un seul point sur la rétine. Cependant à plus courte distance, à partir de 9 mètres et moins environ, le cristallin joue un rôle beaucoup plus actif pour nous aider à faire la mise au point.

Une fois que la lumière a atteint la rétine en y formant une image ni trop sombre, ni trop lumineuse, le système optique de l'oeil a joué son rôle. C'est maintenant à la rétine de jouer le sien.

A.2 La rétine

Pour voir, il faut d'abord que l'oeil forme une image précise de la réalité sur la rétine. Il faut ensuite que l'intensité lumineuse soit transformée en influx nerveux par les cellules photoréceptrices de celle-ci. Le traitement de l'image par le système nerveux devient alors possible et il commence non pas dans le cerveau mais immédiatement dans la rétine elle-même.

A.2.1 La rétine : Une structure multicouche

Concrètement, la rétine est une fine pellicule de tissu nerveux ayant la consistance et l'épaisseur d'un papier à cigarette mouillé (0.1 à 0.5 mm). Les neurones de la rétine sont organisés en trois couches principales séparées par 2 couches intermédiaires où se font surtout des connexions entre les différents neurones.

La première couche située en profondeur contient les photorécepteurs qui sont les seules cellules de la rétine capables de convertir la lumière en influx nerveux. Les photorécepteurs réagissent à différentes longueurs d'ondes et différentes intensités lumineuses. Ils sont séparés en deux grandes familles : les cônes et les bâtonnets. La répartition des cônes et des bâtonnets n'est pas uniforme sur la rétine comme nous l'indique la figure A.2.

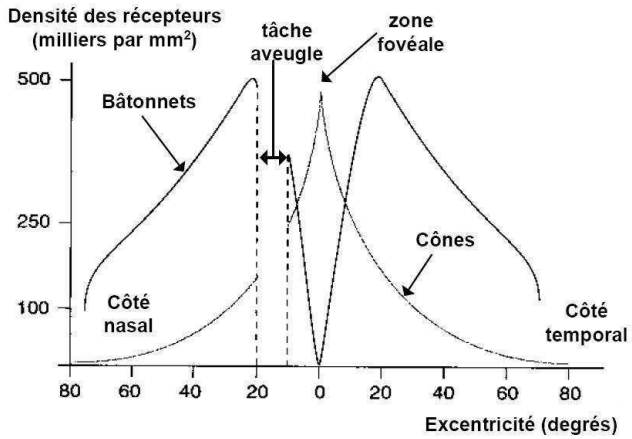


Figure A.2 – Répartition des cellules photoréceptrices sur la rétine.

Les cônes se concentrent au centre de la rétine dans une région appelée la fovéa, alors que les bâtonnets, beaucoup plus nombreux, sont situés dans la rétine périphérique. Le nombre de photorécepteurs connectés à une même cellule ganglionnaire est aussi beaucoup plus grand en périphérie. L'effet combiné de cette organisation est d'accroître la sensibilité à la lumière en périphérie de la rétine. La contrepartie est que la précision de l'image souffre de la convergence de nombreux photorécepteurs sur une même cellule ganglionnaire. Une bonne acuité visuelle comme celle de la rétine centrale demande en effet un faible rapport photorécepteurs/cellules ganglionnaires. Elle est aussi favorisée par les cônes de la fovéa qui sont très petits et tassés les uns contre les autres. Plus on s'éloigne de la fovéa, plus la taille des cônes augmente ainsi que l'espace entre eux, les bâtonnets remplissant l'espace restant. Malgré la grande densité des cônes dans la fovéa, la petitesse de cette région fait en sorte que seulement quelques pourcents des cônes de la rétine s'y trouvent. Des études portant sur les cônes ont permis de les diviser en trois grandes catégories en fonction de leur sensibilité aux longueurs d'ondes lumineuses : les cônes dits S (pour *Small*), les cônes dits M (pour *Medium*) et les cônes dits L (pour *Large*). Ces trois catégories ont une sensibilité maximale située respectivement autour des longueurs d'ondes de 420 nm (proche de la couleur bleue), 531 nm (proche de la couleur verte) et 558 nm (proche de la couleur rouge). Les réponses (normalisées) de ces trois types de cônes en fonction des longueurs d'ondes sont illustrées figure A.3.

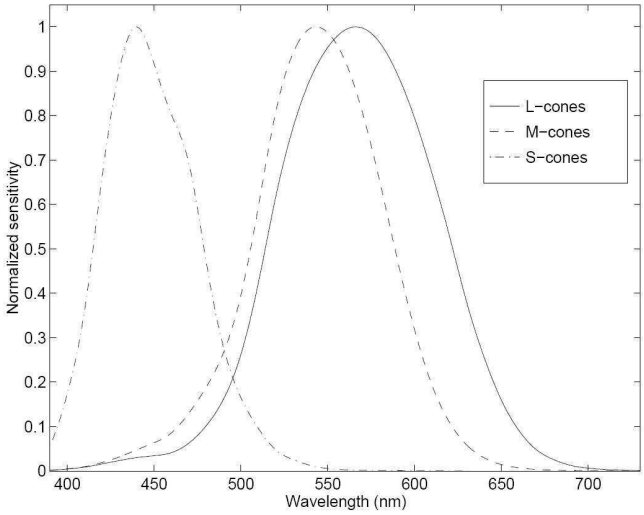


Figure A.3 – Réponse normalisée des cônes L,M et S.

L'influx nerveux issu des photorécepteurs est ensuite transmis aux neurones bipolaires situés dans la deuxième couche, puis aux neurones ganglionnaires situés dans la troisième. Ce sont uniquement les axones de ces neurones ganglionnaires qui vont sortir de l'oeil pour rejoindre le premier relais visuel dans le cerveau.

À côté de cette voie directe qui va des photorécepteurs au cerveau, deux autres types de cellules participent au traitement de l'information visuelle dans la rétine. D'une part les cellules horizontales reçoivent de l'information des photorécepteurs et la transmettent à plusieurs neurones bipolaires environnants. Et d'autre part les cellules amacrines reçoivent leurs « entrées » des cellules bipolaires et procèdent de la même façon avec les neurones ganglionnaires c'est-à-dire activent ceux qui sont dans les environs. Il existe trois types de cellules horizontales présentant une préférence chromatique donnant naissance à des antagonismes chromatiques [Valois58] [Jameson 55] rouge-vert (les signaux des cônes M et L s'opposent) et jaune-bleu (les signaux des cônes S, s'opposent à la somme des signaux des cônes M et L).

A.2.2 Les champs récepteurs

Les neurones des différentes couches de la rétine « couvrent » chacun une région de notre champ visuel. Cette région de l'espace où la présence d'un stimulus approprié modifie l'activité nerveuse d'un neurone est appelée le champ récepteur de ce neurone.

Pour un photorécepteur donné par exemple, on peut dire que son champ récepteur est limité au petit point lumineux qui, dans le champ visuel, correspond à l'emplacement précis du photorécepteur sur la rétine. Mais au fur et à mesure que l'on passe d'une couche de la rétine à l'autre, et à plus forte raison si l'on se rend jusqu'aux neurones du cortex visuel, les champs récepteurs se complexifient.

Ainsi, les champs récepteurs des cellules bipolaires sont de forme circulaire. Le centre et la périphérie de ce disque fonctionnent toutefois en opposition : un jet de lumière qui frappe le centre du champ va avoir l'effet inverse lorsqu'il tombe sur la périphérie. Par exemple, si un stimulus lumineux sur le centre a un effet excitateur

sur la cellule bipolaire, celle-ci subit une dépolarisation. On dit alors qu'elle est à centre ON. Un rayon de lumière qui tombe seulement sur la périphérie du champ de cette cellule aura l'effet opposé, c'est-à-dire une hyperpolarisation de la membrane. D'autres cellules bipolaires, à centre OFF celles-là, vont montrer exactement le comportement inverse : la lumière sur le centre produit ici une hyperpolarisation alors qu'un stimulus lumineux sur la périphérie a un effet excitateur. On distingue donc deux types de cellules bipolaires selon la réponse de leur champ récepteur : à centre ON et à centre OFF.

Tout comme les cellules bipolaires, les cellules ganglionnaires ont également des champs récepteurs concentriques qui possèdent un antagonisme centre-périphérie. Mais contrairement aux cellules bipolaires, ce n'est pas par une hyperpolarisation ou une dépolarisation que répondent les deux types de cellules ganglionnaires, ON ou OFF, mais bien par des potentiels d'action dont la fréquence de décharge est augmentée ou diminuée. Ceci dit, la réponse à la stimulation du centre du champ récepteur est toujours inhibée par la stimulation de la périphérie.

La conséquence de cette organisation est que les zones excitatrices et inhibitrices se neutralisent lorsqu'elles sont excitées par un signal uniforme, alors qu'elles amplifient la réponse à un signal de type contour. La sensibilité de la rétine est donc basée sur l'information de contraste.

On distingue deux types de cellules ganglionnaires, les cellules P et les cellules M, correspondant à deux flux visuels séparés dans le cerveau, appelés respectivement voie parvocellulaire et voie magnocellulaire. La très grande majorité des cellules ganglionnaires sont de type P, elles ont un champ récepteur très réduit et encodent les détails d'une image ainsi que la plupart des informations chromatiques. Les cellules de type M possèdent des champs récepteurs très larges, elles sont insensibles à la couleur mais répondent au mouvement.

La rétine est le premier étage de traitement de l'image par le système nerveux. C'est un élément de prétraitement important permettant de filtrer (spatialement et temporellement) et de décomposer (couleur, mouvement, etc.) l'information lumineuse avant les traitements post-rétiniens.

A.3 De la rétine au cortex

Les axones des cellules ganglionnaires de la rétine se rassemblent pour former le nerf optique. C'est par lui que l'information visuelle, maintenant traduite en influx nerveux se propageant le long du nerf, se rendra jusqu'aux différentes structures cérébrales responsables de l'analyse du signal visuel. Les nerfs optiques quittent donc les yeux au niveau des disques optiques et se réunissent pour former le chiasma optique juste en avant de l'hypophyse. Le chiasma optique permet la décussation d'un certain nombre d'axones en provenance de la rétine, c'est-à-dire leur changement de côté pour assurer le traitement croisé de l'information visuelle. C'est en partie grâce à cette répartition de l'information que nous pouvons percevoir le relief d'une scène visuelle.

Les axones en provenance du côté nasal de la rétine vont changer de côté au niveau du chiasma optique pour faire en sorte que la moitié gauche du champ visuel soit perçue par l'hémisphère cérébral droit, et vice-versa. Comme la partie de la rétine du côté des tempes reçoit déjà son information du champ visuel qui lui est opposé, ses axones n'ont pas besoin de changer de côté et continuent tout droit dans le tractus optique.

La grande majorité des fibres nerveuses du tractus optique projette sur le corps genouillé latéral (CGL)

dans la partie dorsale du thalamus. Le CGL constitue le relais principal de la voie qui mène au cortex visuel primaire. Cette projection du CGL vers le cortex visuel porte le nom de radiation optique. La distribution stratifiée des neurones du CGL indique que des aspects distincts de l'information visuelle en provenance de la rétine pourraient être traités séparément au niveau de ce relais synaptique. Les différentes couches neuronales sont rassemblées en trois types : magnocellulaires, parvocellulaires et coniocellulaires. En fait, il a été démontré que ce sont très exactement les cellules ganglionnaires de type M qui projettent leur réponse dans les couches magnocellulaires du CGL et les cellules ganglionnaires de type P dans les couches parvocellulaires. Le traitement en parallèle de canaux d'information distincts à partir de la rétine semble donc être préservé à travers le CGL, et la spécialisation des cellules du CGL est très similaire à celle des cellules ganglionnaires de la rétine. L'utilité des couches coniocellulaires est encore mal connue.

Le CGL n'est pas un simple relais passif sur la voie qui va de la rétine au cortex. Le cortex visuel primaire exerce une rétroaction importante sur le CGL modifiant en retour ses réponses visuelles. De plus, le CGL peut être activé par des neurones du tronc cérébral dont l'activité est associée à la vigilance et aux processus attentionnels. Ceux-ci agiraient comme modulateur de la réponse des neurones du CGL, renforçant l'idée que le CGL est en réalité le premier endroit de la voie visuelle où des états mentaux particuliers influencent notre perception visuelle.

Les cellules du CGL vont ensuite rejoindre leur cible principale : le cortex visuel primaire. Aussi appelé cortex strié ou simplement V1, le cortex visuel primaire se situe dans la partie la plus postérieure du lobe occipital du cerveau. C'est là que l'image va commencer à être reconstituée à partir des champs récepteurs des cellules de la rétine. Le champ visuel fovéal constitue une zone de projection plus importante que la zone rétinienne correspondante. Ceci étant du à la répartition irrégulière des photorécepteurs sur la rétine comme expliqué section A.2.1. Cette zone joue un rôle important dans la focalisation de l'attention visuelle par une analyse locale plus fine. Ce sont les mouvements oculaires qui permettent de positionner l'image d'un objet particulier sur cette zone de projection. En plus du cortex primaire, près d'une trentaine d'aires corticales différentes contribuant à la perception visuelle ont été découvertes jusqu'à ce jour. Les aires primaires (V1) et secondaires (V2) sont entourées de nombreuses autres aires visuelles tertiaires ou associatives : V3, V4, V5, etc. Dans l'aire V2, les principales caractéristiques de l'aire V1 se retrouve. L'aire V3 traite des informations relatives à la dynamique des formes, mais ne semble pas être sensible aux couleurs. L'aire V4 traite des informations relatives à la couleur et à l'orientation. L'aire V5 est particulièrement sensible au mouvement, mais ne semble pas sensible aux couleurs et aux formes.

Un schéma général émerge toutefois de cette complexité selon lequel il existerait deux grands systèmes corticaux de traitement de l'information visuelle : une voie ventrale qui s'étendrait vers le lobe temporal, et une voie dorsale qui se projette vers le lobe pariétal. La voie ventrale aurait pour mission fondamentale de permettre la perception consciente, la reconnaissance et l'identification des objets en traitant leurs propriétés visuelles « intrinsèques » comme leur forme, leur couleur, etc. La voie dorsale, en revanche, aurait pour mission fondamentale d'assurer le contrôle visuo-moteur sur les objets en traitant leurs propriétés « extrinsèques », celles

qui sont critiques pour leur saisie, comme leur position spatiale, leur orientation ou leur taille.

Annexe B

Implémentation de la DCP

La décomposition en canaux perceptuels met en oeuvre un découpage du plan fréquentiel tel qu'obtenu par une transformée de Fourier. Dans nos travaux, nous utilisons une transformée de Fourier rapide FFT (acronyme anglais : *FFT* ou *Fast Fourier Transform*) afin de calculer la transformée de Fourier discrète (TFD) des images. L'élaboration de la décomposition en canaux perceptuels (figure 2.8) est réalisée à partir d'un ensemble de filtres : les filtres DoM et les filtres Fan. Dans cette annexe nous allons décrire les caractéristiques de ces filtres ainsi que le processus utilisé pour les calculer.

B.0.0.1 Les filtres DoM

Les filtres DoM (*Difference of Mesa*) sont des filtres passe-bande sans sélectivité angulaire : ce sont eux qui permettent de construire les couronnes. Pour construire un filtre DoM, il faut commencer par calculer des filtres Mesa qui sont des filtres 2D passe-bas en fréquences radiales. La fonction de transfert d'un filtre Mesa peut être modélisée par la convolution d'un échelon de Heaviside avec une gaussienne. L'échelon de Heaviside a une valeur unité à l'intérieur d'un cercle de rayon fc et une valeur nulle à l'extérieur. Un filtre Mesa est défini par l'expression suivante :

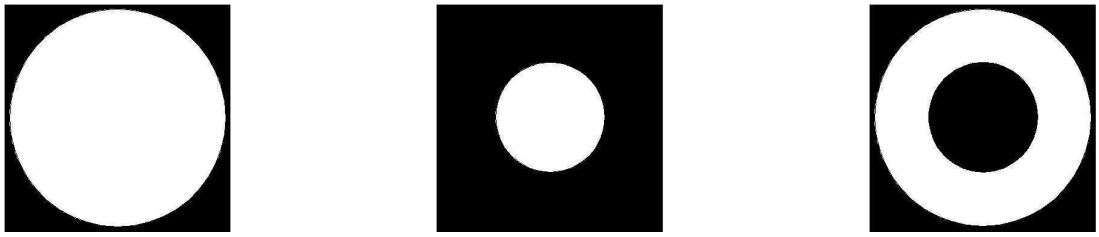
$$Mesa_{f_c}(f) = échelon_{f_c}(f) \otimes \left[\frac{1}{\sigma \times \sqrt{2\pi}}.exp(-\frac{f^2}{2.\sigma^2})\right], \tag{B.1}$$

dans laquelle f est la fréquence spatiale radiale, f_c est la fréquence de coupure de l'échelon et $échelon_{f_c}(f) = 0$ si $f<f_c$, et 1 sinon.

Ensuite, un filtre DoM est construit en calculant la différence entre deux filtres Mesa ayant des fréquences de coupure f_c différentes ($f_{c2} > f_{c1}$). L'équation est la suivante :

$$DoM_{f_{c1},f_{c2}}(f) = Mesa_{f_{c2}}(f) - Mesa_{f_{c1}}(f). \tag{B.2}$$

La figure B.1 illustre la construction d'un filtre DoM à partir de 2 filtres Mesa.



(a) (b) (c)

Figure B.1 – Construction d'un filtre DoM : les 2 filtres Mesa utilisés en (a) et en (b), et en (c) le filtre DoM résultant (correspondant à la couronne IV de la DCP).

B.0.0.2 Les filtres Fan

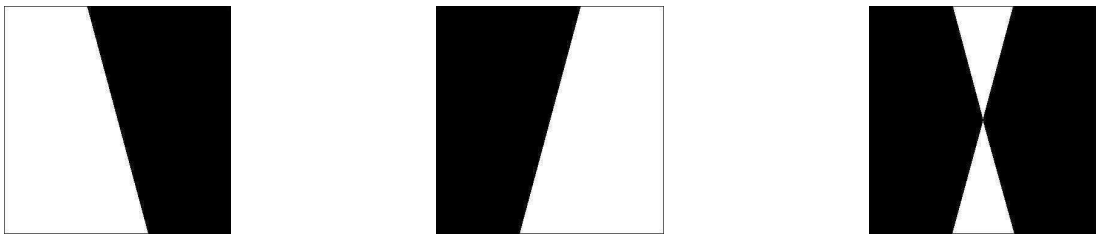
Pour construire un filtre Fan, il faut d'abord obtenir un filtre Step qui est un échelon orienté convolué avec une gaussienne. Un filtre Step ayant une orientation θ est défini par l'équation :

$$Step(u,v) = échelon_\theta(u,v) \otimes \left[\frac{1}{\sigma' \times \sqrt{2\pi}}.exp(-\frac{f^2}{2.\sigma'^2})\right], \tag{B.3}$$

dans laquelle u,v représentent les fréquences spatiales horizontales et verticales respectivement, f est la fréquence spatiale radiale ($f^2 = u^2 + v^2$), et $échelon_\theta(u,v)$ est un échelon orienté selon la direction θ dans le plan 2D (u,v) (comme indiqué sur la figure B.2).

Pour aboutir au filtre Fan, deux filtres Step sont nécessaires avec des angles θ_1 et θ_2 ($\theta_1 > \theta_2$). L'équation du filtre Fan est la suivante :

$$Fan_{\theta_1,\theta_2}(u,v) = |Step_{\theta_1}(u,v) - Step_{\theta_2}(u,v)|. \tag{B.4}$$



(a) (b) (c)

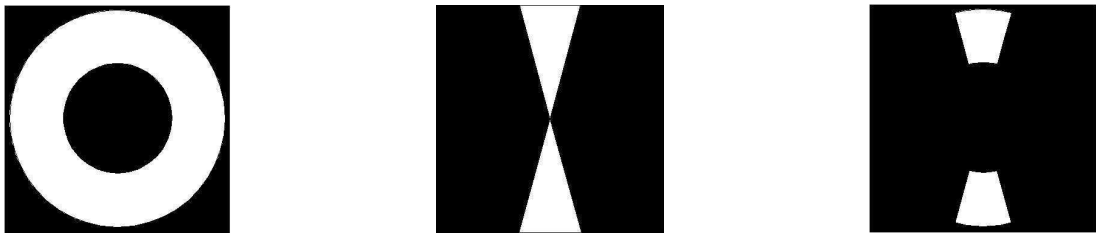
Figure B.2 – Construction d'un filtre Fan : les 2 filtres Step utilisés en (a) et en (b), et en (c) le filtre Fan résultant (correspondant à l'orientation 4 de la DCP).

B.0.0.3 Les filtres Cortex

Un filtre Cortex est le produit d'un filtre DoM et d'un filtre Fan :

$$Cortex_{\rho,\theta}(u,v) = DoM_\rho(u,v).Fan_\theta(u,v). \tag{B.5}$$

Il va donc isoler les fréquences spatiales de l'image qui correspondent à la bande de fréquences ρ et à la gamme d'orientations θ. La construction d'un filtre Cortex est illustrée sur la figure B.3.



(a) (b) (c)

Figure B.3 – Construction d'un filtre Cortex : le filtre DoM utilisé en (a), le filtre Fan utilisé en (b), et en (c) le filtre Cortex résultant.

Annexe C

Résultats par observateur des métriques de qualité d'images basées saillance

Les tableaux suivants (C.1,C.2,C.3,C.4,C.5,C.6,C.7,C.8) présentent les résultats des métriques d'images basées saillance pour huit des observateurs (cf. section 8.4.2.1).

Carte de distorsions	Cumul		FD (saillance)		FN (saillance)	
	Saillance	w_i	CC	RMSE	CC	RMSE
DiffAbs	***Aucune***	***1***	***0.646***	***0.924***	***0.646***	***0.924***
	Réelle	w_1	0.560	1.002	0.571	0.994
		w_2	0.649	0.920	0.649	0.921
		w_3	0.599	0.969	0.644	0.926
		w_4	0.609	0.960	0.660	0.909
	Permutée	w_1	0.469	1.069	0.456	1.077
		w_2	0.645	0.925	0.643	0.926
		w_3	0.530	1.026	0.550	1.010
		w_4	0.546	1.014	0.645	0.925
SSIM	***Aucune***	***1***	***0.741***	***0.814***	***0.741***	***0.814***
	Réelle	w_1	0.711	0.851	0.712	0.850
		w_2	0.741	0.813	0.741	0.813
		w_3	0.711	0.851	0.712	0.850
		w_4	0.732	0.824	0.741	0.813
	Permutée	w_1	0.702	0.863	0.700	0.865
		w_2	0.741	0.813	0.741	0.813
		w_3	0.702	0.863	0.700	0.865
		w_4	0.733	0.824	0.741	0.814

TABLE C.1 – Observateur n°1 : Comparaison des performances des métriques de qualité, lorsque que la différence absolue et la SSIM sont utilisées pour calculer les cartes de distorsion.

Carte de distorsions	Cumul		FD (saillance)		FN (saillance)	
	Saillance	w_i	CC	RMSE	CC	RMSE
DiffAbs	***Aucune***	***1***	***0.752***	***0.749***	***0.752***	***0.749***
	Réelle	w_1	0.561	0.940	0.584	0.923
		w_2	0.751	0.751	0.752	0.749
		w_3	0.709	0.801	0.704	0.807
		w_4	0.723	0.785	0.762	0.736
	Permutée	w_1	0.484	0.994	0.502	0.983
		w_2	0.749	0.753	0.750	0.752
		w_3	0.660	0.854	0.679	0.834
		w_4	0.680	0.833	0.751	0.751
SSIM	***Aucune***	***1***	***0.792***	***0.695***	***0.792***	***0.695***
	Réelle	w_1	0.768	0.728	0.771	0.725
		w_2	0.792	0.695	0.793	0.695
		w_3	0.768	0.728	0.771	0.725
		w_4	0.783	0.708	0.793	0.695
	Permutée	w_1	0.782	0.710	0.783	0.708
		w_2	0.793	0.694	0.793	0.694
		w_3	0.782	0.710	0.783	0.708
		w_4	0.796	0.690	0.794	0.692

Table C.2 – Observateur n°2 : Comparaison des performances des métriques de qualité, lorsque que la différence absolue et la SSIM sont utilisées pour calculer les cartes de distorsion.

Carte de distorsions	Cumul		FD (saillance)		FN (saillance)	
	Saillance	w_i	CC	RMSE	CC	RMSE
DiffAbs	***Aucune***	***1***	***0.704***	***1.057***	***0.704***	***1.057***
	Réelle	w_1	0.579	1.214	0.636	1.148
		w_2	0.709	1.050	0.711	1.048
		w_3	0.644	1.140	0.743	0.997
		w_4	0.662	1.116	0.737	1.007
	Permutée	w_1	0.554	1.239	0.580	1.213
		w_2	0.704	1.057	0.704	1.058
		w_3	0.601	1.191	0.709	1.050
		w_4	0.625	1.163	0.716	1.041
SSIM	***Aucune***	***1***	***0.793***	***0.916***	***0.793***	***0.916***
	Réelle	w_1	0.783	0.932	0.780	0.939
		w_2	0.794	0.915	0.793	0.915
		w_3	0.783	0.932	0.780	0.939
		w_4	0.793	0.914	0.794	0.914
	Permutée	w_1	0.744	1.000	0.767	0.961
		w_2	0.793	0.915	0.793	0.915
		w_3	0.744	1.000	0.767	0.961
		w_4	0.776	0.946	0.793	0.914

Table C.3 – Observateur n°3 : Comparaison des performances des métriques de qualité, lorsque que la différence absolue et la SSIM sont utilisées pour calculer les cartes de distorsion.

Carte de distorsions	Cumul		FD (saillance)		FN (saillance)	
	Saillance	w_i	CC	RMSE	CC	RMSE
DiffAbs	***Aucune***	***1***	***0.677***	***1.035***	***0.677***	***1.035***
	Réelle	w_1	0.565	1.162	0.570	1.156
		w_2	0.675	1.038	0.675	1.038
		w_3	0.607	1.118	0.668	1.048
		w_4	0.601	1.125	0.682	1.029
	Permutée	w_1	0.611	1.115	0.607	1.118
		w_2	0.676	1.038	0.675	1.039
		w_3	0.640	1.081	0.713	0.988
		w_4	0.633	1.089	0.685	1.025
SSIM	***Aucune***	***1***	***0.781***	***0.882***	***0.781***	***0.882***
	Réelle	w_1	0.739	0.954	0.744	0.947
		w_2	0.780	0.886	0.780	0.886
		w_3	0.739	0.954	0.744	0.947
		w_4	0.765	0.911	0.780	0.884
	Permutée	w_1	0.763	0.916	0.770	0.904
		w_2	0.781	0.884	0.781	0.884
		w_3	0.763	0.916	0.770	0.904
		w_4	0.778	0.888	0.782	0.880

Table C.4 – Observateur n°4 : Comparaison des performances des métriques de qualité, lorsque que la différence absolue et la SSIM sont utilisées pour calculer les cartes de distorsion.

Carte de distorsions	Cumul		FD (saillance)		FN (saillance)	
	Saillance	w_i	CC	RMSE	CC	RMSE
DiffAbs	***Aucune***	***1***	***0.670***	***0.939***	***0.670***	***0.939***
	Réelle	w_1	0.458	1.124	0.440	1.136
		w_2	0.670	0.939	0.669	0.940
		w_3	0.595	1.016	0.603	1.009
		w_4	0.608	1.004	0.680	0.927
	Permutée	w_1	0.303	1.205	0.276	1.216
		w_2	0.668	0.942	0.667	0.943
		w_3	0.454	1.127	0.494	1.100
		w_4	0.496	1.098	0.668	0.941
SSIM	***Aucune***	***1***	***0.754***	***0.833***	***0.754***	***0.833***
	Réelle	w_1	0.730	0.865	0.736	0.858
		w_2	0.755	0.832	0.754	0.832
		w_3	0.730	0.865	0.736	0.858
		w_4	0.754	0.832	0.755	0.832
	Permutée	w_1	0.616	0.997	0.651	0.961
		w_2	0.754	0.833	0.754	0.833
		w_3	0.616	0.997	0.651	0.961
		w_4	0.731	0.865	0.754	0.833

Table C.5 – Observateur n°5 : Comparaison des performances des métriques de qualité, lorsque que la différence absolue et la SSIM sont utilisées pour calculer les cartes de distorsion.

Carte de distorsions	Cumul		FD (saillance)		FN (saillance)	
	Saillance	w_i	CC	RMSE	CC	RMSE
DiffAbs	***Aucune***	***1***	***0.685***	***0.840***	***0.685***	***0.840***
	Réelle	w_1	0.522	0.983	0.525	0.980
		w_2	0.684	0.840	0.684	0.840
		w_3	0.696	0.827	0.714	0.806
		w_4	0.705	0.817	0.706	0.815
	Permutée	w_1	0.412	1.050	0.448	1.030
		w_2	0.680	0.844	0.680	0.845
		w_3	0.638	0.887	0.634	0.891
		w_4	0.648	0.878	0.688	0.836
SSIM	***Aucune***	***1***	***0.751***	***0.762***	***0.751***	***0.762***
	Réelle	w_1	0.723	0.798	0.734	0.784
		w_2	0.752	0.762	0.752	0.761
		w_3	0.723	0.798	0.734	0.784
		w_4	0.740	0.776	0.752	0.761
	Permutée	w_1	0.699	0.827	0.705	0.819
		w_2	0.752	0.762	0.751	0.762
		w_3	0.699	0.827	0.705	0.819
		w_4	0.737	0.780	0.752	0.761

Table C.6 – Observateur n°6 : Comparaison des performances des métriques de qualité, lorsque que la différence absolue et la SSIM sont utilisées pour calculer les cartes de distorsion.

Carte de distorsions	Cumul		FD (saillance)		FN (saillance)	
	Saillance	w_i	CC	RMSE	CC	RMSE
DiffAbs	***Aucune***	***1***	***0.734***	***1.015***	***0.734***	***1.015***
	Réelle	w_1	0.599	1.195	0.644	1.142
		w_2	0.735	1.012	0.738	1.008
		w_3	0.703	1.061	0.745	0.997
		w_4	0.715	1.044	0.752	0.984
	Permutée	w_1	0.499	1.294	0.531	1.265
		w_2	0.732	1.017	0.732	1.017
		w_3	0.629	1.160	0.662	1.119
		w_4	0.649	1.136	0.736	1.011
SSIM	***Aucune***	***1***	***0.804***	***0.890***	***0.804***	***0.890***
	Réelle	w_1	0.767	0.959	0.765	0.963
		w_2	0.804	0.890	0.804	0.891
		w_3	0.767	0.959	0.765	0.963
		w_4	0.787	0.922	0.803	0.893
	Permutée	w_1	0.786	0.926	0.799	0.899
		w_2	0.805	0.888	0.805	0.888
		w_3	0.786	0.926	0.799	0.899
		w_4	0.816	0.863	0.807	0.883

Table C.7 – Observateur n°7 : Comparaison des performances des métriques de qualité, lorsque que la différence absolue et la SSIM sont utilisées pour calculer les cartes de distorsion.

Carte de distorsions	Cumul		FD (saillance)		FN (saillance)	
	Saillance	w_i	CC	RMSE	CC	RMSE
DiffAbs	***Aucune***	***1***	***0.665***	***0.974***	***0.665***	***0.974***
	Réelle	w_1	0.570	1.071	0.574	1.068
		w_2	0.666	0.973	0.665	0.973
		w_3	0.628	1.015	0.608	1.036
		w_4	0.640	1.002	0.668	0.970
	Permutée	w_1	0.554	1.085	0.584	1.058
		w_2	0.664	0.974	0.664	0.974
		w_3	0.612	1.031	0.618	1.025
		w_4	0.624	1.018	0.667	0.972
SSIM	***Aucune***	***1***	***0.734***	***0.890***	***0.734***	***0.890***
	Réelle	w_1	0.722	0.904	0.719	0.909
		w_2	0.735	0.889	0.735	0.889
		w_3	0.722	0.904	0.719	0.909
		w_4	0.731	0.893	0.735	0.888
	Permutée	w_1	0.730	0.895	0.729	0.896
		w_2	0.735	0.889	0.735	0.888
		w_3	0.730	0.895	0.729	0.896
		w_4	0.735	0.888	0.735	0.888

Table C.8 – Observateur n°8 : Comparaison des performances des métriques de qualité, lorsque que la différence absolue et la SSIM sont utilisées pour calculer les cartes de distorsion.

Bibliographie

[Ahumada 93] A. J. Ahumada & C. H. Null. *Image quality : A multidimensional problem.* In A. B. Watson, editeur, Digital Images and Human Vision, pages 141–148. MIT Press, 1993.

[Alpert 97] T. Alpert & J.-P. Evain. *Évaluation subjective de la qualité : Les méthodes SSCQE et DSCQE = Subjective evaluation of the quality : The SSCQE and DSCQE methods.* UER-revue technique, vol. 271, pages 12–20, 1997.

[Antonini 92] M. Antonini, P. Barlaud & I. Daubechies. *Image coding using wavelet transform.* IEEE Transactions on Image Processing, vol. 1, no. 2, pages 205–220, 1992.

[Ballard 91] D. Ballard. *Animate vision.* Artificial intelligence, vol. 86, pages 48–57, 1991.

[Barland 06] R. Barland & A. Saadane. *Blind Quality Metric Using a Perceptual Importance Map for JPEG-2000 Compressed Images.* In Proceedings of IEEE International Conference on Image Processing, pages 2941–2944, 2006.

[Barten 99] P. G. J. Barten. *Contrast and sensibility of the human eye and its effects on image quality.* In Proceedings of SPIE Optical Engineering Press, 1999.

[Barten 04] P. G. Barten. *Formula for contrast sensitivity of the human eye.* In Proceedings of SPIE Image quality and system performance, 2004.

[Bedat 98] L. Bedat. *Aspects psychovisuels de la perception des couleurs. Application au codage d'images couleurs fixes avec compression de l'information.* Thèse de doctorat, Université de Nantes, IRESTE, 1998.

[Bekkat 99] N. Bekkat. *Critère objectif de qualité subjective d'images monochromes. Conception du modèle et validation expérimentale.* Thèse de doctorat, Université de Nantes, IRESTE, 1999.

[Bodmann 80] H. Bodmann, P. Haubner & A. Marsden. *A unified relationship between brightness and luminance.* CIE Proceedings of Kyoto Session, pages 99–102, 1980.

[Braddick 78] O. Braddick, F. W. Campbell & J. Atkinson. *Channels in vision : Basic aspects.* In H.-L. Teuber R. Held H. W. Leibowitz, editeur, Perception, vol. 8 of Handbook of Sensory Physiology, pages 3–38. Springer-Verlag, 1978.

[Bradley 99] A. P. Bradley. *A Wavelet Visible Difference Predictor*. IEEE Transactions On Image Processing, vol. 8, no. 5, pages 717–730, 1999.

[Braun 90] J. Braun & D. Sagi. *Vision outside the focus of attention*. Perception and Psychophysics, vol. 48, pages 45–58, 1990.

[Burbeck 80] C. A. Burbeck & D. H. Kelly. *Spatiotemporal characteristics of visual mechanisms : Excitatory-inhibitory model*. Journal of the Optical Society of America, vol. 70, no. 9, pages 1121–1126, 1980.

[Burt 83a] P. J. Burt & E. H. Adelson. *The laplacian pyramid as a compact image code*. Transactions on Communications, vol. 31, pages 532–540, 1983.

[Burt 83b] P. J. Burt & E. H. Adelson. *The Laplacian pyramid as a compact image code*. IEEE Transactions on Communications, vol. 31, pages 532–540, 1983.

[Campbell 68] F. W. Campbell & J. G. Robson. *Application of Fourier analysis to the visibility of gratings*. Journal of Physiology, vol. 197, pages 551–566, 1968.

[Carnec 04] M. Carnec. *Critères de qualité d'images couleur avec référence réduite perceptuelle générique*. Thèse de doctorat, Université de Nantes, École Centrale de Nantes et École des Mines de Nantes, 218 pages, 2004.

[CCIR 94] CCIR. *Projet de révision de la recommandation 500-4 : Méthode d'évaluation subjective de la qualité des images de télévision*. Document commission d'études du CCIR, vol. 11/BL/51-F, 1990-1994.

[Chandler 07] D. M. Chandler & S. S. Hemami. *VSNR : A wavelet-based visual signal-to-noise ratio for natural images*. IEEE Transactions on Image Processing, vol. 16, no. 9, pages 2284–2298, 2007.

[Christopoulos 00] C. Christopoulos, A. Skodras & T. Ebrahimi. *The JPEG2000 Still Image Coding : An Overview*. IEEE Transactions on Consumer Electronics, vol. 46, no. 4, pages 1103–1127, 2000.

[Corriveau 99] P. Corriveau, C. Gojmerac, B. Hughes & L. Stelmach. *All Subjectives Scales are Not Created Equal : The effects of Context on Differents Scales*. Signal Processing, vol. 77, pages 1–9, 1999.

[Daly 92] S. Daly. *The visible difference predictor : An algorithm for the assessment of image fidelity*. In Proceedings of SPIE Human Vision, Visual Processing, and Digital Display III, volume 1666, pages 2–15, 1992.

[Daly 93] S. Daly. *The Visible Differences Predictor : An Algorithm of Image Fidelity*. Digital Images and Human Vision, pages 179–206, 1993.

[De Lange 58] H. De Lange. *Research into the dynamic nature of the human fovea-cortex systems with intermittent and modulated light : I. Attenuation characteristics with white and colored light.* Journal of the Optical Society of America, vol. 48, pages 777–784, 1958.

[De Valois 92] R. De Valois & K. K. De Valois. *A multi stage color model.* Vision Research, vol. 33, no. 8, pages 1035–1035, 1992.

[Faugeras 76] O. D. Faugeras. *Digital color image processing and psychophysics within the framework of human visual system.* Thèse de doctorat, University of utah, 1976.

[Flanagan 90] P. Flanagan, P. Cavanagh & O. E. Favreau. *Independent orientation selective mechanisms for cardinal directions of color space.* Vision Research, vol. 30, no. 5, pages 769–778, 1990.

[Foley 94] J. M. Foley. *Human luminance pattern mechanisms : Masking experiments require a new model.* Journal of the Optical Society of America, vol. 11, pages 1710–1719, 1994.

[Fredericksen 98] R. E. Fredericksen & R. F. Hess. *Estimating multiple temporal mechanisms in human vision.* Vision Research, vol. 38, no. 7, pages 1023–1040, 1998.

[Freeman 91] W. T. Freeman & E. H. Adelson. *The design and use of steerable filters.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, pages 891–906, 1991.

[Gaubatz 05] M. D. Gaubatz, D. M. Chandler & S. S. Hemami. *Spatial Quantization via Local Texture Masking.* In Proceedings of SPIE Human Vision and Electronic Imaging X, volume 5666, pages 95–106, 2005.

[Hammett 92] S. T. Hammett & A. T. Smith. *Two temporal channels or three ? A re-evaluation.* Vision Research, vol. 32, no. 2, pages 285–291, 1992.

[Heeger 92] D. J. Heeger. *Normalisation of cells responses in cat striates cortex.* Visual Neuroscience, vol. 9, pages 181–198, 1992.

[Heeger 95] D. J. Heeger & T. C. Teo. *A model of perceptual image fidelity.* In Proceedings of IEEE International Conference on Image Processing, pages 343–345, 1995.

[Henderson 99] J.M. Henderson, P.A. Weeks & A. Hollingworth. *The effects of semantic consistency on eye movements during complex scene viewing.* Journal of Experimental Psychology. Human Perception and Performance, vol. 25, no. 1, pages 210–228, 1999.

[Hess 92] R. F. Hess & R. J. Snowden. *Temporal properties of human visual filters : Number, shapes and spatial covariation.* Vision Research, vol. 32, no. 1, pages 47–59, 1992.

[Hoffman 98] J. E. Hoffman. *Visual attention and eye movements*. In H. Pashler, editeur, Hove, UK : Psychology Press, pages 119–154. 1998.

[Kelly 79a] D. H. Kelly. *Motion and vision : I. Stabilized images of stationary gratings*. Journal of the Optical Society of America, vol. 69, no. 9, pages 1266–1274, 1979.

[Kelly 79b] D. H. Kelly. *Motion and vision : II. Stabilized spatio-temporal threshold surface*. Journal of the Optical Society of America, vol. 69, no. 10, pages 1340–1349, 1979.

[Kelly 83] D. H. Kelly. *Spatiotemporal variation of chromatic and achromatic contrast thresholds*. Journal of the Optical Society of America, vol. 73, no. 6, pages 742–750, 1983.

[Klein 99] R. Klein & W.J. Mac Innes. *Inhibition of return is a foraging facilitator in visual search*. Psychological Science, vol. 10, pages 346–352, 1999.

[Koenderink 79] J. J. Koenderink & A. J. van Doorn. *Spatiotemporal contrast detection threshold surface is bimodal*. Optics Letters, vol. 4, no. 1, pages 32–34, 1979.

[Krauskopf 82] J. Krauskopf, D. R. Williams & D. W. Heeley. *Cardinal direction of color space*. Vision Research, vol. 22, pages 1123–1131, 1982.

[Larson 08] E. C. Larson, C. T. Vu & D. M. Chandler. *Can visual fixation patterns improve image fidelity assessment ?* In Proceedings of IEEE International Conference on Image Processing, volume 3, pages 2572–2575, 2008.

[Le Callet 01] P. Le Callet. *Critères objectifs avec référence de qualité visuelle des images couleur*. Thèse de doctorat, Ecole Polytechnique de l'Université de Nantes, 216 pages, 2001.

[Le Meur 05] O. Le Meur. *Attention sélective en visualisation d'images fixes et animées affichées sur écran : Modèles et évaluations de performances - Applications*. Thèse de doctorat, Ecole Polytechnique de l'Université de Nantes, 204 pages, 2005.

[Legge 80] G. E. Legge & J. M. Foley. *Contrast masking in human vision*. Journal of the Optical Society of America, vol. 70, pages 1458–1471, 1980.

[Lindh 96] P. Lindh & van den Branden Lambrecht. *Efficient spatio-temporal decomposition for perceptual processing ofvideo sequences*. In Proceedings of IEEE International Conference on Image Processing, volume 3, pages 331–334, 1996.

[Lu 04] Z.A. Lu, X.K. Yang, W.S. Lin, E.P. Ong & S. Yao. *Modelling visual attention and motion effect for visual quality evaluation*. In Proceedings of IEEE International Conference on Image Processing, volume 4, pages 2311–2314, 2004.

[Lu 05] Z.A. Lu, W.S. Lin, X.K. Yang, E.P. Ong & S. Yao. *Modeling Visual Attention's Modulatory Aftereffects on Visual Sensitivity and Quality Evaluation*. IEEE Transactions on Image Processing, vol. 14, no. 11, pages 1928–1942, 2005.

[Lubin 93] J. Lubin. *The use of psychophysical data and models in the analysis of display system performance.* Digital Images and Human Vision (A. B. Watson, ed.), pages 163–178, 1993.

[Lubin 95] J. Lubin. *A visual discrimination model for image system design and evaluation.* Visual Models for Target Detection and Recognition, E. Peli, ed., pages 207–220, 1995.

[Lubin 97] J. Lubin & D. Fibush. *Sarnoff JND vision model.* T1A1.5 Working Group Document ♯97-612, ANSI T1 Standards Committee, 1997.

[Mandler 84] M. B. Mandler & W. Makous. *A three-channel model of temporal frequency perception.* Vision Research, vol. 24, no. 12, pages 1881–1887, 1984.

[Mannan 97] S.K. Mannan, K.H. Ruddock & D. S. Wooding. *Fixation sequences made during visual examination of briefly presented 2d images.* Spatial Vision, vol. 11, no. 2, pages 157–178, 1997.

[Mannos 74] J. L. Mannos & D. J. Sakrison. *The effects of a visual fidely criterion on the encoding of images.* IEEE Transactions of Information Theory, vol. 20, no. 4, pages 525–535, 1974.

[Masry 04] M.A. Masry & S.S. Hemami. *A metric for continuous quality evaluation of compressed video with severe distortions.* Signal processing. Image communication, vol. 19, no. 2, pages 133–146, 2004.

[Milanese 93] R. Milanese. *Detecting salient regions in an image : from biological evidence to computer implementation.* Thèse de doctorat, Université de Genève, 1993.

[Moon 44] P. Moon & D. E. Spencer. *Visual data applied to lighting design.* Journal of the Optical Society of America, vol. 34, pages 605–617, October 1944.

[Nadenau 00] M. Nadenau. *Integration of Human Color Vision Models into High Quality Image Compression.* Thèse de doctorat, École Polytechnique Fédérale de Lausanne, 2000.

[Neisser 67] U. Neisser. Cognitive psychology. New York, Appleton-Century-Crofts, 1967.

[Newhall 43] S. M. Newhall, D. Nickerson & D. B. Judd. *Final report of the O.S.A. subcommittee on the spacing of the Munsell colors.* J. Opt. Soc. Am., vol. 33, no. 7, pages 385–418, 1943.

[Nisbett 05] R.E. Nisbett, A. Norenzayan, E.E. Smith & B.J. Kim. *Cultural preferences for formal versus intuitive reasoning.* Cognitive Science, vol. 7, 2005.

[Odobez 95] J.M. Odobez & P. Bouthemy. *Robust multiresolution estimation of parametric motion models.* Journal of Visual Communication and Image Representation, vol. 6, no. 4, pages 348–365, 1995.

[Osberger 98] W. Osberger, N. Bergmann & A. Maeder. *An Automatic Image Quality Assessment Technique Incorporating Higher Level Perceptual Factors.* In Proceedings of IEEE International Conference on Image Processing, pages 414–418, 1998.

[Parkhurst 04] D.J. Parkhurst & E. Niebur. *Texture contrast attracts overt visual attention in natural scenes.* European Journal of Neuroscience, vol. 19, pages 783–789, 2004.

[Peli 90] E. Peli. *Contrast in complex images.* Journal of Optical Society of America, vol. 7, pages 2032–2040, 1990.

[Peli 93] E. Peli, L. E. Arend, G. M. Young & R. B. Goldstein. *Contrast sensitivity to patch stimuli : effects of spatial bandwith and tmeporal presentation.* Spatial Vision, vol. 7, no. 1, pages 1–14, 1993.

[Pinson 04] M.H. Pinson & S. Wolf. *A new standardized method for objectively measuring video quality.* IEEE Transactions on Broadcasting, vol. 50, no. 3, pages 312–322, Sept. 2004.

[Posner 80] M.I. Posner. *Orienting of attention.* Quaterly Journal of Experimental Psychology, vol. 32, pages 3–25, 1980.

[Posner 84] M. I. Posner & Y. Cohen. *Components of visual orienting.* In D.G. Bouwhuis H. Bouma, editeur, Attention and performance X, pages 531–556. 1984.

[ANSI T1.801.03 03] ANSI T1.801.03. *American National Standard for Telecommunications - Digital Transport of One-Way Video Signals - Parameters for Objective Performance Assessment.* American National Standard Institute, 2003.

[ITU-R Rec. BT.470-6 98] ITU-R Rec. BT.470-6. *Conventional Television Systems.* Recommendations of the ITU, Radiocommunication Sector, 1998.

[ITU-R Rec. BT.470-7 98] ITU-R Rec. BT.470-7. *Conventional Analog Television Systems.* Recommendations of the ITU, Radiocommunication Sector, 1998.

[ITU-R Rec. BT.500-10 00] ITU-R Rec. BT.500-10. *Methodology for the subjective assessment of the quality of television pictures.* Recommendations of the ITU, Radiocommunication Sector, 2000.

[ITU-R Rec. BT.601-5 95] ITU-R Rec. BT.601-5. *Studio Encoding Parameters of Digital Television for Standard 4 :3 and Wide-Screen 16 :9 Aspect Ratios.* Recommendations of the ITU, Radiocommunication Sector, 1995.

[IUT-R Rec. BT.1683 04] IUT-R Rec. BT.1683. *Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference.* Recommendations of the ITU, Radiocommunication Sector, 2004.

[IUT-T Rec. J.144 04] IUT-T Rec. J.144. *Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference.* Recommendations of the ITU, Telecommunication Standardization Sector, 2004.

[Reinagel 99] P. Reinagel & A.M. Zador. *Natural scene statistics at the center of gaze.* Network : Computation in Neural Systems, vol. 10, pages 341–350, 1999.

[Robson 66] J. G. Robson. *Spatial and temporal contrast-sensitivity functions of the visual system.* Journal of the Optical Society of America, vol. 56, pages 1141–1142, 1966.

[Sallio 77] P. Sallio, F. Kretz & J.P. de la Tribonnière. *Méthodologie des tests subjectifs visuels : Tests de dégradation, de qualité et de détection.* Rapport CCETT CTN/T/13/76, 1977.

[Salvucci 99] D.D. Salvucci. *Mapping eye movements to cognitive processes.* Thèse de doctorat, Carnegie Mellon University, 1999.

[Salvucci 00] D. D. Salvucci & J. H. Goldberg. *Identifying fixations and saccades in eye-tracking protocols.* In ETRA '00 : Proceedings of the 2000 symposium on Eye tracking research & applications, pages 71–78, New York, NY, USA, 2000. ACM.

[Sazzad 07] Z. M. P. Sazzad, Y. Kawayoke & Y. Horita. *Spatial features based no reference image quality assessment for JPEG2000.* In Proceedings of IEEE International Conference on Image Processing, 2007.

[Seyler 59] A. J. Seyler & Z. L. Budrikis. *Measurement of temporal adaptation to spatial detail vision.* Nature, vol. 184, pages 1215–1217, 1959.

[Seyler 65] A. J. Seyler & Z. L. Budrikis. *Detail perception after scene changes in television image presentations.* IEEE Transaction on Information Theory, vol. 11, no. 1, pages 31–43, 1965.

[Sheikh 06a] H. R. Sheikh & A. C. Bovik. *Image information and visual quality.* IEEE Transactions on Image Processing, vol. 15, no. 2, pages 430–444, 2006.

[Sheikh 06b] H. R. Sheikh, M. F. Sabir & A. C. Bovik. *A statistical evaluation of recent full reference image quality assessment algorithms.* IEEE Transactions on Image Processing, vol. 15, no. 11, pages 3440–3451, 2006.

[Sénane 96] H. Sénane. *Représentation d'images en sous-bandes visuelles. Applications au codage d'images de télévision sans défauts visibles.* Thèse de doctorat, Université de Nantes, 1996.

[Swift 83] D. J. Swift & R. A. Smith. *Spatial frequency masking and Weber's Law.* Vision Research, vol. 23, pages 495–506, 1983.

[Tam 95] W. J. Tam, L. B. Stelmach, L. Wang, D. Lauzon & P. Gray. *Visual masking at video scene cuts*. In Proceedings of SPIE, Human Vision, Visual Processing, and Digital Display VI, volume 2411, pages 111–119, 1995.

[Tan 98] K. T. Tan, M. Ghanbari & D. E. Pearson. *An objective measurement tool for MPEG video quality*. Signal Processing, vol. 70, no. 3, pages 279–294, 1998.

[Teo 94] P. C. Teo & D. J. Heeger. *Perceptual image distortion*. In Proceedings of SPIE, volume 2179, pages 127–141, 1994.

[Tourancheau 08] S. Tourancheau, F. Autrusseau, Z.M.P. Sazzad & Y. Horita. *Impact of subjective dataset on the performance of image quality metrics*. In Proceedings of IEEE International Conference on Image Processing, 2008.

[Treisman 80] A. Treisman & G. Gelade. *A feature integration theory of attention*. Cognitive Psychology, vol. 12, pages 97–136, 1980.

[Tsotsos 95] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y.H. Lai, N. Davis & F. Nuflo. *Modelling visual attention via selective tuning*. Artificial intelligence, vol. 78, pages 507–545, 1995.

[Turvey 73] M. T. Turvey. *On peripheral and central processes in vision : Inferences from an information-processing analysis of masking with patterned stimuli*. Psychological Review, vol. 80, pages 1–52, 1973.

[van den Branden Lambrecht 96a] C. J. van den Branden Lambrecht. *Color Moving Pictures Quality Metric*. In Proceedings of IEEE International Conference on Image Processing, pages 885–888, 1996.

[van den Branden Lambrecht 96b] C. J. van den Branden Lambrecht. *Perceptual models and architectures for video coding applications*. Thèse de doctorat, École polytechnique fédérale de Lausanne, 1996.

[van den Branden Lambrecht 96c] C. J. van den Branden Lambrecht & O. Verscheure. *Perceptual quality measure using a spatio-temporal model of the human visual system*. In Proceedings of SPIE Digital Video Compression : Algorithms and Technologies, volume 2668, pages 450–461, 1996.

[VQEG 00] VQEG. *Final report from the video quality experts group on the validation of objective models of video quality assessment*, 2000. http ://www.vqeg.org/.

[VQEG 03] VQEG. *Final report from the video quality experts group on the validation of objective models of video quality assessment. Phase 2.*, 2003. http ://www.vqeg.org/.

[Vu 08] C. T. Vu, E. C. Larson & D. M. Chandler. *Visual fixation patterns when judging image quality : Effects of distortion type, amount, and subject experience*. IEEE Southwest Symposium on Image Analysis and Interpretation, pages 73–76, 2008.

[Vuori 04] T. Vuori, M. Olkkonen, M. Pölönen, A. Siren & J. Häkkinen. *Can eye movements be quantitatively applied to image quality studies ?* In NordiCHI '04 : Proceedings of the third Nordic conference on Human-computer interaction, pages 335–338. ACM, 2004.

[Vuori 06] T. Vuori & M. Olkkonen. *The effect of image sharpness on quantitative eye movement data and on image quality evaluation while viewing natural images.* In Proceedings of SPIE International Society for Optical Engineering, 2006.

[Wang 01] Z. Wang. *Rate scalable Foveated image and video communications.* Thèse de doctorat, Dept. Elect. Comput. Eng., Univ. Texas at Austin, 2001.

[Wang 02a] Z. Wang & A. C. Bovik. *A universal image quality index.* IEEE Signal Processing Letters, vol. 9, pages 81–84, 2002.

[Wang 02b] Z. Wang, H. R. Sheikh & A. C. Bovik. *No-reference perceptual quality assessment of JPEG compressed images.* In Proceedings of IEEE International Conference on Image Processing, volume 1, pages 477–480, 2002.

[Wang 03] Z. Wang, E. P. Simoncelli & A. C. Bovik. *Multi-scale structural similarity for image quality assessment.* In Proceedings of IEEE Asilomar Conference on Signals, Systems, and Computers, volume 2, pages 1398–1402, 2003.

[Wang 04a] Z. Wang, A. C. Bovik, H. R. Sheikh & E. P. Simoncelli. *Image Quality Assessment : From Error Visibility to Structural Similarity.* IEEE Transactions on Image Processing, vol. 13, pages 600–612, 2004.

[Wang 04b] Z. Wang, L. Lu & A. C. Bovik. *Video quality assessment based on structural distortion measurement.* Signal Processing : Image Communication, special issue on objective video quality metrics, vol. 19, pages 121–132, 2004.

[Watson 86] A. B. Watson. *Temporal sensitivity.* In J. P. Thomas K. R. Boff L. Kaufman, editeur, Handbook of Perception and Human Performance, volume 1, chapitre 6. John Wiley & Sons, 1986.

[Watson 87] A. B. Watson. *The cortex transform : Rapid computation of simulated neural images.* Computer Vision, Graphics, And Image Processing, vol. 39, pages 311–327, 1987.

[Watson 93] A. B. Watson. *DCTune : A technique for visual optimization of DCT quantization matrices for individual images.* Society for Information Display Digest of Technical Papers, vol. XXIV, pages 946–949, 1993.

[Watson 97a] A. B. Watson. *Model of visual contrast gain control and pattern masking.* Journal of Optical Society of America, vol. 14, no. 9, pages 2379–2391, 1997.

[Watson 97b] A. B. Watson, R. Borthwick & M. Taylor. *Image quality and entropy masking.* In Proceedings of SPIE Human Vision, Visual Processing, and Digital Display VIII, volume 3016, San Jose, CA, USA, 1997.

[Watson 97c] A. B. Watson, G. Y. Yang, J. A. Solomon & J. Villasenor. *Visibility of Wavelet Quantization Noise.* IEEE Transactions on Image Processing, vol. 6, no. 8, pages 1164–1175, 1997.

[Watson 98] A. B. Watson. *Toward a perceptual video quality metric.* In Proceedings of SPIE Human Vision and Electronic Imaging III, volume 3299, pages 139–147, 1998.

[Watson 01] A. B. Watson, J. Hu & J. F. III. McGowan. *DVQ : A digital video quality metric based on human vision.* Journal of Electronic Imaging, vol. 10, no. 1, pages 20–29, 2001.

[Webster 90] M. A. Webster, K. K. De Valois & E. Switkes. *Orientation and spatial frequency discrimination for luminance and chromatic gratings.* Journal of the Optical Society of America, vol. 7, no. 6, pages 1034–1049, 1990.

[Winkler 99] S. Winkler. *A Perceputal Distortion Metric for Digital Color Video.* In Proceedings of SPIE Human vision and electronic imaging IV, volume 3644, pages 175–184, 1999.

[Winkler 00] S. Winkler. *Vision models and quality metrics for image processing applications.* Thèse de doctorat, École Polytechnique Fédérale de Lausanne, 2000.

[Wolfe 04] J.M. Wolfe & T.S. Horowitz. *What attributes guide the deployement of visual attention and how do they do it ?* Nature Reviews Neuroscience, vol. 5, 2004.

[Wooding 02] D. S. Wooding. *Eye movements of large population : II. Deriving regions of interest, coverage, and similarity using fixation maps.* Behavior Research Methods, Instruments and Computers, vol. 34, no. 4, pages 509–517, 2002.

[Yantis 96] S. Yantis & J. Jonidas. *Attentional capture by abrupt onsets and selective attention : evidence from visual search.* Journal of Experimental Psychology. Human Perception and Performance, vol. 20, pages 1505–1513, 1996.

[Zeng 02] W. Zeng, S. Daly & S. Lei. *An overview of the visual optimization tools in JPEG 2000.* Signal Processing : Image Communication, vol. 17, pages 85–104, 2002.

Publications liées à la thèse

Publications dans des revues internationales avec comité de lecture

A. Ninassi, O. Le Meur, P. Le Callet and D. Barba,*Considering temporal variations of spatial visual distortions in video quality assessment*, IEEE Journal Of Selected Topics In Signal Processing : Special Issue On Visual Media Quality Assessment (à paraître).

Publications dans des colloques internationaux avec actes et comité de lecture

A. Ninassi, O. Le Meur, P. Le Callet and D. Barba, *Which Semi-Local Visual Masking Model For Wavelet Based Image Quality Metric ?*, ICIP 2008, San Diego, California, USA, 2008.

A. Ninassi, O. Le Meur, P. Le Callet and D. Barba, *On the performance of human visual system based image quality assessment metric using wavelet domain*, Proc. SPIE Human Vision and Electronic Imaging XIII (HVEI'08), San Jose, California, 2008.

A. Ninassi, O. Le Meur, P. Le Callet and D. Barba, *Does where you gaze on an image affect your perception of quality ? Applying to image quality metric*, ICIP 2007, San Antonio, Texas, USA, 2008.

A. Ninassi, O. Le Meur, D. Barba, P. Le Callet and A. Tirel, *Task Impact on the Visual Attention in Subjective Image Quality Assessment*, EUSIPCO'06 (invited paper), Florence, Italy, 2006.

Brevets

A. Ninassi, O. Le Meur, P. Le Callet and D. Barba, *Wavelet perceptual quality metric*, déposé en EP (Brevet Européen), 2008.

A. Ninassi, O. Le Meur, P. Le Callet and D. Barba, *Method to take into account the temporal variation of quality*, déposé en EP (Brevet Européen), 2008.

Autres publications

P. Le Callet, S. Péchard, S. Tourancheau, A. Ninassi and D. Barba, *Towards the next generation of video and image quality metrics : impact of display, resolution, content and visual attention in subjective assessment*, IMQA (Workshop) 2007, Chiba, Japon.

A. Ninassi, *Évaluation objective de la qualité d'images fixes et attention visuelle*, Journée des Doctorants (JDOC), 2007, Nantes, France.

A. Ninassi, *Impact de la tâche d'évaluation subjective de la qualité d'images sur l'attention visuelle*, Groupement de Recherche Information-Signal-Image-viSion (GdR ISIS), Thème B - Image et Vision, 2006, Paris, France.

Résumé en français :

Cette étude traite de l'évaluation locale des distorsions perceptuelles, de l'évaluation globale de la qualité visuelle, et de l'influence de l'attention visuelle en évaluation de qualité.
Afin d'évaluer localement les distorsions dans les images, nous avons simplifié un modèle existant du système visuel humain en utilisant la transformée en ondelettes et nous avons proposé une meilleure modélisation des effets de masquage par la prise en compte du masquage semi-local. A partir de ces modèles, nous avons conçu et validé des métriques de qualité d'images.
Pour les vidéos, nous avons conçu une méthode d'évaluation locale des distorsions temporelles reposant sur un cumul temporel court terme des distorsions spatiales. Celui-ci simule l'évaluation des distorsions via des mécanismes de sélection de l'attention visuelle. Une métrique de qualité s'appuyant sur cette méthode a été conçue et validée. Celle-ci est basée sur un cumul temporel long terme incorporant un comportement asymétrique et un effet de saturation perceptuelle.
L'influence de l'attention visuelle sur l'évaluation de la qualité a été analysée à partir des données issues de tests oculométriques réalisés sur des images et sur des vidéos, en exploration libre et en tâche de qualité. Les résultats ont confirmé, entre autres, l'influence de la tâche de qualité sur le déploiement de l'attention visuelle. L'impact de l'attention visuelle sur l'évaluation objective de la qualité a également été étudié en utilisant l'information de saillance réelle. Nous avons montré qu'une simple pondération linéaire des distorsions par l'attention visuelle ne permettait pas d'améliorer clairement les performances des métriques de qualité.

Titre et résumé en anglais : From local perception of coding distortions to the overall visual quality evaluation of images and videos. Contribution of visual attention in visual quality assessment.

This study deals with the local evaluation of perceptual distortions, the overall visual quality assessment and the influence of visual attention in visual quality assessment.
To locally evaluate distortions in images, we have simplified an existing human visual system model using wavelet transform and we have proposed an improved visual masking model that takes into account both semi-local masking and contrast masking. From these models, we have designed and tested several image quality metrics.
Regarding videos, we have developed a new method to locally evaluate the spatio-temporal distortions. This method is based on a short-term temporal pooling of spatial distortions which simulates the evaluation of distortions through some selection mechanisms of visual attention. A video quality metric based on this method has been designed and validated. It is based on a long-term temporal pooling incorporating perceptual saturation and asymmetric behavior.
In order to study visual attention in subjective and objective visual quality assessment, eye-tracking experiments on images and videos have been conducted both in free task and quality task. From collected data we have studied the visual attention deployed in the different configurations. The results have confirmed, among others, the influence of the quality task on deployment of visual attention. The impact of visual attention in the construction of the quality judgment has also been studied using the real saliency information. Results show that, both with images and videos, a simple linear weighting of distortions by the visual attention does not clearly improve performances of objective quality metrics.

Mots-clés : qualité visuelle, distorsion perceptuelle, attention visuelle, système visuel humain, cumul temporel, masquage semi-local, codage vidéo.

Discipline : Traitement du Signal et Informatique Appliquée

N° :