Exploitation des statistiques structurelles d'une image pour la prédiction de la saillance visuelle et de la qualité perçue

Michael Nauge

THÈSE

pour l'obtention du Grade de

DOCTEUR DE L'UNIVERSITE DE POITIERS
(Faculté des Sciences Fondamentales et Appliquées)
(Diplôme National - Arrêté du 7 août 2006)
Ecole Doctorale : Sciences et Ingénierie pour l'Information
Secteur de Recherche : Traitement du Signal et des Images

Présentée par :

Michael NAUGE

Exploitation des statistiques structurelles d'une image pour la prédiction de la saillance visuelle et de la qualité perçue

---

Directeur de Thèse : Christine Fernandez-Maloigne
Co-Directeur de Thèse : Mohamed-Chaker Larabi

Soutenue publiquement le 11 décembre 2012
devant la Commission d'Examen

JURY

M. Alain TREMEAU ..... *Rapporteur*
Professeur à l'Université Jean Monnet de Saint-Etienne

Mme Valérie GOUET-BRUNET ..... *Rapporteur*
Directeur de recherche à l'IGN Paris

M. Frédéric MORAIN-NICOLIER ..... *Examinateur*
Professeur à l'Université de Reims-Champagne-Ardenne

Mme Françoise VIENOT ..... *Examinateur*
Professeur au Museum National d'Histoire Naturelle Paris

M. Vincent COURBOULAY ..... *Examinateur*
Maître de conférences à l'Université de La Rochelle

M. Clency PERRINE ..... *Examinateur*
Maître de conférences à l'Université de Poitiers

M. Mohamed-Chaker LARABI ..... *Co-directeur de thèse*
Maître de conférences à l'Université de Poitiers

Mme. Christine FERNANDEZ-MALOIGNE ..... *Directeur de thèse*
Professeur à l'Université de Poitiers

Remerciements

Je remercie tout d'abord Chaker Larabi pour son encadrement de thèse, mais aussi et surtout pour ses valeurs humaines qui ont su amener une ambiance dynamisante, chaleureuse et presque familiale au quatrième étage.

Je remercie également Christine Fernandez-Maloigne, qui a su m'accepter tel que je suis, avec qui j'ai toujours pu dialoguer, partager et débattre de nos différents points de vues, tant sur un plan scientifique que relationnel.

Merci à M. Alain Tremeau, Professeur à l'Université Jean Monnet de Saint-Etienne et Mme Valérie Gouet-Brunet, Directeur de recherche à l'IGN Paris, d'avoir accepté de remplir la fastidieuse tâche de rapporter sur ce mémoire. Merci également à M. Frédéric Morain-Nicolier, Professeur à l'Université de Reims-Champagne-Ardenne, Mme Françoise Vienôt, Professeur au Museum National d'Histoire Naturelle Paris, M. Vincent Courboulay, Maître de conférences à l'Université de La Rochelle et M. Clency Perrine, Maître de conférences à l'Université de Poitiers, qui ont accepté de participer à mon jury de thèse.

Comme de coutume, il m'est maintenant possible de passer à des remerciements plus personnels voire intimes et de citer toutes les personnes qui ont compté pour moi durant cette période de thèse. Certains savent que je suis tête en l'air, qu'il est donc possible que je les oublie. Afin d'éviter cela et étant relativement pudique, je ne vais pas faire une liste exhaustive de vous tous, qui

avez su chacun à votre façon laisser une trace indélébile dans ma mémoire et ma vie. Pour me faire pardonner [1] [2] [3] [4], je vous propose de finir sur une touche d'humour et de philosophie avec un florilège de citations :

«Il semble prudent de remercier un auteur pour son livre avant de le lire. Cela évite d'avoir à mentir.» de George Santayana

«Beaucoup remercier signifie secrètement demander davantage.» Proverbe anglais

«Les idées reçues n'exigent pas de remerciements.» de Ylipe

«Comme tous les hommes, il était beaucoup plus éloquent pour demander que pour remercier.» de Prosper Mérimée

«Quand on voit ce que les pigeons ont fait sur ce banc, il faut remercier Dieu de n'avoir pas donné d'ailes aux vaches.» de Régis Hauser

«Une âme délicate est gênée de savoir qu'on lui doit des remerciements, une âme grossière, de savoir qu'elle en doit.» de Friedrich Nietzsche

«Il faut remercier les hommes le moins possible parce que la reconnaissance qu'on leur témoigne les persuade aisément qu'ils en font trop !» de Benjamin Constant

«Il faut toujours remercier l'arbre à karité sous lequel on a ramassé de bons fruits pendant la bonne saison.» de Ahmadou Kourouma

«L'important pour un homme politique est de vivre assez vieux pour inspirer confiance, avoir eu le temps de se faire appeler, remercier, déboulonner puis panthéoniser... Après quoi on donne votre nom à une rue, ce qui n'est qu'une manière de vous y jeter.» de André Frossard

1. et surtout pour frimer en me la jouant rebel, cool, hipster presque SWAG...
2. Et enfin jouir du plaisir de faire une vraie note de bas de page non contrôlée ;-)
3. « Heureux soient les fêlés, car ils laisseront passer la lumière.» de Michel Audiard
4. Merci Françoise :). Désolé, les autres... mais il fallait bien une exception pour confirmer la règle ;)

Table des matières

Introduction **1**
1 Contexte et objectifs 1
2 Contributions 3
3 Plan du mémoire 4

1 Points d'intérêt **7**
1.1 Des statistiques de l'image aux points d'intérêt 8
1.1.1 Détection de contours 10
1.1.2 Filtrage par convolution 11
1.1.3 Laplacien de Gaussienne 13
1.2 État de l'art des détecteurs de points d'intérêt 14
1.2.1 Points-selle de Beaudet 14
1.2.2 Détecteur de coins de Moravec 19
1.2.3 Détecteur de coins de Harris et Stephens 21
1.2.4 Détecteur de Harris multi-échelles 27
1.2.5 Détecteur de blobs multi-échelle : SIFT 32
1.2.6 Détecteur de Blobs multi-échelle : SURF 35
1.3 Conclusion 39

2 Prédiction de la saillance visuelle par points d'intérêt **41**
2.1 Introduction 41
2.2 Saillance visuelle 42
2.2.1 Explication physiologique 42
2.3 Modèle de prédiction de la saillance 52
2.3.1 Modèle d'Itti et al. 52
2.3.2 Modèle d'Achanta et al. 55
2.3.3 Discussion 56
2.4 Approche proposée : Étude de la relation entre les points d'intérêt et la saillance visuelle 57
2.4.1 Protocole expérimental 57
2.4.2 Mesures oculométriques 57
2.4.3 Paramètres des détecteurs de points d'intérêt 60
2.4.4 Description de la distance EMD 61
2.4.5 Expérimentation et analyse statistique 64

2.5 PINS (Prediction of INterest Points Saliency) 78
2.5.1 Mesure de performance 80
2.6 Conclusion 89

3 Introduction à la qualité de l'expérience 91
3.1 Introduction 91
3.2 Expérience psychovisuelle pour l'évaluation de la qualité 93
3.2.1 Méthodologie d'évaluation subjective de la qualité 93
3.2.2 Conduite de mesures de qualité pour le comité JPEG . . 97
3.3 État de l'art des métriques de qualité d'images 103
3.3.1 Métriques mathématiques 104
3.3.2 Métriques perceptuelles 106
3.3.3 Métriques pondérées par le SVH 107
3.3.4 Métriques à référence réduite et sans référence 108
3.4 Méthodologie d'évaluation des performances des métriques . . . 110
3.4.1 Précision de la prédiction : Corrélation de Pearson et RMSE 111
3.4.2 Uniformité de la prédiction : Outlier Ratio (OR) 112
3.4.3 Monotonie de la prédiction : Corrélation d'ordre de Spearman 112
3.4.4 Bases d'images dédiées à la qualité 115
3.4.5 Mise en place d'une application web dédiée à la comparaison de métriques 117
3.5 Conclusion 121

4 Métriques de qualitée basées sur les points d'intérêt :Application à l'amélioration de la QoE en transmission 123
4.1 Approche proposée : Métrique à référence réduite de prédiction de la qualité par évolution de points d'intérêt 124
4.1.1 Méthode QIP 125
4.1.2 Méthode QIP-HSM 135
4.2 Intégration de QIP dans une chaîne de transmission sans fil . . . 152
4.2.1 Transmissions JPWL à travers un canal MIMO sous monitoring perceptuel 153
4.2.2 Mesure de performance objective 155
4.2.3 Validation de la méthode par expérimentation subjective 158
4.3 Conclusion 169

5 Modélisation de l'évolution des statistiques structurelles de l'image 171
5.1 Migration des propriétés structurelles 171
5.1.1 Extraction d'attributs structurels des pixels 172
5.1.2 Classification des attributs structurels 174

5.1.3 Mesure et modélisation des migrations par graphe multi-étiqueté 180
5.2 Sensibilité des statistiques aux dégradations 184
5.2.1 Prédiction de la qualité perçue 188
5.2.2 Estimation du facteur q de la compression JPEG 193
5.3 Conclusion 209

Conclusion **211**
1 Rappel des contributions 211
2 Perspectives 213

Bibliographie **215**

Bibliographie de l'auteur **231**
1 Revues internationales avec comité de lecture 231
2 Conférences internationales avec comité de lecture 231
3 Conférences nationales avec comité de lecture 232
4 Conférences sans acte 232

Introduction

1 Contexte et objectifs

Nous pouvons observer que les sociétés actuelles et de demain tendent vers des expansions, des articulations et des échanges à l'échelle mondiale. Tous les systèmes, qu'ils soient sociaux, humains, politiques, financiers et technologiques sont donc de plus en plus inter-connectés à grande échelle malgré de grandes contraintes d'hétérogénéités. Dans ce contexte, pourtant très délicat, l'échange de biens, de services et d'informations à grande vitesse et sur tous types de réseaux est de plus en plus sollicité. Ces tendances et évolutions sont en grande partie rendues possibles grâce à l'évolution des systèmes de technologies de l'information, avec entre-autre l'explosion d'internet. A ce titre, une étude récente émanent du Visual Networking Index (VNI) de l'entreprise CISCO, le leader mondial en réseaux de communication, vient de faire état de prévisions et annonce que le trafic internet sera multiplié par quatre d'ici 2016. Elle prévoit un trafic IP annuel mondial de 1,3 zettaoctets (un milliard de milliards de gigaoctets). Cette importante augmentation et utilisation des réseaux peut être expliquée par plusieurs facteurs :

– **De plus en plus d'appareils** : avec la prolifération des tablettes, des téléphones portables et tout autre appareil intelligent (avec l'émergence des communications Machines à Machines : M2M).

- **De plus en plus d'internautes** : c'est 3,4 milliards d'internautes, soit environ 45% de la population mondiale qui sera connectée à Internet d'après les prévisions des Nations Unies.
- **De plus en plus de connexions sans fil** : c'est plus de 50% du trafic internet mondial qui devrait provenir des connexions sans fil.
- **De plus en plus de vidéos** : c'est 1,2 millions de minutes de vidéos qui circuleront sur internet chaque seconde, soit l'équivalent de 833 jours de visionnage.
- **De plus en plus d'images** : actuellement, c'est plus de 2263 photos qui sont mises en ligne sur Facebook chaque seconde, soit plus de 71 milliards par an. C'est bien sûr sans compter les photos partagées sur d'autres plateformes telles que FlickR, Picasa, etc. Il est également important de noter que les performances de nos appareils d'acquisition ne cessent de croître avec des résolutions d'ores et déjà de 8 mégapixels sur smartphones, de 40 mégapixels sur des reflex numériques et enfin de 50 000 mégapixels sur un appareil prototype à faible coût dans les laboratoires de la Duke University et de l'Université de l'Arizona. Sans oublier la prouesse de 2012, présentée par le chercheur Ramesh Raskar, la femto-photographie, qui permet de prendre une image tout les 1 000 000 000 000 $^{\text{ème}}$ de seconde, et permet de voir la lumière en mouvement.

Ces chiffres donnent le vertige et peuvent paraître sur-réalistes. Nous pouvons donc imaginer que ces estimations sont très optimistes et permettent à des multinationales tels que CISCO de séduire plus facilement ses actionnaires. Cependant, en observant nos consommations présentes, il est aisé de se rendre compte que nous sommes de plus en plus nombreux à être connectés du matin au soir en utilisant chacun plusieurs appareils. Nous pouvons passer des appels téléphoniques vidéo, regarder des films sur nos tablettes, sans oublier nos téléviseurs ou réfrigérateurs connectés à internet et bien sûr l'utilisation de vidéo-conférences dans un cadre plus professionnel. Toutes ces pratiques tendent à donner du crédit à ce type de prévisions.

Pour résumer, nous souhaitons acquérir, partager, visualiser et consommer de plus en plus de vidéos et d'images, partout, tout le temps, avec toutes sortes d'appareils, avec les meilleures qualité et résolution possible. Afin de répondre à ces importants besoins de stockage et de transfert, il est impératif de réduire la quantité d'information et donc de compresser les images. Pour cette compression, c'est actuellement la norme JPEG qui est la plus utilisée. Cependant, cette compression a été conçue il y a maintenant plus de 30 ans et est bien loin des contraintes et attentes actuelles. C'est pourquoi la recherche dans le domaine pense activement à son évolution, sous la bannière de AIC pour Avanced Image Coding. Cette nouvelle norme s'articule sur plusieurs volets et a permis de donner naissance au projet ANR CAIMAN (Codage

Avancé d'IMAges et Nouveaux services) qui s'intéresse particulièrement à une approche conjointe, à savoir, de nouvelles transformées, de nouvelles méthodes de transmission sans fil et également la prise en compte de facteurs humains afin de maximiser sa qualité de l'expérience. C'est principalement sur ce troisième volet que porte ce travail de thèse.

Les objectifs de cette thèse s'articulent autour de la qualité perçue et plus particulièrement sur l'exploitation des statistiques structurelles d'une image. Ainsi, nous cherchons à développer des approches permettant d'estimer la qualité à partir des indices d'une image (les points d'intérêt) et s'adaptant aux différents contextes applicatifs comme la transmission ou la compression. La saillance visuelle représente une part importante de ce travail où elle permet de mieux appréhender la qualité et de participer à la définition d'un codeur intelligent exploitant la hiérarchie inhérent à l'image. Enfin, les statistiques structurelles d'une image subissent une variation à chaque fois que le contenu de l'image est modifié ou altéré. Nous cherchons, dans cette thèse, à comprendre ces mécanismes et à en proposer une modélisation fine pour identifier et mesurer l'impact des dégradations en lien avec le jugement humain.

2 Contributions

Étant particulièrement intéressé par les facteurs humains impliqués dans un processus d'appréciation de la qualité d'une image, nous nous devions de les étudier. Pour ce faire, nous avons été amenés à conduire plusieurs campagnes de tests psychovisuels afin d'obtenir des informations sur les capacités sensorielles du système visuel humain et plus particulièrement sur les sensations perçues lors de la visualisation d'images dégradées. Notre première contribution a consisté à mener une campagne d'évaluation de la qualité subjective, en collaboration avec le comité JPEG, dans le cadre d'une procédure de normalisation d'un nouveau codeur d'images. Cette campagne a permis d'obtenir des informations relatives à la visibilité et la sensation de gène occasionnées par quatre types de codeurs dont trois sont des standards.

En ce qui concerne la qualité subjective, nous avons proposé plusieurs méthodes objectives de prédiction de la qualité perçue, couramment appelées métriques de qualité. La première de nos propositions est une métrique de qualité qui extrait une très faible quantité d'attributs d'une image d'origine, afin de prédire la qualité d'une version de cette image potentiellement dégradée. Pour réaliser cette tâche, nous avons orienté nos recherches vers les problématiques de suivi et de reconnaissance d'objets et plus particulièrement les méthodes utilisant les détecteurs de points d'intérêt (*interest points detector*). Cette pro-

position semblait particulièrement adaptée aux contraintes de compression et de transmission d'images sur les réseaux sans fil grâce à son exécution rapide et son faible besoin de référence de l'image avant émission. Nous l'avons donc intégrée dans un schéma de transmission sur réseaux MIMO réalistes en tant qu'organe de décision permettant de garantir le décodage optimal dans des conditions délicates de couverture réseaux. Afin de vérifier l'apport de ce décodage guidé par la prédiction de qualité perçue et en complément des mesures objectives, nous avons mené une campagne de tests psychovisuels pour quantifier le gain en terme d'augmentation de la qualité de l'expérience.

En ce qui concerne les facteurs psychovisuels, il apparaît que l'humain ne regarde pas les images de manière uniforme. En effet, certaines zones attirent particulièrement son regard tandis que d'autres ne sont observées que très partiellement. Ces zones d'intérêt sont également connues sous le nom de zones de saillance visuelle. Nous avons donc proposé et développé des algorithmes à même de prédire cette saillance, avec entre autre l'utilisation de détecteurs de points d'intérêt. Ce type d'outils peut, à terme, être exploité par les méthodes de compression et de transmission afin de garantir et offrir un maximum de qualité sur les régions visuellement pertinentes. Cette saillance peut également se révéler utile dans le développement de métriques de qualité afin de pondérer la visibilité des artéfacts en fonction de leur localisation et leur saillance. C'est ce que nous avons pu mettre en pratique en proposant une version intégrant une saillance hiérarchique dans la métrique de qualité citée précédemment.

Ayant exploité à plusieurs reprises les détecteurs de points d'intérêt, nous avons poussé nos recherches en étudiant finement les statistiques structurelles d'une image. Nous avons proposé un modèle exploitant l'évolution et les changements de ces attributs structurels de l'image pour différents types dégradations. Par ces travaux, nous avons pu modéliser les liens existants entre le facteur q de la compression JPEG et la qualité perçue. Cette modélisation a également permis de produire une toute nouvelle génération de métriques de qualité.

3 Plan du mémoire

Pour atteindre les objectifs et résultats cités précédemment, nous avons choisi d'axer nos recherches sur les méthodologies initialement conçues pour les problématiques de suivi et de reconnaissance d'objets. Dans ces domaines, des méthodes d'extraction de statistiques structurelles de l'images ont particulièrement été développées et portent le nom de *interest point detector*. De ce fait, cette thèse et donc ce mémoire suivent un fil rouge, à savoir les points

d'intérêt.

Le premier chapitre propose d'introduire la notion de points d'intérêt d'un point de vu général, suivi d'explications détaillées sur les théories et méthodes permettant de les détecter. C'est un historique qui est proposé, où l'accent est mis sur les ressemblances et les évolutions des approches de la littérature.

Le second chapitre détaille les mécanismes physiologiques impliqués dans les processus de saillance visuelle ainsi que les modèles proposés pour la prédire. Quelques ressemblances entre les modèles de saillance visuelle et les méthodes d'extraction de points d'intérêt ont pu être observées. Nous proposons donc de mener une étude expérimentale permettant de quantifier dans quelle mesure les détecteurs de points d'intérêt peuvent prédire la saillance visuelle humaine. Les résultats ont permis de développer un nouveau prédicteur de saillance à comparer avec l'état de l'art du domaine.

Le troisième chapitre s'intéresse quant à lui à la qualité de l'expérience (QoE) et particulièrement la qualité visuelle. A ce titre, nous proposons une brève introduction sur la QoE accompagnée des méthodologies d'expérimentation subjective utilisées pour mesurer la qualité visuelle. Nous associons à ce sujet, quelques méthodes de l'état de l'art des métriques de qualité ainsi que les protocoles permettant d'évaluer leur performance. Au vu de la quantité des métriques existantes et de la complexité de leur comparaison, nous détaillons le web-service que nous avons développé pour pallier ces problèmes.

Le quatrième chapitre, propose d'utiliser les détecteurs de points d'intérêt pour estimer la qualité perçue, de manière rapide et avec peu de référence sur l'image d'origine. Ce travail a donné naissance à plusieurs métriques de qualité, dont une exploitant la saillance visuelle. Nous développons ensuite l'intégration d'une de nos métriques dans un contexte applicatif réaliste dans le cadre d'une chaîne de transmission sur réseaux sans fil MIMO. Des mesures objectives et subjectives permettent de quantifier l'apport de notre méthode dans les tâches d'augmentation de la QoE dans ces contextes difficiles.

Enfin, le cinquième chapitre propose d'étudier et d'exploiter les statistiques structurelles de l'image, pour créer un modèle axé sur l'observation des migrations structurelles des pixels engendrées par les dégradations de l'image. Nous détaillons également plusieurs applications de ce modèle permettant de mettre en avant les liens existant entre ces migrations et la qualité perçue et ainsi prouver la pertinence de la proposition.

Chapitre 1

Points d'intérêt

Sommaire

1.1 Des statistiques de l'image aux points d'intérêt **8**
1.1.1 Détection de contours 10
1.1.2 Filtrage par convolution 11
1.1.3 Laplacien de Gaussienne 13
1.2 État de l'art des détecteurs de points d'intérêt **14**
1.2.1 Points-selle de Beaudet 14
1.2.2 Détecteur de coins de Moravec 19
1.2.3 Détecteur de coins de Harris et Stephens 21
1.2.4 Détecteur de Harris multi-échelles 27
1.2.5 Détecteur de blobs multi-échelle : SIFT 32
1.2.6 Détecteur de Blobs multi-échelle : SURF 35
1.3 Conclusion **39**

Commençons par introduire la notion de points d'intérêt de façon générale. Un point d'intérêt est une localisation précise permettant la réalisation d'une tâche ou la détection d'un objet. Typiquement les pharmacies d'une ville sont des lieux particuliers capables de nous aider si l'on souhaite trouver des médicaments, tout comme l'est un oasis au milieu d'un désert. Localiser précisément ces lieux est important si l'on veut éviter de demander de l'aspirine chez le boucher, ou de courir vers un mirage. Pour éviter ces malentendus, il est intéressant de trouver les caractéristiques les plus stables et discriminantes. Dans l'exemple de la pharmacie, il peut être judicieux de repérer les enseignes possédant une "croix verte". Cette caractéristique semble discriminante et fiable, car ce type de croix n'est jamais présent sur les boucheries, et toutes les pharmacies en possèdent une, du moins en France.

Se pose ensuite le problème de la robustesse de détection de cette "croix verte" à différents types de perturbations/variations. Par exemple, est-il diffi-

cile de trouver une pharmacie quand le soleil n'est pas encore levé ? Heureusement pour nous, la réponse est non, car ces enseignes sont lumineuses. La détection de cette "croix verte" est par conséquent invariante à des changements de luminosité ambiante.

Classiquement, il est intéressant d'extraire et d'obtenir des informations supplémentaires sur un lieu d'intérêt. Par exemple, la liste de certains médicaments particulièrement rares, ou les horaires d'ouverture et le numéro de téléphone de ce point de vente. Ces attributs complémentaires peuvent être utilisés afin de trouver et identifier rapidement un lieu d'intérêt particulier parmi les disponibles. Imaginons que nous ayons à rechercher un anti-venin contre la piqure de mygale australienne un lundi matin à 06h45. Trouver rapidement la bonne pharmacie parmi des milliers peut permettre d'éviter de sérieuses complications suite à notre malheureuse rencontre avec un arachnide. La pertinence et l'expressivité des informations complémentaires sont les clefs d'une recherche efficace.

On se rend compte de l'utilité de la détection de points d'intérêt dans la vie courante. La notion de points d'intérêt existe également en vision par ordinateur et analyse d'images. Ce chapitre présente donc les différents types de détecteurs de points d'intérêt de la littérature et leurs genèses, ainsi que les méthodes de calcul de statistiques de l'image nécessaires à leur extraction.

1.1 Des statistiques de l'image aux points d'intérêt

En vision par ordinateur, la notion de points d'intérêt (*Interest Points*) est née des problématiques de reconnaissance et de suivi d'objets 3D à partir d'images numériques 2D. Le suivi d'objet s'apparente à observer son évolution au cours du temps, par exemple son déplacement ou sa déformation. La mesure d'un déplacement revient à mesurer la distance entre deux points. Il est donc indispensable de placer au moins un point de référence sur l'objet que l'on veut suivre. La Figure 1.1 illustre le suivi de la course d'un lapin, avec son œil comme point d'intérêt.

Pour assurer une bonne précision des mesures, il faut être capable de replacer avec précision ce point sur toutes les images d'une séquence. Pour ce faire, on attend d'un point d'intérêt qu'il soit repérable facilement et rapidement. Lors de la phase d'acquisition des images constituant la séquence, il est possible que la caméra ait bougé, que la mise au point ait changé ou qu'un nuage ait changé la luminosité ambiante. Toutes ces modifications ont pu perturber les caractéristiques du point. On souhaite donc également que la détection du

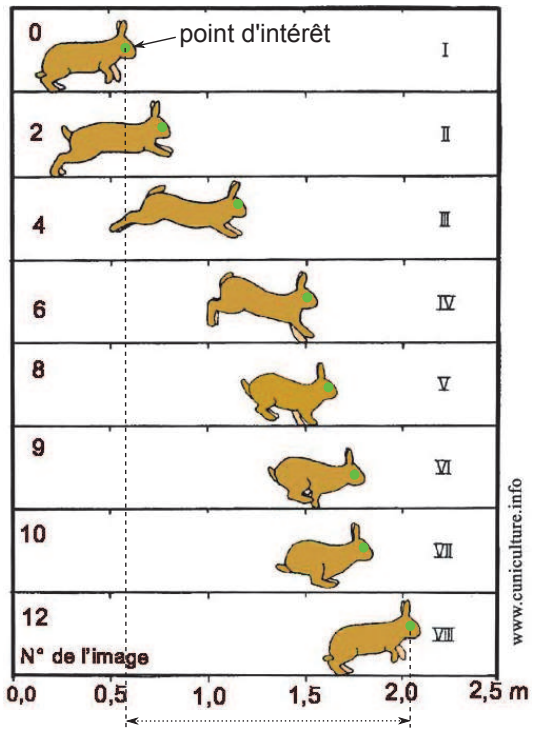


Figure 1.1 : Mesure de la longueur d'un saut de lapin par suivi de point d'intérêt

point d'intérêt soit la plus robuste possible, la plus invariante aux différentes perturbations. Elle se doit d'être stable malgré des perturbations locales ou globales de l'image. Ces perturbations peuvent être des variations d'intensité lumineuse, de contraste, des changements d'angle de vue (transformation affine, rotation, translation, déformation perspective, changements d'échelle), des artéfacts dus au capteur (bruit, flou), des artéfacts de compression (JPEG, JPEG 2000), ...

La précision de la localisation d'un point d'intérêt est un facteur important. Dans le domaine de l'image, cette localisation est la coordonnée spatiale d'un point/pixel.

Étant donné que nous savons que nous cherchons à repérer un point de manière rapide et à le retrouver dans une autre image, la question est de savoir comment. Il est préférable que ce pixel soit particulier, qu'il soit reconnaissable parmi tous les autres pixels de l'image et qu'il soit en rupture avec les autres. La tâche serait aisée si ce pixel était le seul à être vert alors que tous les autres sont rouges, par exemple. Mais dans une image naturelle, la variété des couleurs est nettement plus importante, ce qui rend la tâche bien moins facile. Pour répondre à cette problématique, les démarches adoptées dans la littérature cherchent toutes à exploiter les statistiques de l'image afin de repé-

rer des pixels en rupture ayant des variations "brutales". L'emploi du terme statistiques de l'image est intimement lié au fait qu'une image contient un nombre très important de pixels (plusieurs centaines de milliers. Par exemple, une résolution de 512 représente 262144px), et que si l'on considère une image monochrome, le niveau de gris de chaque pixel peut s'apparenter à une variable aléatoire. Une caractéristique statistique classique peut être le calcul d'histogramme, qui consiste à comptabiliser le nombre de pixel de chaque niveau de gris. Ce qui peut être vu comme la densité de probabilité d'apparition $p(g)$ des niveaux de gris (g) dans une image. On peut également citer les calculs de moyennes ou de moments de divers ordres, de matrice de corrélation, etc. Mais nous allons dans un premier temps nous focaliser sur les détections de contours particulièrement adaptées à la recherche de ruptures dans l'image et exploitées par tous les extracteurs de points d'intérêt.

Afin de mieux appréhender les différentes méthodes d'extraction de contour et de points d'intérêt, certaines bases théoriques et applicatives du traitement du signal et de l'image seront explicitées, telles que les techniques de filtrage, l'utilisation de dérivées et de produit de convolution et la construction d'espaces multi-échelles. Tous ces pré-requis sont la base de la quasi-totalité des traitements sur l'image.

1.1.1 Détection de contours

Considérons une image en niveau de gris, où la couleur noire peut être vue comme une intensité lumineuse nulle, égale à 0 et où la couleur blanche est une intensité maximale, égale à 1. Trouver les contours revient à chercher les évolutions "rapides" du signal intensité, les variations dans l'image où des pixels consécutifs évoluent rapidement entre 0 et 1.

En physique et mathématique, l'étude des variations d'un signal (ou d'une fonction) peut être faite en utilisant le gradient. Le gradient, noté ∇, est un vecteur qui informe de la variation d'une fonction par rapport à la variation de ses différents paramètres. Si l'on considère simplement une ligne de pixels de notre image, nous nous plaçons dans un cas 1D, où l'on peut noter I l'intensité d'un pixel et x la coordonnée horizontale du pixel considéré. En 1D, il y a une équivalence entre les notions de gradient et de dérivée du signal. D'où l'équation 1.1 :

$$\nabla I(x) = I'(x) = \frac{\partial I}{\partial x}(x). \tag{1.1}$$

Afin d'illustrer les notions de dérivée en traitement d'images, la Figure 1.2 donne l'évolution du signal intensité et du module de ses dérivées première et seconde, pour les pixels passant du noir au blanc avec différents niveaux de gris intermédiaires pour une portion de ligne de l'image.

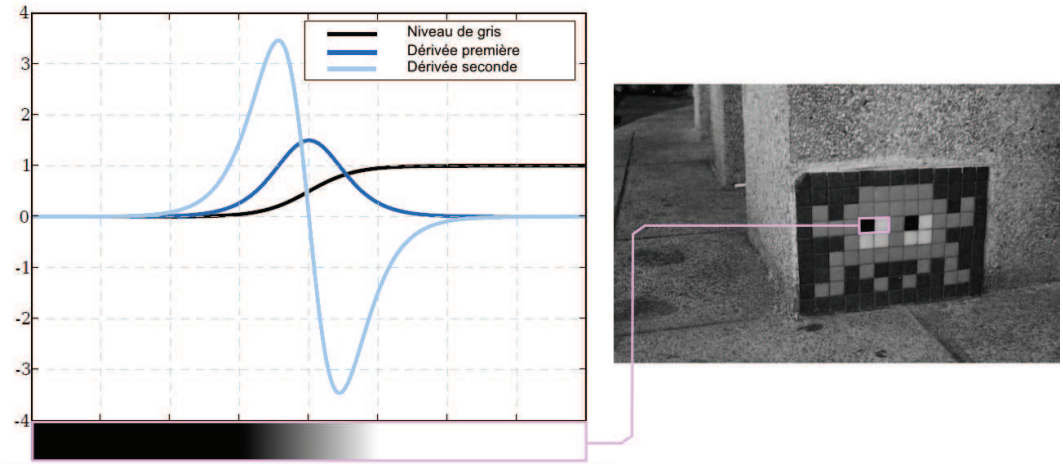


Figure 1.2 : Détection de contours par étude de dérivées

Sur cette figure on remarque que le point de rupture (i.e un contour) peut être trouvé en repérant le maximum local de la dérivée première du signal. On peut également remarquer un passage par 0 de la dérivée seconde au point de rupture. De ce fait, une méthode alternative à l'étude du gradient pour la détection de contour consiste à détecter les passages par zéro, après calcul de la dérivée seconde par l'utilisation de l'opérateur différentiel Laplacien défini par l'équation 1.2. Bien évidemment, ces constats sont également démontrables mathématiquement. Les bases théoriques de la détection de contour par filtrage linéaire sont détaillés dans [MF77, SDG79, TP86].

$$\Delta = \nabla^2 I = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \tag{1.2}$$

1.1.2 Filtrage par convolution

La détection ou suppression de contours peut être abordée comme un problème de filtrage. En effet, isoler les contours revient à supprimer les variations lentes du signal et conserver les variations rapides. On parle dans ce cas d'un filtrage passe-haut (ne laissant "passer" que les hautes fréquences). A l'inverse pour supprimer les contours, il est utile de ne conserver que les basses fréquences. Il s'agit dans ce cas d'un filtre passe-bas. Dans la pratique, et pour

filtrer les signaux discrets et 2D des images, ce sont des filtrages par noyaux de convolutions (*k*ernel) qui sont effectués. Ce sont les valeurs que prend ce noyau qui définissent la type et la puissance du filtrage réalisé.

A titre d'exemple, en traitement d'images, il est classique d'utiliser un noyau Gaussien dont les valeurs sont définies par l'équation :

$$G_\sigma(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -[\frac{x^2+y^2}{2\sigma^2}]. \tag{1.3}$$

La mise en œuvre de Gaussienne permet de tirer profit des différentes qualités de celle-ci telles que : la séparabilité, le noyau auto-reproducteur, la limite de suites de polynômes, approximation par des splines, etc. L'écart type de la Gaussienne défini par son paramètre σ et la taille du filtre considéré permettent de contrôler ce filtrage. Plus le support du filtre est grand (σ élevé), plus l'effet de flou augmente, plus les hautes fréquences sont supprimées, ce qui est illustré sur la Figure 1.3. Ce type de filtre est très utilisé pour supprimer le bruit dans les images ainsi que pour la création d'une représentation multi-échelle de l'image (ce principe est détaillé en section 1.2.4).

image originale résultat par un filtre 5 résultat par un filtre 30

Figure 1.3 : Résultats de convolution d'une image par un filtre Gaussien

Ayant défini le principe de la convolution sur un exemple de suppression de hautes fréquences (contours, bruit), nous pouvons maintenant nous recentrer sur la détection de contours. Comme illustré précédemment, un contour peut être détecté en utilisant la dérivée première ou seconde du signal, suivi d'une détection de maximum local ou de passage par zéro. L'idée est donc d'approximer ces dérivées par un ou plusieurs filtres. Les noyaux de convolution les plus connus pour approximer le gradient sont ceux de *Roberts*, *Prewitt* et *Sobel* [CPF95].

Mais l'utilisation de la dérivée première n'est que la première étape, capable de donner la valeur du gradient en chaque pixel. Il est ensuite impératif de détecter les maxima locaux pour isoler les contours. La détection de ces derniers n'est pas un procédé des plus évidents, car il nécessite souvent le choix de deux seuils lors d'un seuillage par hystérésis.

1.1.3 Laplacien de Gaussienne

L'utilisation de la dérivée seconde (Laplacien) pour la détection de contours apparaît pertinente pour pallier les problèmes d'utilisation de la dérivée première. Cependant, cette méthode est très sensible aux bruits présents dans l'image. C'est pour cela qu'un pré-traitement de lissage du bruit est généralement appliqué à l'image avant la détection de contour. Comme expliqué précédemment (section 1.1.2), il est possible de supprimer le bruit par convolution avec un noyau Gaussien. L'idée est donc de combiner le filtre Laplacien avec celui d'un Gaussien appelé Laplacien de Gaussienne (LoG :Laplacien of Gaussian), filtre de Marr ou même chapeau mexicain. Cette combinaison de filtres peut être vue comme la dérivée d'ordre 2 de la dérivée partielle selon x et y du filtre Gaussien, dont l'équation est donnée par :

$$LoG = \frac{\partial^2}{\partial x^2} G_\sigma(x,y) + \frac{\partial^2}{\partial y^2} G_\sigma(x,y) = \frac{x^2+y^2-2\sigma^2}{\sigma^4} \exp^{-(x^2+y^2)/2\sigma^2}. \quad (1.4)$$

L'utilisation du filtre LoG est considérée comme le prototype du détecteur de contour inspiré des systèmes biologiques, les *p*rimal sketch du neuroscientifique Marr. En effet, il est démontré que le profil des réponses impulsionnelles de ces filtres sont très similaires aux réponses des cellules ON-OFF [Mar74] du système visuel des primates.

On comprend donc l'intérêt d'un tel filtre, autant d'un point de vue calculatoire que d'un point de vue biologique. Dans l'optique de couvrir au mieux le sujet sur les différents filtres utilisés pour la détection de contours, nous pouvons citer le filtre Différence de Gaussienne (*DoG : Difference of Gaussian*). Le filtre DoG est très proche du filtre LoG. Cependant, la construction de celui-ci est légèrement différente. Elle est réalisée par la soustraction de deux fonctions gaussiennes (équation 1.5) d'écarts-type σ légèrement différents. Ce type de filtre a également fait l'objet de plusieurs publications [ECR66, MPD04, You87] pour sa capacité à mimer les procédés neuronaux du système visuel humain impliqués dans l'extraction de détails lors de l'observation d'une image.

$$DoG = G_{\sigma 1} - G_{\sigma 2} = \frac{1}{\sqrt{2\pi}} \left[\frac{1}{\sigma_1} \exp^{-(x^2+y2)/2\sigma_1^2} - \frac{1}{\sigma_2} \exp^{-(x^2+y2)/2\sigma_2^2} \right] \quad (1.5)$$

Pour résumer, les filtres LoG et DoG réhaussent les contours, ils agissent tels des filtres passes-bandes et permettent d'isoler certaines variations ou fréquences particulières de l'image.

1.2 État de l'art des détecteurs de points d'intérêt

La détection de points d'intérêt est une problématique assez ancienne dans le traitement d'images. Dans toutes les approches, anciennes et contemporaines, le mécanisme de détection de ces points est toujours basé sur la mesure de statistiques de l'image. L'idée est d'isoler les pixels ayant des caractéristiques suffisamment en rupture des autres afin de les identifier de la manière la plus robuste et rapide possible.

Les premiers travaux en ce sens ont été introduit en 1976 par Beaudet [Bea76], qu'il a complété en 1978 [Bea78] en utilisant les polynômes orthogonaux discrets de Tchebychev [Bec73].

1.2.1 Points-selle de Beaudet

Beaudet [Bea76, Bea78] propose de détecter les points-selles ou points-cols du signal image I. De manière plus générale et théorique, un point-selle d'une fonction f définie sur un produit cartésien $X \times Y$ de deux ensembles X et Y est un point $(\bar{x}, \bar{y}) \in X \times Y$ tel que :

- $y \mapsto f(\bar{x}, y)$ atteint un maximum en $\bar{y}$ sur Y et,
- $x \mapsto f(x, \bar{y})$ atteint un minimum en $\bar{x}$ sur X.

La Figure 1.4 illustre ce qu'est un point-selle (point rouge) lorsque X et Y $\in \mathbb{R}$. On comprend par cette figure d'où vient le nom point-selle ou point-col grâce à la ressemblance de cette surface avec la forme d'une selle de cheval ou d'un col de montagne.

La détermination du points-selles est en fait bien plus ancienne que celle faite par Beaudet, car elle peut être vue comme un problème général d'op-

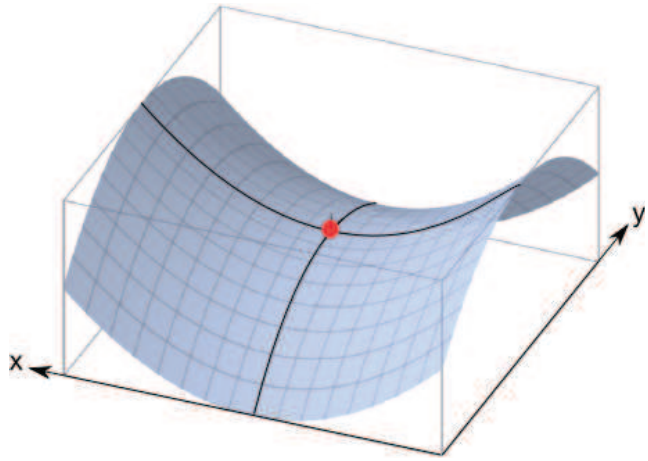


Figure 1.4 :
Illustration d'un points-selles pour $f(x, y) = x^2 - y^2$

timisation de paramètres, correspondant à l'étude des extrema de fonctions à plusieurs variables [1]. Les calculs permettant de déterminer si un point, est un point-selle se basent généralement sur l'utilisation des valeurs propres de la matrice hessienne en ce point. La matrice hessienne d'une fonction f à n variables, notées $(x_i)_{1 \leq i \leq n}$, est la matrice carrée, notée $H(f)$, de ses dérivées partielles secondes :

$$H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \frac{\partial^2 f}{\partial x_2 x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n x_1} & \frac{\partial^2 f}{\partial x_n x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \tag{1.6}$$

Dans le cas de l'image à deux dimensions $I(x, y)$, la matrice hessienne s'écrit :

$$H(I(x, y)) = \begin{bmatrix} \frac{\partial^2 I}{\partial x^2} & \frac{\partial^2}{\partial_x \partial_y} \\ \frac{\partial^2 I}{\partial_y \partial_x} & \frac{\partial^2 I}{\partial y^2} \end{bmatrix} \tag{1.7}$$

Les valeurs propres λ_i de cette matrice permettent de déterminer la nature du point considéré de la façon suivante :

– $\lambda_i > 0$: le point est un minimum local ;

1. L'étude des extrema de fonctions à plusieurs variables commence en fait avec l'étude des formes quadratiques, dont on peut noter les travaux de Fermat (1601-1665), ceux de Euler(1707-1783) et ceux de Lagrange(1763-1814).

– $\lambda_i < 0$: le point est un maximum local ;
– $\prod \lambda_i < 0$ (i.e. les λ sont de chaque signe) : le point est un point-selle.

A noter que ces conditions ne sont pas suffisantes ; puisqu'il faut par exemple que les valeurs des λ soit suffisamment élevées (environ 500 dans l'exemple considéré Figure 1.5) et suffisamment proches pour considérer qu'il s'agisse réellement d'un point-selle :

– $\lambda_{max} \approx |\lambda_{min}|$;
– $\lambda_{max} + |\lambda_{min}| \gg 0$.

La Figure 1.5 permet d'illustrer différents cas de figures où λ_{min} et λ_{max} ont des signes différents. Dans le cas (a), les λ ont des valeurs très différentes, mais il s'agit ici plus d'une vallée que d'un col (selle). Si l'on observe le cas (b), les λ ont bien des valeurs très proches et sont bien de chaque signe, cependant ces valeurs sont trop faibles, et il s'agit ici plus d'un aplat que d'un col. Il n'y a que le cas (c) qui combine à la fois les λ proches et de fortes valeurs.

Le calcul de la matrice hessienne nécessaire à la détermination de points-selles est basé sur des calculs de dérivées secondes de I dont des approximations sont possibles par l'utilisation de noyaux de convolution. Les noyaux de convolution 3×3 utilisés par Beaudet sont les suivants :

$$\frac{\partial I}{\partial x} \Leftrightarrow \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \Leftrightarrow \left(\frac{\partial I}{\partial y}\right)^T, \tag{1.8}$$

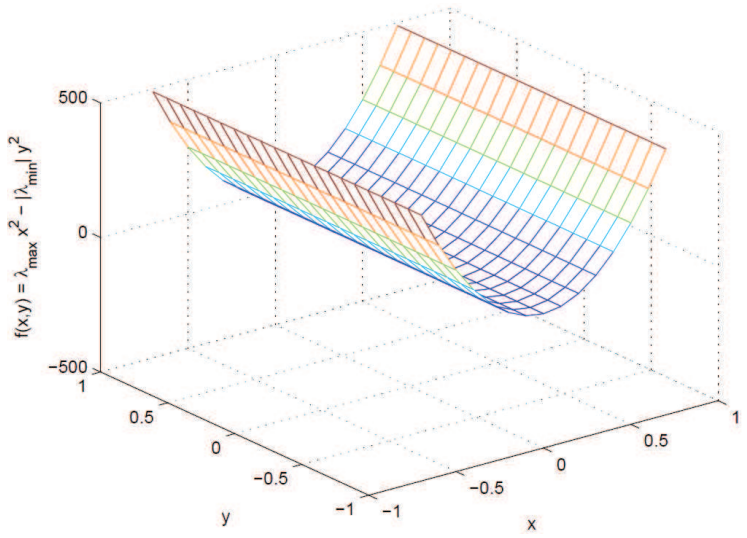
$$\frac{\partial^2 I}{\partial x^2} \Leftrightarrow \begin{bmatrix} 1 & -2 & 1 \\ 1 & -2 & 1 \\ 1 & -2 & 1 \end{bmatrix} \Leftrightarrow \left(\frac{\partial^2 I}{\partial y^2}\right)^T, \tag{1.9}$$

$$\frac{\partial^2 I}{\partial x \partial y} \Leftrightarrow \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix} \Leftrightarrow \left(\frac{\partial^2 I}{\partial y \partial x}\right)^T. \tag{1.10}$$

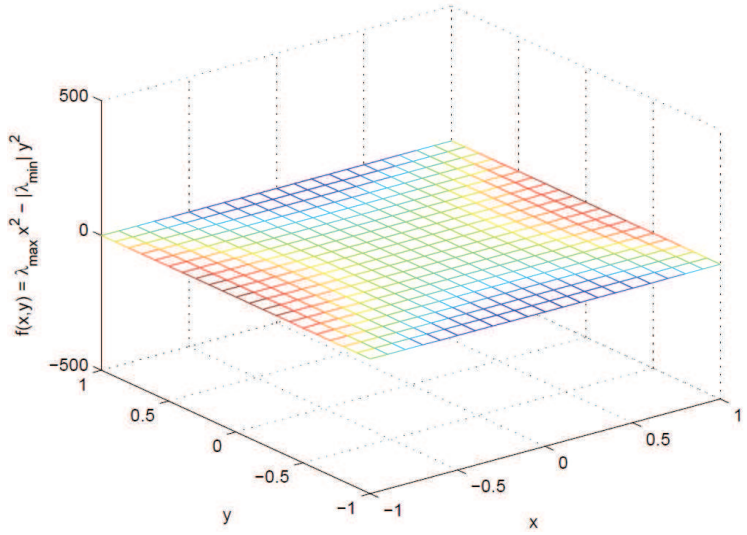
On peut par exemple noter l'utilisation du filtre de Prewitt pour la dérivée d'ordre 1. Nous pouvons également noter l'utilisation de la dérivée seconde dont les problèmes de sensibilité au bruit sont connus. Pour cette raison, Beaudet utilise également un lissage par une gaussienne.

De manière pratique, pour chaque pixel de l'image une mesure de réponse R, proportionnelle au hessien est obtenue par la formule :

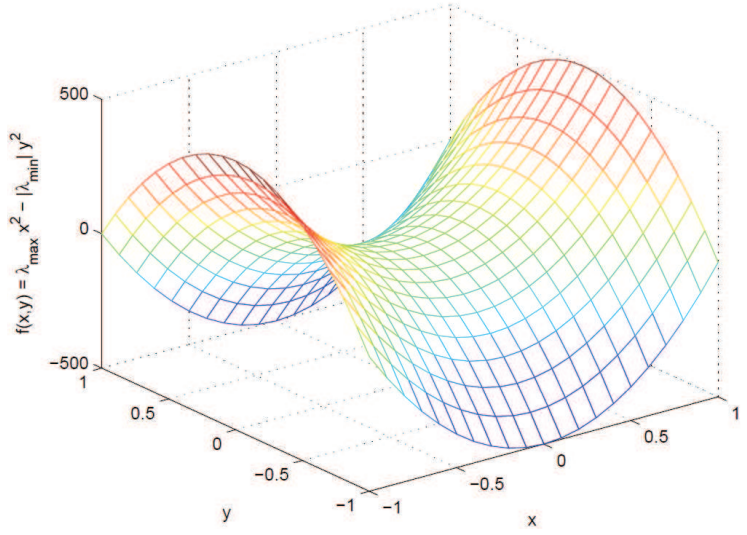
$$R_{Bea78}(x) = C\ det(H(g_\sigma * I(x))), \tag{1.11}$$



(a)λ_{min}=-1,λ_{max}=999



(b)λ_{min}=-1, λ_{max}=1



(c)λ_{min}=-500, λ_{max}=500

Figure 1.5 : Différentes configurations de λ_{min} et λ_{max}

avec C une constante positive et la fonction *det* représente un calcul de déterminant de la matrice hessienne.

En résumé, le calcul de R, basé sur le calcul du déterminant, permet d'avoir des valeurs très négatives (cette valeur est variable car dépendante des plages de valeurs des données traitées) quand les valeurs propres ont conjointement de fortes valeurs et sont en même temps très proches.

De plus, Beaudet fait précéder cette recherche de maxima locaux par un double seuillage :
- $R > T_1$,
- $|\lambda_{max} - |\lambda_{min}|| > T_2$,

afin d'éviter les problèmes de faux points-selles comme le montre la Figure 1.5. Il reste néanmoins le problème de la définition des seuils T_1 et T_2 ; assez récurrent en traitement d'images.

Ces travaux ont été largement repris dans les années suivantes afin d'apporter quelques modifications/améliorations. Nous pouvons par exemple citer les travaux de Dreschler et Nagel [DN82], qui partant de l'équation 1.11, complètent la méthode de recherche de maxima locaux par une recherche de minima locaux. La ligne joignant ces deux extrema peut être utilisée afin de localiser le point d'annulation de la courbure.

En 1983, Nagel [Nag83] démontre que les travaux proposés par Kitchen et Rosenfeld [KR82] sont identiques à son approche de la même année, bien qu'exploitant une méthode légèrement différente par l'utilisation des polynômes bicubiques pour l'approximation de la fonction I :

$$I(x,y) \approx C_1 + c_2x + c_3y + c_4x^2 + c_5xy + c_6y^2 + c_7x^3 + c_8x^2y + c_9xy^2 + c_{10}y^3 \quad (1.12)$$

à partir de laquelle ils calculent :

$$R_{KR82} = \frac{-(c_2^2c_6 - 2c_2c_3c_5 + c_3^2c_4)}{c_2^2c_3^2} \quad (1.13)$$

Nous ne nous attarderons pas sur les détails de ces travaux du fait de la similitude des approches, mais pour être exhaustif dans cet état de l'art nous pouvons citer le travail de Zuniga et Haralick [ZH83] basé sur le travail de Kitchen et Rosenfeld, et dont la principale différence réside sur le mode de calcul de R :

$$R_{ZH83} = \frac{-(c_2^2c_6 - 2c_2c_3c_5 + c_3^2c_4)}{(c_2^2c_3^2)^{\frac{3}{2}}} = \frac{R_{KR82}}{c_2^2c_3^2}. \quad (1.14)$$

Les travaux de Beaudet sont en quelque sorte les précurseurs du développement des points d'intérêt. Ils permettent d'appréhender l'utilité des valeurs

propres (λ) des matrices hessiennes en traitement d'images. Aussi, ils mettent en avant le principe de calcul des réponses R évitant de passer par l'extraction explicite des valeurs propres. Ils permettent également de donner un exemple concret de l'utilité des filtres de convolution pour l'approximation des dérivées de l'image et la détection de contour.

Tout comme les travaux de Beaudet ont inspirés une série de détecteurs de points d'intérêt, nous devons présenter les travaux de Moravec [Mor77] qui ont eux aussi été la base de bon nombre de détecteurs modernes.

1.2.2 Détecteur de coins de Moravec

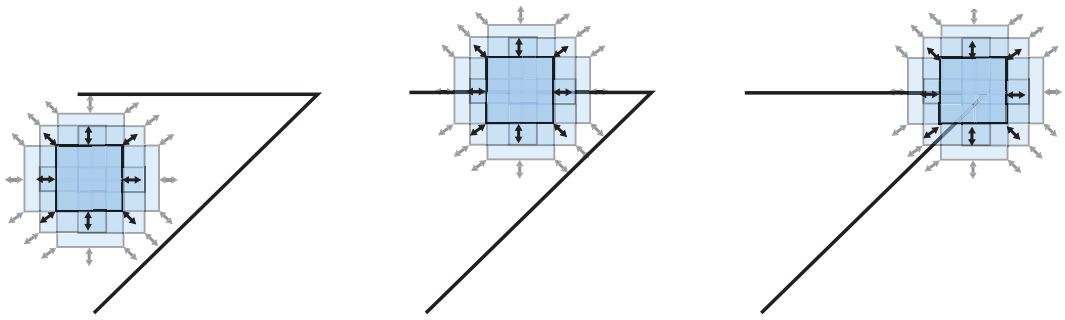
L'idée de base de ce détecteur est d'observer les changements moyens d'intensité locale du signal par des mesures d'[auto][dis]similarité. Plus concrètement, pour chaque pixel, une fenêtre d'observation carrée est considérée autour de celui-ci. Les valeurs d'intensité des pixels de la fenêtre courante sont comparées avec celles de cette même fenêtre, légèrement décalée (d'un pixel) dans toutes les directions d'un voisinage 8 (2 horizontales, 2 verticales, 4 diagonales).

La Figure 1.6 permet d'illustrer plusieurs cas de Figure. Dans le cas (a), la mesure de similarité d'intensité informe qu'il n'y a aucune différence entre la fenêtre considérée et les différents décalages de cette fenêtre. L'intensité apparait donc constante et le pixel considéré appartient à une région uniforme. Dans le cas (b), il y a des changements significatifs quand la fenêtre se décale sur l'axe vertical et diagonal, mais aucun changement sur l'axe horizontal. Dans ce cas précis, le pixel courant appartient à une ligne de contour horizontal. Dans le cas (c), il y a des changements dans toutes les directions considérées. Ce cas est très particulier, et il s'agit d'un pixel de coin. Ce sont ces pixels de coin qui sont considérés comme des points d'intérêt, car un coin est une structure bien définie (au sens mathématique), et ayant une localisation précise.

D'un point de vue plus formel, *Moravec* propose de formaliser la mesure d'[auto][dis]similarité par :

$$E(x,y) = \sum_{u,v} w_{u,v}[I_{x+u,y+v} - I_{u,v}]^2, \tag{1.15}$$

où :
– I représente la composante intensité de l'image,
– w la fenêtre glissante,



(a) uniforme
Aucun changement
dans aucune direction

(b) contour
Pas de changement
dans la direction du contour

(c) coin
Changements significatifs
dans toutes les directions

Figure 1.6 : Mesure du changement local du signal par fenêtre glissante.

– E la variation mesurée pour un pixel en (x, y) produite par un décalage de fenêtre de u, v.

Par cette mesure, ce sont les coins les plus marqués qui ont les dissimilarités les plus importantes. Pour cette raison, un seuillage est appliqué pour supprimer les réponses trop faibles. Cette étape est suivie par une recherche de maxima locaux de E afin d'isoler les meilleurs coins, c'est à dire, les points d'intérêt.

Le principal atout de cette démarche est la faible puissance calculatoire requise, ce qui pour l'époque était primordial. Cependant, ce détecteur souffre de plusieurs problèmes. Le premier est l'anisotropie des réponses et son manque d'invariance aux rotations. La localisation et la quantité de points d'intérêt détectés sur une image et sa variante, ayant subi une rotation, ne sont pas du tout identiques comme l'illustre la Figure 1.7. La raison principale est l'utilisation des décalages discrets de la fenêtre glissante n'opérant que sur des angles de 45°.

Le second problème est lié à sa sensibilité au bruit. En effet, de vrais coins peuvent avoir la même valeur de réponse qu'un pixel de bruit isolé. Ou bien, le bruit a pu légèrement modifier un contour, ce qui peut provoquer l'apparition de fortes valeurs de réponse au milieu de pixels de contour et donc la détection de faux coins.

C'est dans l'objectif de palier aux problèmes du détecteur de Moravec que Harris et Stephens proposent en 1988 [HS88a] leur détecteur de coins et de

(a) image d'origine (b) image après une rotation de 30 °

Figure 1.7 : Résultats de détection des points d'intérêt de Moravec

contours [2].

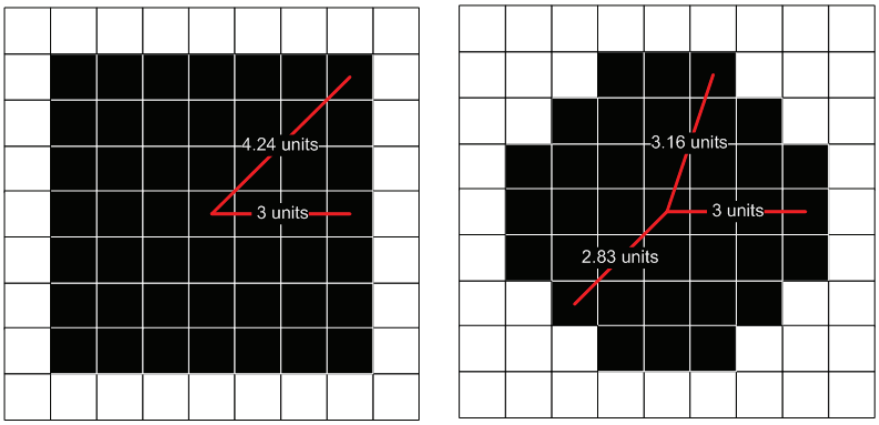
1.2.3 Détecteur de coins de Harris et Stephens

Le détecteur de points d'intérêt de Harris et Stephens [HS88a] (*Harris corner detector*) est l'un des détecteurs les plus connus et les plus utilisés. Bien que relativement ancien, il fait toujours office de référence. Globalement, le principe de base de ce détecteur reste identique au détecteur de Moravec (cf. section 1.2.2), à savoir, l'utilisation de fenêtres glissantes pour des mesures d'[auto][dis]similarité [3] de l'image afin d'isoler les coins.

La première modification apportée consiste à ne pas utiliser une fenêtre glissante carrée binaire, mais plutôt une fenêtre circulaire pondérée. En effet, afin de mesurer avec précision un changement local d'intensité, il est préférable que la distance Euclidienne entre le pixel considéré (au centre) et ceux en bord de fenêtre soit proche dans toute les directions afin de ne pas en favoriser. La Figure 1.8 illustre ce problème de différence de distance Euclidienne entre une fenêtre d'observation carrée, et la version "circulaire". De plus, Nous pouvons

2. Ce détecteur est aussi connu sous le nom d'opérateur de Plessey.
3. Les auteurs parlent d'auto-corrélation, mais de manière pratique il s'agit de la somme des différences carrées et non une réelle corrélation

considérer que les pixels les plus éloignés du centre informent de manière moins précise des variations locales. Pour cette raison, il peut être utile d'effectuer une pondération afin de favoriser et de donner plus de poids aux pixels proches du centre et atténuer ceux en bord de fenêtre. Pour ces deux raisons, Harris et Stephens ont choisi d'utiliser une fenêtre glissante circulaire, pondérée et gaussienne. L'utilisation de cette fenêtre gaussienne a un effet passe-bas et donc atténue la sensibilité au bruit du détecteur.



fenêtre carrée
distance de variation
maximum : 1.24 units

fenêtre "circulaire"
distance de variation
maximum : 0.17 units

Figure 1.8 : Comparaison de distances Euclidiennes de fenêtre glissante de forme carrée et circulaire.

La seconde modification peut être vue dans l'utilisation des développements de Taylor de la fonction d'intensité I à partir desquels il est possible d'effectuer une phase d'identification et une approximation, ce qui conduit à réécrire l'équation 5.1 de Moravec sous la forme :

$$E(X) = x^2(I_x^2 * g_\sigma)(X) + 2xy(I_x I_y * g_\sigma)(X) + y^2(I_y^2 * g_\sigma)(X). \tag{1.16}$$

Ce qui peut également s'écrire sous forme matricielle :

$$E(x, y) = \begin{bmatrix} x & y \end{bmatrix} M \begin{bmatrix} x \\ y \end{bmatrix}, \tag{1.17}$$

où :

$$M = g_\sigma * \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \tag{1.18}$$

Cette matrice M, également appelée tenseur de structure, peut être utilisée afin de mesurer les courbures principales de la fonction E grâce à ses valeurs propres notées λ [4]. Pour Harris et Stephens, trois cas de figures sont notables et dont des exemples graphiques sont donnés par la Figure 1.9 :

- $\lambda_1 \approx \lambda_2 \approx 0$: le pixel appartient à une région uniforme ;
- $\lambda_1 \gg \lambda_2$ (resp. $\lambda_2 \gg \lambda_1$) : le pixel appartient à un contour vertical (resp. horizontal) ;
- λ_1 et λ_2 ont tous deux de fortes valeurs : le pixel est un coin.

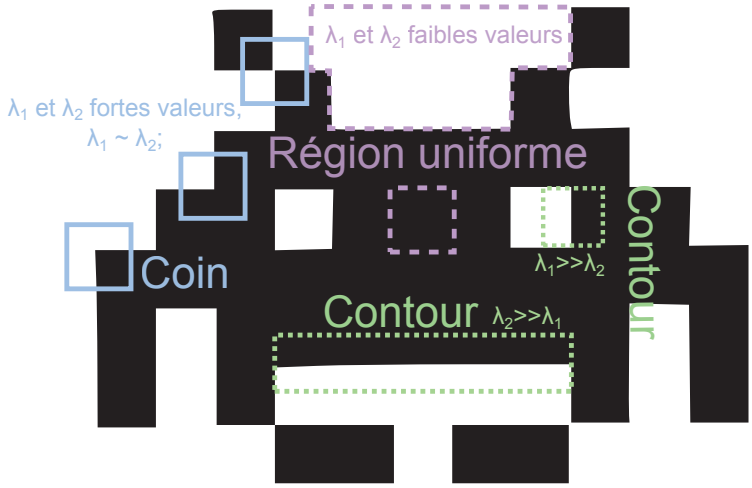


Figure 1.9 : Illustration des différentes valeurs conjointes des λ

Une troisième différence avec les travaux de Moravec se situe dans le calcul d'une réponse R, exprimée par :

$$R_{HS88a} = det(M) - k.trace(M)^2 \tag{1.19}$$

sachant que :
- $det(M) = I_x^2.I_y^2 - I_xI_y.I_xI_y = \lambda_1.\lambda_2$;
- $trace(M) = I_x^2 + I_y^2 = \lambda_1 + \lambda_2$;
- k : un facteur de pondération généralement compris entre 0.04 et 0.06 défini expérimentalement et permettant d'éviter de considérer de très forts contours comme des coins. [5]

4. L'utilisation des valeurs propres pour des mesures de courbures du signal n'est pas sans rappeler les travaux de Beaudet et ses points-selles (section 1.2.1).

5. La valeur 0.04 est très employée dans la littérature. Cependant, Harris lui-même n'a jamais explicitement conseillé cette valeur. La démocratisation de cette valeur est probablement due à celle par défaut utilisée dans l'implémentation Intel pour les applications temps-réel.

Nous pouvons noter que cette formulation de la réponse R permet d'éviter le calcul explicite des valeurs propres en exploitant l'égalité entre le $det(M)$ et $\lambda_1.\lambda_2$ ainsi que celle entre $trace(M)$ et $\lambda_1 + \lambda_2$; ce qui évite la résolution de l'équation du second degré.

La Figure 1.10 permet de juger l'intérêt du calcul de la réponse R. Tout d'abord, il devient très facile d'isoler les pixels de contour ($\lambda_i >> \lambda_j$) en considérant tous les pixels ayant une réponse R inférieure à zéro. De plus, effectuer un seuillage sur les pixels ayant une réponse R inférieure à un seuil T_{uni} permet d'isoler les pixels de régions uniformes ($\lambda_i \approx \lambda_j \approx 0$). Tout comme un dépassement du seuil T_{coin} permet d'isoler les pixels formant un coin suffisamment marqué. Enfin, le réglage de la constante k sert à la définition de ce que l'on considère être un contour, et donc une très grande différence entre λ_1 et λ_2. Le seuil T_{uni} n'a jamais été explicité par Harris qui ne s'intéressait qu'aux coins et non aux régions uniformes. Enfin T_{coin} n'est jamais directement manipulé, car principalement contrôlé par le paramètre k.

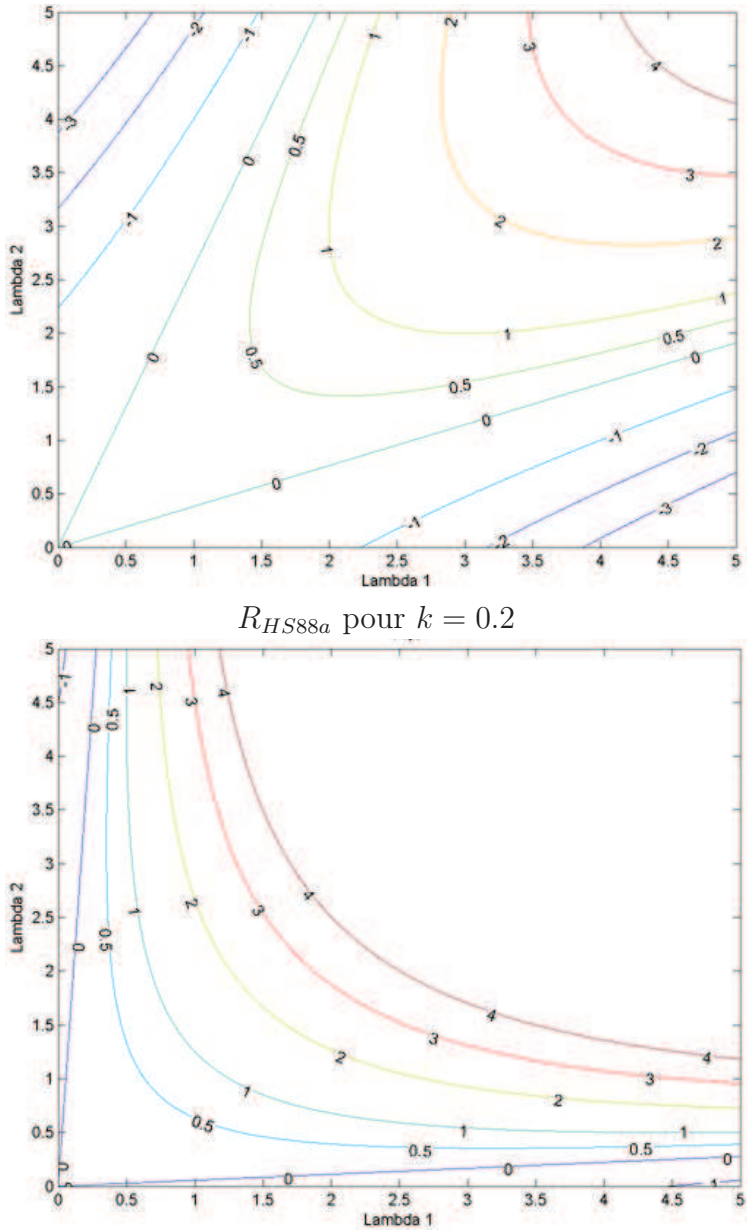
La dernière étape, comme pour tout détecteur de la famille de Moravec, consiste à rechercher les maxima locaux de R après une phase de seuillage par T_{coin} afin d'isoler les meilleurs points d'intérêt.

Pour conclure, les améliorations apportées par Harris et Stephens aux travaux de Moravec ont permis de réduire considérablement les problèmes de sensibilité au bruit et d'invariance aux rotations, sans pour autant les supprimer complètement et bien sûr au prix d'une complexité supplémentaire. De ce fait, de nombreux travaux de la littérature portent soit sur l'augmentation d'invariances ou la réduction de complexité calculatoire. Étant donnée la profusion de travaux en ce sens, il est délicat de faire une liste exhaustive de tous les descripteurs basés sur ce détecteur de coins et d'en expliciter les subtilités. Pour généraliser, les modifications se situent surtout au niveau du calcul de la réponse R. En ce sens, nous pouvons citer par exemple les travaux de Noble [Nob88] ainsi que ceux de Haralick et Shapiro [SH92] qui ont proposés deux variantes du calcul de R :

$$R_{Nob88} = \frac{det(M)}{trace(M)}; \tag{1.20}$$

$$R_{HS93} = 1 - \left(\frac{\lambda_{min}(M) - \lambda_{max}(M)}{\lambda_{min}(M) + \lambda_{max}(M)}\right)^2 = \frac{4\ det(M)}{trace(M)^2}. \tag{1.21}$$

Dans la même lignée, une autre méthode de calcul de R est évoquée par Urban [Urb]. Cependant, elle nécessite le calcul explicite des valeurs propres :



R_{HS88a} pour $k = 0.2$

R_{HS88a} pour $k = 0.05$

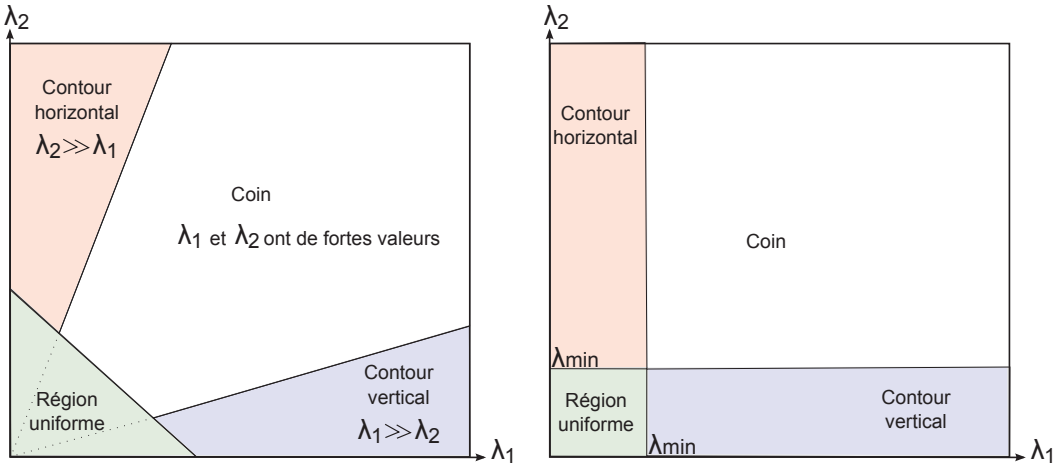
Figure 1.10 : Illustration d'équi-réponse R_{HS88a} pour deux valeurs de k

$$R_{Urb03} = \frac{\lambda_{max}(M) - \lambda_{min}(M)}{trace(M)}. \tag{1.22}$$

Toujours dans les variantes du détecteur de Harris et dans l'objectif de réduire la complexité de calcul et d'augmenter la robustesse du suivi de points, nous pouvons citer les travaux de Shi et Tomasi [ST94] apportant une simplification dans le calcul de R :

$$R_{ST94} = \lambda_{min}. \tag{1.23}$$

Une comparaison des critères de classification de coins peut être faite visuellement entre les critères de Harris et al. et ceux de Shi et al. sur la Figure 1.11. Ce partitionnement par droites orthogonales est bien plus facilement exploitable que le découpage de l'espace de Harris, ne nécessitant que l'utilisation des valeurs max. et min. des valeurs propres.



(a) Harris et Stephens (b) Shi et Tomasi

Figure 1.11 : Classification des configurations du couple (λ_1, λ_2)

Bien que le critère de sélection soit simplifié, les résultats expérimentaux semblent être améliorés en terme de répétabilité et de robustesse de détection. Contrairement aux variantes précédentes et grâce à ses atouts, ce détecteur est assez fréquemment utilisé. [6]

6. La démocratisation de son utilisation est peut être due à son implémentation dans la bibliothèque opencv d'Intel.

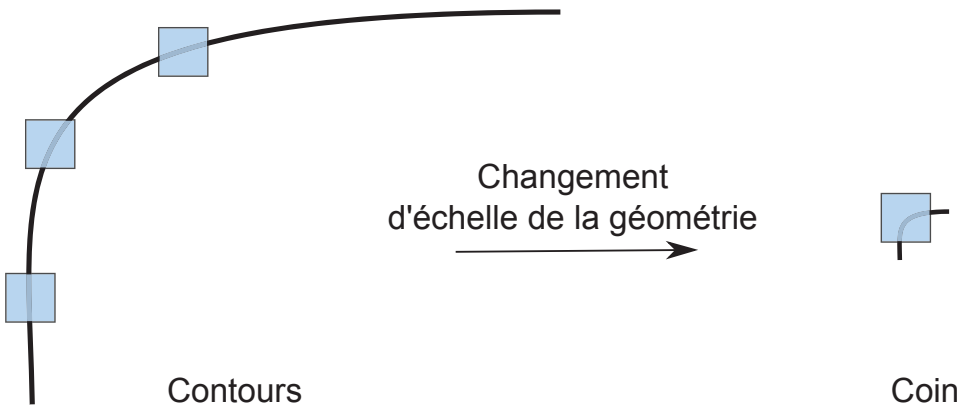


Figure 1.12 : Problème d'invariance aux changements d'échelle

Après cette discussion sur l'éventail de variantes, un problème reste non solutionné, à savoir, l'invariance à l'échelle du détecteur de Harris. La Figure 1.12 illustre un cas où la même géométrie peut être à la fois considérée comme un coin ou une série de contours, selon l'échelle considérée.

1.2.4 Détecteur de Harris multi-échelles

On attend d'un détecteur de points d'intérêt qu'il soit le plus robuste possible, le plus invariant à diverses transformations de l'image, afin de garantir un suivi ou une reconnaissance d'objets dans des configurations variables. Une déformation classique d'un objet dans une image est son changement d'échelle. En effet, si l'on photographie en gros plan une pomme sur une table et que l'on photographie cette même pomme depuis l'entrée de la cuisine, le profil géométrique de la pomme (uniforme circulaire) ne changera pas, c'est simplement l'échelle qui a changé.

Comme l'illustre la Figure 1.12, le détecteur de Harris n'est pas capable de reconnaître un même coin à plusieurs échelles. Il faut une certaine cohérence entre la taille de la fenêtre glissante et la taille du coin.

Pour palier le problème du manque d'invariance du détecteur de coins de Harris, une variante nommé Harris-Laplacian a été proposée par Mikolajczyk et Schmid [MS01]. L'idée est de ne plus considérer l'image dans un plan (x, y) mais dans un espace 3D $(x, y, \textit{échelle})$ où le facteur *échelle* permet de simuler des observations plus lointaines de l'image et ainsi de s'affranchir des problèmes liés à l'échelle. La construction d'espaces multi-échelles de l'image, afin de créer des détecteur de points d'intérêt invariants à ce type de transformation, a été

utilisée par plusieurs auteurs, dont Linderberg et Bay pour les détecteurs SIFT et SURF décrits dans les sections suivantes. C'est pourquoi nous proposons dans la section suivante de poser les bases de la construction de l'espace image multi-échelles. A cette occasion un autre type de détecteur est introduit, à savoir le détecteur multi-échelle de structure blobs.

Construction d'un espace image multi-échelles

La construction d'espaces multi-échelles est récurrente dans de nombreuses problématiques en traitement d'images, aussi bien pour la segmentation, que pour la détection de points d'intérêt, ou même la compression. L'utilité de cet espace vient du fait que le contenu de l'image est de nature multi-échelles. Pour s'en convaincre, il suffit d'observer l'image d'un paysage forestier, dans laquelle nous pouvons distinguer deux masses, le ciel et la forêt. A l'intérieur de cette dernière, il est possible d'identifier des sous éléments tels que de nombreux arbres, chacun d'eux contenant une grande quantité de feuilles. Il y a donc une hiérarchie d'objets et de tailles de structure.

Afin de se faire une idée des traitements devant être appliquées sur une image pour créer cette analyse à plusieurs échelles, il suffit d'observer ce qui se passe lorsque l'on observe une image de plus en plus loin. Tout d'abord, la taille de l'image se réduit de plus en plus, et les détails disparaissent de plus en plus. Cette perte de détails peut s'expliquer physiologiquement par la sensibilité variable du système visuel humain aux différentes fréquences spatiales [7].

Pour reproduire ces effets en traitement d'images, il faut dans un premier temps réduire la taille de l'image en modifiant sa résolution, puis pour supprimer des détails, un filtrage passe-bas est utilisé. Ce filtrage peut bien sûr être réalisé par une convolution avec un noyau Gaussien. La Figure 1.13 permet d'illustrer ce que peut être une image dans un espace multi-échelles, où chaque changement de résolution (division par deux) est appelée octave, et dans chaque octave le filtrage gaussien est de plus en plus important.

Bien que ce concept soit simple, il n'en est pas moins dénué de justification mathématique. En ce sens, nous pouvons citer les travaux de Koenderink [Koe84] et de son espace-échelle gaussien (pour les signaux continus) pour lequel il a posé les axiomes de base nécessaires à ses démonstrations. L'analyse multi-échelle y est définie comme l'application d'un ensemble d'opérateurs T_t qui, appliqués à l'image initiale u_0 définissent un continuum d'images

7. Les études sur ce phénomène sont référencées sous l'appellation de CSF (Contrast Sensitivity Function) [CR68, Dal93]

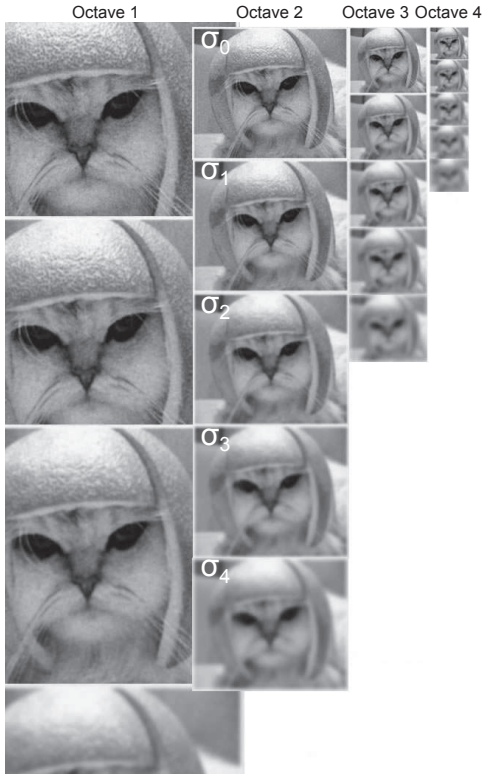


Figure 1.13 : Construction d'un espace multi-échelles

$u(x,t) = T_t(u_0)(x)$. Dans ses travaux, un axiome concerne la récursivité et la structure pyramidale, qui implique que le signal doit être de plus en plus simplifié lorsque l'échelle augmente, et ne doit pas faire apparaître de nouvelles structures. De plus, l'image à une échelle $t+h$, peut être obtenue directement à partir de l'image à l'échelle t sans passer par l'image initiale, par l'application d'un opérateur de transition d'échelle $T_{s,t}$ permettant le passage de l'échelle s à l'échelle t.

$$\forall\ s \leq t, \exists\ T_{s,t}\ /\ T_t = T_{s,t}\ o\ T_s \tag{1.24}$$

Un autre axiome concerne la comparaison locale et implique que si une image I est plus claire qu'une autre image J, alors cet ordre doit être respecté au cours de l'analyse (invariance photométrique). Cet axiome se traduit par l'équation suivante :

$$Si\ I(x) \geq J(x)\ \forall\ x \in V(x_0) \Rightarrow T(I(x)) \geq T(J(x))\ \forall x \in V(x_0) \tag{1.25}$$

où $V(x_0)$ est le voisinage du pixel considéré x_0.

Les invariances gérées par cet espace issu de convolutions gaussiennes ont aussi été démontrées, telles que :

- invariance en translation : $(T(\Delta_x, \Delta_y)g_\sigma * I)(x) = (g_\sigma * T(\Delta_x, \Delta_y)I)(x)$ avec $T(\Delta_x, \Delta_y)$ le vecteur de translation.
- invariance en rotation : $R_\theta g_\sigma(x) = g_\sigma(x\cos\theta + y\sin\theta, x\sin\theta + y\cos\theta)$, $\forall\theta \in \mathbb{R}$ où θ est l'angle de rotation.

D'un point de vue formalisation, cet espace-échelle gaussien est noté L :

$$L(x,y,\sigma) = g_\sigma * I(x,y). \tag{1.26}$$

Pour étudier les évolutions du signal dans cet espace par des dérivées du premier ordre, il peut être utile de décrire les dérivées partielles :

$$L_{x^m y^n}(x,y,\sigma) = \left(\frac{\partial^{m+n}}{\partial x^m \partial y^n} g_\sigma * I\right)(x,y),\ \forall\sigma \in \mathbb{R}, \tag{1.27}$$

sachant que la dérivée du second ordre est souvent elle aussi fort utile pour l'étude du signal, il nous semble important de définir l'espace-échelle laplacien $\nabla^2 L$:

$$\nabla^2 L = L_{x^2} + L_{y^2} \tag{1.28}$$

Par la suite, les travaux de Lindeberg [Lin90],[Lin93] sur la définition d'opérateurs différentiels par différences finies ont permis une approximation discrète de l'espace-échelle, à partir desquels il a établi l'espace-échelle image.

Nous nous devons également de rappeler que cette notion d'espace-échelle de l'image a été propulser par les travaux de Meyer [Mey90] et de ses théories sur les opérateurs et les ondelettes. Nous retrouvons d'ailleurs dans la vie courantes les effets de ses travaux, car c'est sur la base de ses ondelettes que repose aujourd'hui le format de compression JPEG 2000 appliqué au cinéma numérique.

Nous venons de décrire de manière intuitive et formelle, un procédé de création d'espace-échelle par convolution de l'image avec un banc de filtre gaussien. Mais notre but initial est de détecter les structures les plus stables aux changements d'échelles. Il est tout à fait possible de calculer le gradient à chaque échelle de l'image afin de détecter les structures de type contour. Tout comme il est possible d'appliquer le détecteur de Harris pour la détection de coins, ce qui a d'ailleurs été mis en œuvre pour le détecteur Harris-Laplacien. La détection de structure à plusieurs échelles à également été développé par Linderberg, où les structures blobs [8] sont recherchées. C'est également dans ces travaux que Linderberg introduit le concept "d'échelle caractéristique". Son idée a été d'observer l'évolution de la réponse du Laplacien sur l'axe échelle et d'en détecter le maximum local. Son observation a permis de mettre en lumière que les pixels ont une réponse très fortes à une échelle particulière, nommée "échelle caractéristique". Typiquement, les petites structures ont de fortes réponses quand l'écart-type σ est faible, et cette réponse diminue progressivement, à mesure que σ augmente. A l'inverse, les grosses structures ont leur pic de réponse quand σ est important.

Ayant décrit le principe de l'espace-échelle et son emploi avec le Laplacien, il est maintenant possible de résumer plus simplement le détecteur Harris-Laplace. Ainsi, ce descripteur utilise le détecteur de Harris pour son aptitude à localiser les coins de manière précise sur les axes spatiaux. En complément, la robustesse du Laplacien est utilisée dans la détermination de l'échelle caractéristique des structures. Par ces aptitudes combinées, seuls les coins les plus stables sur les différentes échelles sont conservés. Cette méthode en fait l'un des détecteur les plus stables et robustes offrant donc de très bonnes performances en terme d'appariement d'images. Ce détecteur a également été amélioré par la suite en ajoutant des propriétés d'invariance aux transformations affines, dont les détails sont données dans [MS04].

Par ces explications nous avons mis en évidence l'intérêt de l'espace-échelle tout en introduisant les détecteurs de blobs pouvant être considérés comme un nouveau type de points d'intérêt. Dans les sections suivantes, nous proposons d'expliciter les deux versions contemporaines les plus utilisées en vision par

8. Un blob est une région uniforme de l'image de géométrie proche du cercle du fait de la nature et du profil du Laplacien

ordinateur.

1.2.5 Détecteur de blobs multi-échelle : SIFT

La détection de structure blob a été initialement introduite par Linderberg comme expliqué dans la section précédente. Son principe repose sur la création d'un espace-échelle par l'application de noyaux gaussiens de différents écarts-types σ. Par ce traitement, les hautes fréquences sont atténuées, simulant ainsi des distances d'observation dans l'optique de capter différents niveaux de détails dans l'image. C'est la différence entre deux valeurs successives de σ qui définit la finesse de discrétisation de l'espace-échelle. Dans ses travaux, Linderberg utilise un Laplacien dans cet espace afin de faire apparaître les structures blobs de différentes échelles. Le calcul du Laplacien sur des images lissées par des gaussiennes peut être vu comme un filtrage LoG (Laplacien of Gaussian), ce qui est tout de même légèrement coûteux. Le profil de ce dernier est également très similaire à celui d'un filtre DoG (Difference of Gausian). L'équivalence de profils de ces filtres est intéressante car elle peut être utilisée pour réduire la complexité de calcul.

C'est en ce sens que David Lowe [Low99, Low04] a eu l'idée de tirer profit des images calculées en les soustrayant deux à deux pour approximer la DoG obtenant ainsi les résultats de la Figure 1.14.

La formulation du DoG s'écrit sous la forme :

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma), \tag{1.29}$$

où k compris entre 0 et 1, est le paramètre permettant de fixer l'écart de filtrage fréquentiel et ainsi de régler la finesse de discrétisation de l'espace. Dans cet espace, seules les structures comprises entre $k\sigma$ et σ subsistent. A cette échelle, un point d'intérêt défini par (x, y, σ) est un pixel P (représenté par une croix sur la figure 1.15) représentant un extremum local sur son voisinage 26-connexes.

Formellement, l'ensemble est donné par :

$$C_{26} = \{D(x + \delta_x, y + \delta_y, s\sigma), \delta_x \in \{-1, 0, 1\}, \delta_y \in \{-1, 0, 1\}, s \in \{k^{\{-1\}}, 1, k\}\} \tag{1.30}$$

A noter que cette détection d'extrema peut produire un grand nombre de points d'intérêt dont certains sont instables, avec une localisation approximative pour les octaves supérieures, où la résolution devient très faible. Il y a donc

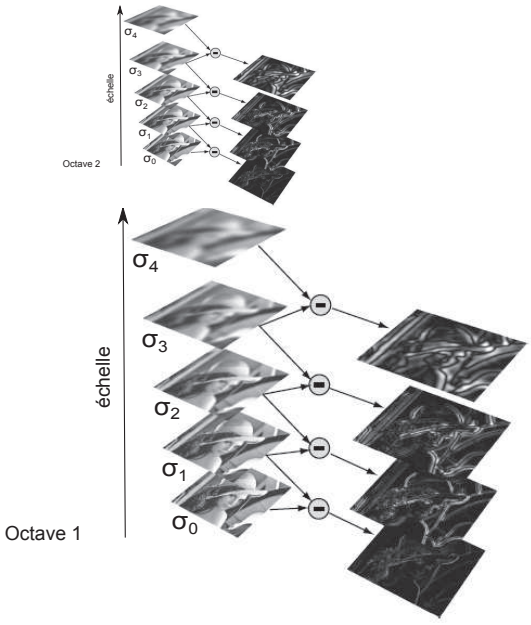


Figure 1.14 : DoG depuis l'espace échelle

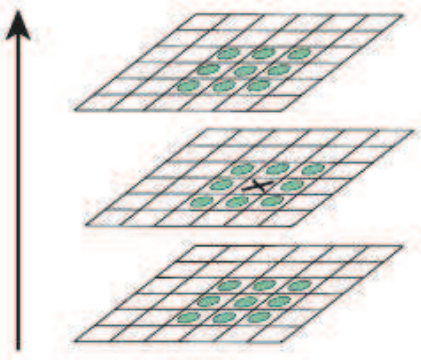


Figure 1.15 : illustration de la 26 connexités

toute une série de traitements permettant de gagner en précision sur (x, y, σ), et d'autres permettant de rejeter les points sur les régions de faible contraste. De par l'utilisation optimisée de l'espace échelle et la stabilité des points obtenus, ce détecteur mérite bien son nom de SIFT pour Scale-Invariant Feature Transform. Nous pourrions nous arrêter là sur ce détecteur, mais son succès vient également du fait des descripteurs qu'il ajoute en chaque point d'intérêt.

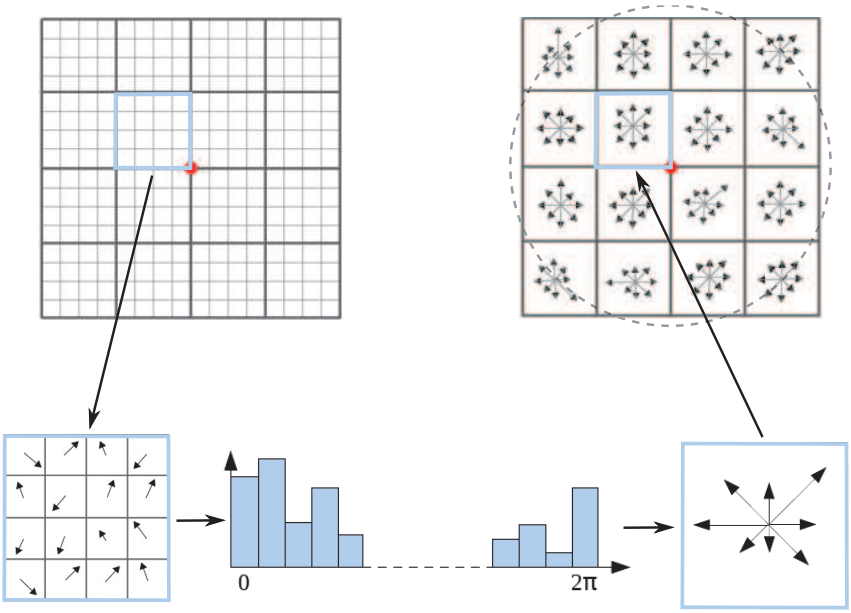


Figure 1.16 : SIFT descripteur par histogramme des orientations de gradient

En effet, afin de garantir un bon appariement des points entre les images et de s'accommoder à un grand nombre de déformations, un vecteur d'information est calculé au voisinage de chaque point d'intérêt. Ce vecteur est basé sur l'orientation des gradients, considérée comme invariante à la rotation. Le voisinage est proportionnel à l'échelle caractéristique du point d'intérêt considéré. Pour illustrer ces propos, la Figure 1.16 donne un exemple de calcul d'histogrammes d'orientations au voisinage du point d'intérêt (cercle rouge) donnant lieu à 4 histogrammes d'orientations de 8 bins, soit un vecteur descriptif de 128 orientations.

La Figure 1.17 permet d'illustrer une détection de points d'intérêt sur une image naturelle. Les cercles jaunes indiquent la localisation des points. Le rayon inscrit dans chaque cercle indique la direction principale des gradients. Les carrés verts permettent de visualiser le voisinage considéré ainsi que le vecteur d'histogrammes d'orientations.

Il peut être également intéressant de noter que ce détecteur se veut bio-

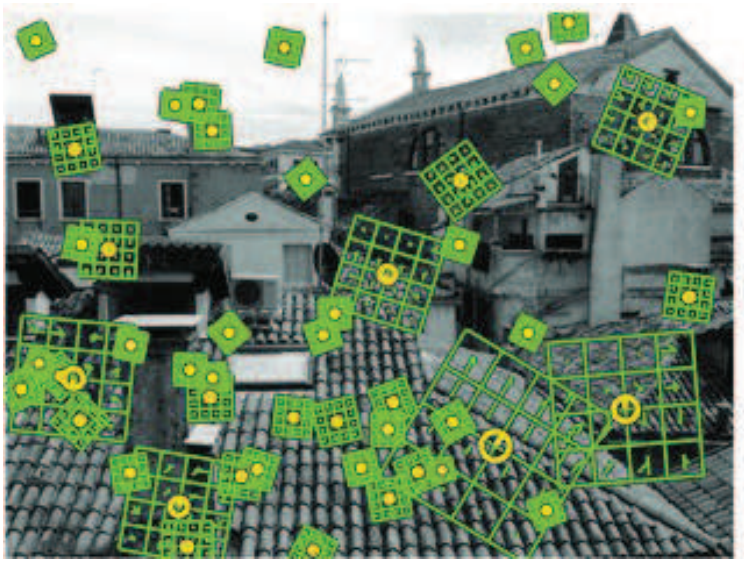


Figure 1.17 : SIFT : exemple de détection de point d'intérêt et de leurs descripteurs sur une image naturelle

inspiré. En effet, Lowe [SKC+05] explique que les structures décrites par son détecteur partagent les mêmes propriétés que certains neurones de la partie inférieure du lobe temporal impliqués dans la reconnaissance d'objets chez les primates.

Grâce à ses nombreux atouts, SIFT s'est rapidement fait un nom parmi les détecteurs de points d'intérêt. Mais son principal défaut est son coût de calcul, quelque peu excessif pour certains traitements en-ligne. C'est principalement ce temps d'exécution qui a été grandement réduit par les travaux de Herbert Bay donnant naissance au récent et très prisé Speed Up Robust Feature, dont les détails sont fournis en section suivante.

1.2.6 Détecteur de Blobs multi-échelle : SURF

Le détecteur de structure blob SURF (Speeded-Up Robust Features) a été introduit en 2008 par Herbert Bay [BTVG06]. Globalement, ce descripteur détecte les mêmes structures que SIFT et se base sur le même principe en utilisant un filtrage DoG. Cependant, l'effort principal dans ces travaux a été mis sur la réduction et l'optimisation du temps de calcul.

La première optimisation concerne justement le filtrage DoG, dans lequel la gaussienne est remplacée par son approximation illustrée sur la Figure 1.18.

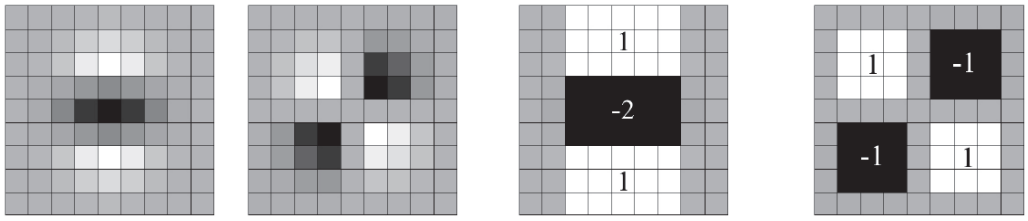


Figure 1.18 : Approximation de dérivée partielle seconde de Gaussienne utilisée dans SURF

Toujours dans le but de réduire le temps de calcul, la construction de l'espace échelle est faite, non pas en modifiant l'image, mais en modifiant la taille du filtre appliqué, comme illustré par la Figure 1.19. C'est donc une pyramide de filtres qui est utilisée est non une pyramide d'images. Mais ce qui permet réellement le gain en temps de calcul dans cette démarche est l'utilisation judicieuse d'images intégrales [9] garantissant un temps de calcul constant quelle que soit la taille du filtre appliqué.

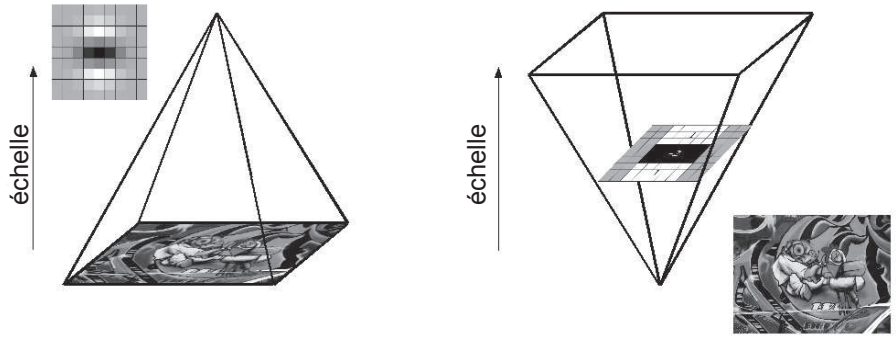


Figure 1.19 : Au lieu de réduire itérativement la taille de l'image (gauche), l'utilisation d'image intégrale permet d'appliquer à temps constant un filtre de taille croissante (droite)

L'image intégrale est une technique algorithmique permettant de calculer rapidement la somme des valeurs de n'importe quelle sous région rectangulaire d'une image. Ce type d'algorithme est très utile pour appliquer des filtres de

9. De manière historique cette méthode a été initialement introduite en 1984 [Cro84], mais son utilisation s'est réellement développée en 2001 par les travaux de re-formulation de Paul Viola et Michael Jones [VJ04].

tailles variables ou pour des calculs d'ondelettes de Harr [NS97]. En pratique, cette technique repose sur deux principes :

Tout d'abord, le calcul initial et unique de ce que l'on nomme image intégrale. Pour ce faire, considérons une image de dimension $L \times H$. L'image intégrale sera de dimension $L+1 \times H+1$ où la première ligne et la première colonne seront de valeurs nulles. Ensuite, chaque pixel de l'image intégrale représente la somme des pixels situés au dessus et à gauche de ce point, comme le montre l'exemple de la Figure 1.20.

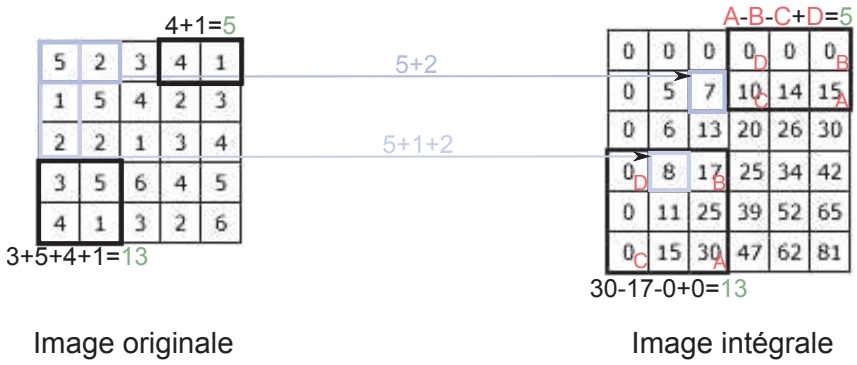


Figure 1.20 : Calculs et utilisation d'images intégrales

Maintenant, il est possible de calculer la somme de n'importe quelle fenêtre rectangulaire à un coût constant, à savoir le coût de trois additions et quatre accès mémoire. Le calcul ($\sum = A - B - C + D$) consiste à utiliser les quatre valeurs de chaque coin de la fenêtre considérée (notés A pour le coin bas droit, B pour le haut droit, C pour le bas gauche et D pour haut gauche) dans l'image intégrale. Deux exemples sont illustrés sur la Figure 1.20 ; permettant de démontrer que la somme des valeurs d'une fenêtre de taille 2×1 nécessite le même nombre d'opérations que pour une fenêtre plus grande (2×2 dans cet exemple). Cet exemple ne valorise pas réellement la méthode, mais cette formule fonctionne avec des tailles de fenêtres quelconques, telles que 1024×2048, ce qui illustrerait sans doute mieux la puissance de l'approche, mais ce qui est plus difficilement représentable sur une figure.

Enfin, tout comme SIFT, un processus de description basé sur l'orientation des gradients au voisinage des points est opéré. Cependant, le descripteur est légèrement différent et exploite lui aussi la puissance des images intégrales. Il repose sur l'utilisation des ondelettes de Haar afin de décrire de la manière la plus invariante possible le voisinage du point d'intérêt, sachant que ce type de filtre peut lui aussi tirer profit de l'image intégrale.

La taille et la pondération du voisinage considérés par SURF sont iden-

tiques à SIFT. Néanmoins, le vecteur descripteur est plus compact (64 valeurs) puisque chaque bloc de 4×4 pixels est décrit par un vecteur v de 4 valeurs issues des réponses des ondelettes de Haar. La Figure 1.21 permet d'illustrer cette description.

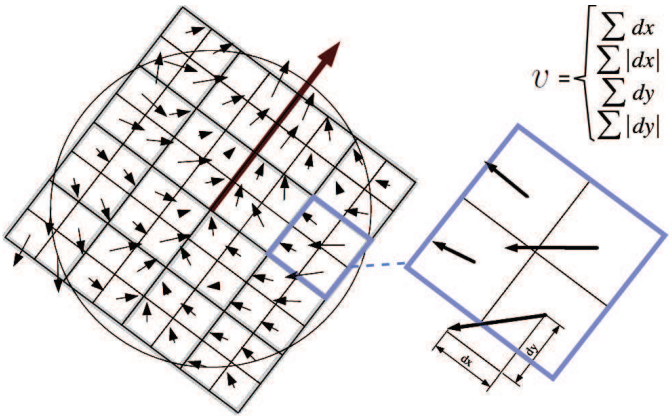


Figure 1.21 : Création du vecteur de description SURF

L'utilisation des ondelettes se révèle très pertinente, dans le sens où des réponses très différentes sont fournies face aux différents types de variation d'intensité rencontrés, comme l'illustre la Figure 1.22.

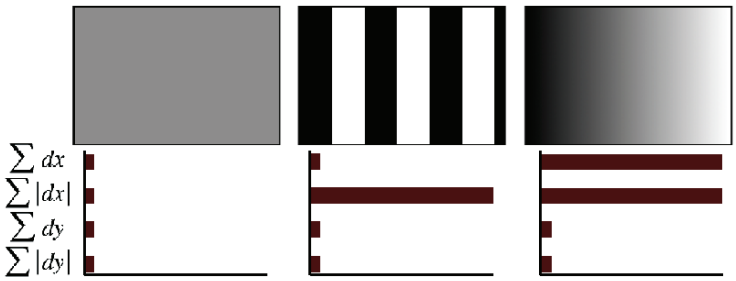


Figure 1.22 : Exemples de réponses caractéristiques du descripteur SURF

Pour conclure, et en complément de la réduction du temps d'exécution de détection et d'appariement (environ 6 fois inférieur), ce détecteur se trouve être plus robuste aux changements photométriques et au bruit, comme l'illustre la Figure 1.23.

De ce fait, le détecteur SURF est actuellement un détecteur de points d'intérêt très utilisé et sans cesse en amélioration, soit en réduisant davantage son temps d'exécution soit en améliorant son invariance.

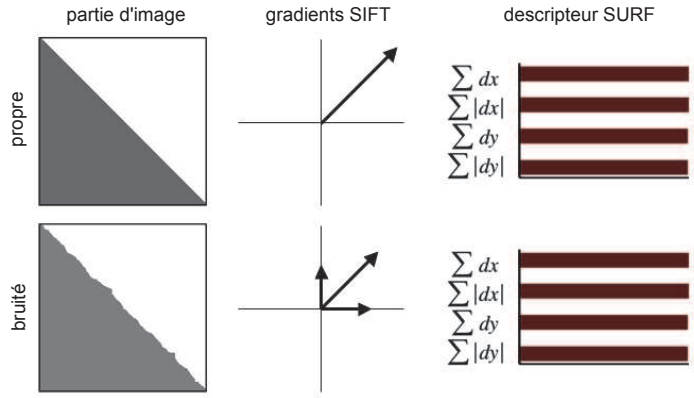


Figure 1.23 : Comparaison de la stabilité des descripteurs

1.3 Conclusion

Nous avons introduit dans ce chapitre la notion d'extraction de caractéristiques structurelles de l'image. Nous avons pu illustrer dans quelle mesure ces informations pouvaient être utilisées pour repérer les pixels les plus pertinents pour répondre aux tâches d'appariement d'images et de reconnaissance d'objets. Les outils développés avec ces objectifs sont regroupés sous l'appellation de détecteurs de points d'intérêt. Nous avons retracé toute la genèse [10] de ces détecteurs allant des points-selles de Beaudet jusqu'aux structures blobs de Bay, en passant par les coins de Harris. Nous avons veillé à mettre en évidence les ressemblances de concepts existant entre les différents travaux de la littérature. La ressemblance la plus flagrante semble être le calcul de réponse R en chaque pixel de l'image accompagnée de sa détection de maxima-locaux.

De nombreux travaux ont porté sur l'optimisation de ces détecteurs afin de les rendre plus rapides et robustes dans les détections et descriptions des pixels singuliers des images. Dans le cadre de cette thèse, nous abordons ces détecteurs sous d'autres regards et de deux manières différentes. Nos premières approches consistent à les exploiter pour leurs fonctionnalités, telle que la détection de points singuliers, afin de prédire la saillance visuelle et estimer la qualité des images. Dans un second temps, nous nous sommes plus focalisé sur leurs capacités descriptives des structures de l'image, et particulièrement l'exploitation des valeurs propres des tenseurs de structure. Par ce biais, nous avons proposé une modélisation des évolutions structurelles de l'image dans le cas d'introduction d'artéfacts associés à leur impact perceptuel.

10. Elle peut être complétée par des travaux récent [GHT11]

En ce qui concerne les performances des détecteurs, elles sont toujours mesurées par leur robustesse d'appariement [MS05]. Nous proposons donc, dans le chapitre suivant, de les évaluer d'une toute autre manière. Notre première contribution vise à quantifier dans quelle mesure ce qui est singulier et important d'un point de vue informatique, l'est également d'un point de vue physiologique. En d'autres termes, les détecteurs de points d'intérêt sont-ils capables de prédire la saillance visuelle humaine.

Chapitre 2

Prédiction de la saillance visuelle par points d'intérêt

Sommaire

2.1 Introduction **41**
2.2 Saillance visuelle **42**
2.2.1 Explication physiologique 42
2.3 Modèle de prédiction de la saillance **52**
2.3.1 Modèle d'Itti et al. 52
2.3.2 Modèle d'Achanta et al. 55
2.3.3 Discussion 56
2.4 Approche proposée : Étude de la relation entre les points d'intérêt et la saillance visuelle **57**
2.4.1 Protocole expérimental 57
2.4.2 Mesures oculométriques 57
2.4.3 Paramètres des détecteurs de points d'intérêt . . . 60
2.4.4 Description de la distance EMD 61
2.4.5 Expérimentation et analyse statistique 64
2.5 PINS (Prediction of INterest Points Saliency) **78**
2.5.1 Mesure de performance 80
2.6 Conclusion **89**

2.1 Introduction

Nous avons précédemment décrit les détecteurs de points d'intérêt comme étant des outils algorithmiques particulièrement adaptés au suivi et à la reconnaissance de formes. Nous avons également soulevé le fait que d'importants

travaux sur ces outils ont porté sur la réduction de leur sensibilité aux différentes déformations de l'image afin d'assurer leurs tâches dans toutes les conditions. Il est intéressant de noter que le système visuel humain est lui même très performant pour ce genre de tâches, car pour l'homme suivre et reconnaître un objet est trivial. Ces objets sont donc très attractifs et saillants pour l'humain. Notre idée est donc d'étudier jusqu'à quelle mesure les points d'intérêt arrivent à décrire la saillance de manière proche de la perception humaine.

Dans cette optique, nous proposons dans un premier temps d'expliquer ce qu'est la saillance visuelle en décrivant les éléments physiologiques impliqués dans son processus. Nous proposons ensuite de décrire les différentes modélisations proposées dans la littérature et servant à mimer et prédire ces phénomènes de saillance. Par ces explications, nous noterons également certaines ressemblances entre les modèles de saillance et les détecteurs de points d'intérêt.

Nous continuerons notre étude en proposant une méthodologie et diverses expérimentations associées à leurs analyses statistiques afin de mesurer et quantifier les liens et ressemblances entre la saillance humaine et les points d'intérêt.

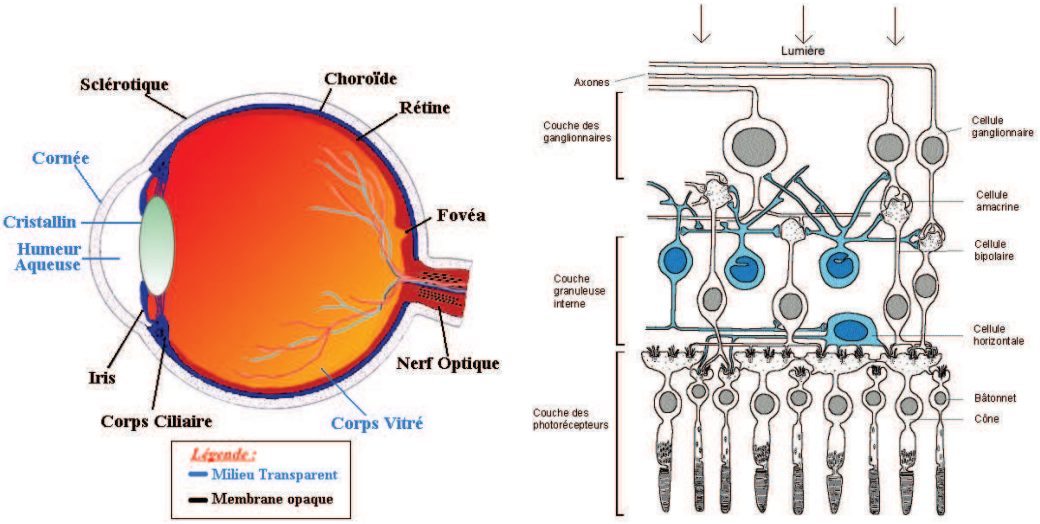
Grâce à ces travaux, nous proposerons également un nouveau modèle de prédiction de saillance utilisant les détecteurs de points d'intérêt tout en fournissant également des mesures de performances détaillées de notre proposition.

2.2 Saillance visuelle

2.2.1 Explication physiologique

Afin d'appréhender au mieux la saillance visuelle, il est nécessaire de prendre connaissance de quelques propriétés et structures de la physiologie humaine impliquées dans le mécanisme de la vision.

Les yeux sont les "capteurs" à même de focaliser et capter la lumière de l'environnement extérieur. Les organes : cornée, cristallin et iris servent à guider la lumière à l'intérieur de l'œil et ont pour objectif d'adapter et focaliser les rayons lumineux vers la rétine, la partie photosensible située au fond de l'œil. Cette rétine est un tissu cellulaire fin, capable de capter les rayons lumineux et de les convertir en signaux nerveux qui, à terme, peuvent être interprétés par le cerveau.



(a) organes de l'œil (b) cellules de la rétine

Figure 2.1 : Schéma structurel de l'œil [Bio]

Rétine

Les cellules de la rétine, sont des neurones organisées en 5 couches comme le montre la Figure 2.1-(b). Bien que les plus profonds, il y a tout d'abord les photorécepteurs, les seuls neurones capables de réaliser la transduction, c'est à dire, la conversion des signaux lumineux en signaux électriques. Ces photorécepteurs se décomposent en deux types, les cônes et les bâtonnets. Les bâtonnets, convergent en grand nombre vers les cellules des couches supérieures. Par ce biais, même en cas de faible luminosité, les cellules supérieures sont capables de fournir des réponses, permettant ainsi la vision crépusculaire et nocturne, mais au coût d'une insensibilité aux longueurs d'onde du spectre visible (et donc à ce qui sera interprété comme une couleur) et d'une faible résolution spatiale. Cependant, les cônes, sont quant à eux adaptés à la vision diurne et sensibles aux différentes longueurs d'ondes des signaux lumineux et donc sensibles aux couleurs. La réponse des cônes est proportionnelle au nombre de photons absorbés, dans une famille de longueur d'ondes particulière. Ils peuvent ainsi être de trois types : les "L" sensibles aux grandes longueurs d'ondes (Long en anglais) qui donneront la sensation dans le rouge, les "M" pour les moyennes longueurs d'ondes (Medium) qui produiront la sensation dans le vert/jaune et les "S" pour les petites longueurs d'onde (Small) qui donneront la sensation dans le bleu. La quantité de chaque type de cônes n'est pas identique. De ce fait, l'humain a des sensibilités et des capacités différentes à distinguer deux couleurs proches. Une différences entre deux bleus est généralement moins bien perçue qu'une différence entre deux verts par exemple. En

ce qui concerne les densités et répartitions, nous pouvons noter que les cônes sont très denses et très nombreux à proximité de la fovéa, tandis que les bâtonnets bien que 20 fois plus nombreux sont réparties sur la périphérie. La fovéa est une petite zone de la rétine située dans le prolongement de l'axe optique de l'œil (cf. Figure 2.1-(a)) où la vision des détails est la plus précise et ce, grâce à la grande densité des cônes en cet endroit et du fait que la convergence vers les couches supérieures y est très faible. Un cône peut être connecté à une seule cellule de la couche supérieure, ce qui garantit une très grande résolution spatiale. La notion de longueur d'onde est perdue après les cônes.

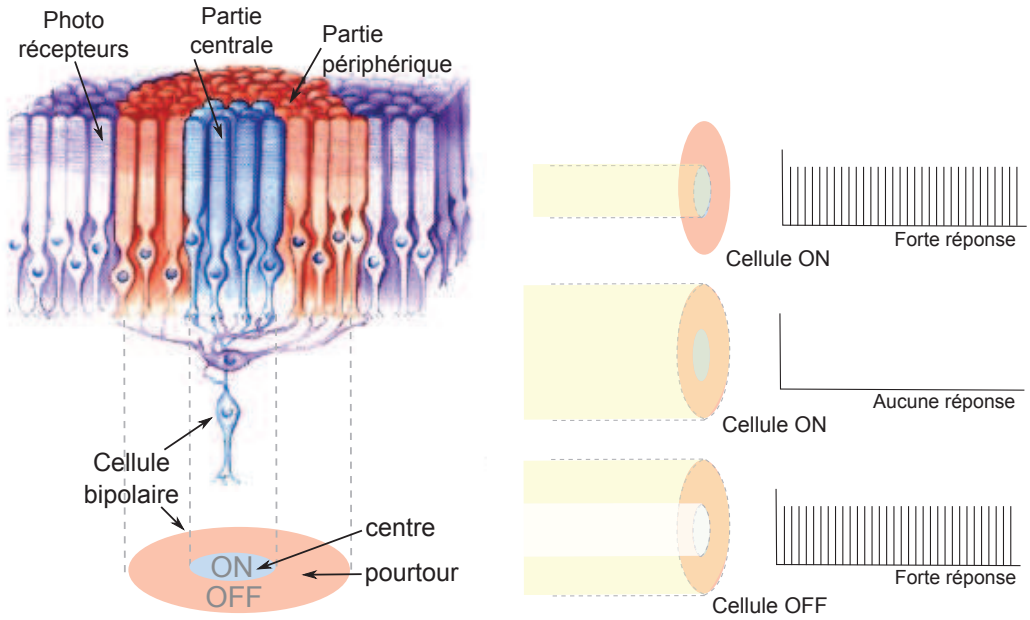
Une fois la transduction effectuée par les photorecepteurs, les informations sont relayées sur les cellules horizontales, amacrines, bipolaires et puis ganglionnaires, comme l'illustre la Figure 2.1-(b). Il y a de nombreux liens de communications entre toutes les couches de la rétine. Nous pouvons noter que les cellules horizontales sont largement connectées entre elles et propagent donc localement les activations des photorécepteurs [M$^+$01]. Par ce biais, des moyennes locales de l'activité des phtotorécepteurs sont effectuées, et peuvent servir de base à des traitements plus complexes pour les couches supérieures.

Autant le fonctionnement des cellules horizontales est actuellement bien compris et identique pour toutes les cellules, autant les cellules amacrines conservent encore beaucoup de mystère. En effet, plus de 30 types ont été identifiés et elles semblent réaliser des traitements complexes et variés [Mas01]. Les connaissances actuelles des cellules bipolaires et ganglionnaires sont également partielles et méritent davantage d'explorations [TS02, MM07] bien qu'étudiées depuis plus longtemps [Kuf53, HW62, HW68].

De manière schématique, les cellules ganglionnaires reçoivent les potentiels électriques des couches précédentes et se chargent d'encoder le résultat de l'ensemble des traitements rétiniens précédents afin de relayer l'information au cerveau, sous forme de potentiels d'action.

D'un point de vue structurel, les cellules ganglionnaires ont des champs récepteurs concentriques et antagonistes, tout comme les cellules bipolaires desquelles elles reçoivent leurs signaux. Ce type de structure est donc constitué d'un centre et d'un pourtour comme l'illustre la Figure 2.2-(a). De par leur réponse aux stimuli, ces cellules sont dites de type ON ou OFF. Une cellule de type ON fournit une réponse maximale quand la partie centrale est stimulée et le pourtour inhibé. Dans le cas d'une stimulation simultanée du centre et du pourtour, aucun signal n'est relayé, à l'opposé des cellules OFF comme nous pouvons le voir sur la Figure2.2-(b).

Certaines de ces cellules sont dites parvocellulaires, avec un champ récep-



(a) champ récepteur concentrique et antagoniste

(b) Réponses des cellules ganglionnaires ON et OFF

Figure 2.2 : Structures et réponses des champs récepteurs ON-OFF (inspiré de [Dub]

teur de petite taille plus adapté à la transmission d'informations chromatiques et des détails de l'image. D'autres, les magnocellulaires, ont des champs récepteurs larges, donc de faibles résolutions spatiales, mais plus impliquées dans la détection de mouvement et, de ce fait, elles sont insensibles à la couleur.

En conclusion, dès la rétine, les signaux visuels sont captés, prétraités, mis en forme et séparés sur différents canaux parallèles afin d'être véhiculés via le nerf optique, jusqu'au corps genouillé latéral pour enfin arriver jusqu'au cortex visuel du cerveau.

Cortex visuel

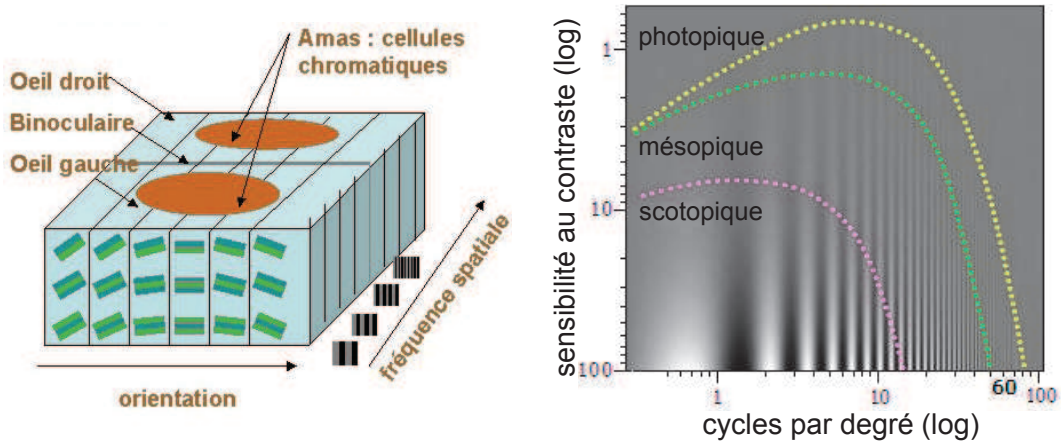
Principalement à l'arrière du cerveau, une zone nommée cortex visuel, pouvant être décomposée en zones fonctionnelles (V1, V2, V3, V4,...), semble être responsable du traitement et de l'interprétation des informations visuelles, allant de la simple détection de contours à la reconnaissance d'une forme pour finir par son interprétation sémantique. La majorité des signaux visuels traités par le corps genouillé latéral arrivent dans l'aire visuelle primaire V1 pour

les premiers traitements. Dans cette aire, la notion de champ récepteur est également importante pour décrire ces neurones. Ces cellules réagissent à des stimuli spécifiques, dont les paramètres sont la taille, la forme, la phase et l'orientation [HW62]. Elles peuvent être classées en différents types :

Les cellules simples : reçoivent les signaux d'une ou plusieurs cellules ganglionnaires ON-OFF. Tout comme ces dernières, elles peuvent être du type ON ou OFF, mais présentent quant à elle des champs récepteurs de forme allongée, avec un centre entouré de deux régions antagonistes. De par leurs structures, elles réagissent particulièrement aux stimuli de la forme d'une barrette orientée dans le même sens que leurs champs récepteurs et sont donc à même de détecter des lignes orientées. En plus d'être sensibles à l'orientation, elles sont également sensibles à la position du signal.

Les cellules complexes : agrègent en entrée les signaux de plusieurs cellules simples. Elles sont sensibles à l'orientation mais de manière indifférente à la position exacte de l'excitation dans le champ visuel. Elles répondent principalement à des lignes de contraste orientées et réagissent souvent à une direction particulière du mouvement.

Les cellules hypercomplexes : appelées également *end-stop*, répondent à des discontinuités comme, créées par la taille d'une ligne, d'une extrémité de contour ou d'une courbure importante [GW89], d'un angle droit éclairé d'un coté et sombre de l'autre, d'un coin... Elles sont donc plus adaptées à la perception des formes. Plus de détails sur les propriétés des différentes cellules de V1 sont disponibles dans [HM06].



(a) Architecture en hypercolonnes (b) Sensibilité aux fréquences spatiales

Figure 2.3 : Perception des signaux et des fréquences spatiales

D'un point de vue organisation, ces différents neurones forment des hypercolonnes en regroupant les différents types d'informations visuelles provenant d'une même région spatiale et ce pour différentes fréquences, comme l'illustre la Figure 2.3-(a). Ces informations différentes et complémentaires sont principalement l'orientation, la position de l'œil (gauche ou droit), l'opposition des couleurs [KSJ00], la direction principale du mouvement et la fréquence spatiale. En ce qui concerne cette dernière, de nombreuses études [Dal93, FMLBR05, RLFM08] ont porté sur la mesure de la sensibilité de l'humain donnant lieu à une modélisation appelée fonction de sensibilité au contraste (CSF : de l'anglais Contrast Sensitivity Function) et variant selon la fréquence, la nature des canaux, l'orientation et les niveaux d'éclairement. Une illustration de la sensibilité aux différentes fréquences et niveaux de clarté est visible sur la Figure 2.3-(b), où l'on peut observer une sensibilité réduite en vision scotopique.

En ce qui concerne la structuration de cette aire visuelle, il est intéressant de noter que la place dédiée au traitement des informations provenant de la fovéa occupe la moitié de l'aire V1 et que l'autre moitié est dédiée quant à elle au reste du champ visuel.

Enfin, les signaux traités par V1, sont ensuite transmis vers de nombreuses autres aires corticales (V2, V3, V4, MT...), largement interconnectées entre elles, pour divers traitements. Cependant, de par la grande complexité des liens et des communications, comme l'illustre la figure 2.4, l'interprétation des signaux visuels devient très délicate à expliquer. Ces parties relèvent d'ailleurs plus des traitements haut niveau, capables à terme, d'associer une sémantique et un sens aux différentes formes détectées.

Nous venons de décrire les éléments biologiques mis en œuvre dans le SVH, partant de la réception de signaux lumineux jusqu'à leur interprétation sémantique par le cerveau. De par la structure même de la rétine avec sa fovéa et des quantités de neurones dans l'aire V1 du cerveau, il apparaît clairement que les aptitudes de la vision ne sont pas égales entre le centre et la périphérie. De plus, la quantité d'informations visuelles d'une scène devant être traitée à chaque instant est trop importante pour être analysée dans son intégralité. Le regard humain est donc contraint de se déplacer vers des régions particulières d'une scène, les régions censées être les plus porteuses d'informations utiles. Ces régions d'importance sont appelées régions saillantes de par leur nature attractive. Afin de comprendre plus en détails le fonctionnement du SVH par rapport à la saillance visuelle, des techniques expérimentales ont été mises en œuvre afin de mesurer et d'interpréter les mouvements oculaires. La section suivante permet de décrire les méthodologies utilisées pour mesurer l'activité oculaire et la saillance visuelle.

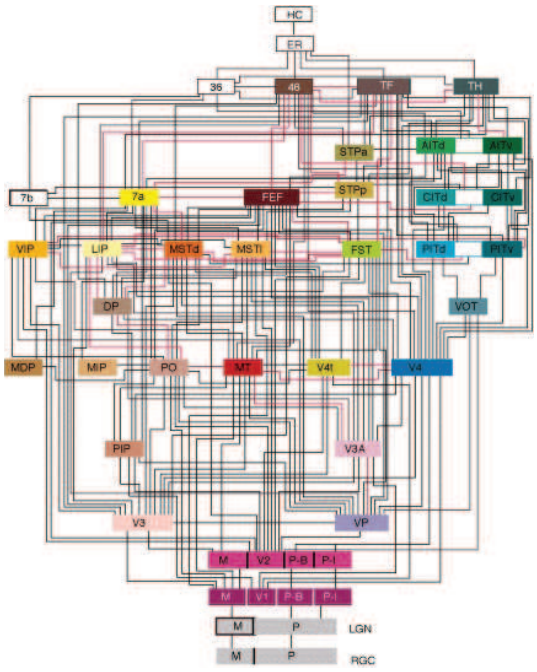


Figure 2.4 : Illustration de la complexité des liens et échanges entres les aires corticales pour le traitement de la vision [FC75]

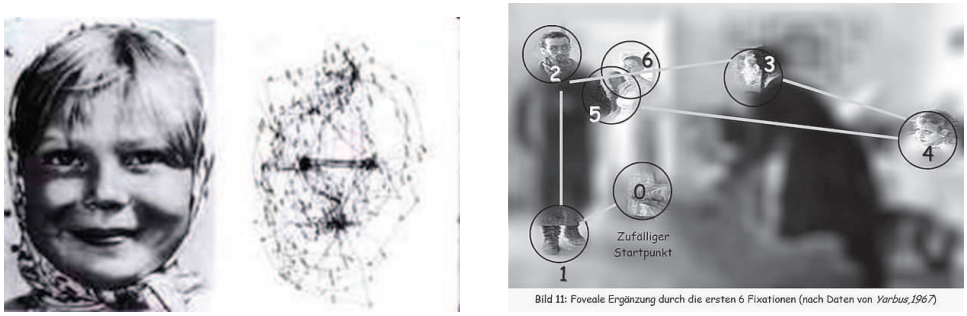
Oculométrie : méthodologie de mesure des mouvements oculaires

La mesure de l'activité des yeux, des mouvements oculaires, est assurée par des techniques d'oculométrie (*eye-tracking* en anglais). Il existe plusieurs technologies capables d'observer ces déplacements du regard. Certaines techniques utilisent des électrodes disposées autour des yeux, afin de mesurer leurs rotations, d'autres utilisent des lentilles spécialement conçues posées sur l'œil. Mais la technique la moins invasive, la moins contraignante et la plus utilisée consiste à filmer le regard des observateurs par l'utilisation d'une caméra. Pour être plus précis, des diodes infrarouges sont placées sur ou sous l'écran, afin d'"éclairer" la pupille du participant. Par ce biais, il est possible de suivre le déplacement du regard lors de la visualisation d'une image, d'une vidéo ou de l'utilisation d'une quelconque application.

Grâce à cette métrologie, différents types de mouvements de l'œil ont pu être décrits [Wid84] :

Les saccades : sont des mouvements très prononcés et rapides qui ont pour but de mettre une région précise de la scène observée au centre de la fovéa. Le mouvement d'une saccade ne prend qu'entre 30 et 80 ms et peut atteindre une vitesse angulaire de plus de 900 deg/s. Après ce type de mouve-

ment, l'œil s'immobilise pendant un temps variable compris entre 250 et 500 ms. Ce temps d'arrêt est appelé fixation. Le point précis où se focalise le regard est généralement appelé **point de fixation** (*gaze point* en anglais, noté **GP**). Ces GP sont extrêmement intéressants et étudiés car ils indiquent la région précise de l'image que l'œil est en train d'analyser, en utilisant toute l'acuité visuelle de l'observateur. La figure 2.5 permet d'illustrer le déplacement oculaire en ne représentant de manière nette que les régions ayant été observées spécifiquement lors de l'analyse d'une scène complexe. C'est l'étude des points de fixation de nombreux observateurs qui permettent de décrire et comprendre le comportement de l'humain vis à vis de la saillance visuelle.



(a) Observation d'un visage par oculométrie

(b) Illustration de parcours oculaire

Figure 2.5 : Visualisation des saccades et fixations (issus des travaux de Yarbus [Yar67]).

Les mouvements de poursuite continue : sont les mouvements qui permettent de suivre un objet en déplacement. Quand les yeux suivent convenablement l'objet, la fovéa reste centrée sur lui, ce qui permet au SVH d'en extraire un maximum d'informations. Contrairement aux mouvements brusques des saccades, ces mouvements sont doux et continus, d'une vitesse maximum de 100 deg/s.

Les micro-saccades : sont des mouvements infimes effectués constamment. C'est par ce biais que le "rafraîchissement" de l'image est effectué en forçant les photorécepteurs à s'activer. Pour s'en convaincre, la figure 2.6 est une illusion d'optique assez connue, mettant en avant le phénomène de non rafraichissement et d'épuisement des cellules.

L'un des premiers à avoir étudié les mouvements des yeux et l'oculométrie est Alfred Yarbus [Yar67]. Dans ces études, il a mis en avant le fait que l'œil humain a tendance à se déplacer vers des zones très spécifiques de l'image lors de l'observation de scènes simples et complexes. Il a également mis en avant le fait que les déplacements oculaires peuvent dépendre des tâches données aux

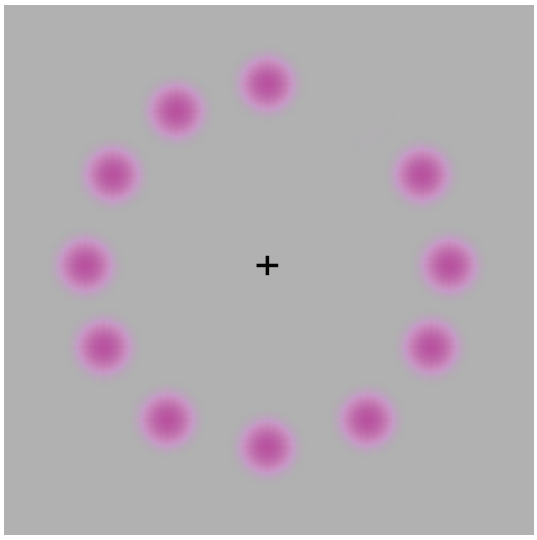


Figure 2.6 : Concentrez-vous seulement sur la croix noire. Les points roses aux alentours disparaissent peu à peu.

observateurs et donc révéler que la saillance oculaire peut être guidée par des processus intentionnels ou non. En observant la Figure 2.7, il est possible de comparer les mouvements oculaires sans et avec plusieurs consignes différentes et ainsi visualiser facilement l'impact de ces dernières.

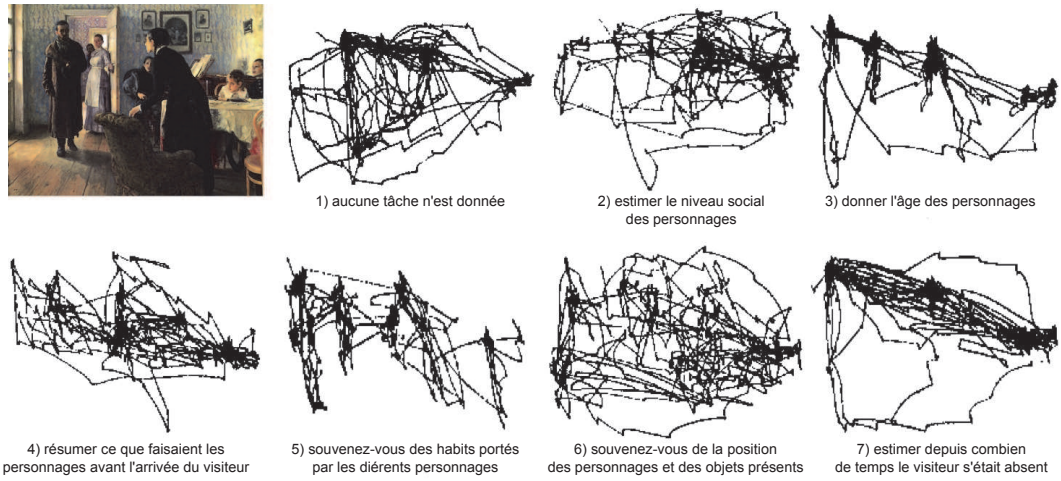


Figure 2.7 : Influence du contexte sur le mouvement oculaire [Yar67].

Ces observations soulèvent le fait que la saillance visuelle peut être considérée sous deux aspects :

Les processus top-down, descendants (du cerveau vers les muscles oc-

culomoteurs), sont des mécanismes endogènes [1] faisant intervenir la volonté du sujet. Ces mécanismes sont liés à un contexte particulier, la réalisation d'une tâche précise et donc fortement orientés par la sémantique du stimulus et les différentes expériences passées des observateurs. Ce sont donc des traitements haut-niveau qui interviennent.

Les processus bottom-up, ascendants (de la rétine vers le cerveau), sont des mécanismes exogènes [2] guidés uniquement par la nature des stimuli présents dans la scène. Il n'y a donc pas réellement de volonté de la part de l'observateur de déplacer sont regard sur ces points précis. Ce sont plus particulièrement les caractéristiques des signaux, tels que le fort contraste, la taille, la forme, la couleur ou la texture qui influencent et guident le parcours des yeux. Ce sont des traitements plus bas-niveau qui interviennent dans ce cas. Ce type de processus est généralement relié à l'idée que le système visuel est guidé par un processus pré-attentif, intervenant avant les mécanismes descendants. Les premiers travaux et également les plus reconnus et servant de base à d'autres plus contemporains, sont les études de Treisman [TS85, Tre85, Tre91] qui défend que certains traits caractéristiques sont détectés de façon très rapide et très fiable par le SVH.

Pour conclure, par sa capacité à mesurer la saillance visuelle, l'étude du déplacement oculaire est utilisée dans de nombreux domaines, tels que la psychologie, les sciences cognitives, le marketing et plus récemment pour des études d'ergonomie de logiciels et de site internet. Cependant, pour disposer de données fiables, de nombreux observateurs sont nécessaires, et comme toute expérience psychovisuelle, d'importants investissements en temps et en argent sont nécessaires. De plus, ce genre de pratiques ne permet pas les traitements en ligne. C'est pourquoi, de nombreux travaux visent à comprendre et à modéliser la saillance visuelle afin de développer des algorithmes capables de prédire en temps réel les déplacements et focalisations de l'œil humain lors de l'observation d'images ou de vidéos.

La section suivante décrira les différents modèles proposées dans la littérature pour la prédiction de la saillance visuelle et ce en s'appuyant sur l'utilisation de données oculométriques.

1. qui prennent naissance à l'intérieur d'un corps, d'un organisme, d'une société, qui est dû à une cause interne

2. ce qui est extérieur à un système

2.3 Modèle de prédiction de la saillance

Grâce à l'observation et à l'étude des mouvements oculaires, de nombreux travaux ont cherché à prédire et à modéliser la saillance visuelle. Nous avons vu qu'il existe deux manières d'appréhender la saillance. Tout d'abord, le top-down implique l'utilisation d'attributs sémantiques, et donc très adaptée à des contextes applicatifs particuliers. De ce fait, peu de modèles descendants sont présents dans la littérature. Nous pouvons tout de même citer les travaux de [HS04, MGPB+05, AHR05] utilisant des ontologies et visant le contexte des réunions en visio-conférence.

Au contraire, les modélisations bottom-up sont relativement nombreuses et se veulent plus invariantes au contexte. Cependant, nous pouvons noter une grande ressemblance entre tous les modèles proposés. La quasi-totalité des travaux, s'appuie sur la Feature Integration Theory (FIT) de Treisman [TG80] et la Guided Search model (GS) de Wolfe [WCF89]. Selon ces théories, la scène observée peut être décomposée en attributs visuels ou traits particuliers simples tels que des variations brutales d'intensités, des oppositions de couleurs, d'orientations, de mouvements, etc. et ce à plusieurs échelles. Chaque type d'attributs est traité en parallèle pour fournir une carte par type de caractéristiques puis combiné pour orienter le regard. C'est d'ailleurs sur le système de fusion qu'intervient la différence entre les approches FIT et la GS. A partir de ces connaissances [KU85], et un besoin de concision, nous proposons de décrire le modèle d'Itti, faisant office de référence dans le domaine.

2.3.1 Modèle d'Itti et al.

Le détecteur de saillance de Itti & Koch [IKN98] est un modèle bio-inspiré du système visuel des primates.

Dans ce modèle, neuf échelles spatiales sont considérées à travers des sous-échantillonnages et filtrages passes-bas de l'image d'entrée. De cet espace, trois différents types de structures locales sont détectées afin de produire différentes cartes de saillance. Ainsi, 42 cartes sont produites, dont 12 consacrées à l'étude des structures colorées, 6 pour la clarté et 24 pour l'analyse des directions des structures. Pour l'étude de la couleur, quatre cartes sont générées par l'extraction du rouge, du vert, du bleu et du jaune. L'analyse des orientations, s'effectue par un banc de filtres de Gabor orientés (0°, 45°, 90°, 135°) qui est utilisé afin de mimer les réponses des champs récepteurs des neurones sélectifs en orientation dans V1. Pour chaque type de caractéristiques, une carte de

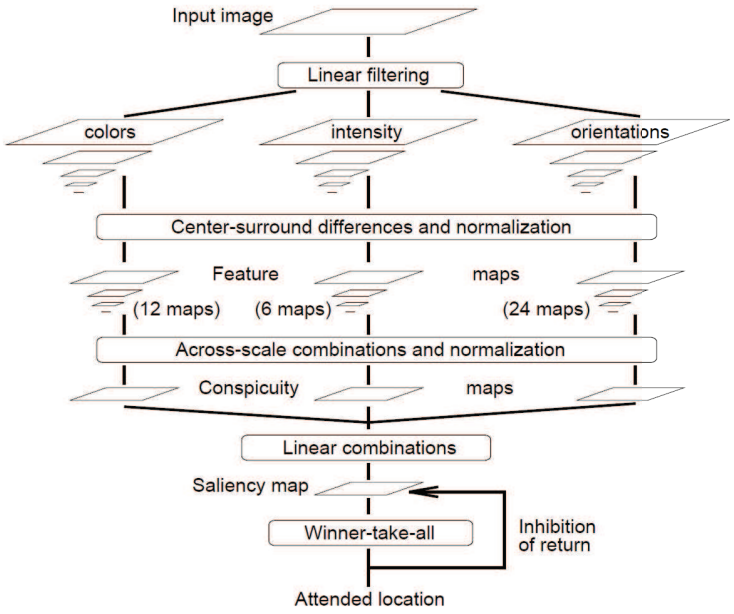


Figure 2.8 : Architecture du modèle de Itti [IKN98].

visibilité (*conspicuity* en anglais) est générée par une normalisation et une détection de maxima locaux, afin de favoriser les régions les plus différentes de leur voisinage reproduisant les mécanismes du "centre-pourtour". Les trois cartes de visibilité sont combinées par une moyenne. Une boucle de détection et suppression des zones les plus saillantes est réalisée dans une phase finale afin de prédire une sorte de parcours oculaire. Cette boucle applique le processus de *"winner-take-all"* du réseau neuronal et l'inhibition de retour, consistant à dire que si une zone a été précédemment visualisée, il ne sera pas prioritaire de s'y re-focaliser. L'architecture globale de cette modélisation est illustrée sur la Figure 2.8. Quelques améliorations sur ce modèle ont été faites par Itti en 2009, en prenant en compte l'extraction d'attributs de mouvement, avec pour principal but d'être capable de traiter les vidéos en plus des images.

Sur les mêmes bases et pour introduire plus de propriétés du SVH, Le Meur [LMLCBT06] dont le modèle est visible en Figure 2.9, propose de remplacer l'espace RVB, par l'espace couleur perceptuel de Krauskopf [DKL84] mimant l'antagonisme des couleurs effectué par les cellules de la rétine. Sur chaque canal couleur, une fonction de sensibilité au contraste est appliquée afin de reproduire la sensibilité spatiale du SVH suite à un passage dans une transformée Cortex [Dal94].

Malgré l'ajout de ces briques bio-inspirées, les gains en terme de performances sont très faibles. Il apparaît que la recherche de mimétisme parfait

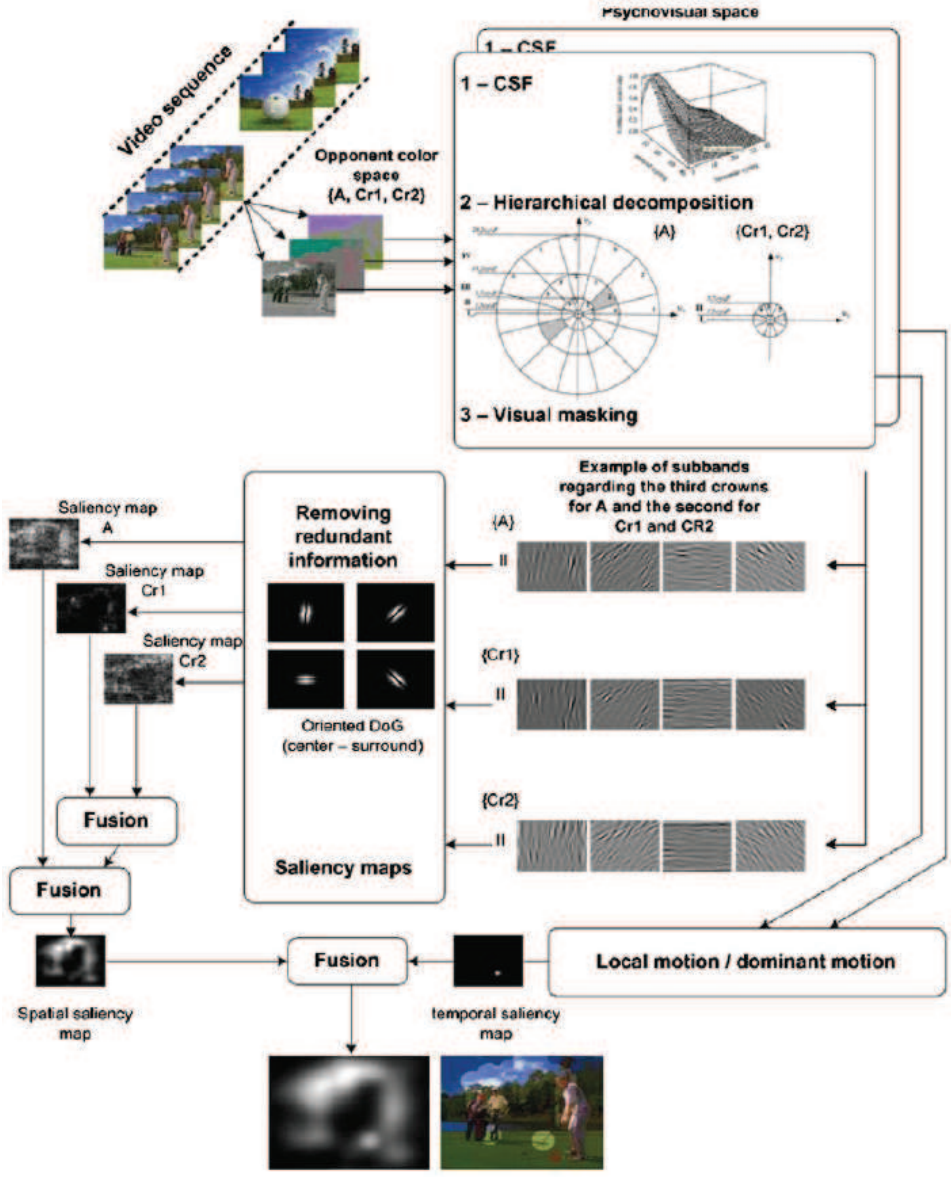


Figure 2.9 : Architecture du modèle de Le Meur [LMLCBT06].

n'est pas toujours justifiée si l'on considère la complexité calculatoire ajoutée en comparaison des gains en performance apportés (+ 0.04% de corrélation).

La complexité et les temps de calcul sont souvent pris en compte lorsqu'il s'agit d'implanter ces mécanismes dans des cas d'usages nécessitant le temps-réel ou presque. Nous proposons donc, dans la section suivante, de décrire le modèle proposé récemment par Achanta et al, dont la notoriété et l'intérêt majeur reposent sur sa faible complexité et sa facilité d'implémentation.

2.3.2 Modèle d'Achanta et al.

Achanta et al. [AHES09a] ont proposé un modèle de saillance basé sur l'utilisation des informations de clarté et de couleur de l'image. Les avantages de ce modèle mis en avant par les auteurs sont : "d'accentuer le plus gros objet saillant, d'isoler l'intégralité de l'objet saillant et non qu'une partie tout en étant précis sur les contours, être peu sensible aux artéfacts de bruit et de compression, et enfin fournir une carte de saillance de même résolution que l'image d'entrée".

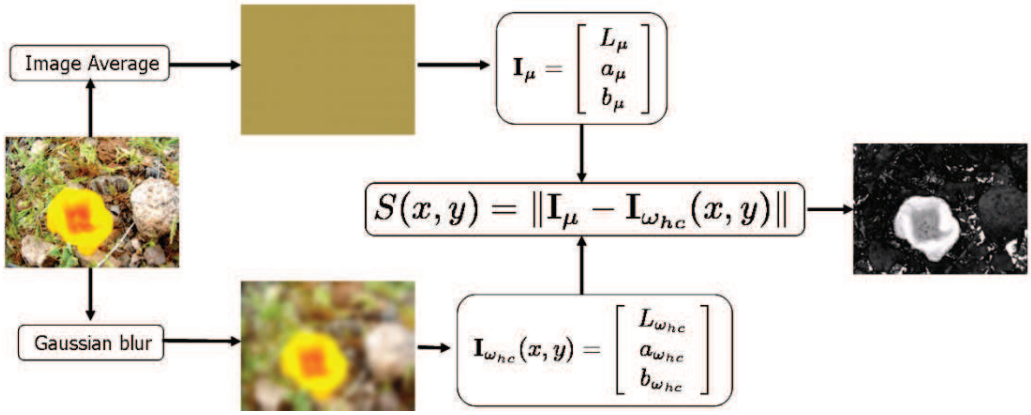


Figure 2.10 : Architecture du modèle de Achanta [AHES09a].

L'architecture de ce modèle est donnée par la Figure 2.10. En considérant l'image d'entrée dans l'espace couleur CIELAB, la carte de saillance S est définie par :

$$S(x,y) = ||I_\mu - I_{w_{hc}}(x,y)|| \tag{2.1}$$

où I_μ est la moyenne sur chaque canal de l'image et $I_{w_{hc}}(x,y)$ est le pixel à la position (x,y) de l'image d'entrée lissée par un filtre gaussien. Bien qu'extrêmement simple, ce modèle apparaît relativement compétitif en comparaison avec l'état de l'art et ce pour une très faible complexité calculatoire.

2.3.3 Discussion

Nous venons de décrire des modèles de prédiction de saillance, dont certains sont très profondément bio-inspirés et complexes tandis que d'autres sont relativement simples. Notre idée d'origine est d'exploiter les détecteurs de points d'intérêt pour prédire la saillance visuelle.

De par leur importante optimisation, ce type d'outils pourrait être classé d'un niveau de complexité calculatoire intermédiaire. Il est également intéressant de noter que certaines ressemblances peuvent être remarquées entre ce type de détecteurs et les modèles de saillance. Par exemple, ces derniers de saillance font des analyses à plusieurs échelles pour imiter la décomposition effectuée par les hypercolonnes de V1. Ceci peut être vu comme équivalent à la construction d'espace-échelle des détecteurs SIFT, SURF et Harris-Laplace. Nous pouvons également noter l'utilisation de détection de maxima locaux et l'utilisation du filtre de Gabor dans les modèles de saillance afin de mimer le comportement des cellules centre-pourtour. Ces détections de maxima locaux sont elles aussi effectuées dans tous les détecteurs de points d'intérêt. De plus, le profil des filtres de Gabor présente la forme du chapeau mexicain, tout comme les filtres LoG ou DoG utilisés dans SIFT et SURF.

Il peut également être intéressant de souligner que les cellules horizontales du SVH présentent une architecture particulièrement adaptée pour répondre à des signaux de type blob, tout comme les types de structures détectées par SIFT et SURF. Rappelons également que les cellules complexes du SVH sont impliquées dans les tâches de détection de contour, et que le cellules hypercomplexes répondent particulièrement aux fortes courbures et structure de coins, ce qui peut être facilement mis en parallèles du détecteur de Harris lui aussi spécialisé dans la détection de coins et de contours.

Enfin, les modèles de saillance intègrent des estimateurs de mouvement pour la prédiction de saillance dans les vidéos. Nous devons rappeler que les détecteurs de points d'intérêt ont été initialement conçus pour des tâches de suivi d'objets en mouvement. Il est donc tout à fait envisageable d'exploiter cette capacité des détecteurs de points d'intérêt pour prendre en compte la dimension temporelle.

2.4 Approche proposée : Étude de la relation entre les points d'intérêt et la saillance visuelle

Nous venons de décrire certains des éléments biologiques du SVH impliqués dans la saillance visuelle tout en mettant en exergue leur modélisation et leur approche algorithmique. Au cours de ces explications, nous avons pu noter certaines ressemblances entre des propriétés du SVH et les détecteurs de points d'intérêt, qui eux-mêmes partagent également certaines similarités avec les modèles de la littérature.

Nous proposons dans cette section de mesurer et quantifier dans quelle mesure les détecteurs de points d'intérêt initialement développés pour être robustes aux déformations des images, pourraient prédire la saillance visuelle dans une image. L'idée sous-jacente est que s'il existe une grande ressemblance, il serait alors possible de produire un nouveau modèle de prédiction de saillance facilement implémentable, et peu coûteux en temps de calcul, car les tâches de développement et d'optimisation ont déjà été réalisées.

2.4.1 Protocole expérimental

Pour mesurer la similarité, il faut disposer de données oculométriques fiables. Il est également utile d'utiliser différents détecteurs, sachant que chacun d'entre eux peut produire divers résultats, car possédant plusieurs paramètres. Enfin, il faut disposer d'une métrique capable de comparer et mesurer la différence/ressemblance entre les points mesurés et les points prédits. La section suivante commence donc par décrire les données oculométriques utilisées pour cette étude.

2.4.2 Mesures oculométriques

Pour cette expérimentation, nous utilisons la base de données oculométriques Visual Attention for Image Quality (VAIQ) [EMZ09] de l'Université de Western Sydney en Australie. Elle contient 53 images de référence extraites du photoCD Kodak, très largement utilisé dans les problématiques de compression et de qualité perçue. Ces images ont été affichées sur un écran 19" Samsung SyncMaster avec une résolution de 1280×1024. Les participants étaient assis à une distance d'environ 60 cm de l'écran. Les données oculométriques des 15 observateurs ont été mesurées par un eye-tracker EyeTech TM3 situé sous

l'écran. La précision des points de fixation enregistrés est d'environ 1 degré d'angle visuel. D'un point de vue vitesse d'acquisition, cet outil est capable d'enregistrer 40-45 points de fixation par seconde. Chacune des images de référence a été présentée pendant une durée de 12 secondes suivie d'une image grise pendant 3 secondes. Le nombre de points de fixation par image et par observateur est donc compris entre 480-560.

La validité de cette base a été testée ([ELZ+10]) par une seconde expérience à l'université de Delft, avec 20 nouveaux observateurs et 29 images identiques. Cette seconde campagne a prouvé la validité de la première expérience en produisant une très bonne corrélation des résultats. De ce fait, la validité et la fiabilité de cette base VAIQ semble donc démontrée.

En plus des 53 images de référence et des points de fixations des 15 observateurs, les cartes de chaleur de chaque image sont disponibles. Les cartes de chaleur sont obtenues par l'application de Gaussienne sur les points de fixation filtrés. L'intensité en chaque point de ces cartes reflète le niveau de saillance et est représenté sous forme d'une « température ». Plus elle est élevée, plus le point est considéré comme "chaud" et donc pertinent. Pour notre étude, nous utilisons les points de fixations d'origines et non les cartes de chaleur afin de rendre possible leurs comparaisons avec les points d'intérêts (points de fixation prédits).

Méthode d'agrégation des points saillants

En analysant les données expérimentales, nous avons remarqué que les points de fixation fournis ont tendance à se concentrer et s'accumuler sur de petites régions, tandis que les détecteurs de points d'intérêt évitent ce genre de comportement. De plus, les données issues d'oculométrie ont généralement pour objectif de produire un modèle du comportement moyen du regard humain. C'est dans cet objectif qu'une méthode de filtrage et d'agglomération de points est adoptée dans [EMZ09]. Cependant, nous avons noté quelques améliorations possibles que nous détaillons dans l'Algorithme 1. L'idée principale consiste à créer une collection de clusters $C_{collection}$ dont chaque cluster C_x contient plusieurs points de fixations GP_x (GP : Gaze Point) partageant une localisation spatiale très proche. Le but est de réduire le nombre de GP en agrégeant tous les GP_x assez proches les uns des autres, en dessous d'un seuil $Tclus$ définissant la distance maximum d'agrégation dans un cluster. Chaque cluster C_x peut être pondéré par le nombre de points GP_x agrégés. Une étape finale supprime tous les clusters n'ayant pas agrégé assez de points, dont le seuil minimum est définit par $Fmin$. Nous fixons $Tclus = 20$ et $Fmin = 4$

comme décrit dans l'algorithme original.

Algorithme 1 méthode modifiée de clustering de points de fixation

Créer une collection de cluster vide $C_{collection}$
{le premier point de fixation GP_1 est un cas particulier}
créer le premier cluster C_1
ajouter GP_1 dans C_1
ajouter C_1 dans $C_{collection}$
pour $i = 2 \rightarrow nombredeGPtotal$ **faire**
  trouver le cluster C_{find} dans $C_{collection}$ qui minimise la distance euclidienne D entre les coordonnées GP_i et les coordonnées de C_{find}
  si D<Tclus **alors**
    ajouter GP_i dans C_{find}
  sinon
    créer un nouveau cluster C_{new}
    ajouter GP_i dans C_{new}
    ajouter C_{new} dans $C_{collection}$
  fin si
fin pour
pour $j = 1 \rightarrow nombredecluster$ **faire**
  si $nombredeGPdans$Cj<Fmin **alors**
    supprimer Cj de $C_{collection}$
  fin si
fin pour

L'algorithme proposé dans [EMZ09] calcule uniquement la distance entre le GP courant et le cluster courant. De ce fait, la méthode de clustering est dépendante de l'ordre d'apparition des GP, tel qu'illustré par la Figure 2.11. Notre objectif est d'obtenir le comportement humain moyen, pour cela nous cherchons d'abord le cluster minimisant la distance avec le GP courant avant de tester son acceptation. De ce fait, le clustering devient relativement invariant à l'ordre d'apparition des GP et autorise l'agrégation des GP des différents observateurs.

La figure 2.12 montre l'effet des différentes méthodes de clustering, où la couleur et l'opacité des cercles informent de l'importance et du nombre de points agrégés dans chaque cluster. Le diamètre de chaque cercle est donné par le GP agrégé le plus éloigné du centre de chaque cluster. Pour résumer, un petit cercle rouge et opaque informe qu'il y a de nombreux points proches les uns des autres agrégés dans ce cluster. A l'inverse, un large cercle jaune transparent indique que peu de points sont agrégés dans ce cluster et qu'ils sont éloignés les uns des autres. Pour les points de fixation mesurés avant l'application de la méthode de clustering (visibles sur la Figure 2.12-(a)), chaque fixation de

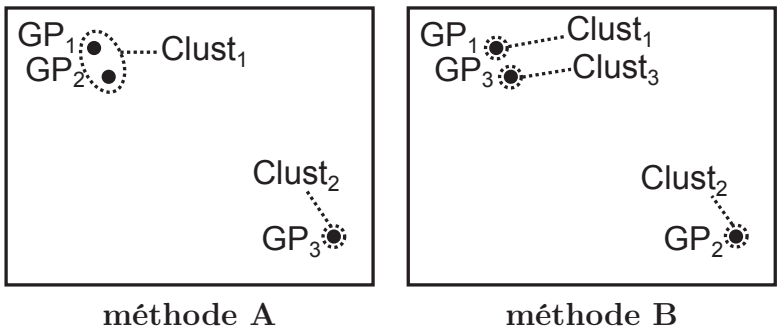


Figure 2.11 : Effet de l'ordre d'apparition des points de fixation (GP) sur la méthode de clustering.

chaque observateur a la même importance, donc la couleur est rouge et nous avons fixé l'opacité a 10%. Quand des régions rouges et opaques apparaissent, cela informe que de nombreux observateurs ont regardés à la même position. Dans cet exemple, notre méthode de clustering (visible sur la Figure 2.12-(c)) réduit le nombre de points d'environ 50%, tout en minimisant la quantité de cercles superposés. Un autre atout est la mise en valeur des régions où tous les humains sont en accord tout en réduisant l'impact des régions où peu d'observateurs ont regardé.

(a) GP brut (7328 pts) (b) clustering original (387 pts) (c) clustering modifié (164 pts)

Figure 2.12 : Points de fixation (GP) des 15 observateurs avant et après application de la méthode d'agrégation.

2.4.3 Paramètres des détecteurs de points d'intérêt

En ce qui concerne les détecteurs de points d'intérêt, nous avons choisi le détecteur de Harris pour sa détection de coins et de contours, ainsi que le

Tableau 2.1 : Paramètres de la fonction Harris

Param.	Description
qualityLevel : **ql**	Fixe la valeur minimum d'acceptation de réponse R_{HS88a} obtenue par la formule 1.19
minDistance : **md**	La distance Euclidienne minimum entre deux points d'intérêt.
blockSize : **bs**	La taille de la fenêtre d'observation mise en œuvre pour le calcul d'auto-corrélation. Sa taille influence la taille des structures détectées.
k	Le paramètre de pondération permettant d'atténuer l'effet des forts contours (utilisé dans la formule 1.19)

détecteur SIFT pour la détection de structures blob et l'utilisation d'espaces multi-échelles. Nous avons également intégré le récent et très utilisé SURF, car il présente les mêmes avantages que SIFT, tout en étant optimisé en terme de temps de calcul.

Nous fournissons dans les Tableaux 2.1, 2.2 et 2.3, la description des paramètres de chaque détecteur (respectivement pour Harris, SIFT et SURF) influençant le nombre et la localisation des points d'intérêt.

Nous avons décrit précédemment la base de données et la méthodologie pour disposer des points de fixation expérimentaux servant de vérité de terrain. Notre objectif est de comparer ces points avec les points d'intérêt dont la détection est pilotée par différents paramètres. Pour cette comparaison, nous avons choisi d'utiliser la métrique Earth Mover's Distance [RTG98] capable de quantifier la ressemblance de deux ensembles de points de cardinalité différente. Nous fournissons les détails de cette métrique dans la section suivante.

2.4.4 Description de la distance EMD

Le nombre de points d'intérêt extraits par chaque détecteur dépend de sa configuration et est malheureusement toujours différent du nombre de GP subjectif. Cependant, notre objectif est tout de même de mesurer leur similarité. Il est donc nécessaire d'utiliser une mesure de distance capable de comparer des jeux de données de cardinalités différentes. Pour ce faire, nous avons choisi la mesure Earth Mover's Distance (EMD). Cette dernière est basée sur un cas particulier de la théorie du transport [Dan51] (de Monge-Kantorovich

Tableau 2.2 : Paramètres de la fonction SIFT

Param.	Description
Number of octaves : **no**	Le nombre d'octaves utilisés pour la construction de l'espace échelle. Par défaut il est conseillé d'avoir le plus d'octaves possibles (i.e. environ log2(min(largeur, hauteur)), ce qui a pour effet de chercher des points d'intérêt de toutes les tailles possibles.
First octave index : **fo**	Indice permettant de fixer le premier octave utilisé dans la détection. Si cette valeur est fixée à 1 toutes les échelles sont considérées. Si une valeur supérieure est utilisée certaines petites structures ne seront pas détectables.
Number of levels per octave : **nol**	Le nombre de couches dans chaque octave, fixé à 3 par défaut. Augmenter ce nombre revient à augmenter le nombre de points détectés, mais en pratique cela a pour effet de rendre le détecteur instable car très sensible au bruit.
Peak threshold : **pt**	Seuil fixant le contraste minimum d'un point d'intérêt
Edge threshold : **et**	Seuil fixant à quelle mesure une structure blob à un profil de contour.

Tableau 2.3 : Paramètres de la fonction SURF

Param.	Description
Hessian threshold : **ht**	Seule les structures avec une réponse hessienne supérieure à ce seuil sont conservées. Les valeurs conseillées sont généralement comprises entre 300-500 (Cette valeur peut dépendre du contraste local de l'image et de sa netteté).
nOctaves : **no**	Le nombre d'octaves utilisés pour la construction de l'espace échelle. Pour chaque nouvel octave, la taille des structures détectables est doublée. Ce paramètre est généralement fixé à 3.
nOctaveLayers : **nol**	Le nombre de couches dans chaque octave, généralement fixé à 4 par défaut.

[Mon81]). Cette distance EMD a été développée par Rubner et al. [RTG98] dans les contextes d'indexation et de recherche d'images. Elle a pour but de mesurer l'effort nécessaire pour déplacer différentes masses de terre dans différents trous dispersés dans un espace afin de les remplir, sachant que le nombre et la taille de chaque distribution peuvent être différents tel qu'illustré par la Figure 2.13. Une unité de travail correspond au transport d'une unité de terre pour une unité de distance dans l'espace considéré.

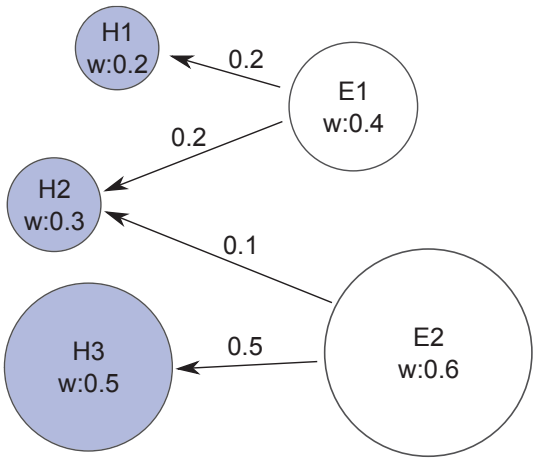


Figure 2.13 : Illustration d'une répartition de deux tas (E) pour trois trous (H)

Dans notre cas, nous considérons que les GP mesurés sont représentés par les trous (les cibles à atteindre) et les points d'intérêt sont les masses de terre à déplacer. De plus, cette métrique peut prendre en compte des facteurs de pondération sur chaque trou et masse de terre. Puisque les GP sont issus d'une méthode de clustering (cf. section 2.4.2), nous utilisons le nombre de points agrégés pour chaque cluster comme facteur de pondération pour les trous afin de refléter la saillance importante de cette région. Sur la même idée, puisque chaque détecteur de points d'intérêt peut fournir des informations additionnelles à la localisation de chaque point détecté, nous pouvons pondérer les masses de terre par ces données supplémentaires. Cela peut être la valeur de la réponse R_{HS88a} (cf. équation 1.19) du détecteur de Harris ou bien la taille et le facteur hessian pour le détecteur SURF.

L'idée, derrière l'utilisation de la distance EMD, est d'être capable d'estimer quel détecteur associé à quelle configuration particulière minimise le coût de transformation entre les points mesurés et les points prédits. L'expérimentation et l'analyse statistique associée sont détaillées dans la section suivante. Cette étude cherche à confirmer qu'un détecteur de points d'intérêt bien configuré est à même de donner des prédictions de zones saillantes en accord avec la vision humaine.

2.4.5 Expérimentation et analyse statistique

A ce niveau, nous avons décrit les détecteurs de points d'intérêt utilisés et leurs différents paramètres (section 1.2.3 et Tableau 2.1 pour Harris, 1.2.5 et Tableau 2.2 pour SIFT et 1.2.6 et Tableau 2.3 pour SURF), notre méthode pour disposer de points mesurés fiables (section 2.4.2), ainsi que la métrique de mesure de similarité choisie (section 2.4.4). Comme discuté précédemment, chaque détecteur de points d'intérêt a plusieurs paramètres influençant la localisation et le nombre de points détectés. Notre but étant de maximiser la ressemblance avec les points de fixation, nous avons essayé de nombreuses configurations de paramètres sur de larges plages de valeurs afin de déterminer le réglage optimal pour chaque détecteur. Cette opération a été effectuée sur l'intégralité des images de la base VAIQ. De ce fait, nous disposons du coût EMD pour chaque image, pour chaque détecteur et pour chaque configuration de paramètres. Pour rappel, un faible score EMD indique une grande similarité entre les points mesurés et prédits.

Cependant, la même configuration ne minimise pas obligatoirement la distance EMD pour toutes les images testées. En effet, la base VAIQ contient une grande variété de contenu avec des visages, des animaux, des paysages, différentes complexités d'arrière-plans, etc., il y a donc vraisemblablement des réglages plus adaptés à un type de contenu. Mais dans cette étude, nous ne nous focalisons pas sur un contexte d'application spécifique. C'est pourquoi, nous restons relativement généralistes et préférons viser la minimisation moyenne en considérant l'ensemble des types de contenu. La Figure 2.14 permet d'avoir une représentation plus visuelle du protocole mis en place pour la génération des données pour le détecteur de Harris.

Sur ce principe, nous avons pu réaliser une première campagne de mesure, détaillée dans la section suivante.

Expérience 1

Pour cette première expérience et au vu de la littérature, nous ne disposons pas d'*a priori* sur les valeurs des différents paramètres à tester en priorité afin de maximiser la ressemblance avec les points de fixation. Cependant, suite à quelques expérimentions et à la compréhension des différents paramètres, nous proposons d'effectuer une première série de tests avec de larges plages de valeurs pour chacun des paramètres. Les valeurs testées sont répertoriées dans les tableaux 2.4, 2.5 et 2.6 respectivement pour les détecteurs de Harris, SIFT et SURF. C'est un total de 560 combinaisons évaluées par image pour Harris

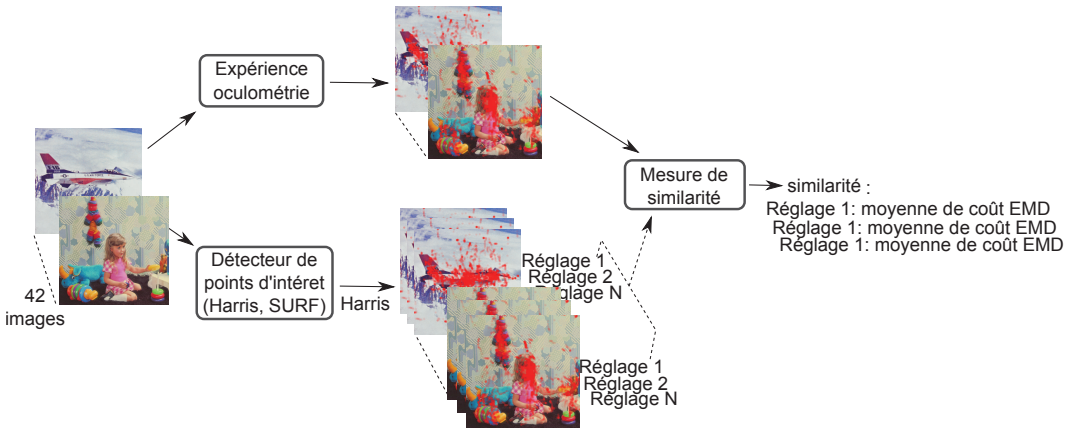


Figure 2.14 : Illustration du protocole expérimental appliqué au détecteur de Harris.

$(5 \times 7 \times 4 \times 4)$, 100 combinaisons pour SIFT $(10 \times 10 \times 1)$ et 4416 pour SURF $(69 \times 8 \times 8)$. Le nombre de combinaisons testées est guidé par la sensibilité et l'influence de chaque paramètre.

Tableau 2.4 : Harris

ql	$\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$
md	$\{20, 40, 60, 70, 80, 100, 120\}$
bs	$\{3, 6, 9, 12\}$
k	$\{0.04, 0.08, 0.12, 0.20\}$

Tableau 2.5 : SIFT

lvl	$\{1:1:10\}$
no	$\{1:1:10\}$
fo	$\{-1\}$

Tableau 2.6 : SURF

ht	$\{100:100:6900\}$
no	$\{1, 2, 3, 4, 6, 9, 12, 18\}$
nol	$\{1, 2, 3, 4, 6, 9, 12, 18\}$

Puisque la distance EMD peut prendre en considération des facteurs de pondération pour chaque point, nous choisissons de réaliser cette étude avec et sans prise en compte de ces poids. Quand ces derniers ne sont pas pris en compte, chaque point extrait a la même influence que les autres. Dans ce cas, seule la localisation est prise en compte. A l'opposé, l'étude de l'utilisation d'informations additionnelles, telles que la taille et la force peuvent être intéressantes pour estimer à quelle mesure elles influencent la ressemblance avec les points de fixation.

Les résultats des meilleurs scores EMD moyens de cette première étude sont compilés dans le Tableau 2.7 pour les mesures sans pondération et le

Tableau 2.8 pour la prise en compte des informations additionnelles. Comme il est question de mesure de moyenne de coût EMD pour chaque configuration de paramètre, nous fournissons également le coût minimum, maximum et l'écart-type de ces mesures afin de mieux appréhender ces scores.

Tableau 2.7 : Coûts EMD des meilleures configurations de chaque détecteur (sans pondération).

Détecteur	moy.	min	max	std
Harris	0,02290	0,01547	0,04148	0,00546
SIFT	0,00023	0,00012	0,00053	0,00008
SURF	0,00132	0,00027	0,00480	0,00092

Tableau 2.8 : Coûts EMD des meilleures configurations de chaque détecteur (avec pondération).

Détecteur	moy.	min	max	std
Harris (Response)	72,313	12,28	175,14	39,23
SIFT (Size)	38,80	9,85	109,44	22,51
SURF (Hessian)	44,79	6,66	130,52	28,38
SURF (Scale)	42,04	2,71	119,28	27,02

En observant les résultats du Tableau 2.7, nous pouvons noter que les meilleurs scores moyens de chaque détecteur sont sur des gammes différentes, avec un facteur ≈ 6 entre SIFT et SURF, et un facteur ≈ 20 entre SURF et Harris. C'est le détecteur SIFT qui minimise de manière importante la distance EMD, suivi de SURF puis Harris. Ce classement est également préservé sur le Tableau 2.8 où différentes pondérations sont prises en compte. Pour le détecteur SURF, deux types de pondérations ont été évaluées, SURF_{scale} et $\text{SURF}_{hessian}$ correspondant à la prise en compte de la taille de la structure détectée et la force de cette structure. C'est SURF_{scale} qui apparait le plus pertinent en comparaison avec $\text{SURF}_{hessian}$, ce qui amène à conclure que la taille des structures détectées a son importance.

Nous pouvons également noter que les scores EMD moyens sont très différents d'un tableau à l'autre. Il semble donc délicat de conclure sur l'impact de l'utilisation des informations additionnelles dans le but de maximiser la ressemblance avec les points de fixation. Afin d'appréhender ce phénomène, nous fournissons les scores des dix meilleurs configurations du détecteur de Harris avec et sans pondération.

Tableau 2.9 : Les dix meilleures configurations moyennes pour Harris.

ql	md	bs	k	mean
1,00E-06	20	3	0,04	0,0229
1,00E-06	20	3	0,08	0,0236
1,00E-06	20	3	0,12	0,0244
1,00E-06	20	6	0,04	0,0244
1,00E-06	20	6	0,08	0,0253
1,00E-06	20	9	0,04	0,0256
1,00E-06	20	6	0,12	0,0259
1,00E-06	20	3	0,2	0,0262
1,00E-06	20	9	0,08	0,0263
1,00E-06	20	12	0,04	0,0266

(a) Harris sans pondération

ql	md	bs	k	mean
1,00E-06	20	9	0,04	72,313
1,00E-05	20	9	0,04	72,314
1,00E-04	20	9	0,04	72,331
1,00E-06	40	9	0,04	72,460
1,00E-05	40	9	0,04	72,461
1,00E-06	40	6	0,04	72,462
1,00E-05	40	6	0,04	72,463
1,00E-04	40	9	0,04	72,468
1,00E-04	40	6	0,04	72,468
1,00E-03	40	6	0,04	72,557

(b) Harris avec pondération

Plusieurs remarques peuvent être faites en observant ces tableaux. Tout d'abord, $ql = 1,00E-06$ et $md = 20$ sont les valeurs retenues pour les dix meilleures configurations sans prise en compte des pondération, ce qui tend à conclure que ces valeurs de paramètres sont très importantes pour la maximisation de la ressemblance dans le cas de Harris. Mais de plus ces valeurs correspondent aux valeurs minimales dans les plages testées. Il semble donc que la minimisation de ces paramètres influe fortement sur la réduction du coût EMD. Pour rappel, le paramètre *md* influe sur la distance minimale autorisée entre deux points détectés, tandis que le paramètre *ql* permet de filtrer les points les moins forts. En résumé, la minimisation de ces deux paramètres maximise le nombre de points détectés. Ceci peut expliquer la minimisation des distances EMD car la grande disponibilité de points peut faciliter la mise en correspondance avec les fixations humaines. Le profil de l'évolution du coût EMD moyen pour chaque paramètre de Harris est visualisable sur la figure 2.15.

Comme mentionné précédemment la minimisation des paramètres *ql* et *md* et donc la maximisation du nombre de points détectés, réduit le coût EMD. Dans le cas de Harris et sur les plages testées, le nombre de points détectés varie dans l'intervalle [13,02-473,73] en moyenne. Cette plage est nettement supérieure pour SURF : [34,11-5515,30] et va même jusqu'à plus de 15000 points détectés pour SIFT pour certaines configurations. La conclusion est donc que la distance EMD est sensible au nombre de points considérés. Par ce fait, il n'est pas équitable de comparer des distances entre deux détecteurs s'ils ne produisent pas le même nombre de points.

Afin de confirmer l'hypothèse de sensibilité de la distance EMD et dans le

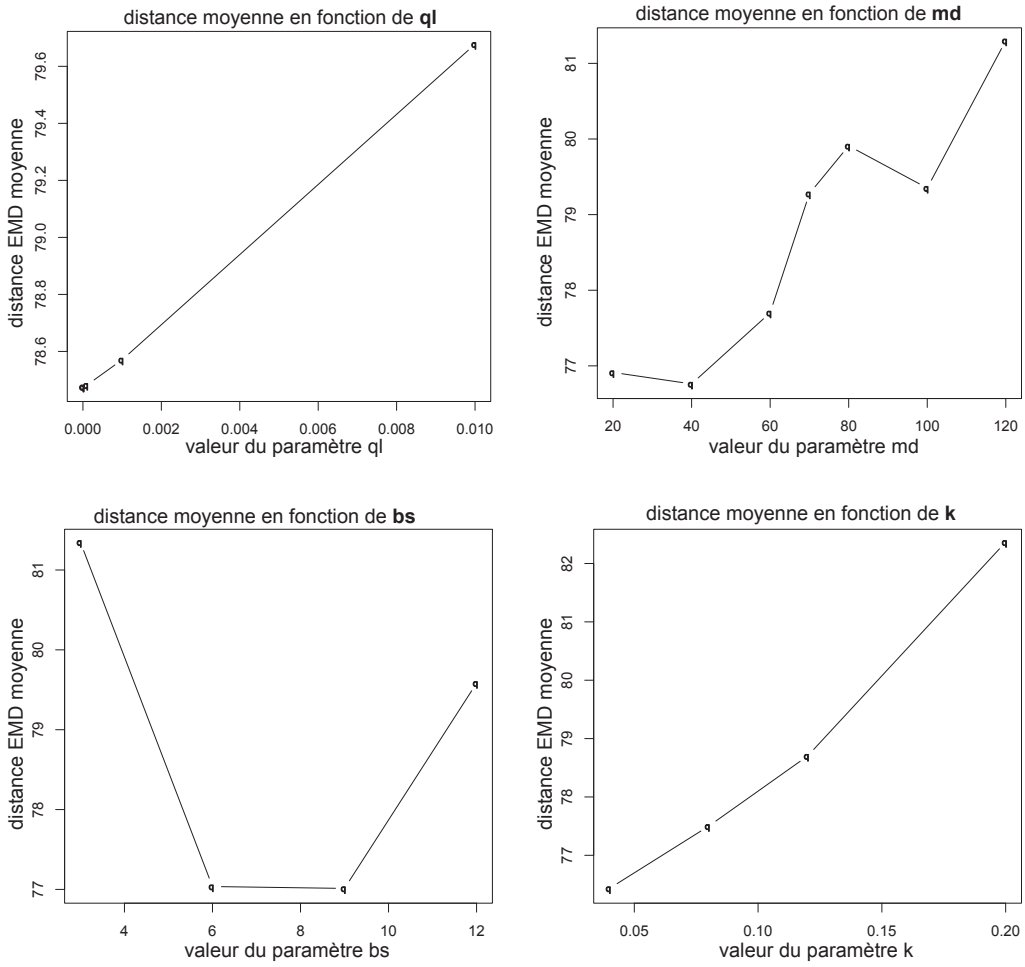


Figure 2.15 : Évolution du coût EMD moyen pour chaque paramètre.

but de juger l'influence de chaque paramètre, nous utilisons le test statistique ANOVA. Comme tout autre test statistique, il compare deux hypothèses :
H_0 (hypothèse nulle) dans le cas de distributions égales. Si cette hypothèse est confirmée, cela informe que le paramètre considéré n'a pas d'influence sur l'évolution de la distance EMD.
H_1 (hypothèse alternative) : dans le cas de distributions différentes. Si cette hypothèse est confirmée, cela informe que le paramètre considéré a une influence notable sur l'évolution de la distance EMD
Le test ANOVA retourne une "p-value" qui détermine l'influence de chaque paramètre. Plus cette valeur est proche de 0, plus le paramètre a une influence. En pratique, quand la p-value est inférieure à 0.05, nous rejetons l'hypothèse nulle et considérons que le paramètre est influent.

Les résultats ANOVA sont donnés dans les Tableaux 2.10 et 2.11 représentant respectivement les mesures d'influence de chaque paramètre avec et sans prise en compte des pondérations. A noter que dans ces tableaux et pour faciliter la lecture, des étoiles sont placées dans la colonne "Influence" avec les significations suivantes : une case vide indique que le paramètre n'a aucune influence, tandis qu'une ou plusieurs $*$ informe d'une influence notable (codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1).

Tableau 2.10 : Influence des paramètres du détecteur de Harris sur les mesures EMD sans pondération.

Paramètre	p-value	Influence
ql	$< 2.2e-16$	$***$
md	$< 2.2e-16$	$***$
bs	0.07815	.
k	$2.686e-05$	$***$
nbPtsMoy.	$< 2e-16$	$***$

Au vu des résultats obtenus, nous pouvons noter que l'hypothèse nulle est

Tableau 2.11 : Influence des paramètres du détecteur de Harris sur les mesures EMD avec pondération.

Paramètre	p-value	Influence
ql	0.4632	
md	$6.985e-07$	$***$
bs	$1.04e-11$	$***$
k	$< 2.2e-16$	$***$
nbPtsMoy.	1	

rejetée pour le paramètre nbPtsMean dans le tableau 2.10 et acceptée pour le tableau 2.11. Cela indique que le nombre de points détectés a une influence sur la mesure EMD sans pondération mais n'en a pas pour les mesures avec pondération. Nous pouvons avancer une explication, en considérant que la pondération va avoir tendance à favoriser grandement certains points, et complètement atténuer et en faire disparaître d'autres. Par ce fait, le nombre de points est donc relativement constant avec peu d'influence. Ce qui est complètement différent du cas sans pondération, où chaque point a une importance égale aux autres, et donc leur nombre est inévitablement influent sur les résultats.

Par ces conclusions, nous pouvons soulever quelques interrogations sur l'interprétation du tableau 2.7, car potentiellement porteur d'un biais de comparaison entre les détecteurs. De surcroît, le coût EMD apparaît comme trop dépendant du nombre de points détectés, minimum et maximum pour chaque détecteur. Nous proposons donc de réaliser une seconde expérimentation afin de limiter les biais de mesure et fournir une étude plus précise.

Expérience 2

Pour cette seconde expérimentation, nous ne remettons pas en cause l'utilisation de la mesure EMD car cette métrique est reconnue et couramment utilisée, et de par son respect de l'inégalité triangulaire, elle est considérée comme une vraie métrique. Cependant, pour l'utiliser dans de bonnes conditions, nous décidons d'ajouter une contrainte sur les plages de valeurs testées pour nos détecteurs en ne conservant que les configurations produisant des nombres de points relativement proches. Toute configuration produisant un nombre de points moyen sur l'ensemble des images en dehors d'un intervalle $filtI$ sera rejetée.

Afin de déterminer $filtI$, nous avons analysé les points de fixation mesurés de la base VAIQ. Dans cette base, le nombre de points par image et par observateur est compris entre [480-540]. Bien que chaque observateur regarde la même image pendant la même durée, le nombre de points de fixation est différent d'un individu à l'autre. Chaque observateur a ses propres saccades et fixations. Nous décidons donc de fixer $filtI$ sur l'intervalle [420-600] afin d'être relativement proche de l'intervalle humain, tout en laissant une certaine tolérance aux paramètres de chaque détecteur.

Les plages des paramètres testés pour cette seconde expérimentation ont dû être modifiées pour respectivement Harris et SIFT car les configurations ne fournissaient pas des nombres de points respectant le $filtI$. Ces nouvelles

plages sont données dans le Tableau 2.12 et le Tableau 2.13. Pour SURF, les paramètres testés précédemment fournissent déjà de nombreuses configurations (511) en accord avec la contrainte $filtI$. C'est un total de 189 nouvelles configurations par image testée pour Harris ($3 \times 7 \times 3 \times 3$) et 3125 pour SIFT ($5 \times 5 \times 5 \times 5 \times 5$).

Tableau 2.12 : Harris exp. 2.

ql	$\{10^{-6}, 10^{-7}, 10^{-8}\}$
md	$\{0.1, 1, 10, 20, 30, 40, 50\}$
bs	$\{3, 6, 9\}$
k	$\{0.005, 0.02, 0.04\}$

Tableau 2.13 : SIFT exp. 2.

lvl	$\{1, 2, 3, 5, 7\}$
no	$\{1, 3, 5, 7, 9\}$
fo	$\{-1, 0, 1, 2, 3\}$
et	$\{2, 5, 7, 10, 15\}$
pt	$\{0, 10, 20, 30, 40\}$

En exploitant ces nouvelles valeurs de paramètres, et le respect de la contrainte $filtI$, le même protocole que pour l'expérience 1 a été utilisé, en se focalisant sur l'étude avec prise en compte des pondérations afin d'être le plus invariant possible au nombre de points détectés. Pour analyser en détails les paramètres de chaque détecteurs, nous fournissons les dix meilleures configurations pour chaque outil. Le Tableau 2.14 pour Harris, le Tableau 2.15 et 2.16 pour SURF pondéré par la taille et le hessian, ainsi que le Tableau 2.17 pour SIFT avec une pondération relative à l'échelle de détection.

Tableau 2.14 : Les dix meilleures configurations pour Harris (expérience 2).

ql	md	bs	k	moy.	min	max	std
1,00E-07	20	9	0,005	71,056	11,874	180,221	39,862
1,00E-08	20	9	0,005	71,057	11,874	180,221	39,862
1,00E-06	20	9	0,005	71,061	11,874	180,222	39,860
1,00E-07	20	6	0,005	71,316	9,963	180,570	39,911
1,00E-08	20	6	0,005	71,316	9,963	180,570	39,911
1,00E-06	20	6	0,005	71,317	9,963	180,570	39,911
1,00E-07	20	9	0,02	71,841	12,598	176,448	39,322
1,00E-08	20	9	0,02	71,841	12,598	176,448	39,322
1,00E-06	20	9	0,02	71,842	12,598	176,449	39,322
1,00E-07	20	6	0,02	72,032	11,496	179,781	39,527

Pour le détecteur de Harris, le paramètre *ql* prend des valeurs comprises entre 1,00E-08 et 1,00E-06 ce qui garantit la conservation des points d'intérêt de forces moyenne et importante. Si le paramètre *ql* est fixé à une valeur trop élevée, trop peu de points d'intérêt seront conservés, ce qui ne permet pas de respecter le $filtI$. Le paramètre md=20 donne les meilleurs résultats et

en accord avec la première expérience. Ce paramètre définit la distance minimum autorisée entre 2 points détectés. Nous avons remarqué que cette valeur est similaire au paramètre $Tclus$ introduit dans le processus de clustering des points mesurés (cf. section 2.4.2), mais ce lien n'a pas été explicitement démontré ou exploré. Le paramètre *bs* prend des valeurs comprises entre 6 et 9 dans ce classement des dix meilleures configurations. Il influe sur la taille de la fenêtre d'observation du détecteur ; cette taille est relativement élevée et a pour effet d'empêcher la détection de coins de trop petite taille. Enfin, le paramètre k prend des valeurs très faibles. Il influence la sélectivité entre les coins et les contours, où de faibles valeurs de ce paramètre permettent à des contours d'être conservés au même titre que des coins.

Tableau 2.15 : Les dix meilleures configurations pour SURF (avec la taille comme pondération).

ht	no	nol	moy.	min	max	std
800	9	1	47,906	5,045	138,739	30,645
800	18	1	47,906	5,045	138,739	30,645
800	6	1	47,906	5,045	138,739	30,645
800	12	1	47,906	5,045	138,739	30,645
800	4	1	48,702	6,408	138,739	30,824
900	6	1	48,873	7,564	140,738	31,285
900	12	1	48,873	7,564	140,738	31,285
900	9	1	48,873	7,564	140,738	31,285
900	18	1	48,873	7,564	140,738	31,285
1000	18	1	48,989	4,738	142,834	32,134

Pour le détecteur SURF, c'est $ht = 800$ qui est sélectionné pour minimiser la distance EMD tel qu'illustré sur la Figure 2.17. Ce paramètre est utilisé pour rejeter les structures les moins prononcées. Le paramètre *no* peut prendre plusieurs valeurs (cf. Tableau 2.15), mais la Figure 2.17 indique une stabilisation de l'EMD à partir de la valeur 4. Nous conseillons donc les valeurs autour de 5 afin d'associer de bons résultats (minimisation de l'EMD) tout en limitant les temps de calcul. Pour le paramètre *nol*, la valeur 1 a été choisie car quand ce paramètre augmente, le nombre de points détectés augmente très rapidement, ce qui ne permet pas de respecter la contrainte $filtI$. En observant les deux Tableaux 2.15 et 2.16, nous pouvons noter que les meilleures configurations de paramètres sont identiques pour les deux types de pondérations. Nous pouvons conclure que la localisation des points d'intérêt est correcte dans les deux cas, mais que c'est la pondération de ces points par l'échelle de détection qui apparaît meilleure que la pondération par les valeurs de réponse hessienne. La Figure 2.16 montre que les informations additionnelles de taille et de force de

Tableau 2.16 : Les dix meilleures configurations pour SURF (avec le hessian comme pondération).

ht	no	nol	moy.	min	max	std
800	9	1	56,992	8,055	153,909	35,123
800	18	1	56,992	8,055	153,909	35,123
800	6	1	56,992	8,055	153,909	35,123
800	12	1	56,992	8,055	153,909	35,123
800	4	1	57,138	8,448	154,162	35,134
900	6	1	57,735	8,541	155,529	35,534
900	12	1	57,735	8,541	155,529	35,534
900	9	1	57,735	8,541	155,529	35,534
900	18	1	57,735	8,541	155,529	35,534
900	4	1	57,887	8,922	155,792	35,548

structure peuvent être complémentaires. En effet, dans cet exemple, un point d'intérêt fort est détecté sur le texte de la casquette jaune en utilisant la pondération par le hessian, mais que la casquette rouge est mise en valeur par la pondération d'échelle de détection. Tandis que la saillance humaine favorise ces deux casquettes, une idée serait donc d'exploiter conjointement ces deux types de pondérations.

points de fixation mesurés | points prédits $\text{SURF}_{hessian}$ IP | points prédits SURF_{scale} IP

Figure 2.16 : Cartes de saillance mesurée et prédites.

Pour le détecteur SIFT, les paramètres optimaux *no* et *lvl* sont similaires aux paramètres *no* et *nol* du détecteur SURF, où le nombre d'octaves doit être suffisamment élevé (supérieur à 5) pour permettre la détection des grosses structures blob, tandis que le nombre de couches par octave doit rester faible (proche de 1) afin d'éviter l'explosion du nombre de points détectés. Le paramètre *pt* est fixé à 0 n'imposant pas de seuil de contraste minimum. Enfin, le paramètre *et* prend des valeurs relativement importantes (supérieures à 5),

Tableau 2.17 : Les dix meilleures configurations pour SIFT (avec l'échelle comme pondération).

no	lvl	fo	et	pt	moy.	min	max	std
9	1	0	5	0	37,944	4,303	131,175	24,861
7	1	0	5	0	37,944	4,303	131,175	24,861
5	1	0	5	0	38,068	4,303	131,175	24,892
5	1	0	7	0	38,126	7,582	129,147	24,625
9	1	0	7	0	38,186	7,582	129,147	24,958
7	1	0	7	0	38,186	7,582	129,147	24,958
3	7	1	15	0	38,324	5,799	113,070	23,758
9	5	1	15	0	38,428	4,803	107,831	22,584
7	5	1	15	0	38,428	4,803	107,831	22,584
5	5	1	15	0	38,452	4,803	107,831	22,560

permettant ainsi de conserver des structures au profil allongé et donc ressemblant plus à des contours qu'à des structures blob. Ce constat est similaire au paramètre k du détecteur de Harris permettant de conserver des structures de type contour en plus des coins.

Avec cette seconde expérience, nous pouvons voir sur le Tableau 2.18 que tous les scores moyens obtenus sont différents de ceux de la première campagne de tests (cf. Tableau 2.8). Cependant, le classement des détecteurs reste inchangé, avec SIFT suivi de SURF_{scale}, $\text{SURF}_{hessian}$ puis Harris.

Tableau 2.18 : Comparaison des meilleures distances EMD avec pondération (exp. 2).

Détecteur	moy.	min	max	std
Harris (Response)	71,056	11,874	180,221	39,862
SURF (Hessian)	56,992	8,055	153,909	35,123
SURF (Scale)	47,906	5,045	138,739	30,645
SIFT (Size)	37,944	4,303	131,175	24,861
Humain (Durée)	48,153	0,984	186,294	28,959

Puisque que ce sont les détecteurs de Harris et $\text{SURF}_{hessian}$ qui ont les scores les plus faibles, nous étudions l'influence des paramètres afin de vérifier si une amélioration est possible (tableau 2.19 et 2.20). En utilisant les résultats de l'ANOVA, aucun paramètre ne semble avoir d'influence significative sur le coût EMD.

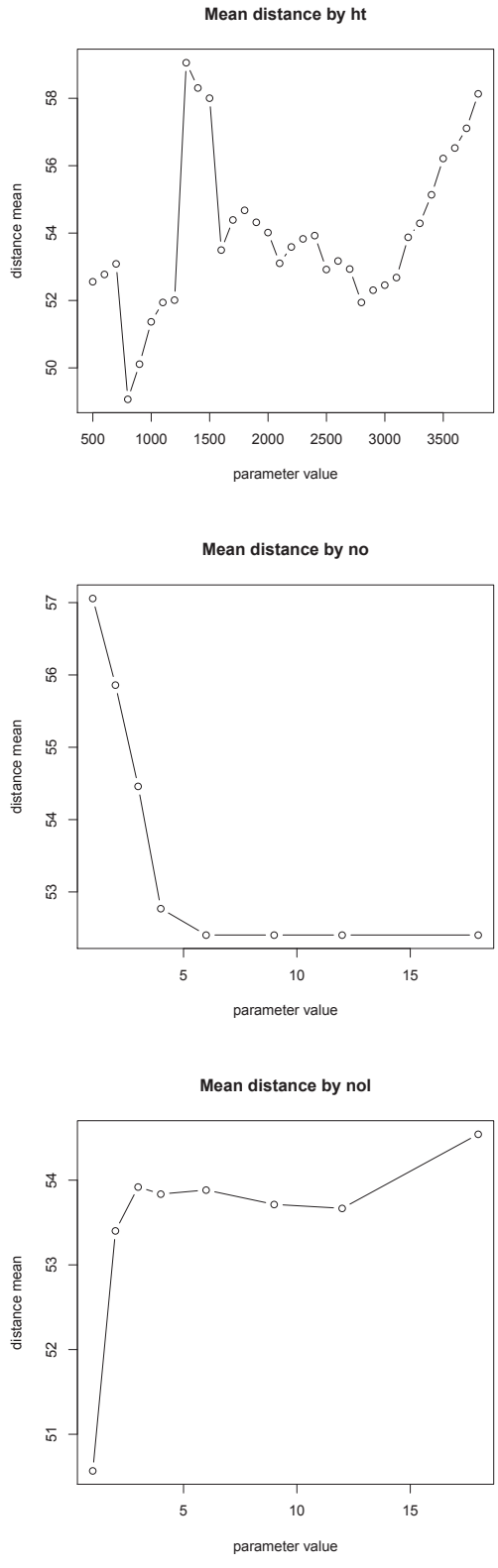


Figure 2.17 : Moyenne de distance pour les paramètres de SURF$_{scale}$.

Tableau 2.19 : $\mathsf{Harris}_{response}$.

Param.	p-value	Influence
ql	1	
md		
bs	0.3778	
k	0.860	
nbPts	1	

Tableau 2.20 : $\mathsf{SURF}_{hessian}$.

Param	p-value	Influence
ht	0.757	
no	0.6201	
nol	0.0098	**
nbPts	1	

Tableau 2.21 : SURF_{scale}.

Param	p-value	Influence
ht	0.04747	*
no	$2.044e-08$	***
nol	0.03675	*
nbPts	1	

Bien que cette étude ait permis de classer les trois détecteurs, se pose encore une question au sujet de l'interprétation des coûts EMD. En effet, un faible coût EMD=47 est-il révélateur d'une réelle similarité avec le comportement humain ? Pour trouver une réponse à cette question, nous avons étudié les données des observateurs disponibles dans la base VAIQ. Pour chaque image et chaque observateur, nous avons calculé la distance EMD entre ses points de fixation et la moyenne des points de fixations de tous les autres observateurs en appliquant la méthode de clustering (cf. section 2.4.2). Afin d'exploiter les pondérations de la distance EMD, nous avons utilisé l'information de durée de chaque point de fixation, en considérant que plus un observateur fixe longtemps une région, plus cette région est riche en information à extraire et donc que cette région est plus saillante qu'une autre. La dernière ligne du tableau 2.18 donne les valeurs du coût EMD avec cette pondération par le temps de fixation. La moyenne du coût EMD est d'environ 48, représentant le coût moyen de différence entre les points d'un observateur en comparaison avec tout le panel. Nous pouvons noter que SURF_{scale} et SIFT_{size} ont des moyennes de coût EMD en dessous de cette valeur. Nous pouvons donc conclure que les résultats de ces détecteurs sont comparables au comportement humain. Nous pouvons également noter que le coût EMD maximum d'un humain comparé à la moyenne des autres humains est nettement supérieure au coût maximum de tous nos détecteurs.

Pour conclure, tous les détecteurs peuvent trouver des points d'intérêt en accord avec le comportement humain. Nous pouvons également noter que tous les détecteurs exploitent l'information du contraste local de l'image pour la détection et la définition de structures de coins/contours/blobs, ce qui semble

un processus également impliqué dans le SVH. Nous remarquons également que SURF et SIFT, avec leur approche multi-échelles, prennent en compte la taille des structures détectées, ce qui est également un procédé reconnu dans le comportement humain. Enfin, indépendamment des tâches fixées aux observateurs, la présence de textes dans une image est très influente sur la saillance perçue, et dans ce cas le détecteur de coins de Harris semble très adapté à la détection de ce type de structures. Par ailleurs, la taille des structures semble importante, il serait intéressant d'évaluer les performances de la version multi-échelles du détecteur de Harris.

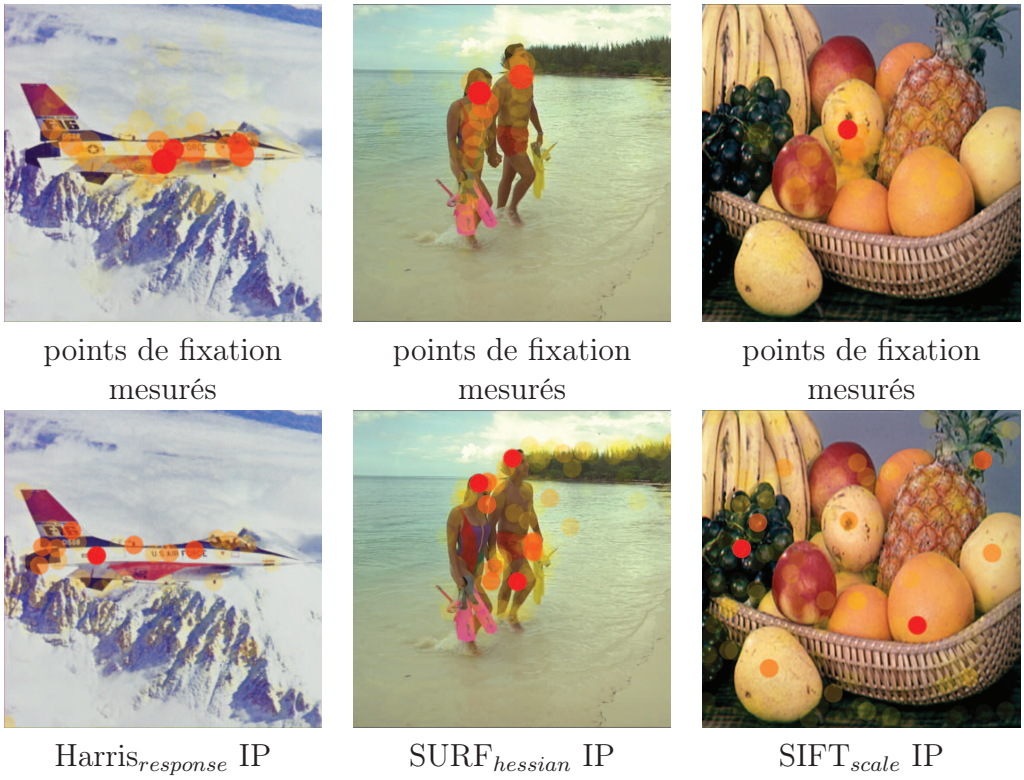


Figure 2.18 : Comparaison visuelle des points mesurés et détectés par Harris, SIFT et SURF.

Ayant mis en avant les capacités des détecteurs de points d'intérêt à prédire la localisation des points de fixations humains, nous proposons, dans la section suivante, un nouveau modèle de génération de carte de saillance basé sur cette étude.

2.5 PINS (Prediction of INterest Points Saliency)

La méthode de génération de notre carte de saillance (SM) est très proche de la construction des cartes de chaleurs utilisant les points de fixation mesurés. Nous remplaçons principalement les points de fixations par ceux des détecteurs de points d'intérêt. Un exemple de génération de carte de chaleur est détaillé dans l'article [EMZ09] associé à la base de données précédemment utilisée VAIQ (cf. section 2.4.2).

Pour notre SM nous utilisons un des détecteurs de points d'intérêt Harris, SIFT ou SURF décrit dans le chapitre 1, pour lesquels nous avons défini la configuration optimale de leurs paramètres respectifs en utilisant les résultats statistiques obtenus en section 2.4.5.

Pour chaque point ainsi détecté nous appliquons un noyau Gaussien d'écart-type $\sigma = 35$ tel que défini pour la construction des cartes de chaleur. Chaque Gaussienne est pondérée par la réponse R pour le détecteur de Harris$_r$, le facteur de taille pour les points SIFT$_s$, le facteur d'échelle pour SURF$_s$ et le facteur hessian pour SURF$_h$ (cf. section 2.4.5 pour plus de détails sur chaque facteur). Un processus de normalisation est appliqué sur chaque pondération et globalement sur la SM de sortie afin que les valeurs de nos cartes soient comprises dans l'intervalle [0-1] et de rendre nos cartes indépendantes du type de détecteur et du facteur de pondération utilisés.

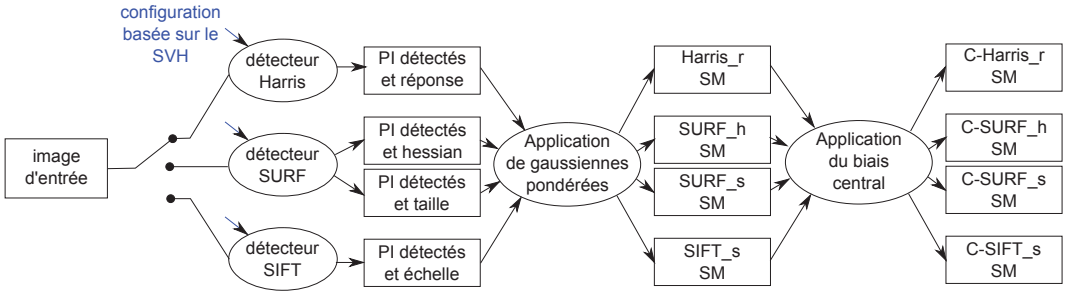


Figure 2.19 : Architecture de notre modèle

Sans ajout de coût de calcul important, nous intégrons également l'application d'un noyau gaussien central afin de mimer le biais centré caractéristique de la saillance humaine. L'ensemble de notre modèle proposé est illustré par la Figure 2.19.

L'évidente simplicité de notre méthode est un réel avantage car elle peut

être ré-implémentée facilement et garantit une faible complexité. De plus, cette méthode peut être très pratique lorsqu'un système existant utilise déjà un détecteur de points d'intérêt pour une toute autre tâche. En effet, l'ajout d'une brique de saillance visuelle est bien souvent très positif et augmente les performances de bon nombre d'algorithmes utilisant initialement les points d'intérêt, comme pour la navigation de robot [BZCM08], que pour la recherche d'images [WJY09] ou les métriques de qualité [TKCT10].

Un autre atout réside dans le fait de pouvoir tirer profit de tous les efforts et travaux passés et à venir sur l'augmentation de la vitesse d'exécution de la détection de ces points. Bien qu'actuellement déjà utilisés sous des contraintes temps réel, des travaux en cours tentent de réduire davantage ce temps d'exécution, avec par exemple une réduction de complexité d'un facteur 9.8 pour le détecteur de coins de Harris [MYL+11] ou une augmentation de vitesse d'exécution d'un facteur de 13 pour le déjà très rapide détecteur SURF [FYZ+11].

Plus de détails sur l'implémentation de notre méthode sont fournis et librement accessibles à l'adresse :

http ://www.sic.sp2mi.univ-poitiers.fr/imagequality/ipsaliency/

En plus du code source multiplate-forme fourni, nous distribuons un exécutable windows pour faciliter son utilisation. Nous tenons tout de même à rappeler que la version disponible n'est en aucun cas l'implémentation la plus rapide, car aucune optimisation n'a été effectuée. A titre d'exemple, le coût le plus important dans notre code est dédié à l'application des gaussiennes pour chaque point détecté. Donc, les ressources disponibles permettent simplement d'illustrer les cartes de saillance obtenues mais ne démontrent pas le faible temps d'exécution.

Les Figures 2.20 et 2.21 permettent de visualiser les cartes de saillance (SM) obtenues pour quelques images en utilisant les différents détecteurs et pondérations (sans l'utilisation du biais centré additionnel). La carte de saillance humaine obtenue par mesures oculométriques et après agrégation et pondération est également affichée à titre de référence. Nous fournissons également les SM de Achanta et Itti afin de comparer nos résultats à l'état de l'art. Des détails sur les performances de notre approche et une comparaison à l'état de l'art est disponible dans la section suivante.

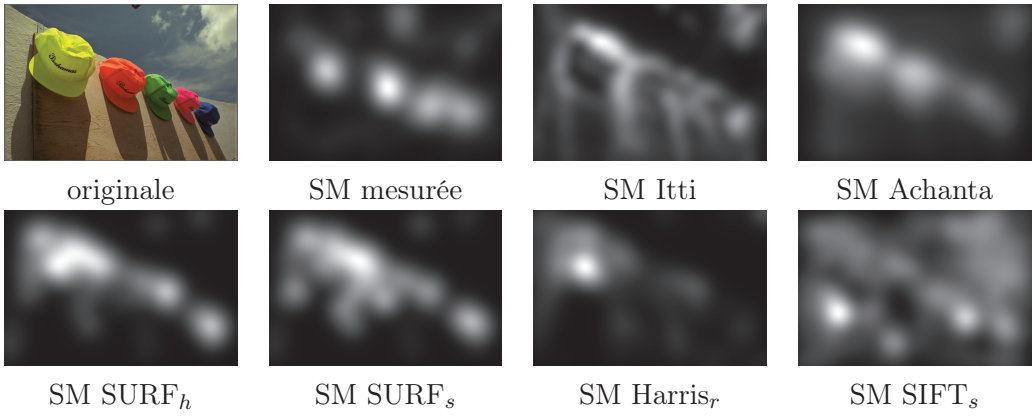


Figure 2.20 : Comparaison visuelle des cartes de saillance obtenues avec les points mesurés, prédits et de l'état de l'art sur l'image caps

2.5.1 Mesure de performance

Cette section est dédiée aux mesures de performances de notre méthode de création de cartes de saillance. Nous proposons ,dans un premier temps, de visualiser plusieurs résultats de cartes obtenues, sur les figures 2.20 et 2.21 pour plusieurs images et à travers l'utilisation des différents détecteurs de points d'intérêts et pondérations. Pour cette comparaison visuelle, nous fournissons également les cartes des deux méthodes de l'état de l'art (le modèle de Itti [IKN98] et celui d'Achanta [AHES09a]). Plus précisément, les cartes de saillance de Achanta ont été générées par l'application SalientRegionDetectorCVPR09 [AHES09b] et celles de Itti via son implémentation matlab [Har11]. Les cartes de saillance faisant office de vérité de terrain, obtenues par oculométrie sont également disponibles et notées SM mesurées.

Sur ces figures, nous pouvons noter les capacités de nos méthodes à prédire des cartes de saillance en accord avec les SM mesurées. C'est plus particulièrement notre approche utilisant les détecteurs Harris et SURF qui semblent les plus pertinentes. Tandis que la méthode exploitant SIFT apparaît moins précise, en fournissant des zones saillantes diffuses (avec un faible écart-type de pondérations) et donc peu de différences de saillance entre chaque zone. Nous pouvons à ce titre noter que les SM mesurées ont en moyenne seulement deux ou 3 zones très particulièrement saillantes, et donc un fort écart-type de pondérations. Ce type de distribution est donc plus comparable aux résultats obtenus par les pondérations des détecteurs de Harris et SURF.

En complément de l'analyse visuelle, nous utilisons deux méthodes objec-

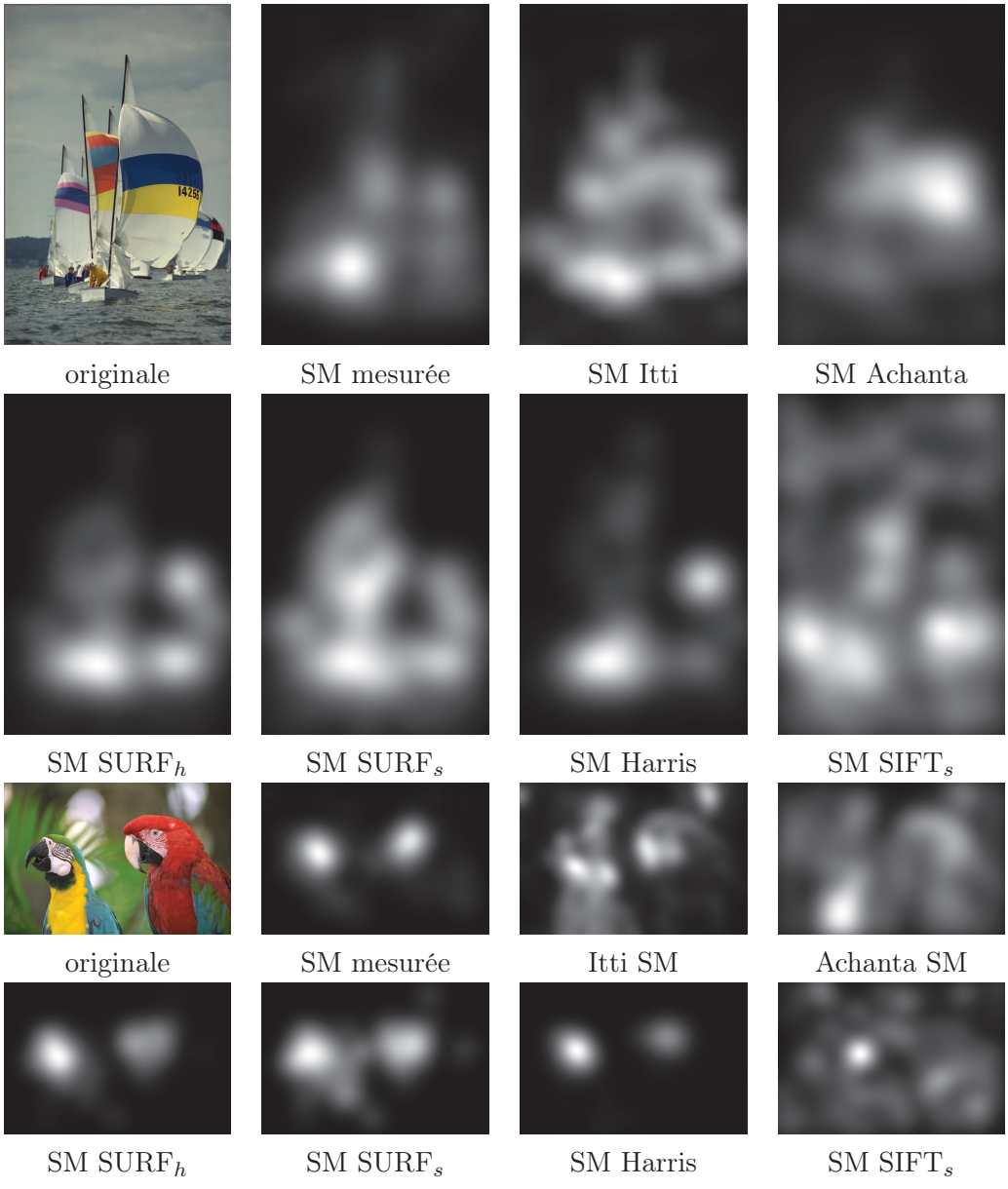


Figure 2.21 : Comparaison visuelle des cartes de saillance pour les images "sailing2" et "parrots".

tives de mesure de performance. La comparaison des modèles est effectuée sur trois bases de données de référence pour un total de plus de 1208 images testées. Nous utilisons la base VAIQ (cf. section 2.4.2) avec ses 42 images contenant une riche variété de contenus (visage, personne, animaux, gros plan, large plan, scène naturelle, paysage, objets manufacturés, images avec faible et forte différence entre premier et arrière plan, images avec et sans réel objet d'intérêt). Nous utilisons également la base TUD [ALRH, ALRH10], contenant 40 images différentes où chaque image contient une zone particulièrement attractive pour le regard humain avec des fonds plus ou moins complexes. Chacune de ces images est également disponible sous quatre versions compressées par JPEG. L'intérêt de cette base est de pouvoir évaluer la robustesse des approches en cas d'introduction de plusieurs niveaux d'artéfacts de compression (c'est donc un total de 160 images testées). Enfin, nous utilisons également la base MIT [JEDT09] qui contient 1006 images extraites de manière aléatoire de l'application internet flickR. Cette base a une importante quantité d'images avec un fort contenu sémantique. Elle est donc plus adaptée pour la saillance top-down, mais il peut être intéressant d'évaluer les capacités des approches bottom-up dans ce genre de contextes difficiles.

Pour ces mesures, nous évaluons également l'impact de l'ajout du biais centré sur chaque type de nos SM. Ces cartes sont identifiées par le préfixe C-, ce qui produit les cartes C-SIFT, C-Harris, C-SURF.

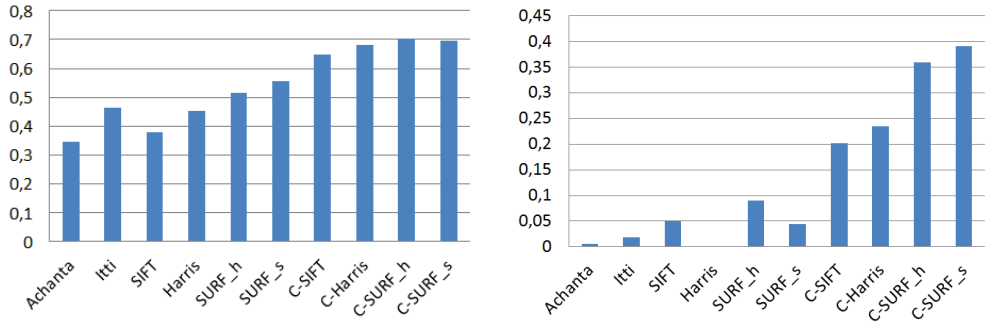
La première mesure de performance objective est une mesure de coefficient de corrélation entre les SM mesurées et chaque carte prédite. Le coefficient de corrélation entre deux cartes est déterminé par l'équation 2.2.

$$\rho = \frac{\sum_x [(M_h(x) - \mu_h) \times (M_c(x) - \mu_c)]}{\sqrt{\sum_x (M_h(x) - \mu_h)^2 \times \sum_x (M_c(x) - \mu_c)^2}}, \tag{2.2}$$

où $M_h(x)$ et $M_c(x)$ représentent respectivement la SM mesurée et la carte calculée à évaluer. μ_h et μ_c sont les valeurs moyennes des deux cartes $M_h(x)$ et $M_c(x)$.

Cette mesure de corrélation est réalisée sur toutes les images des bases citées précédemment. Nous fournissons les scores de résultats pour la base VAIQ dans le Tableau 2.22, TUD dans le Tableau 2.23 et dans le Tableau 2.24 pour la base MIT. Chaque tableau indique la corrélation moyenne, la corrélation minimale, la corrélation maximale et l'écart type des scores de corrélation.

Au regard de ces résultats, nous pouvons noter que nos propositions (sans biais centré) sont comparables voire meilleures que les modèles de Achanta et Itti en terme de corrélation, min, max et moyenne. Nous pouvons noter le bon



Corr. moyenne Corr. minimum

Figure 2.22 : Comparaison de corrélation sur base VAIQ.

comportement moyen de SURF$_s$ sur la variété des bases testées. Bien que cette méthode soit supérieure aux autres, nous pouvons tout de même souligner des scores de corrélation relativement faibles, autour de 55% sur la base TUD et de 35% sur la base MIT. Cependant pour chacune de ces bases délicates contenant des artéfacts de compression ou de forts contenus sémantiques, nous pouvons noter que notre méthode peut prédire plusieurs SM d'importantes corrélations, avec un maximum autour de 91% sur la base TUD et de 95% sur la base MIT.

En intégrant le biais centré, notre approche augmente ses performances d'environs 10% de corrélation. Ce biais centré est très connu et largement observé pour de nombreuses images. C'est pourquoi imiter ce biais garantit une bonne corrélation, tout en augmentant l'invariance aux artéfacts de compression (car calculé indépendamment du contenu de l'image). En complément des tableaux précédemment décrits, nous fournissons des représentations graphiques sur les figures 2.22 et 2.23 afin de mieux percevoir les performances de nos propositions.

Tableau 2.22 : Corrélation sur la base VAIQ (42 images).

	Acha.	Itti	SIFT	Harr.	SURF$_h$	SURF$_s$	C-SIFT	C-Harr.	C-SURF$_h$	C-SURF$_s$
moy.	0,345	0,464	0,380	0,453	0,514	**0,554**	0,647	0,682	**0,703**	0,697
min	0,005	0,018	0,050	0,001	**0,089**	0,044	0,202	0,235	**0,359**	0,390
max	0,728	0,832	0,733	0,918	**0,928**	0,907	0,874	**0,909**	0,892	0,903
std	0,218	0,198	0,165	0,270	0,235	0,190	0,149	0,148	0,128	**0,121**

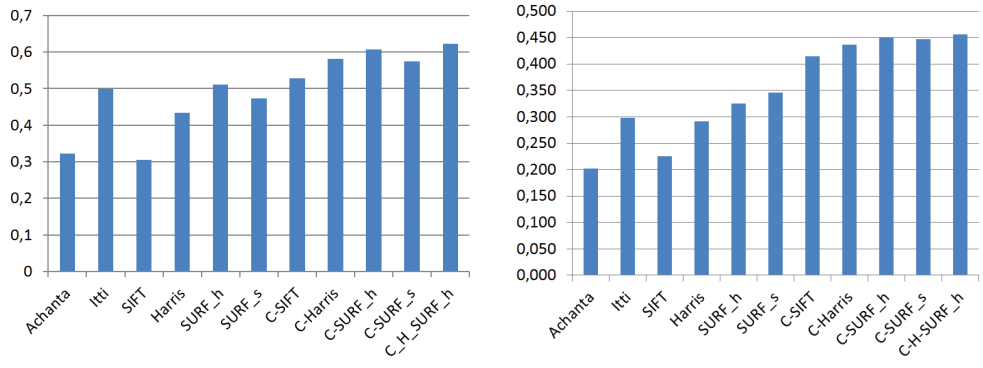
En complément des mesures de corrélation, nous utilisons le coefficient DICE ([PBT09]). Cette méthode consiste à binariser les deux cartes que nous cherchons à comparer par l'application d'un seuillage. Une fois les cartes binarisées et donc contenant des pixels pouvant être classés comme saillants ou non,

Tableau 2.23 : Corrélation sur la base TUD observation libre (160 images).

	Acha.	Itti	SIFT	Harr.	$SURF_h$	$SURF_s$	C-SIFT	C-Harr.	$C\text{-}SURF_h$	$C\text{-}SURF_s$
moy.	0,322	0,499	0,306	0,433	**0,511**	0,474	0,528	0,581	**0,607**	0,574
min	0,000	**0,076**	0,011	0,005	0,060	0,075	0,192	0,254	**0,332**	0,317
max	0,833	0,785	0,693	0,905	**0,915**	0,847	0,745	0,805	0,818	**0,825**
std	0,216	0,170	0,175	0,242	0,205	0,156	0,111	0,117	0,104	**0,099**

Tableau 2.24 : Corrélation sur la base MIT (1006 images).

	Acha.	Itti	SIFT	Harr.	$SURF_h$	$SURF_s$	C-SIFT	C-Harr.	$C\text{-}SURF_h$	$C\text{-}SURF_s$
moy.	0,202	0,298	0,225	0,292	0,325	**0,346**	0,415	0,437	**0,450**	0,447
min	0,000	0,001	0,000	**0,003**	0,000	0,002	0,003	0,019	**0,020**	0,009
max	0,892	0,847	0,676	**0,947**	0,935	0,935	0,697	0,748	**0,754**	0,737
std	0,163	0,172	**0,132**	0,212	0,206	0,181	**0,119**	0,136	0,128	0,121



Moyenne de corrélation sur la base TUD (160 images)

Moyenne de corrélation sur la base MIT (1006 images)

Figure 2.23 : Comparaison de corrélation.

il devient possible de quantifier le nombre des pixels bien classés. Si un pixel est classé saillant et qu'il l'est aussi sur la carte mesurée, ce pixel est considéré comme un Vrai Positif (sachant que nous considérons la SM mesurée comme la vérité de terrain). Si ce pixel n'est pas saillant dans la carte mesurée, il sera considéré comme un Faux Positif. A l'inverse, si ce pixel n'est pas considéré comme saillant, alors qu'il l'est réellement, ce pixel est un Faux Négatif. Le coefficient DICE est basée sur la quantification des vrais positifs, faux positifs et faux négatifs. Plus concrètement, ce score est déterminé par :

$$s = \frac{2|A \cap B|}{|A| + |B|} \tag{2.3}$$

$$s = \frac{2 \times VP}{2 \times VP + FP + FN} \tag{2.4}$$

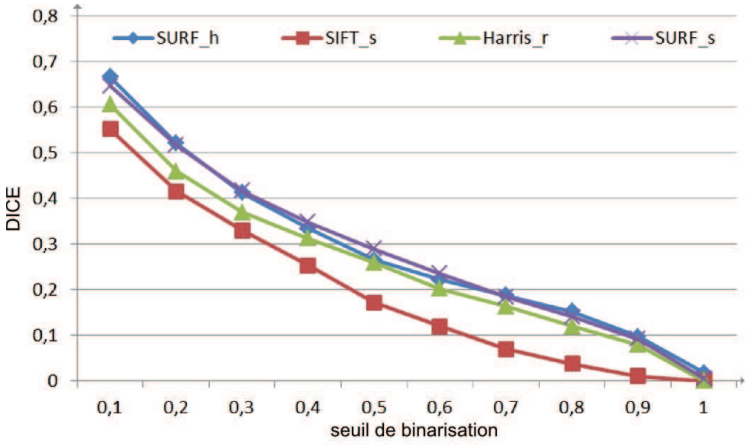
où VP = Vrai Positif, FP = Faux Positif et FN = Faux Négatif. Plus le coefficient est élevé, plus les images binaires sont similaires.

Nous fournissons l'évolution des scores DICE pour différents seuils de binarisation sur les trois bases de test, respectivement sur les figures 2.24, 2.25 et 2.26.

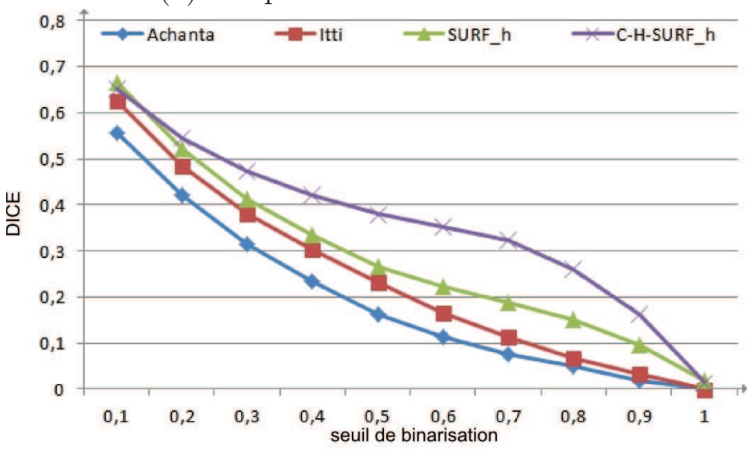
Dans un premier temps, nous pouvons noter que nos propositions Harris, SURF_h and SURF_s fournissent des résultats très proches, contrairement à SIFT qui a généralement des scores plus faibles, tels qu'illustré par les Figures 2.24-(a) et 2.26-(a). Cela peut s'expliquer par la différence d'écart-type des valeurs de pondération, comme déjà évoqué dans l'analyse par corrélation.

Par cette mesure, c'est SURF_h qui apparait comme la meilleure de nos propositions. C'est donc elle qui est utilisée afin d'effectuer une comparaison avec les méthodes de l'état de l'art comme le montre les Figures 2.24-(b), 2.25-(b) et 2.26-(b). Notre SURF_h est toujours nettement supérieure à la méthode de Achanta et légèrement supérieure à la méthode de Itti, ce qui est plus significatif par l'ajout du biais centré et principalement pour les hautes valeurs de seuillage.

Bien que nos résultats soient intéressants au vu de l'état de l'art, nous pouvons noter les faibles performances de tous les modèles sur la base TUD et particulièrement sur la base MIT qui contient une importante quantité de photo de visages et de personnes. De ce fait, nous proposons une extension à notre modèle C-H-SURF_h en intégrant un détecteur haut-niveau de reconnaissance de visage [VJ01], yeux, bouche et corps, à notre méthode C-SURF_h. Les effets de cet extension sur les bases VAIQ, TUD et MIT sont visible sur les

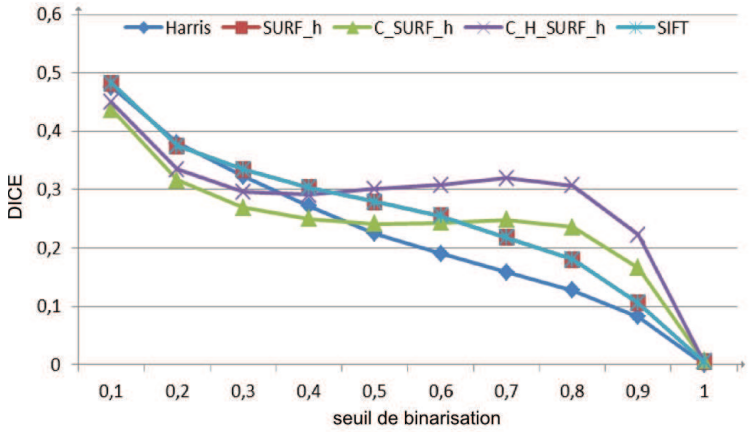


(a) comparaison de nos modèles

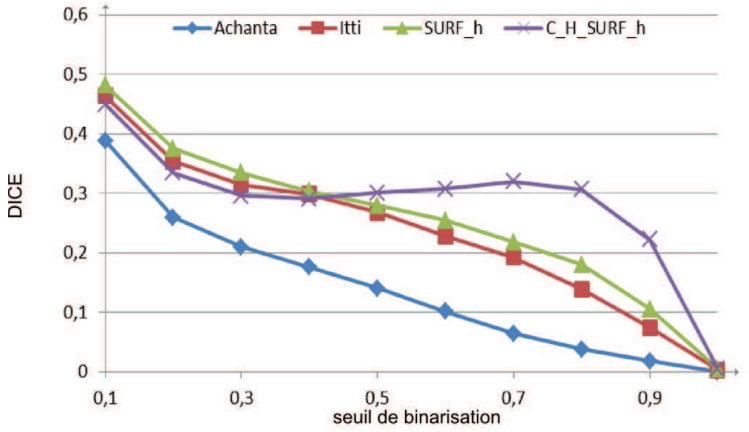


(b) comparaison avec Achanta et Itti

Figure 2.24 : Comparaison de DICE sur la base VAIQ.



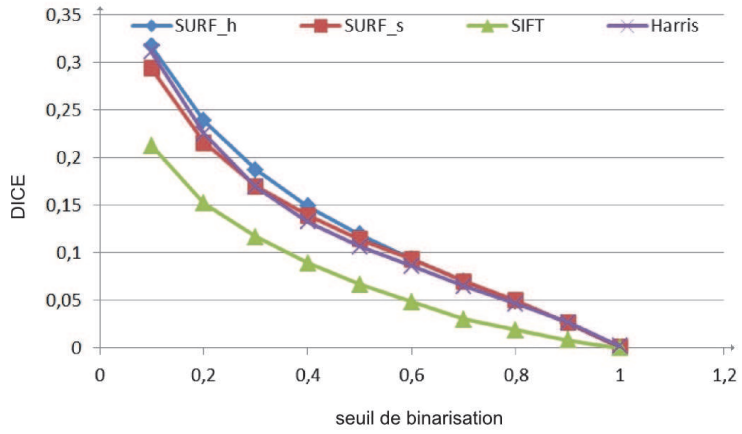
(a) comparaison de nos modèles



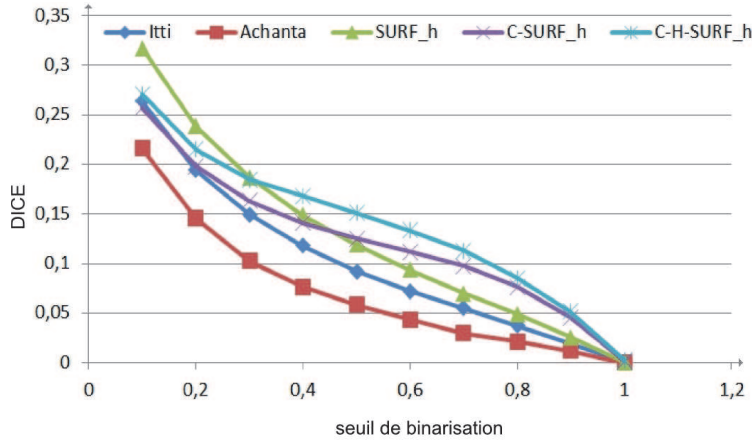
(b) comparaison avec Achanta et Itti

Figure 2.25 : Comparaison de DICE sur la base TUD (160 images).

Figures 2.24, 2.25 et 2.26. Avec cet ajout, nous pouvons noter une amélioration des performances, mais qui n'est pas justifiable au vu de la complexité de calcul ajoutée. Ce faible apport est probablement dû à la faible quantité de visages correctement détectés et l'importante quantité de fausses détections. Bien qu'ayant utilisé une méthode de reconnaissance classique et reconnue pour ce test, nous savons qu'il existe actuellement bon nombre de travaux récents très performants pour la détection de visages. Pour preuve, il suffit d'observer le comportement de tous les appareils photo récents capables de détecter correctement les visages, nous pensons donc que ce type d'algorithmes pourrait clairement changer et augmenter significativement les résultats. Mais cet objectif n'étant pas notre but initial, nous ne développerons donc pas plus d'études en ce sens dans cette thèse.



(a) comparaison de nos modèles



(b) comparaison avec Achanta et Itti

Figure 2.26 : Comparaison de DICE sur la base MIT (1006 images).

Pour conclure, nous avons démontré les performances de nos propositions à travers des analyses subjectives et objectives, et plus particulièrement noté

que C-SURF$_h$ fournit des résultats comparables et meilleurs que les modèles calculatoires et bio-inspirés de la littérature.

2.6 Conclusion

Au cours de ce chapitre, nous avons mis en avant l'existence de ressemblance entre les détecteurs de points d'intérêt et les points de fixation du regard humain. Pour ce faire, nous avons mené une étude statistique afin de déterminer les paramètres optimaux des différents détecteurs de points d'intérêt et avons révélé qu'avec un paramétrage spécialisé, la saillance visuelle est très corrélée aux points des détecteurs objectifs. Une importante rigueur a été portée sur cette étude afin de disposer de données fiables et de mesures non biaisées.

Grâce à la preuve d'existence d'une forte ressemblance, nous avons pu développer un nouveau modèle de prédiction de saillance, à la fois précis et peu coûteux en temps de calcul, car tirant profit des nombreuses optimisations d'exécution des détecteurs de points d'intérêt. Les performances de cette proposition ont également été évaluées en détail et comparées à l'état de l'art en révélant les avantages de notre méthode.

Par cette contribution, notre prédiction de saillance peut être intégrée très facilement et rapidement dans bon nombre d'applications, assurant ainsi des gains de performance notables prenant en compte des paramètres psychovisuels dans des chaînes de traitement souvent contraintes à ne pas pouvoir respecter ce genre de considérations.

Au vu des résultats encourageants par l'utilisation des points d'intérêt pour la prédiction de la saillance visuelle, nous proposons dans toute la suite de cette thèse, d'évaluer et exploiter les capacités et intérêts de ces détecteurs dans les tâches de quantification et d'estimation de la qualité perçue et de l'augmentation de la qualité de l'expérience.

Chapitre 3

Introduction à la qualité de l'expérience

Sommaire

3.1 Introduction **91**
3.2 Expérience psychovisuelle pour l'évaluation de la qualité . **93**
3.2.1 Méthodologie d'évaluation subjective de la qualité . 93
3.2.2 Conduite de mesures de qualité pour le comité JPEG 97
3.3 État de l'art des métriques de qualité d'images **103**
3.3.1 Métriques mathématiques 104
3.3.2 Métriques perceptuelles 106
3.3.3 Métriques pondérées par le SVH 107
3.3.4 Métriques à référence réduite et sans référence . . . 108
3.4 Méthodologie d'évaluation des performances des métriques **110**
3.4.1 Précision de la prédiction : Corrélation de Pearson et RMSE 111
3.4.2 Uniformité de la prédiction : Outlier Ratio (OR) . . 112
3.4.3 Monotonie de la prédiction : Corrélation d'ordre de Spearman 112
3.4.4 Bases d'images dédiées à la qualité 115
3.4.5 Mise en place d'une application web dédiée à la comparaison de métriques 117
3.5 Conclusion **121**

3.1 Introduction

Nous proposons dans ce chapitre d'introduire la notion de qualité de l'expérience, tout en détaillant particulièrement une de ses composantes dans notre domaine, à savoir, la qualité des images.

La qualité de l'expérience, accompagnée de tous les artifices servant à l'améliorer, n'est pas un concept réellement nouveau, puisque l'on peut en trouver une genèse dès l'antiquité, avec par exemple Epicure et la naissance de l'hédonisme. L'hédonisme est une doctrine philosophique dont l'élément fondamental est la recherche du plaisir et l'évitement du déplaisir. De manière plus contemporaine et technologique, nous pouvons retrouver ces préceptes dans la norme ISO 9241 [ISO98] qui définit l'utilisabilité comme "le degré selon lequel un produit peut être utilisé, par des utilisateurs identifiés, pour atteindre des buts définis avec efficacité, efficience et **satisfaction**, dans un contexte d'utilisation spécifié". Il est également possible de comparer cette définition à celle donnée par l'IEA (International Ergonomics Association) [IEA00] qui décrit l'ergonomie (*Human Factors*) comme "la discipline scientifique qui vise la compréhension fondamentale des interactions entre les humains et les autres éléments d'un système, et la profession qui applique les principes théoriques, les données et les méthodes en vue d'optimiser le **bien-être** des personnes et la performance globale des systèmes". Mais c'est finalement dans les applications plutôt orientées télécommunications et transmissions des signaux que le terme QoE (Quality of Experience) est apparu [EZ11]. Dans ces domaines, il était courant de ne considérer que la QoS (Quality of Service) basée sur des mesures purement matérielles et logicielles, tels que la quantification du taux d'erreur binaire, le rapport signal sur bruit, etc. Cependant, ces mesures ne reflètent pas toujours le ressenti de l'utilisateur humain en fin de chaîne. C'est donc pour effectuer des mesures de performance de bout en bout, en incluant l'humain dans la boucle que la notion de QoE est apparue. Finalement, cette notion est pluridisciplinaire, et à ce titre nous pouvons citer l'effort du consortium QUALINET (European Network on Quality of Experience in Multimedia Systems and Services) [oQoEiMSS] qui tente de rassembler et joindre les efforts de tous les acteurs de la recherche dans ce domaine fragmenté.

Pour résumer, la QoE vise à estimer le degré de satisfaction et de plaisir que ressent l'utilisateur d'un système. Cette notion est très large et englobante, et peut aussi bien être utilisée pour mesurer l'apport de la nouvelle interface graphique d'un site internet, que pour évaluer le niveau de satisfaction d'un utilisateur lors de l'affichage d'une photographie sur un smartphone malgré une faible couverture réseau et un fort ensoleillement. De ce fait, pour la suite de ce chapitre, nous proposons de nous focaliser uniquement sur les mesures liées à la qualité des images.

Nous proposons, dans une première section, de décrire les règles et méthodologies théoriques permettant de mesurer la qualité perçue des images, associées à une application pratique dans le cadre de la conduite d'une campagne de tests pour le comité JPEG. Nous décrivons dans un second temps les modèles et méthodes logicielles, couramment appelées métriques de qua-

lité, à même d'estimer et de prédire la qualité perçue. Nous détaillons également les procédures existantes permettant de mesurer les performances de ces métriques en terme de ressemblance et donc de corrélation avec le jugement humain. C'est à ce titre que nous développons également une contribution, en proposant un service-web spécialement conçu pour le "benchmark" (en français l'étalonnage, à savoir l'établissement d'un indicateur de performance) et l'utilisation des métriques de qualité. Ce service doit permettre de faciliter la comparaison des métriques et de démocratiser leur utilisation.

3.2 Expérience psychovisuelle pour l'évaluation de la qualité

3.2.1 Méthodologie d'évaluation subjective de la qualité

Afin d'obtenir les informations issues du système sensoriel humain, pour à terme augmenter sa sensation de confort, diverses méthodes et protocoles ont été proposés. Cependant, dans le cadre de ce chapitre, nous allons nous focaliser uniquement sur les méthodes permettant de mesurer l'opinion de qualité moyenne, nommée MOS (Mean Opinion Score), lors de l'observation d'images dégradées par l'introduction d'artéfacts de compression, de transmission ou tout autre traitement algorithmique ou physique ayant pu perturber l'intégrité des images.

En ce qui concerne ce type d'évaluations, nous pouvons les classer en deux grandes familles : les test comparatifs et les tests de mesures absolues.

Évaluation comparative

Les tests comparatifs permettent à l'observateur de juger une image en la comparant à une autre. Dans cette catégorie, nous pouvons décrire la méthode DSIS (Double Stimulus Impairment Scale, soit en français : méthode à double stimulus utilisant une échelle de dégradation) [ITU02, ITU08]. Cette méthode formalisée et normalisée par l'ITU, permet d'appréhender facilement le concept d'évaluation de la qualité. Dans ce type d'expériences, une première image de référence est affichée à l'écran pendant dix secondes et doit être considérée comme étant de qualité parfaite. Après ce laps de temps, elle disparaît pour laisser apparaître une image grise et neutre pendant trois secondes. Puis, l'image à juger apparaît pendant dix secondes à l'écran. Enfin, l'observateur

doit donner son avis sur la qualité de la seconde image au regard de la première. On peut parler de mesure de fidélité à la place de qualité puisque le jugement est effectué en mode comparatif, où la qualité maximum est atteinte quand aucune différence n'est observée. Afin d'exprimer son ressenti perceptuel sur la fidélité de l'image, une échelle graduée est fournie à l'observateur, sur laquelle il peut choisir entre cinq catégories afin de juger la dégradation perçue (1 : Très dérangeant, 2 : Dérangeant, 3 : Peu dérangeant, 4 : perceptible mais pas dérangeant et 5 : imperceptible).

Des variantes existent, tels que la DSCQS (Double Stimulus Continuous Quality Scale, en français, échelle continue de la qualité sur double stimuli)[ITU02, ITU05] où l'échelle n'est pas discrète, mais continue [1], permettant ainsi plus de flexibilité et de finesse dans les mesures. Une autre différence peut être notée dans ce test, l'observateur ne sait jamais quelle image est la référence puisqu'elles apparaissent dans un ordre pseudo-aléatoire. De ce fait, il a pour consigne de juger la qualité des deux images présentées. Le score de l'image de test est donc obtenu par différence entre la note donnée à l'image de référence et celle de l'image testée, dont une illustration est fournie en Figure 3.1

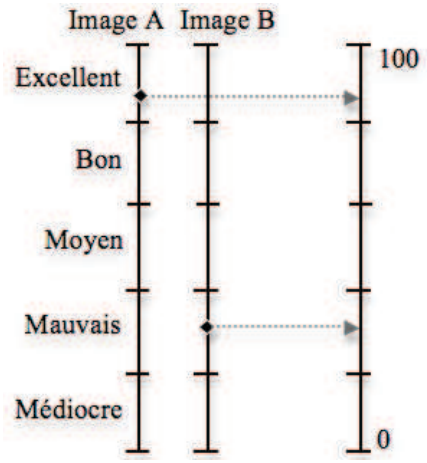


Figure 3.1 : Echelle de notation de la qualité

Il existe bien sûr d'autres méthodes de ce type, dites à double stimuli, où il est par exemple question de classement d'images ou bien d'afficher simultanément (et non les unes à la suite des autres,) les couples d'images. Mais dans tous les cas, l'observateur a toujours affaire à une évaluation comparative.

1. L'échelle ne peut pas être continue au sens strict du terme, mais elle fournit tout de même des valeurs allant de 0 à 100

Mesure absolue

Dans le cadre des mesures absolues, telles que la méthode ACR5 [ITU02, ITU08], il n'est pas question de jugement par couple. L'observateur doit juger toutes les images les unes à la suite des autres sans jamais avoir à les comparer à une image dite de référence. Les images sont donc jugées dans l'absolu, sur une échelle allant de 1 à 5 comme présenté précédemment. Il existe également des variantes où l'échelle n'est pas la même telle que la méthode ACR11 avec ses 11 catégories. Bien que l'utilisateur ne le sache pas, il est possible de placer des images de référence à évaluer. Dans ce cas, on parle alors de méthode ACR5-HR (Hidden Reference ou à référence cachée)/ACR11-HR, permettant ainsi l'obtention de scores par mesures de différence.

Les différentes méthodes présentées sont décrites et normalisées par l'Union Internationale des Télécommunications (UIT ou ITU en anglais pour *International Telecommunication Union*). Cet organisme a également émis des recommandations concernant les observateurs et l'environnement dans lequel se passe ce type de campagnes de test, afin de garantir fiabilité et reproductibilité.

Recommandations ITU

En ce qui concerne les observateurs, l'ITU [ITU09] recommande un panel d'un minimum de 15 personnes. Ce nombre d'observateurs minimum est également appuyé par les travaux de Winkler [Win09] qui définit une stabilisation et un passage sous le seuil d'intervalle de confiance de 95% à partir de 10 observateurs. Pour ces observateurs, il est également conseillé d'avoir testé leur acuité visuelle grâce à l'échelle de Snellen et leur capacité à distinguer les couleurs en utilisant le test d'Ishihara. Enfin, certaines recommandations conseillent de choisir un panel d'observateurs relativement jeunes pour leurs aptitudes visuelles [KOD09], elles suivent en cela les désirs de nombreux industriels du domaine de l'imagerie numérique qui eux souhaitent à terme satisfaire leurs futurs utilisateurs.

Après la sélection des observateurs répondant aux critères des tests visuels, on conseille de prévoir, avant chaque séance de tests, une explication sur le type de méthodologie employée, le système de notation, le protocole de présentation et sur tout élément que le conducteur des tests juge utile. Les conditions psychologiques dans lequel se situe l'observateur sont à la fois difficilement définissables et très influentes sur son évaluation, ce qui donne une grande importance à ses explications préliminaires. Il est également recommandé de débuter les présentations par quelques cas typiques permettant d'ancrer le ju-

gement des observateurs. Ces présentations d'entrainement ne seront pas prises en compte dans les résultats finaux. En fin de séance, il est possible de réaliser un bilan individuel en vue de détecter d'éventuelles mauvaises appréciations.

Au sujet de l'environnement dans lequel se passe les campagnes d'évaluation, l'ITU préconise de nombreuses contraintes à respecter ; le but étant toujours de mettre les observateurs dans les meilleures conditions en favorisant au maximum leur concentration et leur confort visuel.

Tout d'abord, l'éclairage doit être constant pendant toute la période de l'évaluation, avec un éclairage ambiant de 200 lux par exemple. Il est également conseillé d'éviter l'éclairage direct sur les écrans et d'éviter les reflets.

En ce qui concerne les écrans, leurs caractéristiques doivent être suffisamment élevées afin de garantir un contraste d'affichage suffisant et d'une valeur de luminance d'au moins 200 cd/m^2. Le réglage de la correction Gamma des couleurs doit également être effectué sur chaque écran, par l'utilisation de sondes de calibrage. Certaines indications concernent également le type et la couleur des revêtements des murs de la salle, afin de s'assurer de leur niveau de réflexion de la lumière, et de ne pas être attractif pour le regard ou d'engendrer la fatigue visuelle.

La distance entre l'observateur et l'écran est également un facteur important, car elle influence la visibilité de certaines fréquences spatiales. Il est donc conseillé que la distance permette à un pixel d'occuper une minute d'arc du champ de vision de l'observateur ; ce qui dans la pratique revient à considérer trois fois la hauteur de l'écran s'il est de format $16/9^{\text{ème}}$. Enfin, dans le but de garantir une concentration maximum et d'éviter la lassitude, la durée des tests ne doit pas excéder 30 minutes.

Exploitations des résultats

Une fois que tous les observateurs de l'expérience sont passés, dans de bonnes conditions d'évaluation, il est maintenant question d'extraire et d'analyser les résultats. Pour ce faire, il est tout d'abord nécessaire de calculer le score de qualité moyen pour chaque image de l'évaluation. Cette moyenne d'opinion, nommée MOS s'obtient par la formule :

$$MOS_{jk} = \frac{1}{N} \sum_{i=1}^{N} s_{ijk}, \tag{3.1}$$

où S_{ijk} est le score de l'observateur i pour la dégradation j de l'image k. Le nombre d'observateurs est noté N. Il est également conseillé d'ajouter au calcul de chaque MOS, la mesure de l'intervalle de confiance à 95% définie par :

$$IC_{jk} = 1.95\frac{\sigma_{jk}}{\sqrt{N}}, \tag{3.2}$$

avec

$$\sigma_{jk} = \sqrt{\sum_{i=1}^{N} \frac{(s_{ijk} - MOS_{jk})}{N-1}}. \tag{3.3}$$

Si l'on considère que les scores respectent la contrainte de distribution normale, ils se trouvent dans l'intervalle $[MOS_{jk} - IC_{jk}, MOS_{jk} - IC_{jk}]$ avec une probabilité de 95%.

Comme toute mesure, certaines données peuvent parasiter les résultats. Dans le cas présent et puisque les données sont issues d'observateurs, il se peut qu'ils aient mal compris les consignes, ou que leur concentration et vigilance aient pu chuter au cours du test. Il est donc également conseillé d'appliquer un processus de filtrage sur les données collectées afin d'éviter de considérer des scores aberrants, dans le sens où ils s'écartent trop de la moyenne des observateurs. Cette tâche est généralement assurée par un test de Kurtosis pour chaque MOS_{jk} :

$$\beta_{jk} = \frac{\frac{1}{N}\sum_{i=1}^{N}(MOS_{jk} - S_{ijk})^4}{(\frac{1}{N}\sum_{i=1}^{N}(MOS_{jk} - S_{ijk})^2)^2}, \tag{3.4}$$

pour lequel la valeur β_{jk} doit être comprise dans l'intervalle $[2, 4]$. Dans le cas inverse, un processus de filtrage doit être appliqué sur les scores ayant constitué ce MOS_{jk} afin d'isoler et de rejeter les scores des observateurs aberrants.

3.2.2 Conduite de mesures de qualité pour le comité JPEG

Ayant décrit de manière théorique les concepts et préceptes de l'évaluation subjective, nous proposons dans cette section un exemple concrêt de campagne de tests subjectifs menée au sein du laboratoire Xlim-SIC. Cette expérimentation s'inscrit dans une démarche de validation de nouveaux algorithmes de compression pour le compte du comité ISO/JPEG [com86] responsable de la normalisation de ce type de technologie.

Plus précisément, nous sommes intervenus dans le cadre du *core-experiment CO-LAR-02* et du projet CAIMAN dont l'objectif était d'évaluer les performances d'un nouveau codec nommé LAR [DBN+04] proposé à la normalisation en réponse à l'appel à proposition du groupe AIC (Advanced Image Coding). Cette expérimentation met en compétition le LAR avec trois autres codecs actuellement normalisés que sont JPEG [JPE91], JPEG 2000 [JPE00] et JPEG XR [JPE09]. Le but étant de tester de manière subjective la visibilité et la gêne occasionnées par une compression de type LAR et de la comparer aux autres codecs existants.

Procédure d'évaluation

Pour cette campagne, les observateurs ont pour objectif de juger la qualité perçue des images présentées sur l'écran par une méthodologie de type double stimuli. L'image sans dégradation est toujours présentée à gauche de l'écran et l'image dégradée à évaluer est, quant à elle, toujours à droite. Un logiciel spécialement conçu pour ce genre d'expérimentation, dont une représentation est fournie sur la Figure 3.2, permet à l'observateur de donner son score de qualité perçue.

Figure 3.2 : Capture d'écran du logiciel d'évaluation subjective utilisé pour le *core-experiment CO-LAR-02*

L'observateur fournit un score de qualité en utilisant une échelle graduée

continue allant des évaluations « très mauvaise » à « très bonne ». Il s'agit donc d'un processus d'évaluation avec échelle continue de la qualité sur double stimuli.

Voici une version résumée des paramètres de l'évaluation :
- Images originale et dégradée présentées côte à côte (Double stimulus) ;
- Score de qualité sur une échelle continue (0 à 100) ;
- Pas de répétition des images ;
- Session d'entrainement permettant de présenter l'amplitude des dégradations ;
- Session d'évaluation d'une durée moyenne de 18 minutes.

Nous proposons dans la suite de cette section de fournir de manière détaillée tous les paramètres que nous avons appliqués pour ce *core-experiment COLAR-02*.

Observateurs

Cette campagne de test a nécessité de disposer d'un panel d'observateurs respectant le nombre minimum de 15 personnes préconisé par l'ITU. C'est au final 17 observateurs qui ont participé à cette évaluation, d'une moyenne d'âge de 29 ans. Tous ces observateurs ont prouvé leurs aptitudes en réussissant le test d'acuité de Snellen et de perception des couleurs de Ishihara. Ils ont également eu une phase d'entraînement et d'explication pour ce test spécifique d'évaluation de la qualité.

Codecs

Les paramètres et codecs suivants ont été utilisés afin de pouvoir comparer le codec LAR avec les autres standards JPEG (dont certaines fonctionnalités ont été bridées afin de garantir une certaine équité entre les codeurs, sauf pour JPEG baseline qui n'existe que dans cette version) :

JPEG 2000 (Kakadu version 6.3)
- configuration : 4 : 4 : 4
- 64×64 : taille de code-bloc
- 1 couche, sans precincts
- 1 tuile
- ondelette 9×7
- 5 niveaux de décomposition

– sans pondération visuelle

JPEG IJG-codec (version 8a)

– configuration :YCrCb 4 : 2 : 0
– optimisation visuelle de la matrice de quantification (spécif. standard)
– codage de Huffman

JPEG XR (reference software wg1n5233)

– configuration : 4 : 4 : 4
– sans pondération visuelle
– un niveau de chevauchement
– sans répétition

Nous pouvons noter que mis à part JPEG, les autres codeurs n'ont aucune optimisation dans les profils utilisés. Différents taux de compression ont également été testés afin de quantifier la performance de compression de chaque codec dans différents cas d'utilisations.

Débit binaire en bpp (bits par pixel)

– $[0.25, 0.5, 0.75, 1, 1.25, 1.5]$

Salle de tests psychovisuels

En ce qui concerne l'environnement d'évaluation, les tests subjectifs se sont déroulés dans une salle spécialement conçue pour ce type d'expérimentations au sein du laboratoire Xlim-SIC et dont voici les principales caractéristiques :

Écran : Dell monitor

– Résolution : 2560×1600
– Largeur : 63,5 cm
– Hauteur : 39,5 cm
– Luminosité : 370 cd/m^2

Outils de calibration

– Mesure de l'énergie lumineuse : Spectrophotomètre PR-650
– Ajustement des couleurs et luminosité de l'écran : Eye-one Display2

L'éclairage

– Tubes à néon D65
– 65 lux de lumière indirecte

Distance d'observation

– maintenue à 1,5 fois la largeur de l'écran

Images de référence

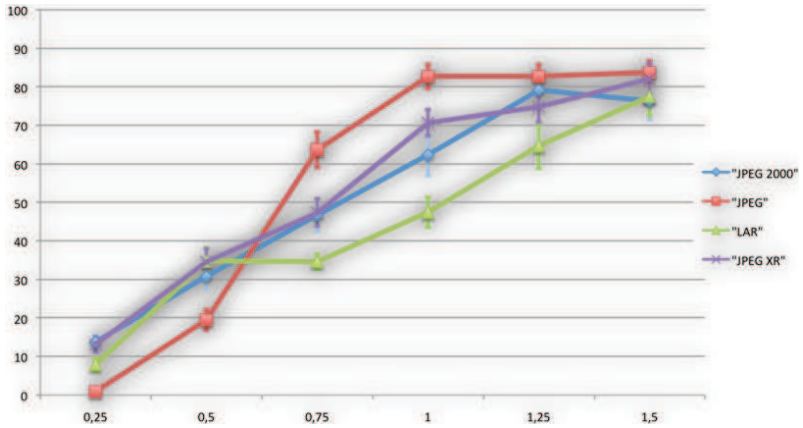
Les images utilisées pour cette évaluation sont issues de la base de test du comité JPEG (bike, woman, P10, rokuonji, green, northcoast, P26, P22, honolulu-zoo). Afin de pouvoir afficher simultanément deux images sur l'écran d'évaluation d'une résolution de 2560×1600, chaque image est découpée pour atteindre une résolution de 1280×1600. De ces neuf images, trois sont utilisées pour la phase d'entrainement, et les six autres pour l'évaluation. Disposant de six images, six taux de compression et quatre codecs, c'est un total de cent quarante quatre images qui ont été jugées par chaque observateur.

Ayant décrit la procédure et les différents paramètres mis en œuvre pour cette campagne de test, nous proposons de détailler quelques résultats.

Résultats

Une fois les scores de qualité des observateurs recueillis, en ayant veillé à appliquer les procédures de rejet de valeurs aberrantes, nous pouvons présenter quelques résultats de cette campagne d'évaluation. Nous les fournissons pour deux images sur les Figures 3.3 et 3.4 où l'axe des abscisses représente l'évolution du débit binaire et les MOS sur l'axe des ordonnées.

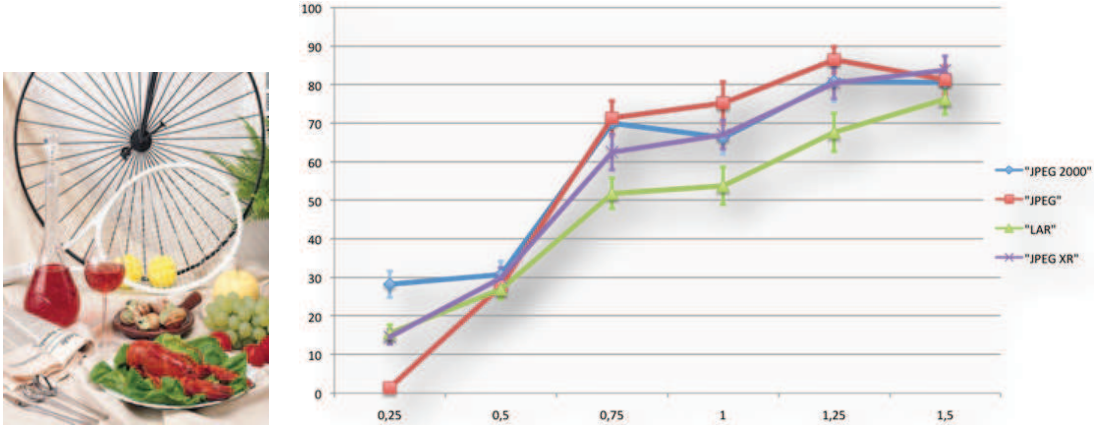
(a) woman



(b) Scores subjectifs des codecs

Figure 3.3 : Scores subjectifs des codecs sur l'image "woman"

En observant ces figures, il est possible de constater que le codec LAR introduit des dégradations visibles sur la totalité des taux de compression testés.



(a) bikes (b) Scores subjectifs des codecs

Figure 3.4 : Scores subjectifs des codecs sur l'image "bikes"

Nous pouvons remarquer qu'au vu de la qualité ressentie par les observateurs pour les forts taux de compression (0.25 bpp), c'est JPEG 2000 qui apparait comme le plus performant et JPEG comme le pire, tandis que le LAR est situé entre les deux. Nous pouvons ensuite observer une très forte ressemblance des résultats pour le taux de 0.5 bpp où tous les codecs sont jugés de manière équivalente. Cependant, pour les faibles taux, à partir de 0.75 bpp, les différences de qualités sont plus visibles et accentuées et font apparaître le codec LAR en dessous des codecs déjà normalisés. Par ces mesures subjectives le LAR ne semble pas démontrer sa pertinence. Par la suite, ce codec a également fait l'objet de mesures de qualité objectives par diverses métriques. Pour conclure, le comité ISO/JPEG a statué sur ce codec en considérant qu'il n'était pas encore assez mature pour être standardisé, mais qu'il offrait tout de même quelques avantages en terme de services disponibles.

Cette campagne de tests illustre le fait que l'introduction de l'humain dans la boucle des développements technologiques est de plus en plus prise en compte par les communautés scientifiques et industrielles. A ce titre, notre laboratoire a également participé à l'évaluation de nombreux codecs vidéo 3D dans le cadre d'une campagne d'évaluation internationale MPEG d'une grande ampleur, ou plus d'une dizaine de codecs ont été évalués par de nombreux laboratoires à travers le monde, afin de garantir la meilleure qualité de l'expérience pour la transmission et la diffusion des futures vidéos 3D.

Bien que très utiles, les campagnes de tests subjectifs sont délicates car elles nécessitent le respect d'une méthodologie stricte et un équipement spécialisé. En plus du temps conséquent nécessaire à leur réalisation, d'importants moyens humains et financiers doivent être dégagés. C'est pourquoi, des efforts

de recherche ont été consacrés pour produire des modèles et algorithmes capables de prédire la qualité perçue. Nous proposons dans la section suivante de décrire certaines de ces méthodes, également appelées métriques de qualité.

3.3 État de l'art des métriques de qualité d'images

Cette section a pour objectif de compléter le tour d'horizon des méthodes d'évaluation de la qualité, en s'intéressant ici, aux méthodes objectives que sont les métriques de qualité [2]. Nous avons vu précédemment qu'il existait plusieurs types d'évaluations subjectives. Certaines permettent à l'observateur de se référer à tout instant à l'image de référence pour juger la qualité de l'image testée, tandis que pour d'autres, l'image d'origine n'est pas ou plus accessible. Nous pouvons tout d'abord faire un parallèle de ce manque de disponibilité de l'image de référence pour classer les métriques objectives. Elles peuvent être de trois types : les métriques avec référence (FR - *Full Reference* en anglais), avec référence réduite (RR - *Reduced Reference*) et sans référence (NR - *No Reference NR*).

Les métriques FR utilisent l'intégralité de l'image sans dégradation pour effectuer les comparaisons. Tandis que les métriques NR utilisent seulement l'image dégradée. C'est une tâche facile pour l'homme mais très complexe pour une machine. Enfin, les métriques RR extraient un minimum d'attributs de l'image sans distorsion, puis la comparaison s'effectue sur l'image dégradée avec ce minimum d'information. La Figure 3.5 permet d'illustrer ces trois catégories.

D'un point de vue applicatif, les métriques FR sont généralement utilisées pour mesurer la fidélité des images ayant subi des traitements de compression, de tatouage, de restauration, etc. Elles sont utilisées pour leur grande sensibilité à de nombreux types de dégradations. Ces précision et robustesse sont possibles grâce à la disponibilité à tout instant de tous les pixels de l'image. Dans ces problématiques, le temps d'exécution n'est pas non plus une réelle contrainte.

Ce n'est pas le cas des métriques NR et RR qui sont quant à elles, le plus souvent, destinées à être utilisées pour des traitements en ligne et généralement sur des systèmes embarqués. Cela peut être le cas pour du monitoring TV [MVF12, CO03], ou de la transmission sur terminaux mobiles [Eng08, Kus05].

Hormis les contraintes de temps d'exécution et de disponibilités de réfé-

2. Par abus de langage, énormément d'algorithmes d'estimation de la qualité portent le nom de métriques, alors que les modèles produits ne le sont pas au sens strict et mathématique du terme car ne respectant pas l'inégalité triangulaire.

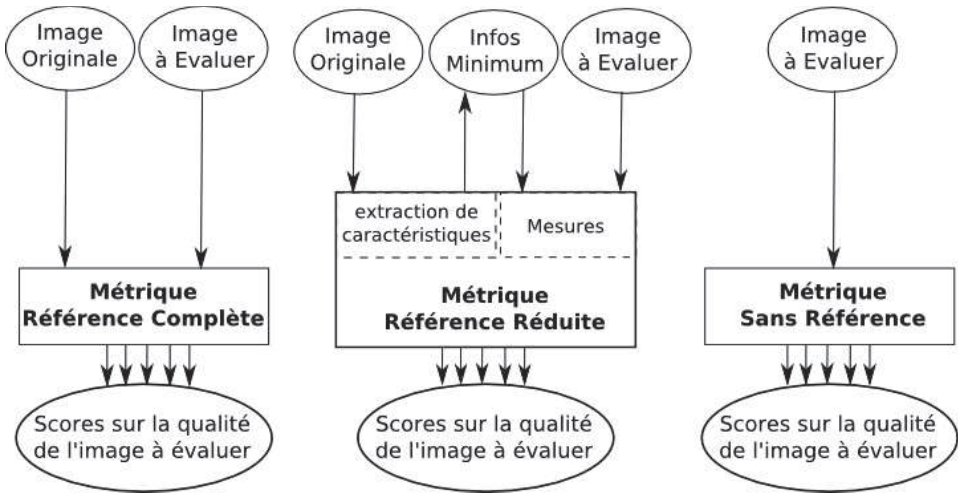


Figure 3.5 : Types des métriques de qualité

rences, les métriques peuvent également être classées en fonction du type de statistiques de l'image qu'elles utilisent ainsi que leurs prises en compte des aspects perceptuels et bio-inspirés.

3.3.1 Métriques mathématiques

Il y a tout d'abord les métriques purement mathématiques, directement issues du monde du traitement du signal 1D. Dans ces métriques, le facteur 2D des images et les relations existant entres les pixels ne sont pas pris en compte. Il s'agit dans ce cas d'analyse pixel à pixel.

Tout d'abord, la métrique EQM pour Erreur Quadratique Moyenne est définie par :

$$EQM = \frac{1}{M.N} \sum_{m=1}^{M} \sum_{n=1}^{N} ||I_o(m,n) - I_d(m,n)||^2, \tag{3.5}$$

avec I_o et I_d représentant respectivement l'image d'origine et l'image dégradée, toutes les deux de dimension $M \times N$.

De cette mesure, une variante nommée PSNR (Peak Signal-to-Noise Ratio) introduit l'utilisation d'un logarithme permettant de s'adapter à la dynamique du signal étudié :

$$PSNR = 10log_{10}(\frac{a^2}{EQM}), \tag{3.6}$$

avec a, l'amplitude maximale du signal considéré. Si l'on considère une image monochrome codée sur 8 bits par pixel, a sera égale à 255.

Cette métrique PSNR est une des métriques les plus anciennes, mais aussi la métrique la plus connue et la plus utilisée, tous domaines confondus. Cependant, par son approche pixel à pixel, elle ne semble pas réellement être la plus adaptée aux signaux images, qui par nature, ont des relations entre les pixels. Ce sont ces liens entre pixels, qui font apparaître tout type de structures, telles que les contours, les textures, etc.

C'est avec l'idée de mesurer les différences de structures, que la métrique SSIM [WBSS04] s'est également fait connaître. Pour étudier les variations structurelles, cette métrique considère des fenêtres d'observations locales, desquelles sont extraits trois critères :

- la valeur moyenne locale pour l'étude de la luminance, notée μ ,
- l'écart-type local pour l'information de contraste, noté σ,
- la covariance pour l'information structurelle, notée $\sigma_{v_o v_d}$,

d'où :

$$SSIM = \frac{(2 \times \mu_{v_o}\mu_{v_d} + C_1)(2 \times \sigma_{v_o v_d} + C_2)}{(\mu_{v_o}^2 + \mu_{v_d}^2 + C_1)(2 \times \sigma_{v_o}^2 + \sigma_{v_d}^2 + C_2)}, \tag{3.7}$$

où les constantes C_1 et C_2 ont été introduites par les auteurs afin de garantir la stabilité des calculs.

Les scores retournés par cette métrique varient de 0 à 1, où un score proche de 1 indique une très grande fidélité, tandis qu'un score proche de 0 informe de la présence de nombreuses dégradations. Il est intéressant de disposer de scores normalisés car cela facilite l'utilisation, la compréhension et l'interprétabilité des résultats. Contrairement à la métrique PNSR qui fournit des scores infinis en cas d'égalité parfaite des images. Ce n'est qu'avec l'expérience qu'il est possible d'interpréter les valeurs retournées, il est donc courant de considérer qu'un score supérieur à 40dB décrit une image de bonne qualité et ayant des dégradations quasiment imperceptibles.

En ce qui concerne les métriques de qualité FR, nous pourrions nous arrêter là, car ce sont réellement les deux seules métriques largement connues et acceptées dans les communautés scientifiques et industrielles. Cependant, dans les communautés de spécialistes de la qualité, il est possible de distinguer plus de cent autres métriques [PH09], dont une grande partie exploitent des modèles et propriétés directement issus du système visuel humain (SVH). Nous proposons donc dans la section suivante une synthèse des travaux sur les métriques de qualité bio-inspirées.

3.3.2 Métriques perceptuelles

Afin de maximiser la corrélation entre les métriques et les avis subjectifs, certains auteurs ont entrepris le pari de modéliser le plus finement possible les diverses propriétés du SVH, de la rétine au cortex. Dans cette direction, nous pouvons citer les métriques VDP [MMS04], Pdiff [Yee04], etc.

Il est possible de décrire ce type de métriques de manière générique, car elles partagent toutes de nombreux points communs, dont nous donnons l'architecture fonctionnelle sur la Figure 3.6. Les différences se trouvent généralement dans des techniques d'implémentation ou dans quelques choix de simplification et d'optimisation algorithmique.

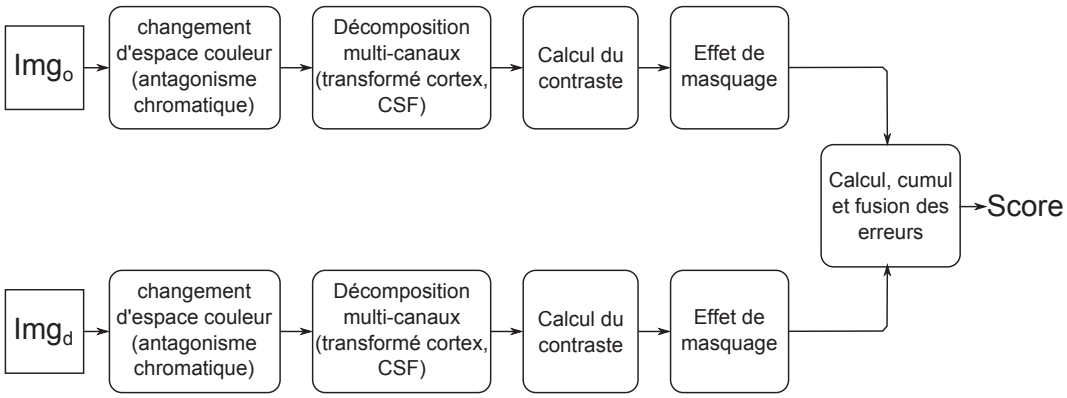
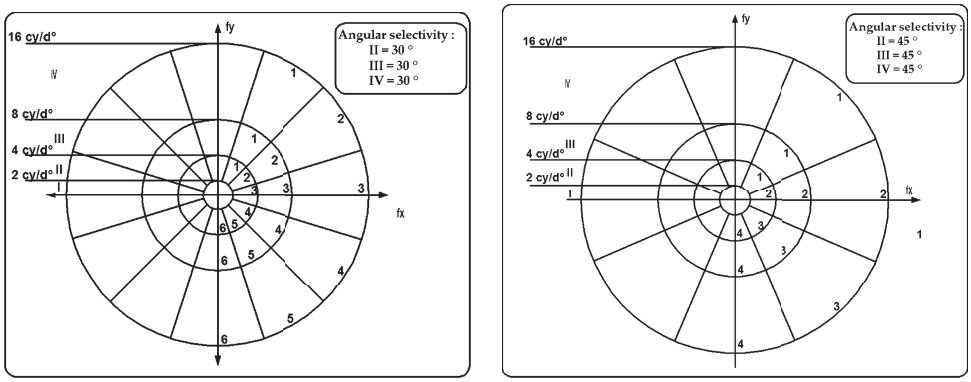


Figure 3.6 : Architecture fonctionnelle générique des métriques de qualité pereptuelles

Tout d'abord l'espace RVB est généralement remplacé par un espace couleur plus proche du codage effectué par le système visuel humain (CIE L*a*b [Con04], YCrCb [ITU84], Krauskopf [DKL84]) mimant le codage par antagonisme de couleurs effectué par les cellules de la rétine à la suite des cônes. Ensuite, l'image est décomposée en différents canaux spatio-fréquentiels orientés, appelés aussi transformée Cortex. Ce passage est en pratique réalisé par la décomposition de Daly [Dal94] ou celle de Watson [Wat87]. Ces deux décompositions sont très similaires et se caractérisent par une sélectivité radiale dyadique. La différence se situe au niveau de la largeur des bandes qui est de 30 degrés pour Daly et de 45 degrés pour Watson, comme l'illustre la Figure 3.13.

Dans cet espace, un contraste local est calculé pour chaque point de chaque sous-bande après l'application d'une fonction de sensibilité au contraste adaptée. C'est ensuite qu'intervient l'effet de "masquage" [3][Lub95, Dal93] suivi de

3. la visibilité d'un stimulus ou artéfact peut être affectée par la présence d'autre stimulus



(a) Daly [Dal94] (b) Watson [Wat87]

Figure 3.7 : Décomposition spatio-fréquentielle orientée

la phase de seuillage permettant de ne conserver que les erreurs au dessus du seuil humain de visibilité. Pour finir un score de qualité est produit en combinant les valeurs des différentes cartes d'erreur de chaque sous-bande.

Il est également intéressant de noter que le schéma que nous avons décrit est finalement extrêmement proche des modèles bio-inspirés pour la prédiction de la saillance visuelle que nous avons précédemment détaillés dans la section 2.3. En effet, nous retrouvons les mêmes types de conversion d'espace couleur, de transformée cortex, d'application de CSF, d'effet de masquage, et enfin de fusion d'informations issues des différents canaux. Tout comme pour la prédiction de la saillance visuelle, leur coût et leur complexité en font des modèles généralement peu utilisés en pratique. On leur préfère généralement les méthodes ayant trouvé un compromis entre la simplicité des méthodes mathématiques et l'exploitation de quelques propriétés issues du SVH. Nous proposons donc, dans la section suivante, de décrire le principe des métriques pondérées par le SVH (appelée également métriques simple canal).

3.3.3 Métriques pondérées par le SVH

L'idée de ces métriques est généralement de se baser sur des approches très rapides à calculer, telles que le PSNR, et d'y introduire des pondérations permettant la prise en compte de quelques propriétés du SVH. Nous pouvons citer les métriques VSNR [CH07], PSNR-HVS et PSNR-HVSM [DVKG+00] qui effectuent des mesures proches du PSNR après avoir effectué une pondération par une fonction de sensibilité aux contraste et un effet de masquage simplifié.

avoisinants

Le but étant de mimer les capacités variables de perception de stimuli en fonction de la fréquence, l'orientation, la couleur et la présence d'autres stimuli sur une image. Encore une fois les principales différences entre les métriques reposent sur les méthodes mises en œuvre pour effectuer ces traitements ; certaines travaillant directement dans le plan image, d'autres dans l'espace de fourrier ou bien l'espace ondelette.

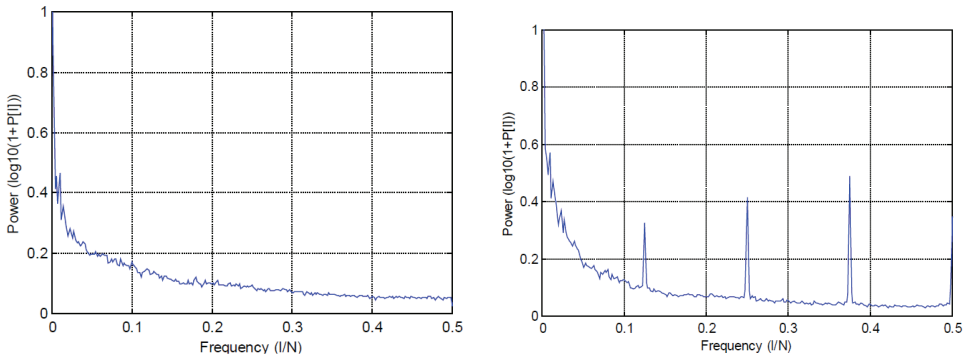
Dans cette catégorie de métriques pondérées par le SVH, exploitant les CSF et effets de masquage, nous pouvons citer les métriques IFC [She04], VIF et VIFP [SB06]. Ces métriques se distinguent des modèles proposés précédemment car elles exploitent une toute autre manière de concevoir la problématique d'observation des dégradations. Plus précisément, elles partent de l'hypothèse que malgré la variabilité des images naturelles, elles partagent tout de même certaines caractéristiques au sens statistique du terme. Typiquement et hormis quelques cas particuliers, toutes les images naturelles possèdent une certaine constance de quantité de contours, d'orientations, de textures et de régions uniformes à différentes échelles. Cependant, ces constantes statistiques sont altérées et modifiées lors de l'introduction d'artéfacts. Intuitivement, des images lissées vont perdre des quantités de détails et de textures, tandis que des images contenant des effets de blocs vont posséder des quantités importantes de forts contours horizontaux et verticaux. De ce constat, ces métriques exploitent des mécanismes d'apprentissage (par l'utilisation de réseaux de neurones ou de SVM - *Support Vector Machines*) sur les données statistiques (dans le domaine de Fourier ou des ondelettes) des images naturelles afin des les caractériser et d'identifier celles ne respectant pas ces constantes. Ces mécanismes de modélisation par apprentissage de statistiques d'images sont relativement récents et fournissent de très bons résultats, ils commencent a s'étendre pour les métriques à référence réduite et sans référence.

3.3.4 Métriques à référence réduite et sans référence

Nous venons de détailler différentes approches d'estimation de la qualité des images. Cependant elles ont toutes la même limitation car elles requièrent toutes la présence des pixels de l'image d'origine, ce qui les rend inopérantes pour les problématiques et attentes actuelles. En effet, l'image de référence n'est pas disponible dans les contextes de manipulation de systèmes portables (tablette, smartphone,...) sous des contraintes de transmission sans fil. Afin de répondre à ces attentes et garantir la meilleure qualité de l'expérience dans ces situations difficiles, des métriques à référence réduite et sans référence ont été développées.

L'idée commune derrière ces métriques est de considérer un *a priori* sur le type de dégradations rencontrées afin de mettre en place une mesure spécialement conçue pour la quantifier sans utiliser (ou partiellement) les pixels de l'image d'origine. Les artéfacts les plus considérés par ce type de métriques sont le flou et l'effet de blocs. C'est pourquoi, nous proposons de décrire les approches les plus connues pour ce type de dégradations.

En ce qui concerne l'effet de bloc, tel qu'introduit par la compression JPEG (ou AVC/H264 pour la vidéo), l'idée est de considérer que des contours horizontaux et verticaux vont apparaître de manière périodique sur toute l'image. Les approches consistent donc à isoler et quantifier ce type de contours afin d'estimer un taux d'effet de bloc. Une manière relativement intuitive et ne nécessitant pas de référence est proposée dans [WBE00] où la transformée de fourrier discrète est utilisée sur les lignes et les colonnes afin d'étudier les pics périodiques se produisant par l'apparition des blocs 8×8, comme l'illustre la Figure 3.8.



(a) Spectre pour l'image lena sans dégradation

(d) Spectre pour l'image lena avec effet de bloc

Figure 3.8 : Observation des pics périodiques introduits par l'effet de bloc [WBE00].

Pour l'effet de flou, la métrique NR [MDWE02] illustre relativement bien le concept généralement utilisé [CWB+04, N.R09] pour ce type d'artéfact, à savoir une mesure d'étalement des contours. Pour ce faire, un filtrage de sobel est appliqué pour détecter les contours, puis une recherches d'extrema-locaux permet d'enregistrer les positions de début et de fin de contours, suivi d'une quantification de l'étalement des contours.

Les métriques à référence réduite exploitent globalement le même type d'approches que les métriques sans référence décrites précédemment, hormis le fait qu'elles s'autorisent une phase d'extraction de ce type d'attributs sur l'image

d'origine afin de disposer d'une référence pour estimer le niveau de dégradation. Afin de réduire l'*a priori* de la dégradation rencontrée, nous pouvons citer la métrique à référence réduite [KZ03], robuste aux dégradations de flou, de blocs et d'aliasing. Pour garantir cette robustesse, la métrique exploite les travaux de [WBE00] pour l'effet de blocs, ceux de [MDWE02] pour le flou et enfin ceux de [SV00] pour l'effet d'aliasing.

Dans la section précédente, nous avions noté une nouvelle tendance dans les métriques FR, par l'exploitation de mécanismes d'apprentissage sur les statistiques d'images naturelles. Ce type d'approches semble également des plus prometteurs pour les métriques RR et NR, si l'on en juge par les bonnes performances de la métrique à référence réduite [WS05] et de celles sans référence [GPRZ07, SBC10].

En résumé, les métriques ont évolué, en passant de mesures pixel à pixel vers des mesures considérant leur relation et leur structure au travers des échelles. Les tendances actuelles visent également à baser leur modélisation par des mécanismes d'apprentissage des statistiques des images naturelles. Enfin, la visibilité des dégradations détectées peut être pondérée par des facteurs issus du SVH. Cependant, malgré toutes ces avancées et travaux, la métrique pixel à pixel PSNR reste indéniablement la métrique de référence dans tous les domaines malgré ses limitations.

3.4 Méthodologie d'évaluation des performances des métriques

Ayant décrit les méthodes subjectives permettant d'étudier la qualité perçue des images ainsi que les modèles prédictifs à même de l'estimer, il est indispensable d'estimer les performances de ces métriques au regard du jugement humain. Pour ce faire, nous proposons de résumer le plan de test proposé par le VQEG (Video Quality Experts Group) [4] [ITU02] qui est un groupe constitué d'experts de divers horizons, aussi bien académiques qu'industriels, ayant tous une expertise en évaluation de la qualité vidéo. Ce groupe dépend de l'ITU dont nous avons cité les principaux documents de référence en ce qui concerne les évaluations subjectives [ITU09].

Afin de quantifier les performances des métriques et donc comparer les MOS subjectifs avec les MOSp prédits par des métriques, il est conseillé de mesurer plusieurs facteurs tels que la précision, la monotonicité et l'uniformité

4. Le site internet du VQEG (http ://www.its.bldrdoc.gov/vqeg/)

des prédictions.

3.4.1 Précision de la prédiction : Corrélation de Pearson et RMSE

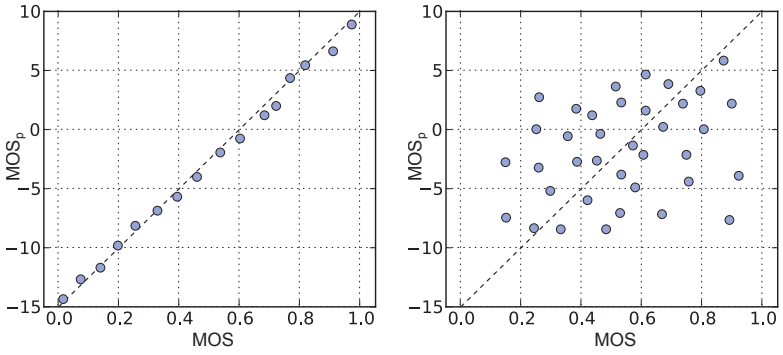
La mesure de la précision des prédictions est effectuée en utilisant à la fois la corrélation de Pearson et une mesure d'erreur quadratique moyenne (RSME de l'anglais Root-Mean Square Error).

La mesure de corrélation de Pearson estime ici un taux de ressemblance et plus particulièrement de corrélation linéaire, existant entre l'évolution des scores subjectifs et l'évolution des scores prédits.

Si l'on considère une collection de N images, pour lesquels nous disposons des MOS_i et $MOSp_i$, le coefficient de Pearson est obtenu par :

$$C_{Pearson} = \frac{\sum_{i=1}^{N}(MOS_i - \overline{MOS})(MOS_{pi} - \overline{MOS_p})}{\sqrt{\sum_{i=1}^{N}(MOS_i - \overline{MOS})^2 \sum_{i=1}^{N}(MOS_{pi} - \overline{MOSp})^2}}, \tag{3.8}$$

Les scores obtenus sont compris en -1 et 1. Plus les scores sont proches des valeurs extrêmes, plus les données sont corrélées. Au contraire, des coefficients de Pearson proches de 0 indiquent qu'aucune relation linéaire n'existe entre les données considérées. La Figure 3.11-(a) permet d'illustrer une très bonne corrélation linéaire positive entre les MOS et MOSp. Tandis que dans le cas (b), les MOSp ne semblent pas prendre des valeurs évoluant en accord avec les MOS ; il n'y a donc pas de corrélation dans ce cas.



(a) Pearson ≈ 1 (b) Pearson ≈ 0

Figure 3.9 : Illustration de la corrélation de Pearson

Bien que la corrélation de Pearson soit sensible à la dispersion des valeurs, il peut être utile de quantifier plus finement les écarts de prédiction quand deux

méthodes ont des corrélations élevées et très proches par exemple. Ces écarts de prédiction sont obtenus par une mesure de RMSE, définie par l'équation suivante :

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(MOS_i - MOS_{pi})^2}, \tag{3.9}$$

où des scores proches de 0 indiquent la bonne qualité des prédictions car minimisant les erreurs. Nous matérialisons sur la Figure 3.10-(a) les écarts de prédictions par rapport à la référence par l'intermédiaire de droites bleu reliant les cercles à la diagonale centrale. C'est la moyenne de ces écarts qui est quantifié par cette mesure.

Sur cette figure, nous remarquons également une prédiction particulièrement éloignée et aberrante, ce qui nous permet d'introduire la mesure d'Outlier Ratio décrite dans la section suivante.

3.4.2 Uniformité de la prédiction : Outlier Ratio (OR)

Le but de cette mesure est de quantifier le taux de points prédits trop éloignés de la valeur de référence. Un point est considéré comme aberrant s'il a une erreur de prédiction au dessus d'un certain seuil, considéré comme l'intervalle de confiance. En général, ce seuil est défini comme étant le double de l'écart-type σ obtenu en étudiant les données subjectives. D'où :

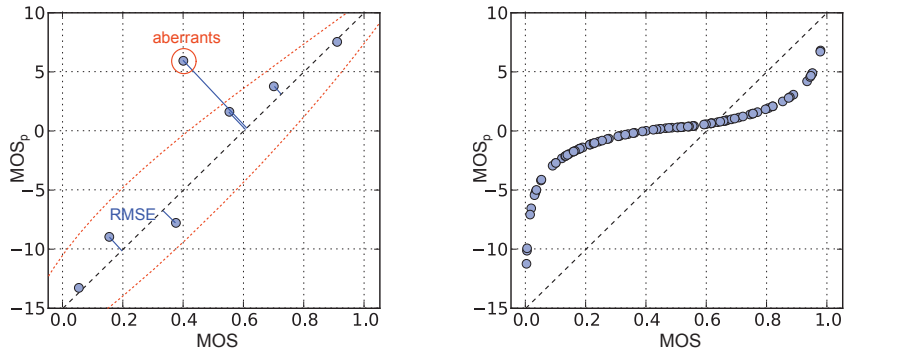
$$P_{aberrant} = |MOS_i - MOSp_i| > 2\sigma, \tag{3.10}$$

$$OR = \frac{N_P}{N}, \tag{3.11}$$

où N_P est le nombre de points aberrants. Sur la figure, 3.10-(a), il est possible de visualiser l'intervalle de confiance définissant les seuils d'erreurs de prédiction tolérés, ainsi qu'une prédiction ne respectant pas cette contrainte.

3.4.3 Monotonie de la prédiction : Corrélation d'ordre de Spearman

Dans certains cas applicatifs, comme l'ordonnancement d'images par leur qualité, les valeurs exactes des prédictions et leurs écarts à la référence n'a que



(a) RMSE et OR (b) Spearman = 1 (Pearson ≈ 0.88)

Figure 3.10 : Illustration des mesures de RMSE, OR et Spearman

peu d'importance. Ce qui est important, c'est l'ordre de classement devant rester identique. La quantification de cette aptitude est obtenue par des mesures de monotonie et plus précisément par la mesure du coefficient de corrélation de Spearman, défini par l'équation :

$$C_{Spearman} = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)}, \tag{3.12}$$

où d_i est la différence de rang entre le score prédit $MOSp_i$ et le score de référence MOS_i de la i ème image. Comme pour la corrélation de Pearson, les meilleures scores sont -1 et 1, tandis que les pires sont pour les scores proches de 0. Afin de mieux appréhender les différences entre ces deux mesures, nous fournissons sur la Figure 3.10-(b) une illustration de Spearman = 1, car l'ordre est parfaitement respecté, tandis que la corrélation de Pearson n'est pas a son maximum, puisque les valeurs n'évoluent pas de manière linéaire.

Nous venons de décrire différentes mesures de performances. Cependant, avant de les effectuer, le VQEG recommande d'appliquer une adéquation des données MOSp et MOS par l'utilisation d'une fonction de régression non linéaire. Les fonctions les plus recommandées sont de la forme polynomiale :

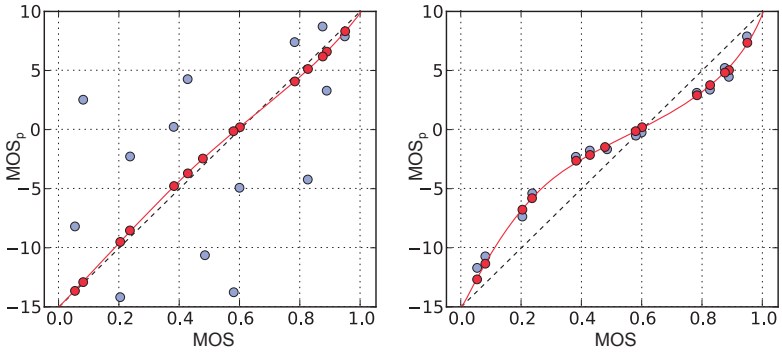
$$y = \beta_1 x^3 + \beta_2 x^2 + \beta_3 x + \beta_4, \tag{3.13}$$

ou bien logistique :

$$y = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp\left(\beta_2 (x - \beta_3)\right)} \right) + \beta_4 x + \beta_5, \tag{3.14}$$

Il n'existe malheureusement aucune justification ou conseil permettant de choisir l'une ou l'autre des méthodes. De plus, de nombreux algorithmes (solveur)

permettent de réaliser ce type d'adéquation. Cependant, ils exploitent tous divers paramètres tels que le degré de précision minimum, le nombre d'itérations maximum, les valeurs d'initialisation des paramètres β, etc. Par la variété des paramètres et la dynamique des scores MOS et MOSp, variants d'une expérience subjective à l'autre et d'une métrique à l'autre, il est très délicat de trouver un paramétrage et une uniformisation des valeurs. Il est donc quasiment impossible de garantir une équité des performances entre différentes métriques en utilisant cette adéquation des données.



(a) Métrique A Métrique B

Figure 3.11 : Fort impact de l'adéquation des données en cas de faible corrélation initiale

Enfin, il peut être intéressant de rappeler que dans la section "4.7 Costs and Benefits of the logistic transformation" d'une version antérieure du plan de test du VQEG [VQE08], il est expliqué que l'utilisation de telles fonctions peut avoir une forte influence sur les scores de corrélation de Pearson, si la corrélation n'était pas supérieure à 80% sans leur utilisation. Ce fort impact peut être appréhender en observant la Figure 3.11, où la métrique A est initialement peu corrélée, mais après l'application d'une adéquation aux données, les prédictions MOSp (cercles rouges) se retrouvent être extrêmement proches des MOS. Si l'on compare cette métrique à une seconde métrique B, dont les prédictions initiales étaient meilleures, nous pouvons observer qu'après ce traitement, elle sera malheureusement jugée moins performante que la métrique A. Ceci illustre les possibles biais de mesure pouvant être introduits par l'utilisation des adéquations aux données.

Nous venons de décrire les mesures objectives à même de quantifier les performances des prédicteurs de qualité. Toutes ces mesures s'appuient sur des vérités de terrain issues de tests subjectifs. Dans la section suivante, nous proposons donc de détailler les différentes bases d'images disponibles dans la littérature fournissant des notes subjectives (MOS) pour diverses dégradations.

3.4.4 Bases d'images dédiées à la qualité

Il y a peu de bases d'images permettant de vérifier les performances des métriques. Ces bases sont LIVE [SWCBa, SWCBb], Toyama [HKS], IVC [CA05], TID2008 [PLE+08] et CSIQ [LC10]. Le Tableau 3.1 présente les caractéristiques de chacune des bases de manière condensée.

Tableau 3.1 : Caractéristiques des bases de données

Car.	LIVE	Toyama	TID2008	IVC	CSIQ
Distors.	JPG, J2K bruit, flou transmission	JPG, J2K	JPG, J2K 8 bruits, flou 2 transmissions, etc.	JPG, J2K Lar, flou	JPG, J2K bruit, flou contraste
Codeur Jpg	Matlab imwrite	cjpeg	Non précisé	Non précisé	
Codeur jp2k	Kakadu v2.2	Jasper v1.7	kakadu	Non précisé	Non précisé
Méthode	Single stimulus	Single stimuli	Double stimuli	Double stimuli	
Image	Couleurs RVB avec 24 bits/pixels				
Résolution	entre 768x512 et 634x438	768x512	812x384	512x512	512x512
Nb img.	29	14	25	10	30
écran	CRT 21-inch (1024x768) (non calibrés)	CRT 17-inch (1024x768)	LCD et TFT 17 et 19 inches 1152x864	Non précisé. (1920x1200)	LCD
Distance déobservation	2-2.5H (hauteur écran)	4H (hauteur image)	Trés varié	6H (hauteur écran)	70cm
Eclairage ambiant	Bureau	Faible	Trés varié	Normalisé	
Nb. observateurs	20-25 (JP2K) 20 (JPEG)	16	654	15	35
Type observateur	Etudiants Univ.Texas	Non expert, étudiants	Non précisé	Non précisé	Non précisé
Ecart type et IC	Calculable	Fournis	Incalculable	Incalculable	Fournis

Tout d'abord, nous remarquons une différence au niveau de la disponibilité des informations entre les bases. On peut critiquer l'absence de l'écart-type des MOS qui rend impossible certains calculs, comme le taux de rejet (OR) nécessaire à l'évaluation des performances des métriques.

Bien que LIVE ne les fournisse pas, les notes de chaque observateur sont données, ce qui permet de calculer les informations manquantes. Nous pouvons remarquer que les recommandations de l'ITU-R relatives aux protocoles d'expérimentation ne sont pas toujours respectées. Par exemple, les distances d'observation ne sont pas suivies. Les conditions d'éclairage ne sont pas toujours contrôlées. Les dispositifs d'affichage ne sont ni identiques ni réglés pour chaque expérimentation. Dans de telles conditions, allons-nous réellement tester les performances des métriques ou les conditions de l'évaluation subjective ?

En ce qui concerne le respect des recommandations, certaines études se veulent rassurantes et montrent que les différences ne sont pas notables. La

base TID2008 qui a eu recours a une très large expérimentation (3 laboratoires de pays différents, ainsi que des dispositifs d'affichages TFT et CRT mélangés, associés à des distances d'observation et des conditions d'éclairage variées) affirme que les résultats obtenus entre les laboratoires sont corrélés à 97%. Une autre étude a tenté de vérifier l'impact de la différence de culture (Japon/France) ainsi que l'incidence du type d'affichage (CRT/LCD). Les résultats numériques démontrent une corrélation à plus de 95%. Donc, la combinaison de toutes ces contres-indications semble être négligeable.

Abordons maintenant le choix des images et la magnitude des distorsions. Pour tester les performances des métriques de qualité, il est important d'avoir des images variées et représentatives de la diversité des images échangées dans les applications réelles. Nous pouvons noter que les 3 bases LIVE, Toyama et TID2008 utilisent les mêmes images sources (12 images communes). Ce panel d'images est tout de même intéressant car il contient des images d'objets manufacturés, de visages, d'animaux, de paysages naturels, différentes prises de vue avec des premiers et arrières plans plus ou moins distincts. Bien que les images sources soient les mêmes, la magnitude des distorsions et les codeurs sont différents. Par exemple, la base TID2008 a des distorsions qui rendent le contenu des images indiscernable tandis que toutes les images de Toyama restent très correctes en terme de qualité. LIVE quant à elle propose des distorsions réparties de manière plus homogène en magnitude. On se rend compte que les échelles, ayant été proposées aux utilisateurs, n'ont pas le même sens pour une base ou pour une autre. Quand les valeurs donnent un état « Bad » pour une image de la base Toyama, il s'agit finalement d'un état « Good » de la base TID2008. Une métrique performante avec la base Toyama, est une métrique très sensible, capable de quantifier des artéfacts sur des images très peu dégradées. Mais si cette métrique donne de mauvais résultats avec la base TID2008 qui crée d'importantes dégradations, cela sous entend qu'elle n'est pas très robuste aux importantes distorsions. L'idée est d'exploiter la complémentarité des bases de données pour juger les métriques.

Si certains sont sceptiques sur la diversité des images de ces bases de données et qu'ils espèrent trouver d'autres images avec la base IVC, il faudra être très prudent. Bien que cette base affirme respecter de manière rigoureuse le protocole d'évaluation (environnement normalisé), le choix des images sources peut laisser perplexe. Les images semblent éloignées des images actuelles. Leur dynamique des couleurs et leur résolution sont très basses, bien en dessous des capacités d'acquisition des capteurs grand public. Il est très difficile de disposer de bases de données alliant respect des protocoles, qualité de contenu et détails sur les résultats.

Par la description des bases d'images disponibles, nous avons soulevé l'im-

portance de pouvoir disposer de plusieurs d'entre elles. Il semble qu'une vague de prise en compte de l'aspect subjectif de la qualité des traitements déferle sur le monde scientifique. De plus en plus de laboratoires envisagent de disposer d'une salle permettant de réaliser des tests subjectifs. Nous conseillons de veiller à respecter les standards existant afin de minimiser les inquiétudes des futurs utilisateurs. Il est également important de veiller à fournir des informations détaillées des résultats obtenus afin d'assurer une transparence et permettre plus de flexibilité pour de futures analyses.

3.4.5 Mise en place d'une application web dédiée à la comparaison de métriques

Nous venons de décrire les méthodologies relatives à l'évaluation de la qualité des images en soulevant l'utilité des métriques objectives. Afin de répondre à ces attentes, de nombreuses métriques de qualité ont vu le jour ces dernières années et nous pouvons en dénombrer plus d'une centaine [PH09]. Bien sûr, toutes ces métriques n'ont pas les mêmes performances ; chacune ayant ses atouts et faiblesses. Cependant, il est délicat de comparer réellement cette grande quantité de métriques. De plus, la comparaison est valable à un instant donné, mais devient obsolète dès lors qu'une nouvelle métrique ou qu'une nouvelle base d'images est publiée. Afin d'essayer de garantir une comparaison pérenne de métriques, nous avons eu l'idée de développer un service-web nommé ImQual [NLF11b] entièrement dédié au benchmark de métriques de qualité tout en veillant également à faciliter le choix et l'utilisation des métriques existantes. Ce service-web a été pensé pour différents types d'utilisateurs, des débutants aux experts en évaluation de la qualité des images.

La première catégorie d'utilisateurs regroupe les chercheurs et concepteurs de nouveaux algorithmes de traitements d'images tels que la compression, le tatouage ou la transmission d'images. Ces types d'algorithmes introduisent des artéfacts ou des dégradations. La performance de ces outils est liée à la qualité visuelle des images obtenues. L'utilisation de métriques de qualité est indispensable pour mesurer l'apport des algorithmes proposés au regard de la littérature. Malheureusement, les chercheurs utilisent bien souvent et uniquement la métrique basée sur le signal qu'est le PSNR, ce qui peut en grande partie s'expliquer par la simplicité et la faible complexité de cette mesure. Cependant, il existe actuellement d'autres métriques de qualité ayant de meilleures corrélations avec le jugement humain, mais elles sont bien moins connues et donc moins exploitées. Pour remédier à ce problème, le service-web ImQual a une partie entièrement dédiée à l'évaluation de la qualité perçue des images dégradées. Cette partie offre à l'utilisateur la possibilité d'utiliser facilement

les métriques de la littérature sur son propre ensemble d'images. Il suffit d'envoyer ses images dégradées ainsi que l'image originale pour obtenir les scores de qualité prédits par les métriques de son choix. Les prédictions de qualité de chaque métrique sur chaque image peuvent être sauvegardées sous formes textuelle ou graphique et ainsi directement intégrables au sein de sa future publication. La Figure 3.12 illustre cette partie du service-web.

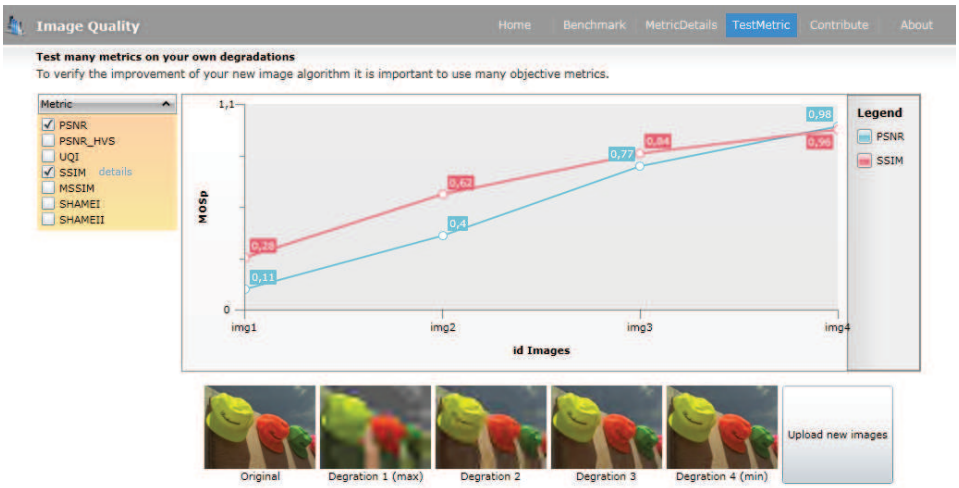


Figure 3.12 : Test de différentes métriques sur un set d'images envoyé par l'utilisateur

En utilisant cet outil, l'utilisateur a la garantie d'utiliser les dernières versions des métriques disponibles et ceci très facilement sans avoir à passer un temps conséquent à obtenir et à appréhender les détails de chaque métrique.

Le seconde catégorie d'utilisateurs visée regroupe les chercheurs et concepteurs de métriques de qualité. Après le développement d'une nouvelle métrique, il est important et impératif d'évaluer ses performances en termes de cohérence avec le jugement humain. La méthode classique consiste à comparer les scores prédits par la métrique avec ceux obtenus par des évaluations subjectives comme expliqué précédemment.

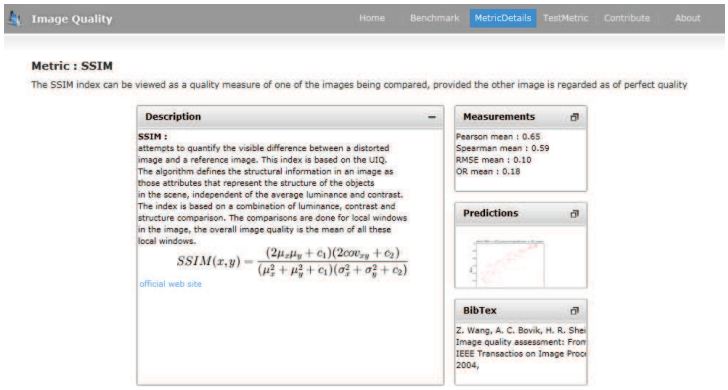
Grâce à ces mesures, il est possible de se faire une idée des performances de la métrique développée et de la comparer à l'état de l'art. C'est dans cet objectif qu'ImQual met à disposition des concepteurs une section destinée à l'envoi de nouvelles métriques. Il est alors possible d'envoyer une description accompagnant le code de sa métrique (sous forme de fichiers binaires ou de script). Après cet envoi, le service-web ImQual prend en charge l'exécution de la métrique sur toutes les images de toutes les bases dont les MOS sont connus. Après ces exécutions, les performances de la métrique sont mesurées et comparées à toutes les métriques de la littérature déjà présentes dans le service. Le chercheur s'affranchit donc des tâches délicates consistant à obtenir

les différentes bases d'images, les différentes métriques existantes, ainsi que les mesures performance. ImQual permet donc de faire économiser les nombreuses heures nécessaires à l'exécution de ces tâches et ainsi fournir des scores de performance standards et normalisés pour toutes les métriques afin de garantir des comparaisons équitables. De plus, en enrichissant ImQual avec une nouvelle métrique, le laboratoire ou l'entreprise à l'initiative de cette production pourrait éventuellement augmenter la visibilité de ses travaux par la publication de liens, de références et de documentations sur sa métrique, tel qu'illustré sur la Figure 3.13.

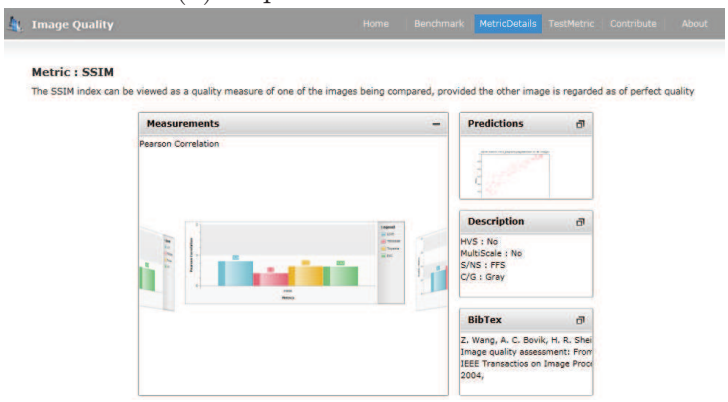
Toute personne intéressée par l'évaluation de la qualité des images est considérée comme appartenant à la troisième catégorie d'utilisateurs des services ImQual. Cette plateforme permet d'obtenir facilement une liste exhaustive des métriques existantes. Chaque métrique est détaillée et ses performances mesurées et comparées comme illustré sur la Figure 3.14. Grâce à ces descriptions et scores de performances, il devient possible de sélectionner une métrique adaptée aux contraintes des utilisateurs.

La quatrième catégorie d'utilisateurs vise le spécialiste de l'évaluation subjective en charge de la création de bases d'images. De plus en plus de laboratoires sont équipés pour effectuer des campagnes d'évaluation subjective sous environnement contrôlé. Ces évaluations sont très chronophages ; il est donc important de valoriser ce travail et de partager les résultats obtenus au sein de la communauté. ImQual peut accueillir ces différentes bases de données et ainsi fournir à tous un service encore plus complet. Pour chaque nouvelle base, toutes les métriques présentes dans le système sont exécutées sur chaque nouvelle image, afin de calculer les nouveaux scores de performance.

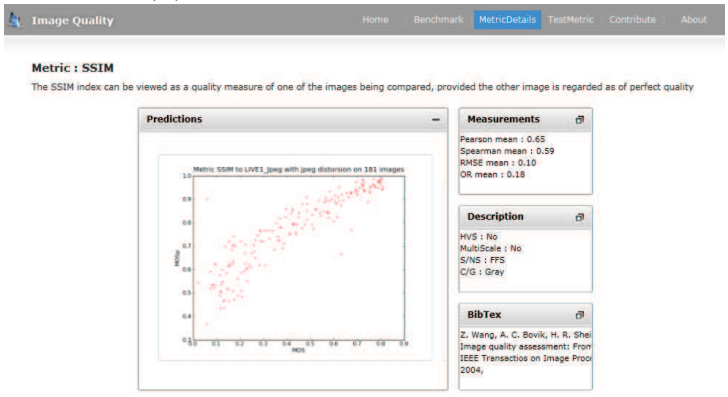
Pour résumer, ImQual est un web-service dédié à l'évaluation de la qualité des images et au benchmark des métriques de qualité. Comme mentionné précédemment, il combine plusieurs outils d'aide aux scientifiques intéressés par l'utilisation de métriques de qualité, et ce peu importe leur niveau d'expertise dans ce domaine. En ce qui concerne l'avancement de ce service, nous avons d'ores et déjà validé les phases de conception détaillée et de prototypage. C'est actuellement une phase de déploiement impliquant une architecture de traitement distribuée et massivement parallèle qui est réalisée. De ce fait, le service n'est à l'heure actuelle pas encore disponible au grand public dans son intégralité. Cependant, ce n'est qu'une question de temps, car ce projet est actuellement encouragé et soutenue par le comité JPEG et la division 8 de la CIE, pour à terme, en faire l'outil de référence pour les mesures de performance.



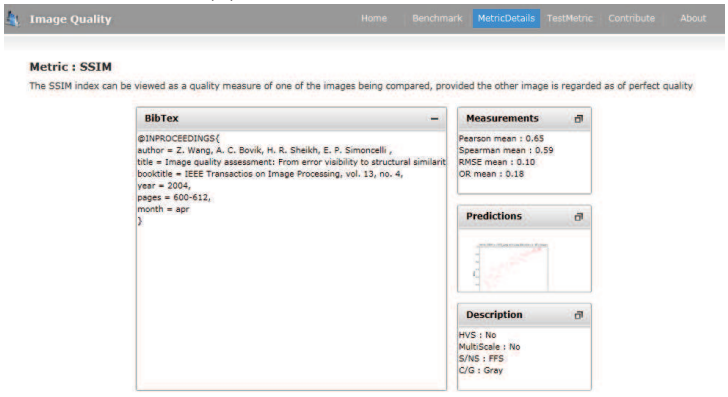
(a) Explications et formules



(b) Graphiques de performances



(c) Nuage de prédiction



(d) Bibtex

Figure 3.13 : Description détaillé de la métrique SSIM

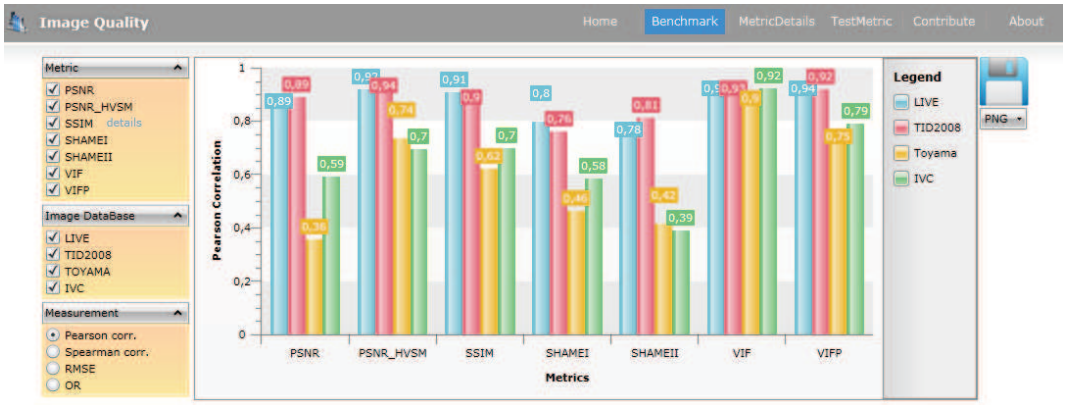


Figure 3.14 : Mesure de performance des métriques

3.5 Conclusion

Ce chapitre a été l'occasion de faire un tour d'horizon des méthodologies permettant de prendre connaissance du ressenti des observateurs lors de l'évaluation d'images. Nous avons également illustré ces méthodes en les mettant en pratique lors d'une campagne d'évaluation d'une proposition d'un nouveau codeur dans le cadre du processus de standardisation par le comité JPEG. Bien qu'indispensable, ce type de tests est très coûteux ; c'est pourquoi ils servent généralement de vérité de terrain fournissant des références au développement de métriques de qualité. Nous avons décrit les principales approches de la littérature en détaillant les méthodes permettant de mesurer les performances de leurs prédictions. Au vu de la quantité de métriques existantes, de la variété de bases d'images et de l'aspect délicat des mesures de performances de ces outils, nous avons proposé à la communauté scientifique un service-web spécialement conçu pour le benchmark et la démocratisation de l'utilisation des métriques. Bien qu'il soit encore en phase de déploiement, ce web-service ImQual [5] est d'ores et déjà encouragé et soutenue par le comité JPEG et la division 8 de la CIE, pour à terme, en faire l'outil de référence pour les mesures de performances.

Ce chapitre a également été l'occasion de prouver que la recherche et les organismes en charge du développement des technologies de demain sont réellement intéressés par l'introduction du ressenti humain dans les évolutions technologiques. L'objectif visé est d'augmenter la qualité de l'expérience des utilisateurs finaux. Dans cette direction, nous avons également soulevé l'intérêt de ces considérations dans les chaînes de transmission sans fil. C'est en ce sens que nous proposons de détailler, dans le chapitre suivant, notre méthode de

5. accessible à l'adresse : http ://www.qualimage.net/

prédiction de la qualité à référence réduite, utilisant les points d'intérêt, appliquée aux problématiques de transmission sans fil sur canaux MIMO réalistes.

Chapitre 4

Métriques de qualitée basées sur les points d'intérêt :Application à l'amélioration de la QoE en transmission

Sommaire

4.1 Approche proposée : Métrique à référence réduite de prédiction de la qualité par évolution de points d'intérêt . . **124**
4.1.1 Méthode QIP 125
4.1.2 Méthode QIP-HSM 135
4.2 Intégration de QIP dans une chaîne de transmission sans fil **152**
4.2.1 Transmissions JPWL à travers un canal MIMO sous monitoring perceptuel 153
4.2.2 Mesure de performance objective 155
4.2.3 Validation de la méthode par expérimentation subjective 158
4.3 Conclusion **169**

Dans le contexte actuel, l'échange de contenus multimédia et particulièrement les images, est devenu monnaie courante et les besoins et attentes en ce sens ne cessent de croître. A titre d'exemple, il est courant d'envoyer (par email via un ordinateur de bureau) les photos de ses enfants aux grands-parents. Ou bien de partager sur les réseaux sociaux les plus beaux paysages immortalisés durant les vacances. Dans un futur proche, il pourrait être possible de se faire diagnostiquer une pathologie par un spécialiste en médecine via un smartphone avec un protocole de communication sécurisé. On se rend facilement compte

de la diversité des technologies utilisées autant pour la visualisation que pour le transport de l'information. Dans ces contextes, la compression des images est impérative afin de garantir un transfert rapide et/ou de limiter l'espace de stockage.

Cependant la compression peut réduire la qualité des images et introduire des défauts plus ou moins visibles et désagréables. Mais pour notre plus grand plaisir, la prise en compte du ressenti de l'humain, notre rapport vis à vis de la qualité de l'expérience entre nous et la machine, est de plus en plus pris en compte, afin de développer des machines qui nous comprennent enfin.

Dans les exemples cités, on note l'émergence des tablettes et smartphones, deux types de technologies en pleine progression et dont le transfert de données est la plupart du temps sans fil. C'est dans ce contexte d'échange aérien et de mobilité que ce chapitre tend à s'inscrire. Dans ce cadre, la qualité de l'expérience est intimement liée à deux facteurs. Il y a tout d'abord la vitesse à laquelle le média est envoyé ou réceptionné et également la qualité de l'image reçue. Si l'on considère un temps de transfert maximum acceptable, il ne reste plus qu'à maximiser la qualité des images échangées. Dans cet objectif, il est utile de disposer de métriques capables de mesurer la qualité des images en temps réel, afin d'adapter dynamiquement et au mieux, la stratégie de transfert en fonction de l'état instantané du canal.

La qualité perçue des images est dépendante de la visibilité des artéfacts introduits. La compression JPEG, avec l'effet de bloc ou de pixelisation la caractérisant, fait à la fois perdre des détails, des textures et des petites structures de l'image, tout en introduisant de nouvelles structures, ces fameux blocs, ces carrés, dont les contours sont visibles selon le niveau de compression. La compression JPEG 2000, quant à elle, peut faire apparaître des oscillations autour des contours, voire même les atténuer fortement. Globalement, la compression fait apparaître ou disparaître les structures de l'image. L'idée est donc de trouver une méthode capable d'extraire et de quantifier les statistiques structurelles de l'image afin de juger leur perte ou apparition.

4.1 Approche proposée : Métrique à référence réduite de prédiction de la qualité par évolution de points d'intérêt

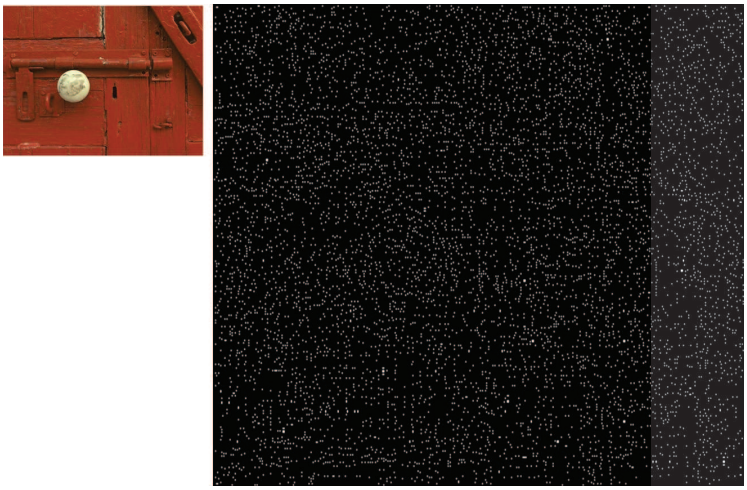
Note approche part du principe que l'évolution des statistiques structurelles de l'image est un bon indice de la qualité perçue. Les structures classiques de

l'image sont les contours, les coins... Bien sûr, l'extraction de ce type de statistiques est une problématique qui a déjà été étudiée dans beaucoup d'autres domaines, dont certains avec de fortes contraintes de temps de calcul. C'est en fouillant dans cette direction que nous nous sommes orientés vers les détecteurs de points d'intérêt (cf. Chapitre 1). Pour rappel, les points d'intérêts ont pour objectif de repérer les structures de l'image les plus prononcées, les plus stables malgré les déformations et ce en un minimum de temps. Il existe de nombreux détecteurs, certains détectent les bords, d'autres les coins ou encore les structures blobs et sont plus ou moins invariants aux variations affines, photométriques, d'échelles et de rotation. L'idée consiste à tirer profit des travaux sur les détecteurs en tant qu'extracteurs de statistiques structurelles de l'image dans le but d'estimer la qualité perçue. La section suivante décrit une proposition de métrique basée sur ce principe (Quality by Interest Points : QIP).

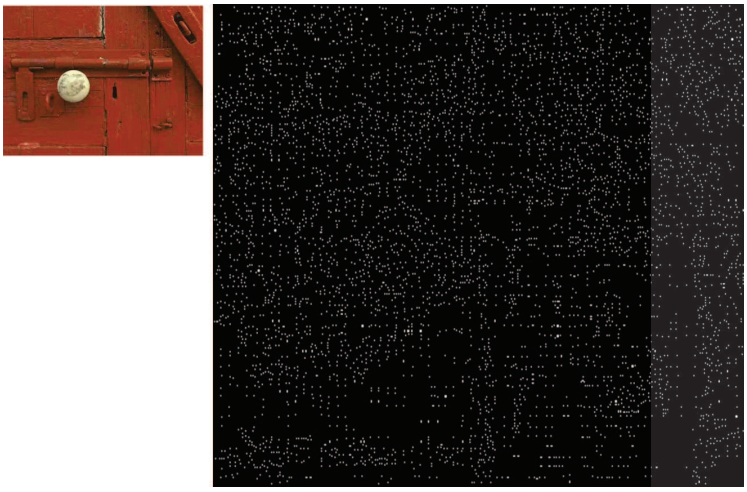
4.1.1 Méthode QIP

La métrique que nous proposons s'oriente vers les problématiques d'amélioration de la qualité de l'expérience sous des contraintes de transmission d'images sur les réseaux sans fil réalistes. Plus une image peut être compressée, plus elle peut être diffusée efficacement sur ces réseaux. Cependant, la compression et la transmission peuvent affecter la qualité perçue et c'est cette dernière qu'il faut être capable de quantifier. Au vu des dégradations introduites, à savoir la perte ou l'apparition de structures, de contours, de coins, nous nous sommes orientés vers la famille des détecteurs de Moravec (section 1.2.2) et plus particulièrement l'évolution proposée par Harris et Stephens (section 1.2.3). En effet, cette méthode a été spécialement conçue pour repérer si un pixel est dans une région uniforme, de contours ou de coins.

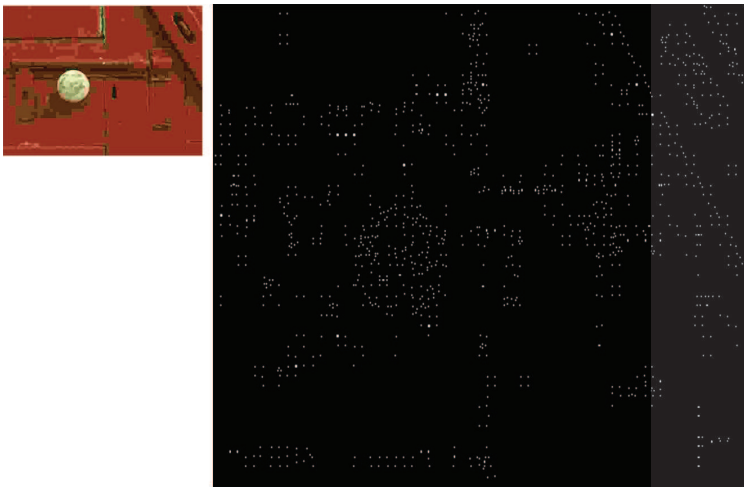
Notre démarche vise également à mesurer des changements, des variations de structure. Il ne faut donc pas que le détecteur choisi soit trop invariant aux modifications de l'image. En se basant sur l'étude [MS05], il semble que tous les détecteurs de points d'intérêt voient leurs performances d'appariement diminuer à mesure que les dégradations augmentent. Nous pouvons noter qu'ils sont tous lourdement affectés par les dégradations de type JPEG et particulièrement par l'effet de flou (phénomène apparaissant également lors de la compression JPEG 2000). Mais c'est le détecteur de structure blob SIFT, qui dans cette étude semble être le plus stable. Cependant, tous les détecteurs sont sensibles aux dégradations introduites par les outils de compression et donc potentiellement à même de mesurer des dégradations.



Sans compression



Compression JPG faible



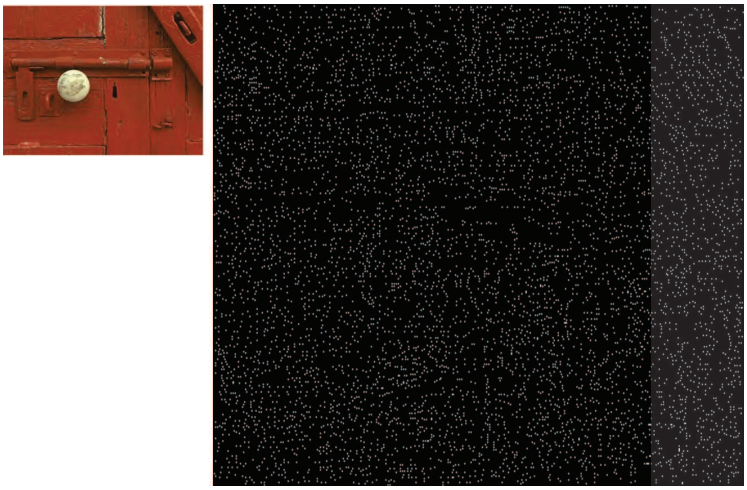
Compression JPG importante

Figure 4.1 : Évolution des points d'intérêt en fonction de l'augmentation de la compression JPEG

La Figure 4.1 permet d'illustrer la sensibilité de détection de points d'intérêt liée à une dégradation de type JPEG à travers l'utilisation d'un détecteur de coins. Il est aisé d'observer les changements des points d'intérêt détectés et leur répartition. On remarque que, plus la force de la compression augmente, plus le nombre de points diminue. Ceci peut s'expliquer par le fait que la compression utilise un découpage en bloc de 8×8 ou 16×16 pixels suivi d'un processus de sous-échantillonnage de la couleur et d'une quantification. Cette succession de traitements a tendance à réduire la quantité de variations dans chaque bloc. De ce fait, des blocs de plus en plus importants et uniformes apparaissent alors. Le détecteur de points d'intérêt ne va donc plus extraire des structures intéressantes qu'aux coins de chaque bloc. Il y a donc réduction du nombre de points et un glissement de leur localisation vers les coins. La Figure 4.2 illustre les effets de la compression JPEG 2000. Au même titre que JPEG, le nombre de points d'intérêt détectés diminue à mesure que la compression augmente, car les hautes fréquences disparaissent progressivement en laissant apparaître de larges régions uniformes, donc de moins en moins de coins détectables.

Paramétrage du détecteur

Sur les Figures 4.1 et 4.2, on peut s'interroger sur la quantité importante de points d'intérêt détectés sur l'image sans dégradation. Cette quantité importante est souhaitée et favorisée dans notre démarche. En effet, nous souhaitons mesurer des variations structurelles, aussi bien sur les objets principaux de la scène que sur les textures qui les composent. Les textures sont finalement constituées de coins et de contours mais dont la taille et la puissance sont faibles. Pour capturer ces infimes variations, nous utilisons un paramétrage très particulier du détecteur de Harris. Pour rappel, le détecteur de Harris se base sur un calcul de réponses R (cf. équation 1.19), suivi d'un processus de filtrage sur ces réponses visant à ne garder que les structures les plus stables. Cependant, la stabilité n'est pas notre objectif, au contraire, nous voulons jouer de ce manque de stabilité et de ce manque d'invariance, car c'est ce que nous exploitons pour s'adapter à l'évolution ou à l'augmentation du taux de compression. C'est en minimisant le paramètre de filtrage des réponses de R que la maximisation des points d'intérêt est effectuée, en le fixant à une valeur très proche de 10^{-7}. En effet, ce sont en général les coins et les contours de faible réponse qui vont être les premiers à être affectés, même pour de faibles déformations. Toujours dans l'optique de maximiser le nombre de points détectés (et donc la sensibilité du détecteur), le seuil supprimant les points d'intérêt trop proches les uns des autres a été fixé à son minimum, afin d'autoriser la présence de deux points d'intérêt espacés simplement d'un pixel.



Sans compression

Compression JPEG2000 faible

Compression JPEG2000 importante

Figure 4.2 : Évolution des points d'intérêt en fonction de l'augmentation de la compression JPEG 2000

Méthode de prédiction de la qualité

Grâce aux réglages particuliers du détecteur et les illustrations des effets de la compression sur la détection de points d'intérêt, il semble intéressant de mesurer la qualité par une quantification d'évolution de ces points. La Figure 4.3 permet d'illustrer, à travers un schéma fonctionnel, une méthode de mesure de changements structurels de l'image somme toute simple et économique mais cohérente avec les observations précédentes.

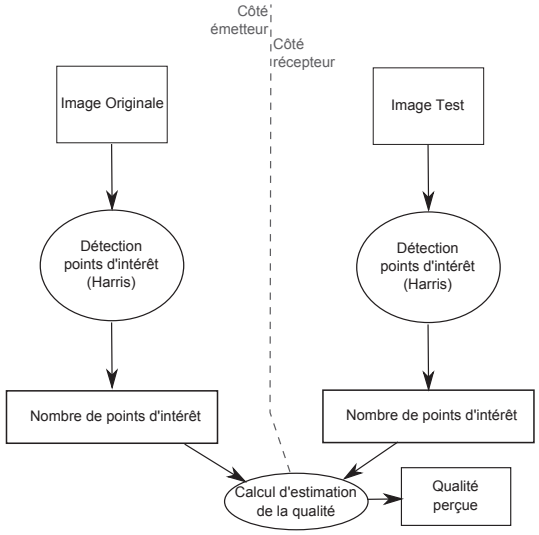
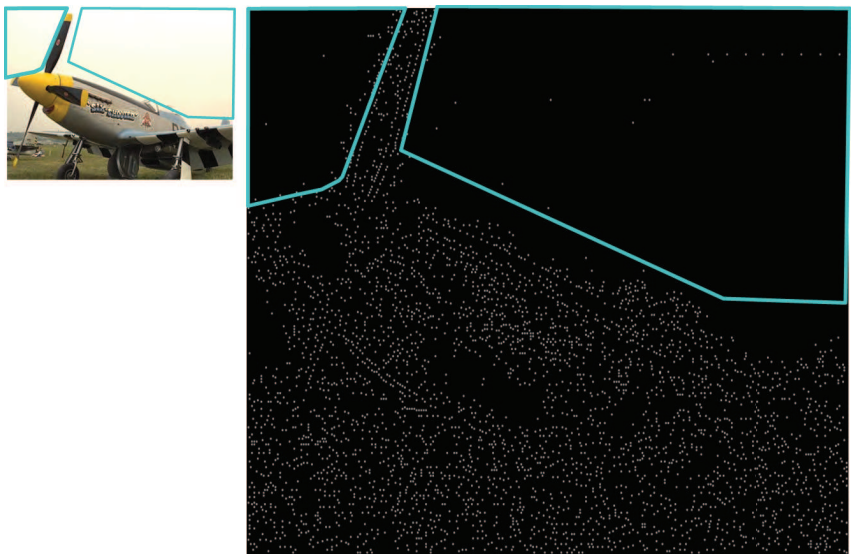


Figure 4.3 : Schéma fonctionnel de la métrique QIP version 1

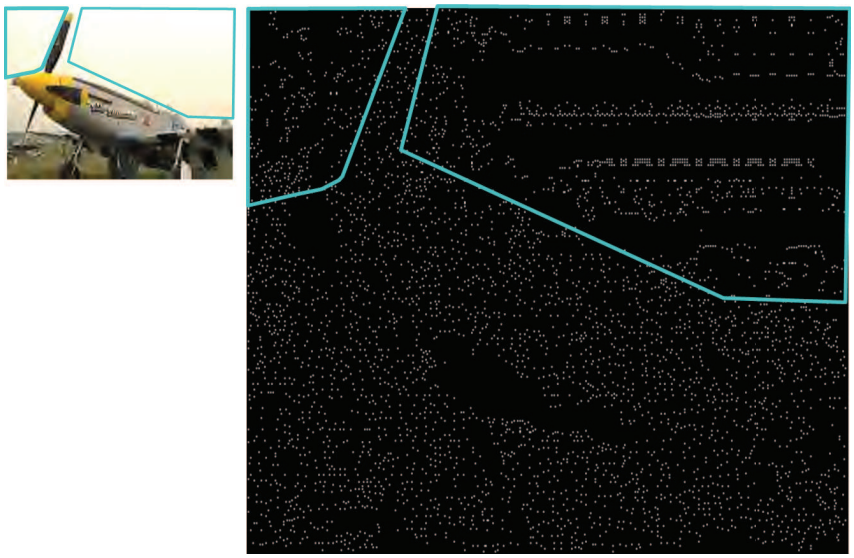
La première étape consiste à extraire les points d'intérêt de l'image sans dégradation et à les comptabiliser. Cette même comptabilisation de points est effectuée sur l'image de test, potentiellement dégradée. La mesure de qualité, de changement structurel, est réalisée en comparant les deux entiers, représentant le nombre de points avant et après dégradation. Si le nombre de points est identique, l'image n'a pas été modifiée. Au contraire, si le nombre de points est nettement moins important, l'image a sans doute été très altérée par le processus de compression, l'image est donc de mauvaise qualité. En effet, cela indique que de nombreuses textures et contours marqués ont dû être supprimés.

Cependant, l'observation de la perte du nombre de points d'intérêt n'est pas réellement une vérité absolue. En effet, certaines dégradations vont faire apparaître des points et non les supprimer, et ce principalement dans les régions initialement uniformes, tout en continuant à les supprimer dans les régions texturées comme l'illustre la Figure 4.4.

Mais dans tous les cas, que ce soit l'apparition ou la disparition du nombre de points, une mesure de différence semble pertinente. C'est pourquoi, afin



Sans compression



Compression JPEG 2000 importante

Figure 4.4 : Évolution des points d'intérêt dans les régions de faibles activités

Figure 4.5 : Illustration du partionnement de l'image en fonction du niveau d'activité du contenu

d'être sensible à ces deux phénomènes, nous proposons de subdiviser l'image en deux classes. L'idée est de partitionner l'image en régions, par exemple des blocs rectangulaires, et d'étiqueter chaque bloc en fonction de son activité. Nous notons les régions de faible activité (régions uniformes) S_{LA} et les régions de forte activité (régions très texturés ou contenant de forts contours) S_{HA}. La Figure 4.5 permet d'illustrer ce partitionnement, où les régions noires indiquent celles de type S_{LA}, et les autres régions indiquent celles de type S_{HA}.

Afin de tirer profit de ce partitionnement (stocké dans un masque d'activité), une seconde variante de la métrique QIP [NLF10b] est proposée. La Figure 4.6 illustre les modifications apportées afin d'intégrer la prise en compte de l'activité initiale de l'image garantissant la sensibilité aux différents types de variations structurelles. Nous pouvons voir que le masque d'activité est calculé côté récepteur et est utilisé pour la quantification des points d'intérêt dans chaque type de classe. Ce masque ainsi que deux entiers représentant le nombre de points dans chaque classe sont ensuite envoyé en tant que référence réduite. Il peut donc être utilisé côté récepteur pour la nouvelle quantification de points d'intérêt.

De ce fait, la mesure de quantité de variation structurelle n'est maintenant plus obtenue en utilisant deux mais quatre entiers pour estimer la qualité perçue. Le premier couple d'entiers permet de mesurer l'évolution des points dans les régions de forte activité (S_{HA}), le second quant à lui est dédié à la mesure dans les régions de faible activité (S_{LA}). Le score de qualité de la métrique est obtenu par la formule :

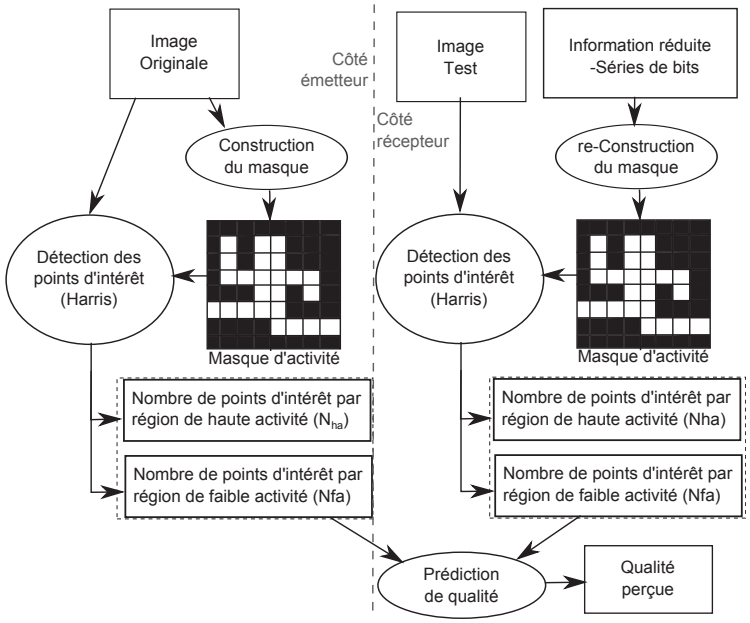


Figure 4.6 : Schéma fonctionnel de la métrique QIP version 2

$$Q_{qip} = w_{HA} \times S_{HA} + w_{LA} \times S_{LA} \tag{4.1}$$

où w_{HA} est le taux d'occupation que représente la région de forte activité par rapport à la surface totale de l'image et respectivement $w_{LA} = 1 - w_{HA}$ pour la proportion de faible activité. S_{HA} et S_{LA} sont respectivement les mesures de qualité pour les deux types de régions. Ils sont calculés comme suit :

$$S_{**} = \begin{cases} 1 - \frac{VDiff}{maxP} & si\ maxP > 0 \\ 1 & sinon \end{cases} \tag{4.2}$$

Avec $VDiff$ la différence en valeur absolue du nombre de points d'intérêt entre l'image d'origine et l'image de test et $maxP$ le nombre maximum de points détectés (dans l'image d'origine ou après dégradation).

Par cette formulation du calcul de la qualité, les valeurs sont comprises entre 0 et 1. Un score de 1 indique une qualité parfaite, aucune différence n'est observée entre l'image d'origine et l'image dégradée ; l'image est donc visuellement de bonne qualité. A l'inverse, une image avec un score proche de 0 indique une qualité visuelle médiocre car une importante différence a été observée avec l'image d'origine. La Figure 4.7 permet d'illustrer diverses prédictions de la qualité en comparaison à une métrique à référence complète parmi les plus utilisées, à savoir SSIM [WBSS04] pour des puissances de compression allant de très dégradée à légèrement dégradée. Les images et leurs scores de qualité

subjective MOS (Mean Opinion Score) sont issus de la base d'images LIVE [SWCBa].

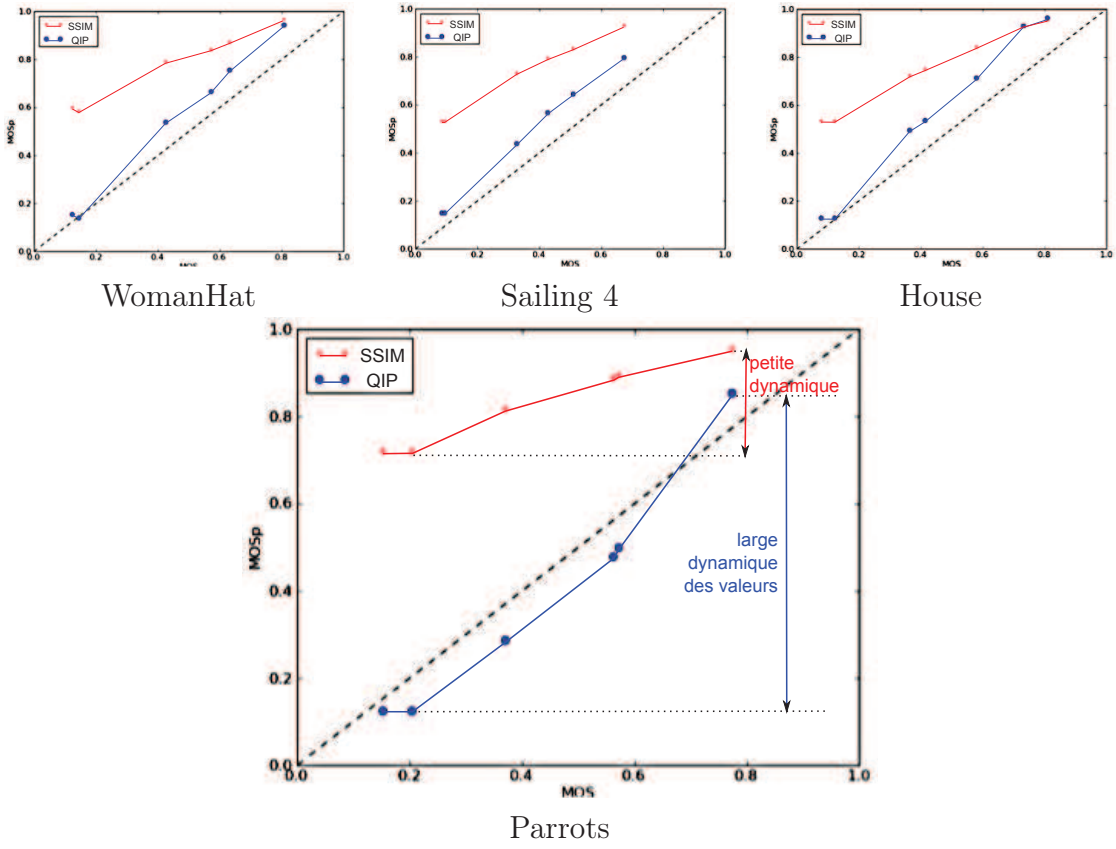


Figure 4.7 : Comparaison des dynamiques des scores prédits

Sur ces graphiques, on peut noter que les deux métriques prédisent la qualité dans un sens concordant avec le jugement humain. Les prédictions de SSIM sont plus stables que celles de QIP et varient moins d'une image à l'autre si l'on observe le comportement sur les 29 images de référence de la base. Cependant, la plage des valeurs est également plus limitée et aucune prédiction de qualité n'est en dessous de 0.5. Ainsi, même pour des images très dégradées pour lesquelles les observateurs jugent la qualité comme inacceptable, SSIM quant à lui, prédit une qualité moyenne et donc acceptable. La plage des valeurs minimale et maximale de cette métrique parait inappropriée en comparaison à la métrique proposée qui fournit des scores exploitant toute la dynamique et donc plus en accord avec l'humain. De ce fait, les valeurs prédites par notre métrique paraissent comme plus cohérentes car ne nécessitant pas de ré-interprétation de l'échelle des valeurs.

Pour continuer l'analyse des valeurs de qualité prédite par QIP, nous proposons le tableau 4.1. Sur ce tableau, nous pouvons voir deux images différentes,

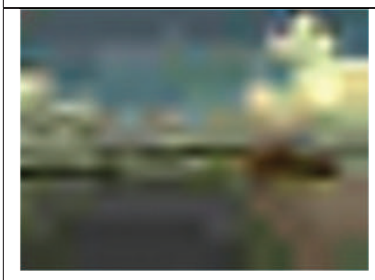
Images \Metriques	PSNR	PSNR-H	SSIM	QIP2
	24	24.45	0.78	**0.99**
	24	19.8	0.58	**0.09**

Tableau 4.1 : Comparaison des dynamiques des scores prédits

avec deux niveaux de dégradation très différents. L'une apparaît d'une assez bonne qualité pour un observateur, tandis que l'autre est quasiment inexploitable. Cependant, on peut noter que les scores de qualité des métriques ne reflètent pas toujours cette différence notable. En effet, le très connu PSNR affiche deux valeurs identiques. On note une légère amélioration pour la version PSNR-HVS [EAP+06] qui prend en compte certains comportements du système visuel humain. Mais c'est finalement QIP qui fournit les valeurs de qualité les plus facilement compréhensibles et exploitables, en donnant un score proche de 0, qui est son minimum, pour l'image de très mauvaise qualité, et un score proche de 1, qui est son maximum, pour l'image visuellement très acceptable. Rappelons que ces résultats sont obtenus sans faire appel à la référence de manière complète.

Au vu des résultats encourageants de cette approche d'estimation de la qualité par mesure d'évolution des points d'intérêt, nous nous sommes penchés sur l'adoption d'un partitionnement en régions plus intelligent en introduisant des critères perceptifs en se focalisant sur la saillance visuelle à plusieurs niveaux, ce qui est l'objet de la prochaine section.

4.1.2 Méthode QIP-HSM

Nous venons de présenter une nouvelle méthode d'extraction de caractéristiques structurelles capable de quantifier l'apparition d'artéfacts dans l'image. Cependant, il est reconnu que la localisation des artéfacts a une influence sur leur perception. Typiquement, des défauts placés dans une région saillante de l'image, par exemple au milieu d'un visage, vont être plus facilement détectés et jugés plus lourdement que s'ils sont placés dans une région moins attractive. En ce sens, les travaux de Tong et al. [TKCT10] ([NLMLCB07], [MB09]) ont montré qu'ajouter une carte de saillance à des métriques existantes permet d'augmenter la corrélation avec le jugement humain. C'est dans cette direction et dans le but de prendre en compte la localisation spatiale des artéfacts et leur niveau de saillance que nous avons proposé une évolution de la métrique QIP.

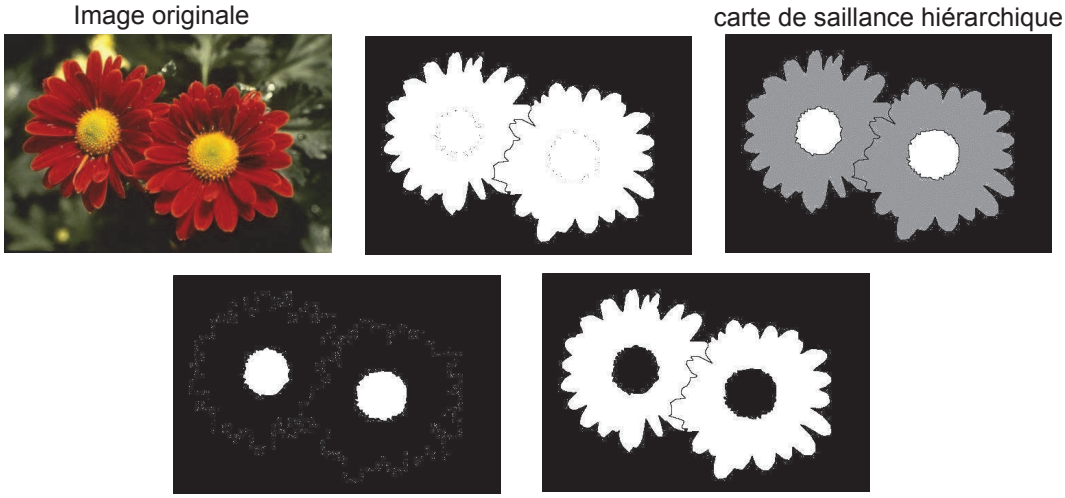


Figure 4.8 : Illustration de la hiérarchie de saillance

L'idée est donc de continuer à mesurer les défauts par l'évolution des points d'intérêt, mais d'ajouter dans la formulation de prédiction de la qualité un processus de pondération, en utilisant une carte de saillance informant de l'importance et de la visibilité de chaque région. Concrètement, nous souhaitons disposer d'une carte qui partitionne l'image en régions et que chaque région soit étiquetée par son niveau de saillance. Il y a dans une image une sorte de hiérarchie d'importance du contenu. Si l'on considère la photographie des fleurs de la Figure 4.8, on se rend compte que c'est principalement les fleurs rouges qui attirent notre attention par rapport à l'arrière plan. Mais qu'au sein de ce couple de fleurs, le cœur jaune de chaque fleur est lui même plus attractif que les pétales. Nous pouvons également noter à travers cet exemple, que le partitionnement en régions saillantes fait également sortir des régions

similaires d'un point de vue richesse de contenu. Par exemple, l'arrière plan est une région de faible activité, les pétales rouges contiennent de forts contours, les cœurs jaunes ont des textures très fines. Finalement, ce partitionnement par niveaux saillants ne semble pas non plus si éloigné du partitionnement par deux niveaux d'activité utilisé dans la méthode QIP définie précédemment. Il peut également favoriser la sensibilité à l'évolution des points d'intérêt. Nous reviendrons sur la contribution de la carte de saillance hiérarchique.

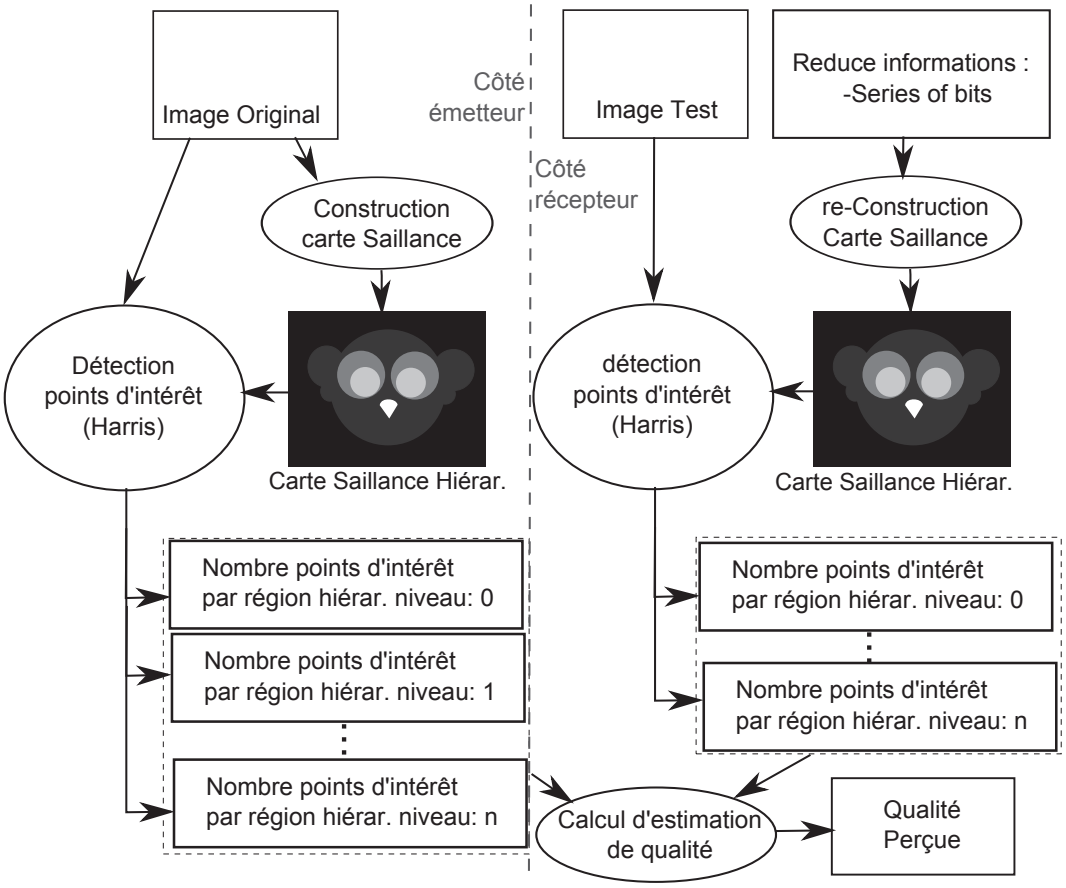


Figure 4.9 : Schéma fonctionnel de la métrique QIP-HSM

La Figure 4.9 illustre les modifications apportées à la métrique QIP pour en faire la version QIP-HSM [NLF11a] (Quality by Interest Point using Hierarchical Saliency Map) tirant profit de la carte de saillance hiérarchique. Nous pouvons noter le remplacement de la carte d'activité par la carte de saillance hiérarchique. Cette carte de saillance peut contenir jusqu'à n couches c'est-à-dire n niveaux de saillance, dépendant des éventuels *a priori* des images à partitionner et de la finesse d'analyse souhaitée. Sur le même principe que précédemment les points d'intérêt sont extraits dans chaque couche permettant les mesures de qualité S_* à chaque niveau. Disposant de n scores de qualité S_*, une pondération w_* peut être affectée à chacune afin de donner plus ou moins

d'importance aux dégradations détectées en fonction du niveau de saillance de chaque région [1]. Le score de qualité prédit est donc obtenu par :

$$Q_{qip-hsm} = \sum_{i=1}^{n} w_i \times S_i. \tag{4.3}$$

Par cette méthode, différentes constantes doivent êtres fixées. Il y a tout d'abord le nombre de couches n souhaité dans la hiérarchie de saillance. Ainsi que la pondération w_* associée à chacun des niveaux. Cependant, cette pondération est également dépendante de la méthode d'extraction de saillance et éventuellement de la surface qu'occupe chacune des régions. Nous proposons dans la section suivante notre approche de création de cartes de saillance hiérarchique et d'analyser par la même occasion les différents aspects mentionnés.

Construction de carte saillance hiérarchique

Nous avons précédemment introduit la notion de saillance visuelle dans le Chapitre 2, et avons fait un état de l'art des différents modèles de prédiction de cartes de saillance. Nous avons distingué deux propositions. La première, proposée par Achanta [AHES09a], se veut relativement minimaliste et très orientée vers la réduction du temps d'exécution ; contrairement à la seconde, dans laquelle s'inscrivent les travaux de Itti [IKN98] et Le Meur [LMLCBT06], qui quant à elle, vise à modéliser et intégrer de nombreuses propriétés du système visuelle humain (SVH). La méthode que nous proposons, cherche à tirer profit des deux approches. Pour ce faire, nous basons nos développements sur la méthode proposée par Achanta, dans laquelle nous intégrons plusieurs caractéristiques bio-inspirées issues des modèles de Itti et Le Meur.

Pour la création de notre carte de saillance, nous commençons par convertir l'image dans l'espace couleur perceptuel CIE L*a*b* [Con04] spécialement étudié pour que les distances calculées entre couleurs soient mieux corrélées avec les différences perçues par l'œil humain. Puis, pour chaque composante nous appliquons un filtrage par une fonction de sensibilité au contraste (CSF

1. La pondération de la visibilité des artéfacts en fonction de propriétés psycho-visuelles est un processus couramment exploité dans les métriques perceptuelles comme détaillé dans la section 3.3.2. Dans cet objectif, les métriques exploitent classiquement un filtrage issu de modèles de sensibilité au contraste (CSF [RRFM07]) et non la saillance comme c'est le cas dans notre approche. Cependant, ce type filtrage est tout de même utilisé dans nos travaux, mais au niveau de la construction de nos cartes de saillance, dont les détails sont disponibles dans la section suivante

Figure 4.10 : Différence achromatique et chromatique

L a b

Figure 4.11 : Carte de saillance par différence sur les canaux achromatique et chromatique

[Dal93, FMLBR05, RLFM08]) spécifique pour chaque canal (issu des travaux de Nadenau [Nad00]) afin de mimer la sensibilité variable du SVH aux différentes fréquences spatiales sur les canaux achromatique et chromatique. Une fois ce filtrage effectué, nous appliquons la même formulation que dans les travaux de Achanta et al., qui consiste à comparer chaque pixel à la valeur moyenne des pixels de la composante considérée. La Figure 4.11 permet de visualiser le type de carte de saillance pouvant être obtenue si l'on considère chaque canal indépendamment. Par cette observation, nous pouvons juger de l'utilité de considérer tous les canaux car ce qui paraît saillant sur un canal, ne l'est pas forcément sur les autres.

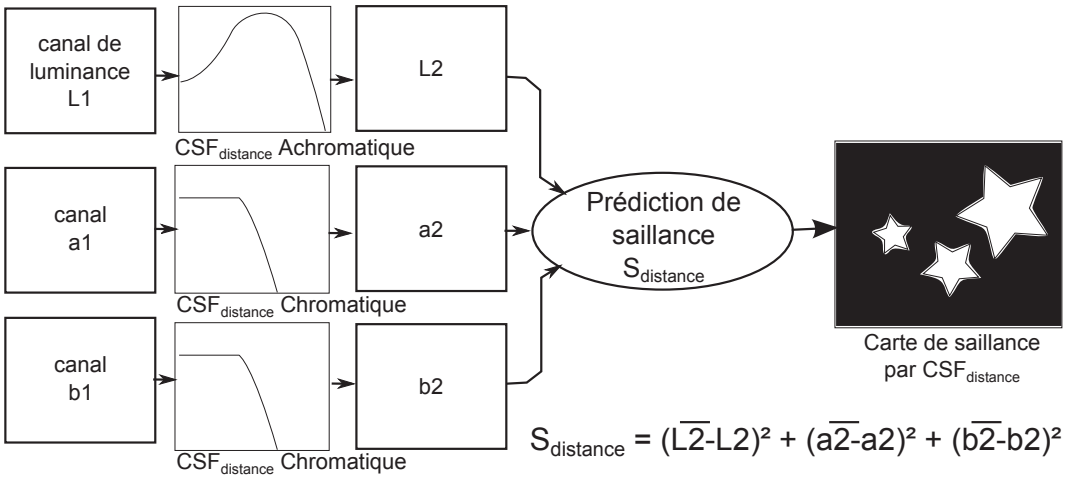


Figure 4.12 : Prédiction de saillance simulant une distance d'observation au travers du paramétrage des CSF

L'intérêt d'avoir intégré la CSF réside dans sa capacité à filtrer l'image en fonction d'une distance d'observation [RCFM09]. Par ce biais, il nous est possible de filtrer l'image de manière à simuler une observation lointaine (ou observation peu attentive), ou une observation de près, optimale (ou observation attentive). Ceci permet de générer plusieurs cartes de saillance dépendant du niveau d'observation considéré. La Figure 4.12 illustre le processus de prédiction de saillance, utilisant les modèles de CSF paramétrés pour une simulation de distance considérée. Notre idée est de prédire la saillance en simulant une observation lointaine afin d'extraire les objets saillants principaux et de gagner en précision de localisation et particulièrement de contours d'objets par la prédiction de saillance en simulant une distance d'observation optimale. Une autre spécificité du SVH est prise en compte. Il s'agit du biais centré [Tat07] qui fait que notre regard se pose en premier au centre de l'image. Nous avons simulé ce phénomène par l'application d'une gaussienne centrée sur les résultats de saillance. Nous proposons de fusionner, dans un premier temps, la carte de saillance simulant une vision lointaine avec le biais centré, car ces deux processus se rapportent à une vision pré-attentive. Puis, dans un second temps, nous

gagnons en précision en ajoutant la simulation de saillance attentive. La figure 4.13 permet de visualiser le processus de fusion nécessaire à la construction de la carte de saillance de niveau 1.

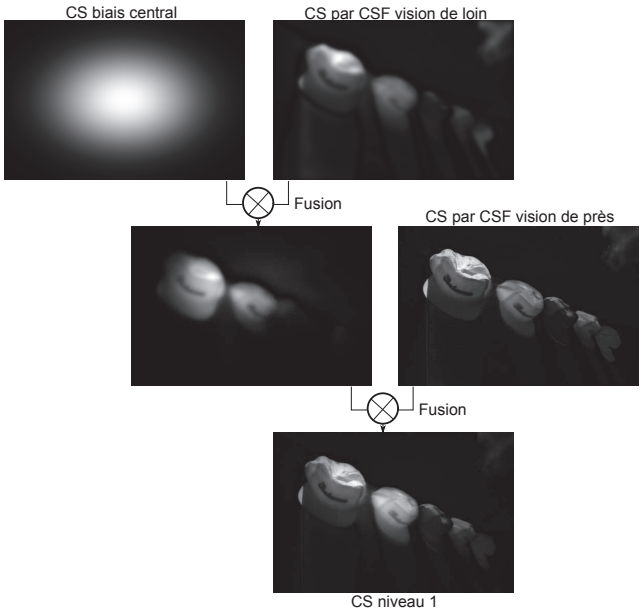


Figure 4.13 : Fusion des cartes de saillance (CS) pour la construction de la carte de niveau 1

Enfin nous binarisons la carte obtenue en appliquant un seuillage T_{bin} dont nous proposons la valeur 0.5 après avoir normalisé la carte de saillance entre 0 et 1, en considérant que les régions réellement attractives sont celles dont la saillance est supérieure à 50% de la saillance maximale [2].

Par cette démarche nous avons extrait les régions saillantes de premier niveau. Dans le but de construire notre hiérarchie, nous choisissons de simuler l'inhibition temporelle ou de retour, en considérant qu'une région déjà observée n'est plus réellement attractive. C'est pourquoi nous proposons d'utiliser la carte de saillance binarisée obtenue précédemment pour masquer les pixels déjà détectés comme saillant dans le calcul de moyenne par canal. Par ce biais, nous donnons la possibilité à d'autres régions de devenir saillantes à leur tour. Ce processus peut être itéré n fois en fonction du nombre de niveaux hiérarchiques souhaités et tant qu'il reste des pixels à traiter. L'ensemble de la procédure de création de nos cartes hiérarchiques est résumé sur la Figure 4.14 accompagnée d'un exemple complémentaire (Figure 4.15) illustrant les cartes obtenues à chaque niveau.

2. Il est également possible d'ajouter une contrainte sur le taux d'occupation de la surface saillante. Tpiquement, nous pouvons choisir que le taux d'occupation de la saillance de premier niveau soit la plus proche possible de 20% de la surface totale de l'image.

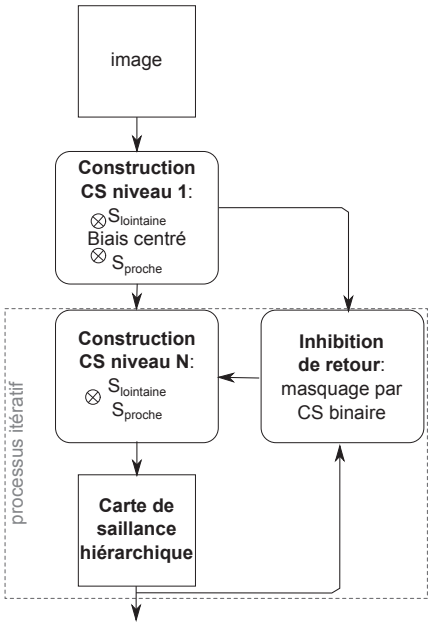


Figure 4.14 : Architecture du modèle de construction de cartes de saillance hiérarchique

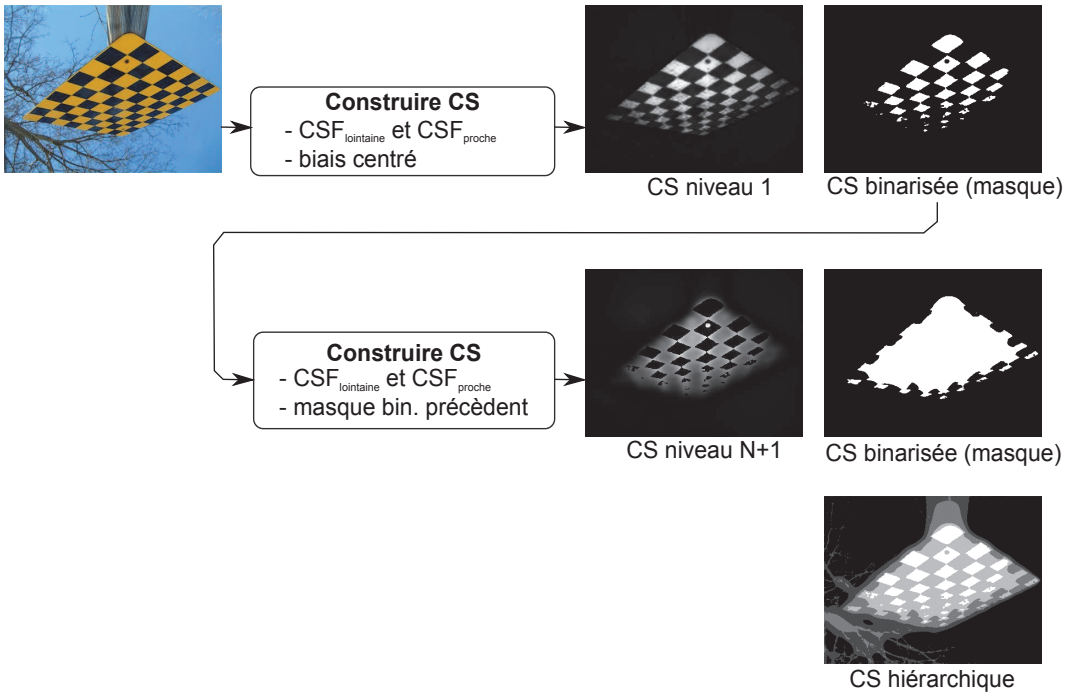


Figure 4.15 : Illustration de cartes des saillances obtenues le long de la hiérarchie

Création d'une application flexible pour la création de cartes de saillance hiérarchique

La démarche présentée précédemment illustre la création de cartes de saillance hiérarchique basées sur des propriétés visuelles de bas-niveau, telles que le biais centré et les contrastes chromatique et achromatique. Cependant, la saillance visuelle est aussi influencée par d'autres propriétés de haut et de bas-niveaux. En ce sens, nous avons également fait évoluer notre modèle en développant une application de prédiction de saillance flexible et adaptable en fonction des contextes d'application. A ce titre, nous avons intégré la prédiction de saillance par richesse de contenu à travers la quantification du gradient par bloc, en considérant que les régions de faible activité sont moins attractives que les régions fortement texturées ou contenant de forts contours (net/flou-premier - plan/arrière plan). La Figure 4.16 permet d'illustrer visuellement des cartes obtenues par cette approche, où les régions saillantes sont de plus forte intensité.

Figure 4.16 : Carte de saillance par richesse de contenu local

Toujours en utilisant les statistiques de bas-niveau de l'image nous avons intégré la prédiction de saillance par analyse des structures locales détaillée dans le Chapitre 2.

Nous avons également intégré la génération de cartes de saillance de haut-niveau par la détection/reconnaissance de visages, yeux, bouche, corps... Tout en respectant la notion hiérarchique au sein de cette détection en favorisant les yeux, suivi de la bouche, suivi du visage. Nous pouvons citer l'hypothèse selon

laquelle les yeux et la bouches sont plus attractifs que le reste du corps, car très utilisés pour transmettre ses émotions de manière non verbale. A ce titre, ce sont ces deux régions particulières qui sont majoritairement utilisées pour la reconnaissance des émotions humaines en vision par ordinateur [Pic00, Pic01]. Nous illustrons sur la Figure 4.17 quelques détections de saillance haut-niveau obtenues.

Figure 4.17 : Carte de saillance haut-niveau (visage, yeux, bouche)

Disposant de toutes ces modalités de saillance, se pose un problème récurrent, à savoir, la fusion des informations. Nous pensons qu'il n'existe aucune solution optimale capable de gérer tous les cas de figure. Nous pensons qu'il existe plutôt diverses méthodes, très dépendantes de l'utilisation visée, du contexte et de l'expérience ciblée. Pour cela, nous ne proposons pas non plus un type de fusion mais plutôt de laisser la possibilité et un maximum de flexibilité pour agréger ces cartes avec un accès à des pondérations variables et le choix d'activer ou désactiver tel ou tel type de détection. A titre d'exemple, nous pouvons illustrer (Figure 4.18) la fusion de détection de la saillance bas-niveau (chromatique/achromatique/biais centré) avec la saillance de haut-niveau (Visage/yeux), qui pourrait être particulièrement adaptée à la diffusion de vidéos où la présence humaine et les visages sont nombreux.

En ce qui concerne la hiérarchie de saillance, nous avons intégré un processus de maximisation de surface cible par couche. En effet, dans bon nombre

Figure 4.18 : Fusion de saillance bas-niveau (chromatique/achromatique/biais centré) avec la saillance haut-niveau (visage/yeux)

d'utilisations, il est intéressant de pouvoir fixer la taille de chaque couche, par exemple pour de la compression, du cropping, du redimensionnement, de la transmission... La Figure 4.19 illustre le résultat de création d'une carte de saillance hiérarchique avec quatre couches et une contrainte de taux d'occupation spatiale fixée à 25% par couche.

Pour finir, nous pouvons tout de même donner quelques recommandations sur la fusion des cartes de saillance. D'après la littérature, ce sont des propriétés relativement de haut-niveau qui semblent les moins invariantes au contexte, telles que la détection de visages, suivie de la détection de textes [EMZ09, ELZ+10]. Viennent ensuite les propriétés de bas-niveau toujours accessibles sur tout type de contenu, contrairement aux deux premières. Pour respecter ces contraintes, nous conseillons de pondérer les différentes cartes générées par notre outil, en affectant le plus de poids au détecteur de visage, suivi du prédicteur de saillance utilisant le détecteur de coins de Harris (favorisant la détection de textes).

Expérimentations et validation

Nous avons décrit précédemment la métrique de qualité QIP-HSM, en explicitant diverses méthodes de construction de cartes de saillance. Au vu des nombreuses variables pouvant influencer la qualité prédite $Q_{qip-hsm}$, nous avons dû faire des choix dans la démarche d'expérimentation et de validation de l'ap-

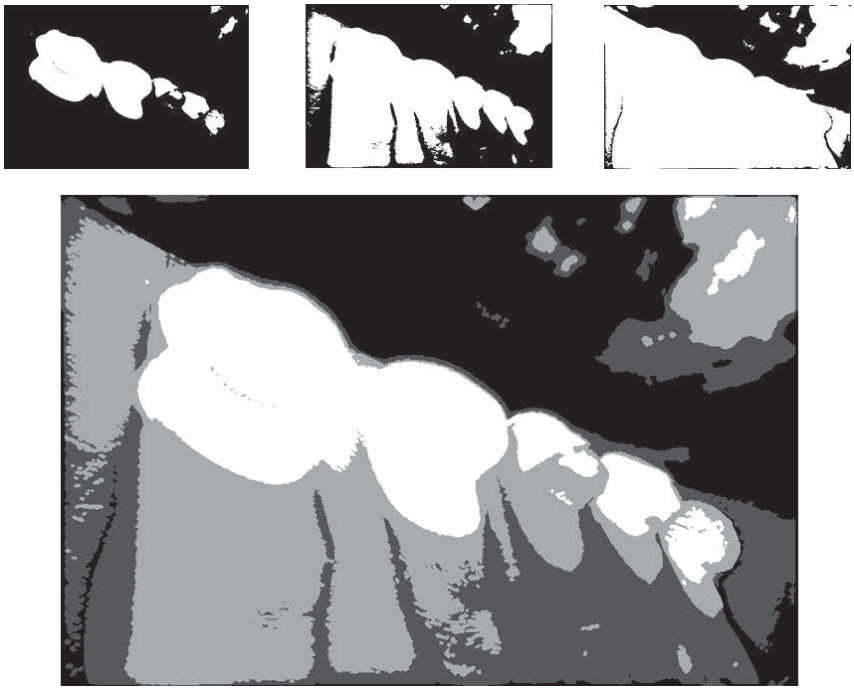


Figure 4.19 : Carte de saillance hiérarchique à 4 couches, dont les contraintes de taux d'occupations cibles sont fixées (niveau 1 : 25%, niveau 2 : 25%, niveau 3 : 25%, niveau 4 : 25%). Les taux d'occupation obtenus : (18.3%,31.2%,22.6%,27.8%)

proche en fixant un maximum de variables. Dans un premier temps, notre but est de prouver qu'au moins une méthode de construction de cartes de saillance associée à un choix de constante w_i permet de prédire la qualité en corrélation avec le jugement humain. Dans un second temps, c'est un travail d'optimisation qui doit être effectué afin d'identifier la démarche optimale de création de cartes de saillance hiérarchique associées à la meilleure pondération de chaque couche. Dans de nombreux processus d'optimisation, les pondérations optimales sont obtenues par l'utilisation de bases d'apprentissage et de réseaux de neurones. Cependant, ces démarches sont toujours sujettes à de nombreux questionnements, avec par exemple le choix de la base d'apprentissage et du modèle utilisé pour le processus d'optimisation. Ne cherchant pas spécialement à focaliser nos travaux sur ces problématiques, ce travail d'optimisation ne sera pas traité. C'est donc une démarche relativement empirique et expérimentale qui est détaillée dans cette section. Néanmoins, les résultats obtenus par cette démarche sont très encourageants, tout en gardant en tête qu'un gain de performance notable est bien souvent possible par un processus d'optimisation, et ceci pourrait faire l'objet de travaux à la suite de cette thèse.

Afin de prouver la pertinence de notre approche, nous avons choisi de focaliser cette étude expérimentale sur la prédiction de qualité d'images affectées

par la compression JPEG, du fait de la large adoption de ce format de compression. N'ayant pas d'*a priori* sur le nombre optimal de couches de saillance, nous avons fait le choix de le fixer à six, permettant à la fois d'avoir un partitionnement relativement fin, garantissant ainsi une certaine précision de la mesure, tout en ne paraissant pas spécialement excessif. Nous avons également choisi d'effectuer nos mesures sur la base d'images LIVE 2 [SWCBb] très utilisée et reconnue dans le domaine de l'évaluation de la qualité d'images, permettant d'avoir accès aux valeurs subjectives de qualité (MOS) sur le type de compression choisi. Nous avons également extrait deux images (caps et sailing3) de cette base pour illustrer notre approche afin de disposer d'images avec un contenu assez différent.

Dans un premier temps, nous proposons d'étudier le comportement de chaque couche de manière indépendante afin d'évaluer les éventuels apports et performances de chacune d'elles. Sur les Figures 4.20 et 4.21, nous pouvons visualiser l'évolution des prédictions de qualité (S_*) de chaque couche en fonction de l'évolution du débit binaire des images après compression. Sur chaque figure, la courbe à approcher et servant de référence est la courbe d'évolution des MOS, matérialisée en bleu et sans marque.

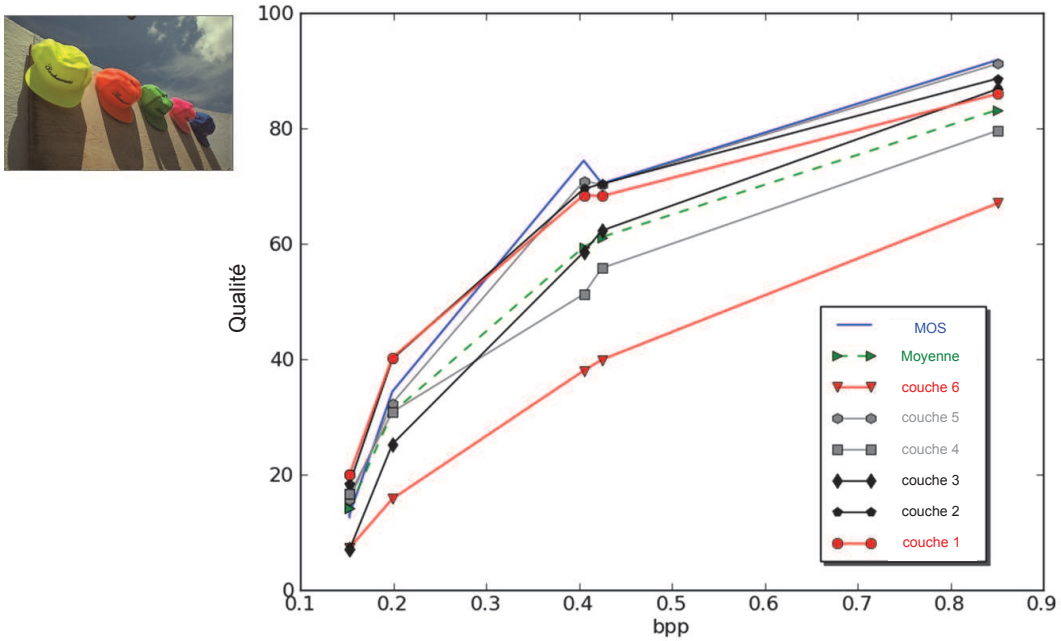


Figure 4.20 : Qualité par couche sur l'image "caps" en fonction du débit binaire en bits par pixel (bpp)

Au vu de ces courbes, nous pouvons noter que l'estimation de qualité pour chaque couche est en accord avec le jugement humain et l'évolution du débit binaire. Cependant, nous pouvons observer que la couche 1 (correspondant à

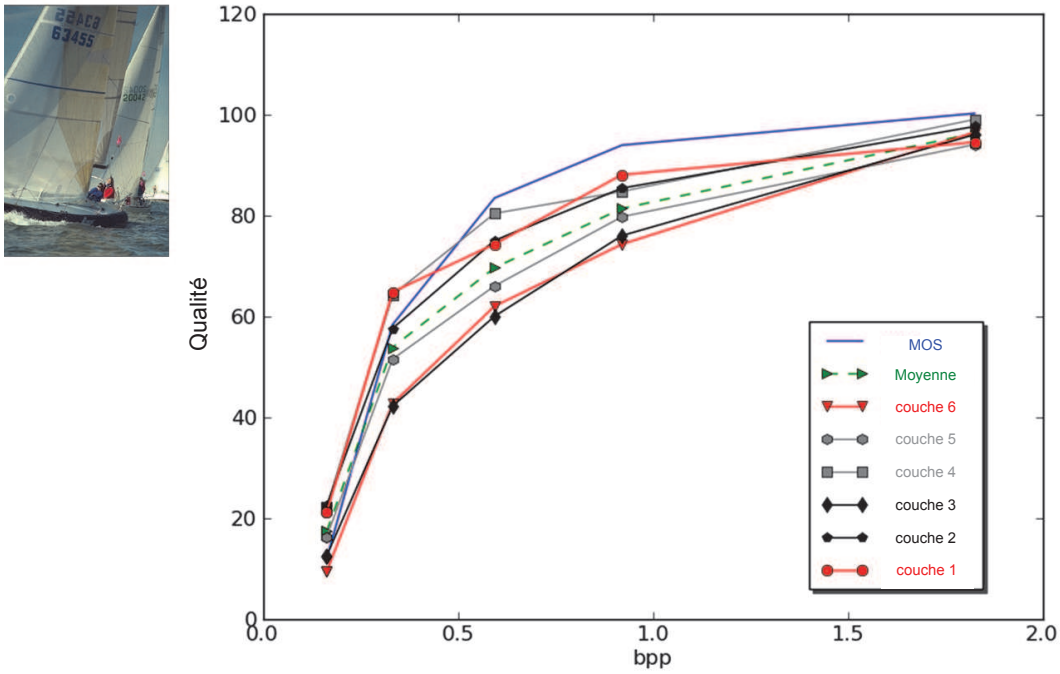


Figure 4.21 : Qualité par couche sur l'image "sailing3" en fonction du débit binaire en bits par pixel (bpp)

la saillance de premier niveau) est la plus proche des notes subjectives. Cette remarque semble se confirmer sur les deux images. Les couches intermédiaires (de 2 à 5) sont globalement assez proches, mais avec plus de variabilité. Enfin, la dernière couche est également bien en adéquation avec l'évolution du MOS, mais présente l'écart le plus important. Une première conclusion consiste à constater que toutes les couches permettent de prédire la qualité perçue, mais que certaines d'entre-elles paraissent comme plus pertinentes et précises. Cependant, n'utiliser qu'une unique couche ne semble pas judicieux car il est envisageable qu'une dégradation puisse n'affecter qu'une localisation particulière, et si celle-ci est dans une autre région que celle étudiée, elle pourrait ne pas être quantifiée du tout. Comme les quatre couches intermédiaires présentent une proximité des résultats, nous décidons de fusionner ces informations. De ce fait, l'équation 4.3 peut être réécrite sous la forme :

$$Q_{qip-hsm} = w_f \times S_f + \sum_{i=2}^{n-1} w_{mi} \times S_{mi} + w_b \times S_b \tag{4.4}$$

où S_f et S_b sont respectivement les prédictions de qualité pour la couche 1 (foreground/ premier plan) et la couche n (background/ arrière plan). Tandis

que les couches intermédiaires sont regroupées sous S_m.

D'après les analyses précédentes, nous proposons trois types de pondération : une première maximisant l'influence de la couche S_f, une autre maximisant S_b et une dernière pour la maximisation des niveaux intermédiaires. Le récapitulatif des pondérations testées est fourni sur le tableau 4.2 et les résultats graphiques sont visualisables sur les Figures 4.22 et 4.23. En observant ces courbes, nous pouvons conclure que la maximisation de la pondération $MOSpF$ est pertinente en minimisant l'écart avec les notes subjectives. Cependant, nous pouvons noter également l'intérêt de la pondération w_m, extrêmement précise sur les faibles débits binaires, et dont le profil semble plus stable que la pondération $MOSpF$. C'est pourquoi nous proposons une dernière pondération $MOSpOpt$ dite "optimale" favorisant la première couche pour sa proximité avec le subjectif, en donnant un poids important aux couches intermédiaires garantissant une meilleure stabilité, tout en considérant que la superficie occupée par ces couches intermédiaire est importante en comparaison à la première.

Tableau 4.2 : facteurs de pondération (valeurs w_*)

Score prédit	w_f	w_m	w_b
MOSpF	0,8	0,1	0,1
MOSpB	0,1	0,1	0,8
MOSpMean	1/6	4/6	1/6
MOSpOpt	0,4	0,5	0,1

Ayant déterminé une pondération pour chacune des couches de la métrique QIP-HSM, nous proposons de vérifier (Figure 4.24 et 4.25) l'apport de cette méthode en la comparant avec nos deux autres propositions utilisant des points d'intérêt. Les courbes les plus claires et grises, nommées MOSpIP correspondent à la prédiction de la qualité de la métrique QIP version 1 (cf. Figure 4.3), les courbes vertes sont consacrées aux prédictions de QIP version 2 exploitant le masque d'activité (cf. Figure 4.6) et enfin les courbes rouges représentent la métrique QIP intégrant la saillance hiérarchique (cf. Figure 4.9) et la pondération définie précédemment.

Sur ces graphiques, nous pouvons noter l'apport de l'utilisation des masques d'activité et des cartes de saillance hiérarchique. L'apport du masque d'activité est prouvé, mais faible en comparaison de l'apport de la saillance. La première conclusion consiste à ce que la version la plus simple de notre métrique, ne nécessitant finalement l'envoi que d'un seul entier peut être envisagée dans un cadre où la recherche de minimisation de données à transmettre et de

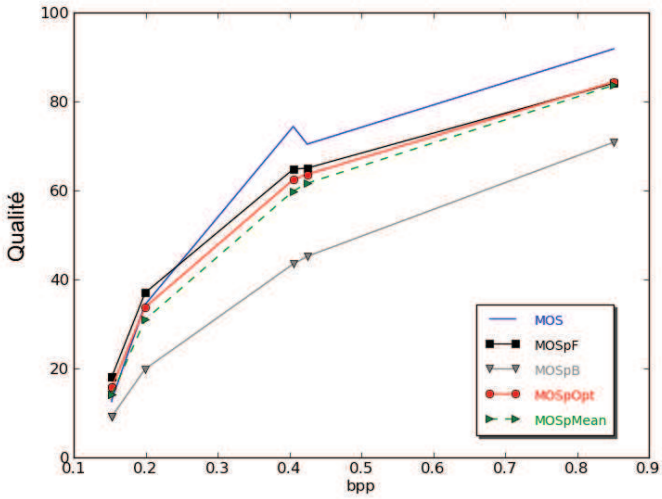


Figure 4.22 : Qualité estimée par différentes pondérations des couches sur l'image "caps3"

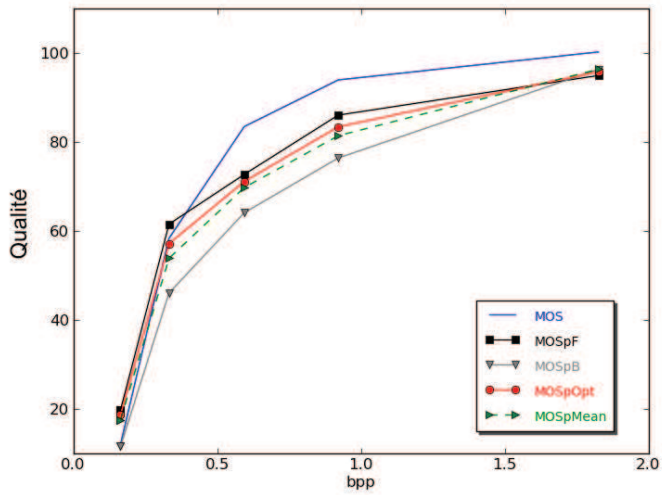


Figure 4.23 : Qualité estimée par différentes pondérations des couches sur l'image "sailing3"

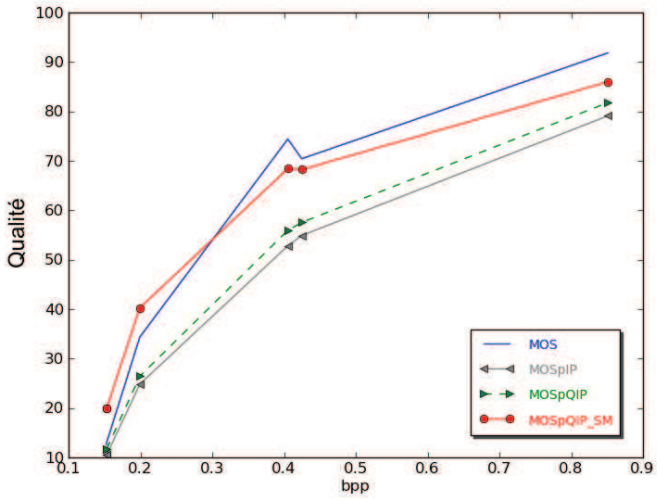


Figure 4.24 : Qualité estimée par les différentes variantes de la métrique QIP sur l'image "caps"

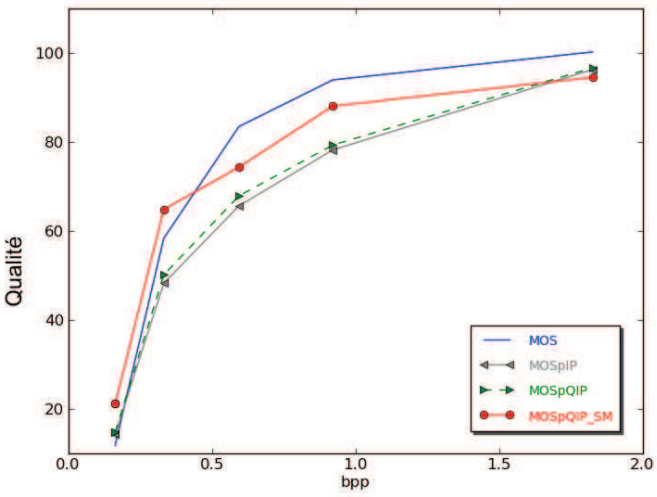


Figure 4.25 : Qualité estimé par les différentes variantes de la métique QIP sur l'image "sailing3"

temps de calcul est extrême, garantissant une prédiction certes moins précise mais acceptable. Ensuite, nous pouvons également conclure que l'utilisation des cartes de saillance hiérarchique, intégrant quelques propriétés du SVH, est la plus pertinente de nos propositions. Cependant, un léger surcoût de calcul et de référence est tout de même nécessaire. En ce qui concerne le poids des références, c'est principalement le poids de la carte hiérarchique qui est supérieur au poids de la carte d'activité binaire utilisant un découpage en blocs. Cependant, le codage d'une carte hiérarchique à n couches n'est pas élevé en comparaison d'une image couleur (RVB) 8 bits (sachant que $n << 3 \times 255$) et en comparaison aux métriques à référence complète. Il est de plus envisageable d'utiliser diverses techniques de compression pour coder et transmettre cette carte. Néanmoins, au même titre que le choix optimal de pondération, le choix optimal de compression de cette carte n'est pas développé.

Notre dernière validation consiste à comparer notre métrique avec sept métriques de la littérature (VSNR [DVKG+00] ,PNSR, Pdiff [Yee04], PSNR HVS [EAP+06], SSIM [WBSS04] et IFC [She04], chacune d'elles étant détaillée dans la section 3.3) sur les 29 images de référence de la base d'images LIVE 2 [SWCBb]. Les résultats en terme de corrélation de Pearson et de RMSE sont compilés dans le Tableau 4.3 pour les dégradations de type JPEG et le Tableau 4.4 pour JPEG 2000.

Tableau 4.3 : Performances sur la base de données LIVE 2/JPEG

D\M	VSNR	PSNR	VIF	Pdiff
Corr.	0,951	0,905	0,949	0,937
RMSE	0,208	0,354	0,203	0,167
D\M	PSNR HVS	SSIM	IFC	QIP HSM
Corr.	0,956	0,976	0,915	0,979
RMSE	0,203	0,203	0,262	0,243

Tableau 4.4 : Performances sur la base d'image LIVE 2/JPEG2000

D\M	VSNR	PSNR	VIF	Pdiff
Corr.	0,960	0,917	0,959	0,972
RMSE	0,208	0,354	0,203	0,167
D\M	PSNR HVS	SSIM	IFC	QIP HSM
Corr.	0,970	0,961	0,938	0,978
RMSE	0,203	0,203	0,262	0,243

Au vu de ces résultats, nous pouvons conclure que la métrique proposée

offre de bons résultats en terme de corrélation et de RMSE, sachant que notre proposition peut être vue comme une métrique à référence réduite utilisant une toute nouvelle manière de prédire la qualité.

Finalement, avec les résultats encourageants de nos propositions, nous avons intégré l'une de nos métriques de qualité dans un contexte applicatif réel, dans une collaboration dans le cadre du projet ANR caiman dans un contexte de transmission.

4.2 Intégration de QIP dans une chaîne de transmission sans fil

Au cours de la dernière décennie, les transmissions d'images à travers des canaux sans fil sont apparues comme un service multimédia très populaire, en particulier avec le développement des terminaux mobiles (smartphones, tablettes). Cependant, le canal de transmission sans fil varie au cours du temps d'une manière aléatoire à cause de la mobilité de l'environnement et des utilisateurs. La nature instable et la bande passante limitée des liens sans fil représentent le problème central devant être pris en compte afin de garantir des services multimédias offrant la meilleure qualité d'expérience aux utilisateurs.

Dans cette optique, le laboratoire XLim-SIC, propose de nouvelles approches de transmissions d'images en exploitant les canaux MIMO [3] couplés à divers stratégies d'adaptation des signaux, de codes correcteurs d'erreurs et d'adaptation des puissances d'antennes.

Également impliquée dans le codage et la compression des images sur canaux sans fil, l'entreprise Thales travaille quant à elle, sur la compression JPWL (JPeg WireLess [JPW91]) robuste, afin d'apporter des réponses et un décodage des images malgré les nombreuses pertes de paquets ou d'erreurs introduites par la transmission.

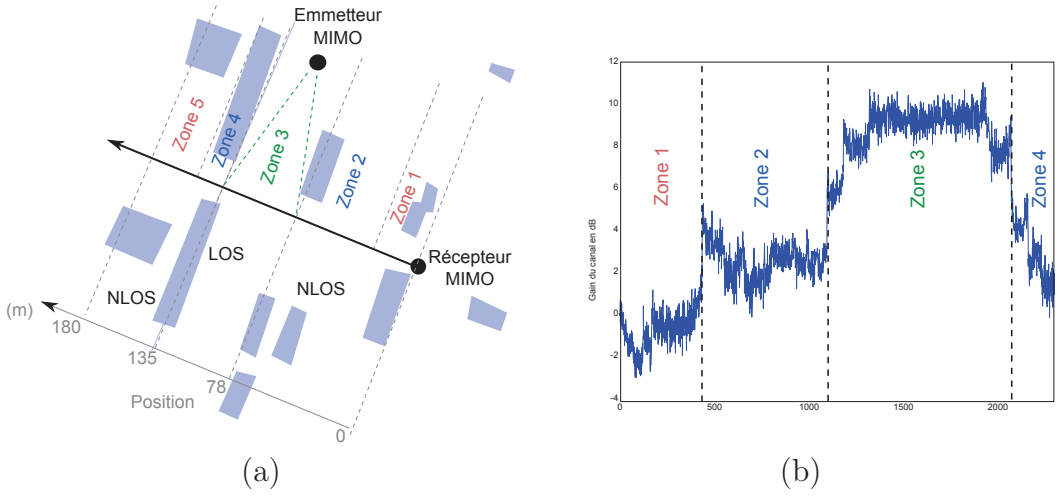
Dans ces domaines, les travaux de la littérature s'intéressent principalement à l'augmentation de la Qualité de Service [CGK+11] (QoS :Quality of Service), basée sur des mesures objectives, tels que le taux d'erreur binaire (TEB) et le rapport signal sur bruit (SNR). Finalement, la QoS est seulement considérée du point de vue de la transmission et ne reflète en aucun cas le jugement de l'utilisateur final. C'est pourquoi nous intervenons dans ces problématiques en intégrant l'humain dans la boucle avec pour objectif d'augmenter la Qualité

3. plusieurs antennes à l'émission, plusieurs antennes à la réception

de l'Expérience (QoE :Quality of Experience).

4.2.1 Transmissions JPWL à travers un canal MIMO sous monitoring perceptuel

Le travail proposé consiste à tirer profit des domaines d'expertise de chacun des acteurs afin de fournir un schéma de transmission global avec une démarche conjointe dans des conditions d'utilisation réalistes. Pour ce faire, un environnement suburbain réaliste propice aux erreurs a été choisi et modélisé comme le montre la Figure 4.26-(a) où les bâtiments sont représentés en rouge. L'émetteur MIMO est fixe et le récepteur MIMO se déplace sur une distance de 180m à une vitesse de 5 m/s. Le système atteint un débit global de transmission de 24Mbits/s. La qualité de la transmission alterne successivement entre mauvaise (NLOS :Non Line Of Sight, signifiant qu'aucun trajet directe du signal n'est possible) dans la zone 1, moyenne (NLOS) dans les zones 2 et 4 ou bonne dans la zone 3. L'évolution du gain du canal MIMO est présentée sur la Figure 14-(b), où les réponses impulsionnelles sont fournies par un simulateur de canal basé sur la technique du tracé de rayons 3D [CPV05].



(a) (b)

Figure 4.26 : (a) Topologie de l'environnement de transmission et (b) évolution du gain du canal MIMO pour un utilisateur mobile

Dans les conditions les plus délicates, il est impossible d'assurer une transmission sans erreur de l'intégralité de l'image. Cependant, le but visé est d'obtenir la meilleure image possible du côté du récepteur. Pour ce faire, un schéma de transmission novateur a été proposé. L'idée est d'associer un codage JPWL (ISO/IEC 15444-11) [JPW91] permettant le découpage de l'image en couches

de qualité hiérarchisées [4] à une transmission MIMO permettant le découpage en sous-canaux SISO hiérarchisées associés à une stratégie d'allocation inégal de puissance des antennes (UPA :Unequal Power Allocation [SHJB05]). Ce choix est inspiré du fait qu'il a été prouvé que l'usage des systèmes multi-antennaires MIMO (Multiple Input Multiple Output) dans un environnement riche en multi-trajets, améliore significativement la fiabilité et/ou le débit de transmission, en comparaison avec des systèmes SISO (Single Input Single Output) [CRT01]. Ainsi, dans cette étude, nous nous concentrons sur la transmission d'images à travers un système MIMO en boucle fermée (CL-MIMO). Ce type de schéma exploite l'information sur l'état du canal (CSI pour Channel State Information) du côté de l'émetteur afin d'ajuster de manière dynamique la puissance d'émission sur chaque antenne, en prenant en compte l'état instantané du canal et l'importance du flux de données correspondant. Une stratégie d'adaptation dynamique des quantités de codes correcteurs d'erreur (UEP :Unequal Error Protection) est également appliquée afin d'adapter la quantité de code de redondance apportée à chaque couche, dans le but de garantir le meilleur compromis entre quantité d'information et robustesse aux erreurs. Cette redondance est gérée par des codes Reed-Solomon[JBZ95] (RS [5]) intégrés dans la norme JPWL. Une Modulation d'Amplitude en Quadrature [RW98] (MAQ) adaptative est également mise en place afin d'adapter au mieux la taille des symboles pour la transmission. La Figure 4.27 permet de visualiser le schéma de transmission proposé.

Sur ce schéma, nous pouvons visualiser les étapes de codage source, de codage canal, de modulation et de réception, le tout étant configuré de manière adaptative et optimisée en fonction de l'état instantané du canal.

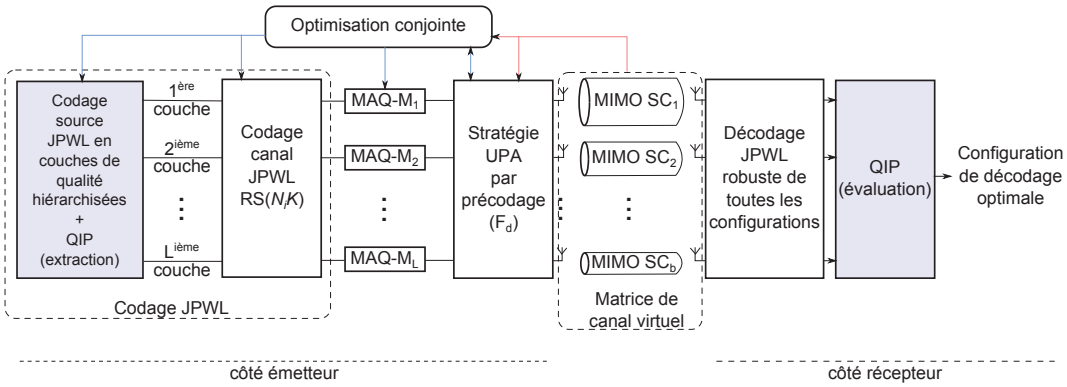


Figure 4.27 : Chaîne de transmission utilisée.

4. La première couche de qualité étant celle qui garantit la structure globale de l'image et les suivantes apportent des détails et de la netteté

5. Un code RS est défini par un nombre de blocs de symboles en entrée et un nombre de blocs de symboles en sortie notés respectivement K et N.

Bien que l'adaptation se fasse de manière optimale, il arrive que certaines couches de qualité arrivent avec des erreurs. Dans les implémentations classiques JPWL, en cas d'erreurs trop importantes, le décodage échoue ou se met à boucler. C'est là que le décodage robuste intervient afin de garantir dans tous les cas l'obtention d'une image en sortie. Cependant, en ajoutant une couche de qualité, normalement porteuse de détails, se sont également des dégradations qui sont ajoutées. L'idée est donc de disposer d'un outil capable de quantifier le gain en qualité perçue qu'apporte chaque nouvelle couche, afin de déterminer s'il est préférable de l'ajouter, ou de s'arrêter à la couche précédente. C'est là qu'intervient l'utilisation d'une métrique de prédiction de la qualité. Rappelons que la majorité des métriques de qualité sont à référence complète, or côté récepteur l'image d'origine n'est pas accessible. C'est pourquoi, le choix doit se porter sur une métrique de qualité sans référence ou à référence réduite. Pour rappel, l'intégration de la métrique QIP, présentée précédemment, dans la chaîne de transmission paraît comme la solution naturelle dans notre cas, car elle offre à la fois un faible temps de calculs et une taille de référence négligeable.

L'idée est donc d'utiliser QIP en tant qu'outil de monitoring capable de mesurer la qualité de chaque image décodée ayant chacune une couche de qualité supplémentaire. Ceci permettra de statuer sur la version de l'image offrant la meilleure qualité de l'expérience, sachant qu'il est possible que certaines couches ajoutent plus de défauts que de qualité. Afin de tester l'apport de cette métrique dans la chaîne, un protocole expérimental objectif a été mis en place.

4.2.2 Mesure de performance objective

Afin de quantifier l'apport de la brique d'estimation de la qualité perceptuelle dans le schéma de transmission, nous comparons la qualité des images décodées avec et sans utilisation de ce monitoring dans le contexte de simulation réaliste défini précédemment. La Figure 4.28 permet de visualiser, par un cercle, le PSNR d'une image décodée le long de la trajectoire. Les cercles pleins et rouges représentent les images décodées en utilisant QIP. Tandis que les cercles bleus sont pour les images sans QIP.

Sur ce graphique, nous pouvons visualiser l'impact de l'utilisation de QIP et son aptitude à maximiser le PSNR dans certains cas. Le Tableau 4.5 permet de quantifier de manière plus fine les résultats obtenus et de les comparer en terme de PSNR moyen.

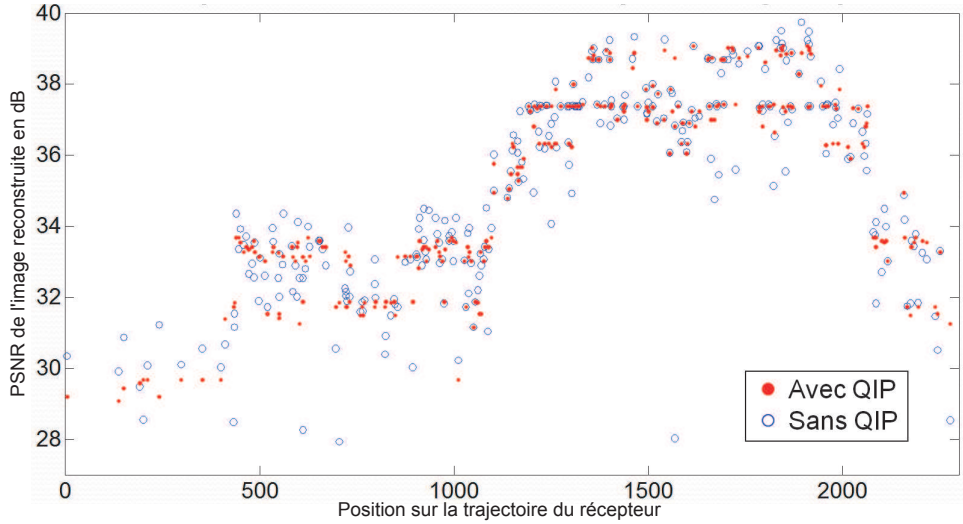


Figure 4.28 : Comparaison PSNR avec/sans QIP sur l'image "caps"

Images transmises	Canal « mauvais » Zone 1		Canal « moyen » Zone 2 et 4		Canal « bon » Zone 3	
	∅	QIP	∅	QIP	∅	QIP
Caps	**30,15dB**	29,99dB	32,76dB	**32,84dB**	37,26dB	**37,38dB**
House	25,67dB	**26,75dB**	28,43dB	**28,84dB**	31,61dB	**31,85dB**
Monarch	25,12dB	**25,52dB**	29,23dB	**29,53dB**	**34,78dB**	34,37dB

Tableau 4.5 : Comparaison des performances en terme de PSNR moyen

Par l'étude de ce tableau, nous pouvons constater que l'apport de QIP est notable dans toutes les conditions de transmission "moyenne", mais que l'apport en terme de PSNR peut paraître peu significatif. Cependant, il ne faut pas oublier que la métrique a décidé de conserver une couche de moins, mais que la qualité est tout de même maximisée. Il a été également observé que dans plusieurs cas, l'image ayant le meilleur PSNR n'est pas celle qui apparaissait visuellement la plus acceptable. C'est pourquoi une seconde métrique plus récente et reconnue comme plus performante que le PSNR en terme de corrélation avec le jugement humain, à savoir SSIM, a été utilisée et dont les résultats sont disponibles sur la Figure 4.29 et le Tableau 4.6.

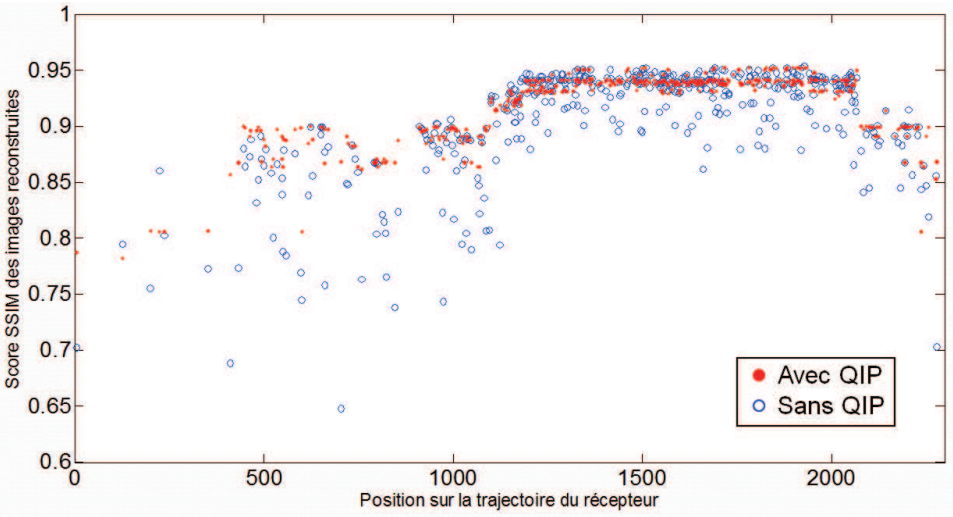


Figure 4.29 : Comparaison SSIM avec/sans QIP sur image Caps

Images transmises	Canal « mauvais » Zone 1		Canal « moyen » Zone 2 et 4		Canal « bon » Zone 3	
	∅	QIP	∅	QIP	∅	QIP
Caps	0,769	**0,815**	0,857	**0,887**	0,931	**0,938**
House	0,593	**0,653**	0,727	**0,748**	0,836	**0,852**
Monarch	0,789	**0,813**	0,865	**0,887**	0,929	**0,937**

Tableau 4.6 : Comparaison des performances en termes de SSIM moyen

Avec la métrique SSIM, l'intérêt de QIP est plus facilement interprétable, en fournissant des gains en qualité plus marqués.

Cette expérimentation a permis de prouver l'intérêt de l'approche proposée à travers l'augmentation du PSNR ou du SSIM, et surtout d'éviter de décoder une couche contenant des erreurs. Cependant, rien ne prouve réellement que l'augmentation du PSNR reflète réellement un gain de qualité de l'expérience.

Nous ne savons pas davantage en quoi un gain de quelques dB influe effectivement sur la qualité perçue. De plus, il arrive que notre algorithme soit en conflit avec les résultats PSNR, à savoir que notre métrique informe d'une augmentation de qualité alors que le PSNR diminue. Afin de donner des réponses à ces questions, nous avons décidé de mettre en place un protocole expérimental pour la mesure de la qualité subjective.

4.2.3 Validation de la méthode par expérimentation subjective

Afin de mesurer l'augmentation de la qualité d'expérience, nous avons eu recours à des tests subjectifs. Par ce biais, il devrait être possible de vérifier que l'idée initiale, qui consiste à considérer, qu'il peut être préférable de ne pas ajouter une couche de qualité à l'image si elle contient trop d'erreurs de transmission, peut permettre une augmentation de la qualité perçue. De plus, il sera également possible d'estimer dans quelle mesure notre métrique QIP fournit des réponses en adéquation avec le jugement humain.

Environnement d'évaluation

La campagne d'évaluation subjective s'est déroulée dans une salle dédiée construite selon les recommandations de l'ITU [ITU08], équipée d'un écran LCD 30 pouces avec une résolution native de 2560×1600 pixels. Le rapport de la luminance de l'écran inactif et de la luminance crête a été maintenu au dessous d'une valeur de 0,02. L'éclairage était assuré par 4 néons offrant une lumière D50, contrôlés et orientés pour obtenir 64 lux à l'écran tout en évitant un éclairage direct. Le calibrage de l'écran a été réalisé en utilisant un dispositif d'étalonnage EyeOne de Gretag Macbeth. La distance d'observation a été maintenue entre 2 et 4 fois la hauteur des images présentées.

Base d'images

Pour les besoins de simulation et de test, nous avons utilisé trois images de la littérature à savoir Caps, House et Monarch données par la Figure 4.30. Le nombre faible d'images utilisées s'explique par la nature de l'expérience générant des milliers de résultats par image d'entrée (comme décrit en section 4.2.3), ce qui est incompatible avec des tests psychophysiques par nature déjà très chronophages.

Caps

House

Monarch

Figure 4.30 : Images de références utilisées

Bien que n'utilisant que trois images de référence, c'est au final un ensemble de 1512 images qui a été utilisé pour cette expérimentation et ce afin de refléter les différents cas de figure se produisant durant les diverses réceptions dans le contexte de mobilité dans un environnement réaliste.

Panel d'observateurs

Dix-sept observateurs (non-experts de la thématique) ont participé à la campagne d'évaluation. Ce chiffre est conforme à l'exigence minimale mentionnée dans [ITU08] après le rejet des observateurs/observations aberrants.

Le panel de participants est constitué de 7 femmes et 10 hommes, pour la plupart des étudiants, présentant une moyenne d'âge de 25 ans. Sachant que les jeunes représentent la majorité des utilisateurs des nouvelles technologies, il est souvent recommandé par les différentes instances d'évaluation d'avoir un panel relativement jeune hormis pour quelques études spécifiques.

Présentation des images

L'expérience psychovisuelle s'est déroulée en trois étapes :

- Tout d'abord, une collecte des informations personnelles de l'observateur, accompagnée de différents tests permettant de vérifier l'acuité visuelle et la vision des couleurs (test d'Ishihara).
- Par la suite, le test est expliqué aux observateurs afin de comprendre sa finalité. Ainsi, un discours unique est lu lors de la phase d'entrainement, indispensable pour appréhender le test. Les participants sont invités à poser toutes les questions en cas de besoin avant de commencer l'expérience réelle.
- Enfin, l'évaluation proprement dite consiste à faire évaluer 765 couples d'images, dans un ordre totalement aléatoire aux participants (pour limiter le biais de passage). La présentation est faite selon un protocole strict. Les images sont affichées par couple comme le montre les captures d'écran de la Figure 4.31.

Concrètement, il est demandé à l'observateur de comparer deux images placées l'une à côté de l'autre et de donner son niveau de préférence pour l'une ou l'autre. Il lui est expliqué que l'une ou l'autre peut arriver sur son smartphone ou sa tablette, et bien qu'il soit possible qu'aucune des deux ne soient parfaite, il doit choisir laquelle des deux il aurait préféré recevoir. Par exemple, une des images est issue du décodage de la première et la deuxième couches, qu'il doit comparer au décodage des trois premières couches de la même image de référence, mais dont des erreurs de transmissions ont pu perturber le contenu de la dernière couche. Bien évidemment, la position (droite ou gauche) des images est aléatoire (afin d'éviter un phénomène d'habitude) et l'observateur ne dispose pas des informations concernant le nombre de couches et l'existence d'erreurs de transmission. Le temps d'évaluation de chaque paire d'images n'est pas fixé. Cependant, afin d'éviter la fatigue visuelle, l'application comporte un compteur obligeant l'observateur à s'arrêter toutes les vingt minutes. Pour donner son avis, son ressenti, la notation se fait sur une échelle discrète composée de deux parties (une pour chaque image). Si les deux images ont une qualité similaire, l'observateur peut choisir le qualificatif «équivalente». Dans le cas où une des images présente une meilleure qualité, trois graduations sont possibles : «légèrement meilleure», «meilleure» et «bien meilleure». La Figure 4.31 permet de visualiser l'application ayant été développée pour ce test ainsi que différents cas de figure pouvant se présenter.

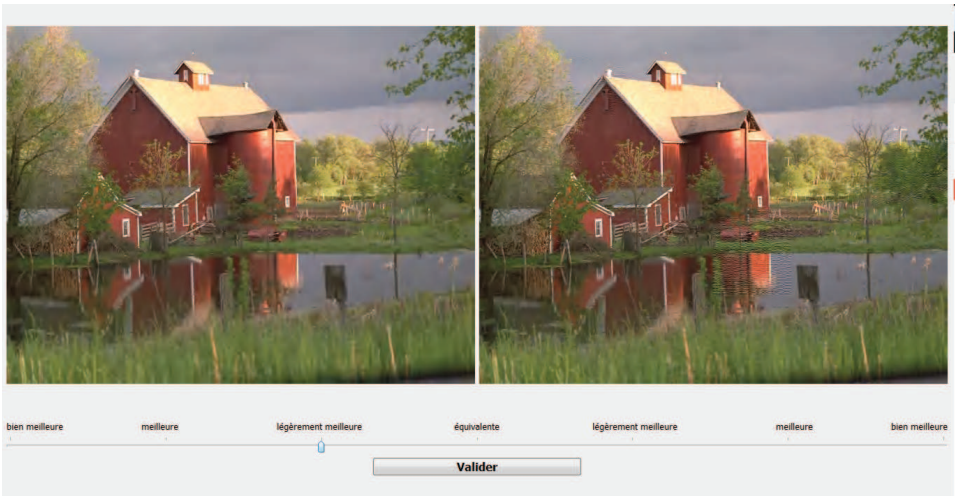


Image 2 couches (à gauche) jugée "légèrement meilleure" à la version à 3 couches

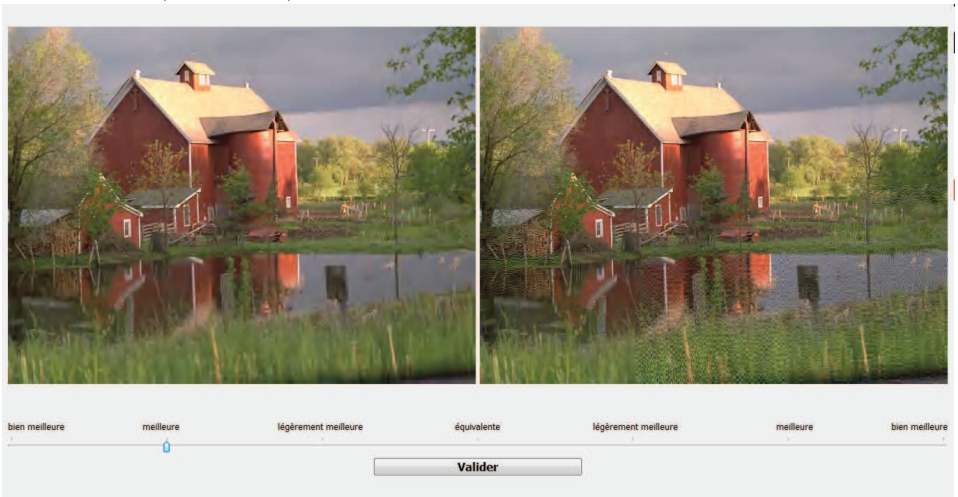


Image 2 couches (à gauche) jugée "meilleure" à la version à 3 couches

Image 2 couches (à gauche) jugée équivalente à la version 3 couches

Figure 4.31 : Interface d'évaluation subjective

Résultats et discussion

Dans cette section, nous présentons l'ensemble des résultats issus de la campagne d'évaluation subjective.

Le tableau 4.7 et sa représentation graphique sur la Figure 4.32 présentent les résultats de manière condensée et synthétique. Pour ce faire un point est donné à la métrique QIP quand elle est en accord avec le choix humain (et ce indépendamment du niveau qualitatif de préférence de l'humain allant de légèrement à bien meilleure). Dans le cas contraire, c'est le cas sans QIP qui se voit attribuer le point. Un troisième cas se présente quand l'humain juge que les deux images sont identiques, ce qui peut finalement être sommé par la suite avec l'un ou l'autre des cas précédents.

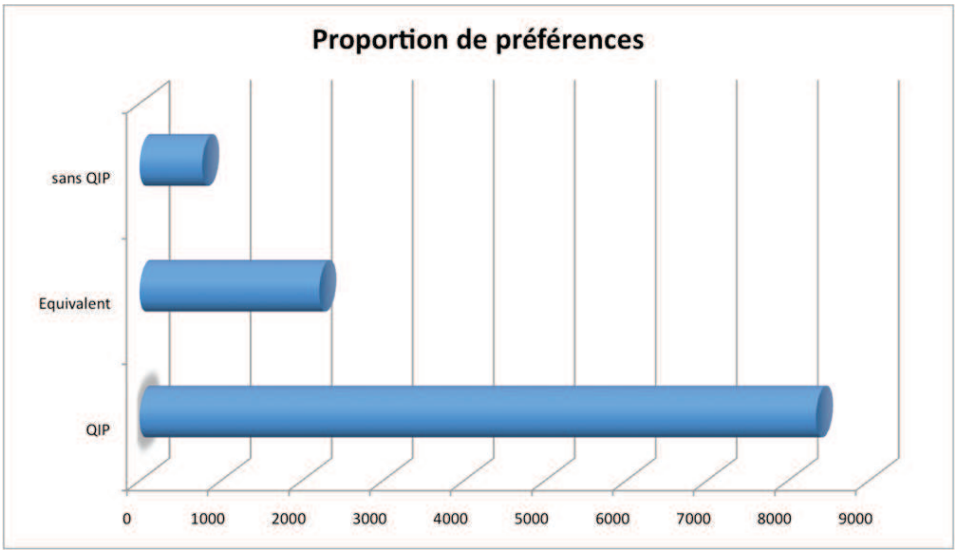


Figure 4.32 : Scores fusionnés pour toutes les images et tous les observateurs

	QIP	Equivalent	Sans QIP
Occurrence	8349	2218	773
%	73,63%	19,55 %	6,82 %

Tableau 4.7 : Scores fusionnés pour toutes les images et tous les observateurs

De ces mesures, nous remarquons clairement, que dans la majorité des cas (74% (QIP)+20% (équivalent) = 94%), les observateurs sont allés dans le sens de la décision prise par QIP. Ceci confirme que la métrique exploitée dans cette chaîne de simulation est capable d'estimer l'expérience de l'utilisateur et de donner la meilleure configuration possible de décodage de l'image.

Afin d'analyser plus finement ces résultats globaux, nous représentons dans la Figure 4.33 les scores des observateurs en donnant les différents qualificatifs de l'échelle de notation. QIP- (resp. sans QIP -) correspond au qualificatif «légèrement meilleur», QIP (resp. sans QIP) correspond au qualificatif «meilleur» et enfin QIP+ (resp. sans QIP +) correspond à «bien meilleur».

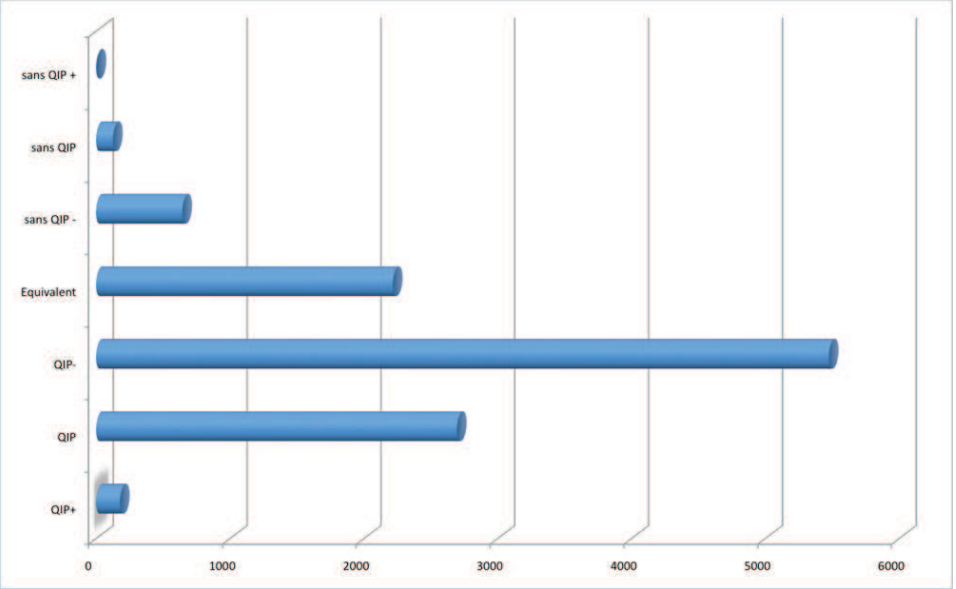


Figure 4.33 : Scores détaillés pour toutes les images et tous les observateurs

	QIP+	QIP	QIP-	Equivalent	Sans QIP-	Sans QIP	Sans QIP+
Occ.	178	2696	5475	2218	644	126	3
%	1,57%	23,78%	48,28%	19,55%	5,68%	1,11%	0,03%

Tableau 4.8 : Scores détaillés pour toutes les images et tous les observateurs

Il est possible de remarquer que le qualificatif le plus utilisé est celui « légèrement meilleur » pour QIP. Pour rappel, les choix fait par QIP consiste à favoriser une image avec moins de détails, car exploitant une couche de qualité de moins, mais ne possédant pas d'artéfacts dus aux erreurs de transmission. L'humain semble en accord avec ce principe, et préférera une image avec moins de détails mais sans erreur localisée.

Afin de mieux comprendre le processus, nous détaillons les résultats par image et par couche. En effet, les résultats globaux permettent de comprendre les tendances mais ne relatent ni l'influence du contenu de l'image, ni la variation entre les paires de couches de qualité.

Analyse détaillée sur l'image Caps

Les résultats globaux pour l'image Caps sont donnés par le tableau 4.9. Un comportement similaire au comportement global peut être retrouvé dans le cas de cette image. En effet, le cas prédominant est celui de QIP-. De plus le nombre de cas en faveur de QIP approche les 90%.

	QIP+	QIP	QIP-	Equivalent	Sans QIP-	Sans QIP	Sans QIP+
Occ.	20	441	1474	858	299	57	1
%	0,63%	14,00%	46,79%	27,24%	9,49%	1,81 %	0,03%

Tableau 4.9 : Scores détaillés pour l'image Caps

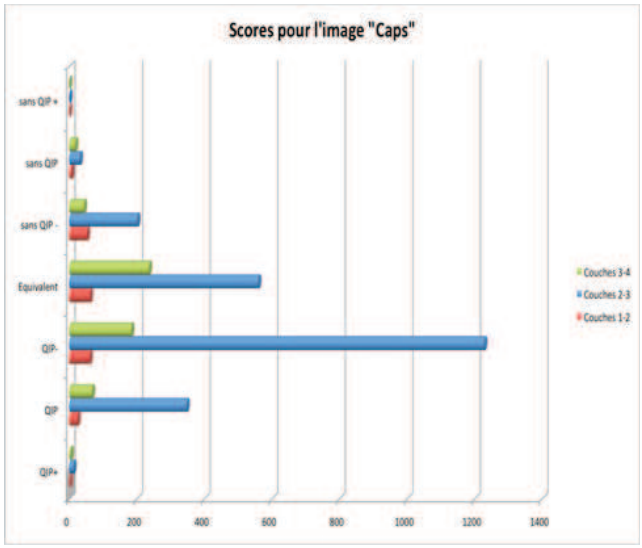
La Figure 4.34 présente les résultats par paire de couches. Il est à rappeler que l'observateur avait à choisir entre deux images ayant une couche de qualité de différence. Il est possible de remarquer que le nombre de cas ayant été présentés à l'observateur est plus souvent entre la couche 2 et la couche 3. En effet, c'est à cette étape que les différences sont les plus notables. Comme mentionné précédemment, la couche 1 est celle qui va apporter la structure de l'image et la couche 2 apporte un niveau de texture grossière. La couche 2 est plus exploitable que la couche 1. Cela explique la répartition uniforme des scores de part et d'autre du qualificatif « équivalent ».

Entre la couche 3 et la couche 4, les différences deviennent plus difficiles à percevoir. C'est pourquoi, le nombre de cas équivalents est le plus important proportionnellement. Ceci s'explique également par le contenu de l'image qui est assez sommaire. L'image est caractérisée par des zones uniformes multiples et le contenu est concentré au niveau des casquettes et plus particulièrement sur le texte brodé sur ces dernières.

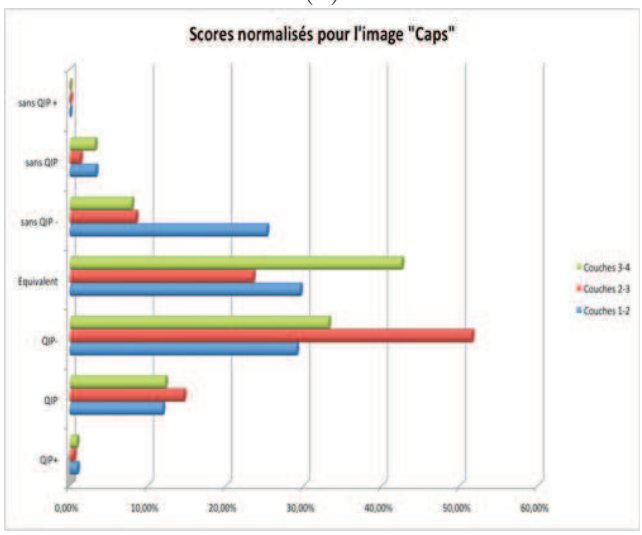
Analyse détaillée sur l'image House

Dans cette section, nous allons discuter des résultats pour l'image « House ». Sur cette image, plusieurs zones sont susceptibles de présenter une gêne pour l'observateur si elles contiennent des erreurs de transmission. De la Figure 4.35 et du tableau 4.10, nous pouvons remarquer que la proportion de votes allant dans le sens de QIP est proche de 95%. Nous pouvons conclure que les observateurs préfèrent avoir une image avec moins de détails plutôt que d'avoir une image avec des patchs d'erreur. Comme pour l'image précédente, le cas prépondérant est celui de QIP mais cette fois-ci avec une proportion très forte du cas QIP correspondant à presque 30% du jugement des observateurs.

La présentation des résultats par couche confirme le constat global et nous pouvons ainsi remarquer que pour les 3 histogrammes de la Figure 4.37-(b), le taux des QIP est plus important pour l'image « House » que pour les autres



(a)



(b)

Figure 4.34 : Scores par couche : a) nombre d'occurrences et b) scores normalisés sur l'image Caps

	QIP+	QIP	QIP-	Equivalent	Sans QIP-	Sans QIP	Sans QIP+
Occ.	84	1009	1804	479	147	29	2
%	2,36%	28,39%	50,76%	13,48%	4,14%	0,82%	0,06%

Tableau 4.10 : Scores détaillé pour l'image House

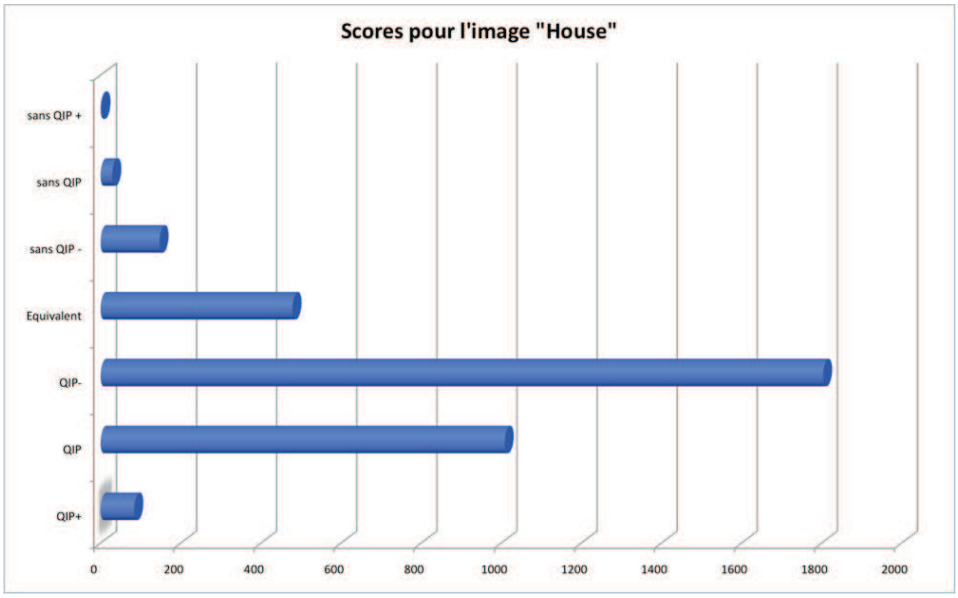


Figure 4.35 : Scores détaillés pour l'image House

images.

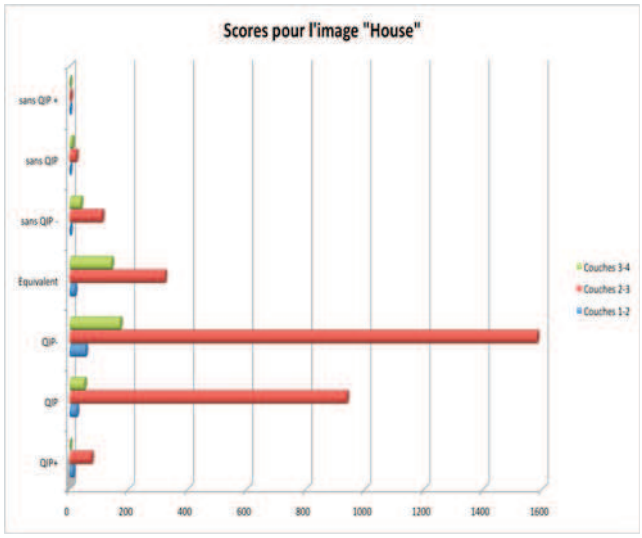
Analyse détaillée sur l'image Monarch

L'image « Monarch » est proche en terme de construction de l'image « House » puisqu'elle est constituée d'un objet d'intérêt centré sur un fond uniforme/flou. Cela veut dire qu'une dégradation sur l'objet d'intérêt va susciter immanquablement le rejet de l'observateur. Le tableau 4.11 confirme ce constat puisque 95% des votes sont dans le sens de la métrique QIP.

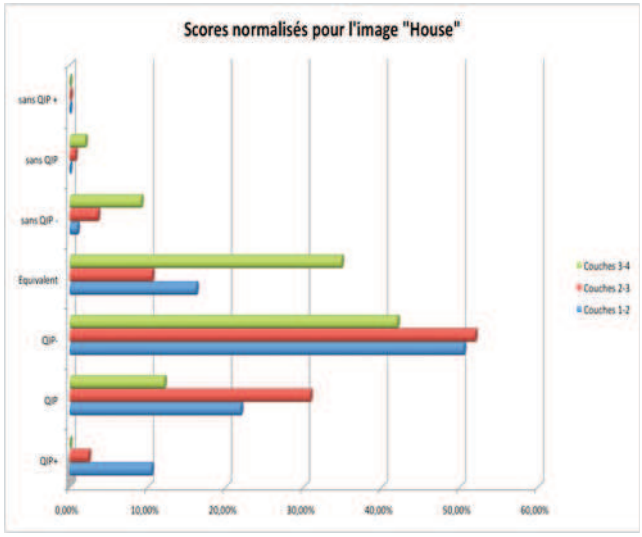
	QIP+	QIP	QIP-	Equivalent	Sans QIP-	Sans QIP	Sans QIP+
Occ.	74	1246	2197	880	198	40	0
%	1,60%	26,88%	47,40%	18,99%	4,27%	0,86%	0,00%

Tableau 4.11 : Scores détaillés pour l'image Monarch

Par cette expérience nous avons également mis en avant le fait que la localisation des erreurs a une importance dans la perception de la qualité. En effet, dans le cas où l'image contient un élément d'intérêt, tel que pour l'image "Monarch", le moindre patch d'erreur sur cette zone sera sanctionné durement. L'humain préférera une image avec moins de détails mais sans erreur localisée. Ceci tend à dire qu'utiliser une carte de saillance et éventuellement la métrique QIP-HSM pourrait améliorer les résultats obtenus. Cependant, le prix de cette augmentation de corrélation avec le jugement humain se verra sur la taille de la référence à transmettre.

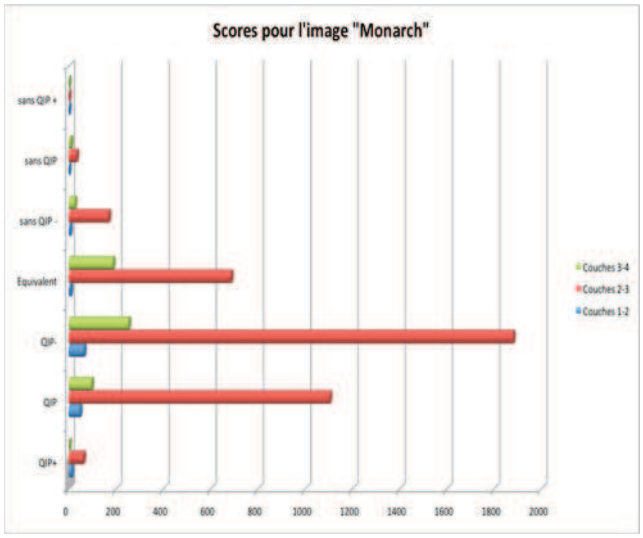


(a)

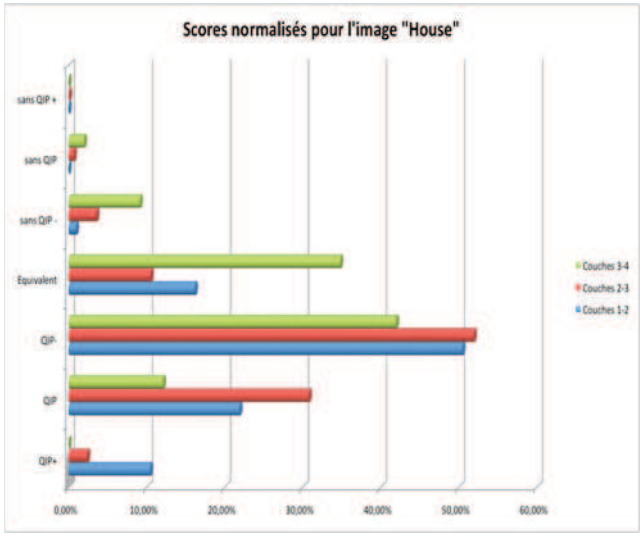


(b)

Figure 4.36 : Scores par couche : a) nombre d'occurrences et b) scores normalisés sur l'image House



(a)



(b)

Figure 4.37 : Scores par couche : a) nombre d'occurrences et b) scores normalisés sur l'image Monarch

4.3 Conclusion

Dans ce chapitre, nous avons démontré les capacités d'un détecteur de points d'intérêt pour estimer et quantifier les artéfacts dans les images. Ce travail a donné naissance à trois métriques de qualité exploitant plus ou moins de référence de l'image. De ce fait, une des métriques proposées a été introduite dans un schéma de transmission sans fil, dans le but de maximiser la qualité de l'expérience. Diverses méthodes de mesure de performances ont été mises en œuvre afin de prouver la validité de l'approche proposée. La validation ayant le plus de poids est celle utilisant l'humain comme référence. Elle a montré que plus de 90% des votes vont dans le sens de la métrique qui recommande d'arrêter le décodage à la couche précédente afin d'éviter les erreurs de transmission.

Ce travail permet d'amorcer la prise en compte de la qualité de l'expérience (et donc un passage de la QoS à la QoE) dans les chaînes de transmission sans fil pour un coût relativement faible puisque la métrique QIP ne nécessite que 22 octets pour mesurer la dégradation engendrée par la transmission.

Chapitre 5

Modélisation de l'évolution des statistiques structurelles de l'image

Sommaire

5.1 Migration des propriétés structurelles **171**
5.1.1 Extraction d'attributs structurels des pixels 172
5.1.2 Classification des attributs structurels 174
5.1.3 Mesure et modélisation des migrations par graphe multi-étiqueté 180
5.2 Sensibilité des statistiques aux dégradations **184**
5.2.1 Prédiction de la qualité perçue 188
5.2.2 Estimation du facteur q de la compression JPEG . 193
5.3 Conclusion **209**

Au cours des précédentes parties, nous avons démontré les performances des détecteurs de points d'intérêt dans le cadre de la prédiction de zones saillantes mais également pour l'estimation de la qualité perçue. Tous ces détecteurs exploitent des statistiques structurelles de l'image. C'est pourquoi, dans ce chapitre nous approfondissons notre démarche, en étudiant et modélisant explicitement la sensibilité et l'évolution des statistiques structurelles de l'image pour différentes dégradations tout en les comparant avec la sensibilité humaine.

5.1 Migration des propriétés structurelles

Notre objectif est de modéliser l'évolution des statistiques structurelles des images à différents types et magnitudes de dégradations. L'idée sous-jacente

est de considérer que les dégradations classiques, telles que celles issues de la compression ou la transmission introduisent des changements structurels dans l'image. Typiquement un pixel appartenant à une région uniforme dans une image sans dégradation, peut devenir un pixel de contour, s'il se situe au bord d'un effet de bloc introduit par une compression JPEG. A l'inverse, un pixel de contour dans l'image d'origine peut devenir un pixel de région uniforme, suite à une compression JPEG 2000 qui a tendance à lisser certains contours. Ce changement de propriétés de structure semble particulièrement identifiable et perceptible par l'humain. En effet, la reconnaissance de contours, de textures, de régions uniformes, de coins sont les clefs de traitements haut-niveaux dans les processus bottom-up (cf. section 2.2.1) telle que la reconnaissance de formes et d'objets. De ces *a priori* nous pensons qu'un modèle basé sur des évolutions structurelles peut être pertinent dans nos problématiques.

L'approche que nous proposons peut être résumée et décomposée en 3 grandes étapes dont une illustration est proposée sur la Figure 5.1 :

- Extraire les attributs structurels des pixels de l'image d'origine et de sa version dégradée
- Affecter chaque attribut à une classe (bord, contour horizontal, contour vertical, coin, région uniforme,...)
- Quantifier les migrations de classes entre la version avec et sans dégradation

5.1.1 Extraction d'attributs structurels des pixels

La méthode choisie pour extraire les statistiques structurelles de l'image repose sur des mesures d'auto-corrélation par fenêtre glissante. Pour chaque pixel de l'image, on considère une fenêtre d'observation que l'on déplace légèrement dans plusieurs directions. Le but est d'analyser le voisinage du pixel considéré dans ces directions, afin de déterminer le changement moyen d'intensité. S'il n'y a aucun changement d'intensité, dans aucune direction, le pixel considéré appartient à une région uniforme. Au contraire, s'il y a de forts changements dans une ou plusieurs directions, le pixel appartiendra à un contour ou à un coin. (Cette approche d'extraction d'informations structurelles a été introduite par Moravec [Mor80] et améliorée par Harris [HS88b] dans les années 1980).

Donnons maintenant une formulation mathématique à ce concept. Considérons I la composante intensité de notre image et w la fenêtre glissante. La variation mesurée E pour un pixel en (x, y) produite par un décalage de fenêtre de u, v est donnée par l'équation 5.1.

$$E_(x, y) = \sum_{u,v} w_{u,v}[I_{x+u,y+v} - I_{u,v}]^2. \tag{5.1}$$

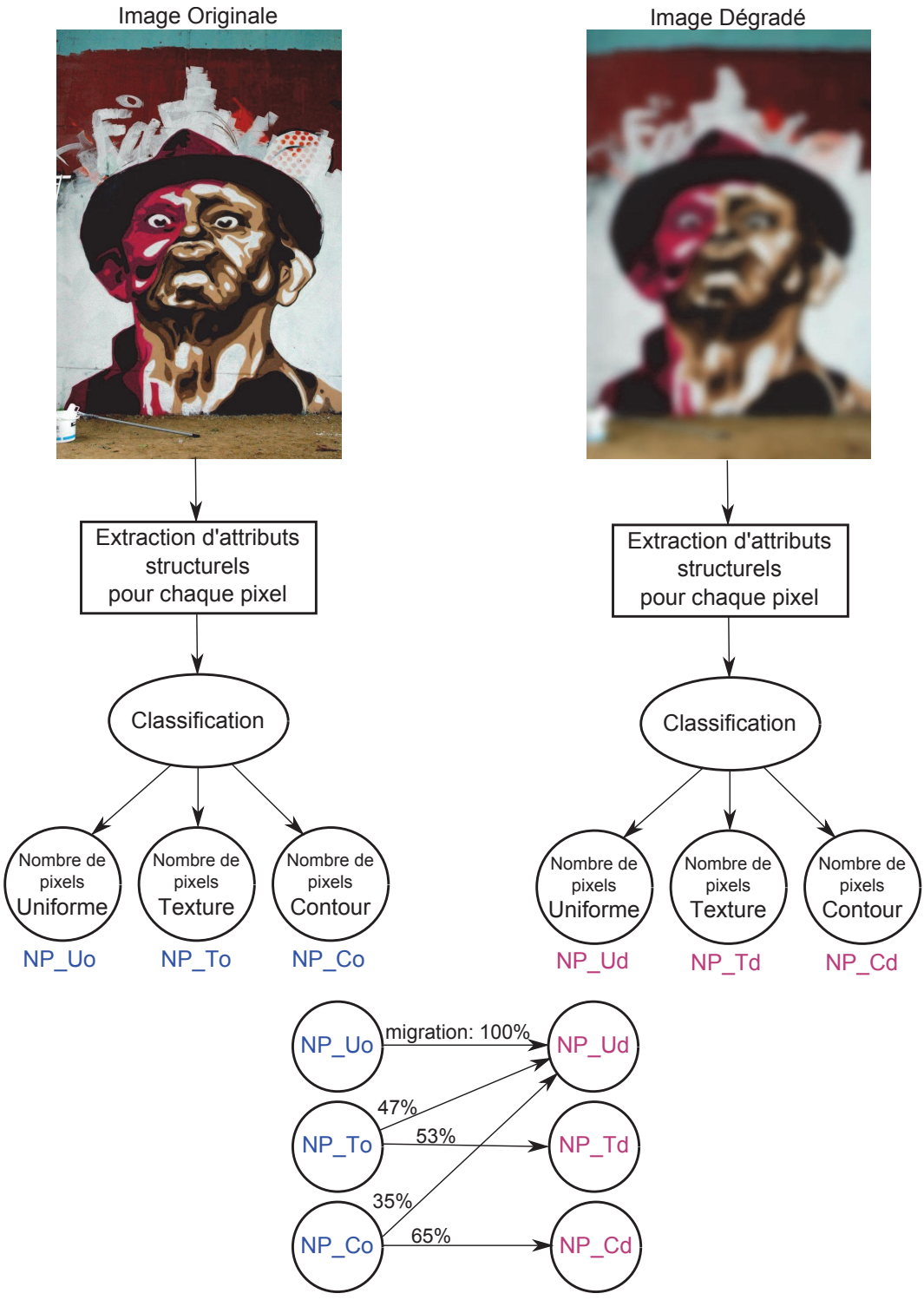


Figure 5.1 : Mesure de migrations des propriétés structurelles. Dans cet exemple, un flou gaussien dégrade l'image, ce qui a un effet sur les propriétés structurelles des pixels. Une importante quantité de pixels contours et textures sont devenus des pixels de régions uniformes.

On peut noter que cette version est discrète et ne peut donc considérer que des directions multiples de 45 degrés. Afin de couvrir la totalité des directions, pour de faibles déplacements, l'extension analytique donnée par l'équation 5.2 peut être utilisée :

$$E_(x,y) = \sum_{u,v} w_{u,v}[xX + yY + O(x^2, y^2)]^2, \tag{5.2}$$

où la dérivée première (gradient) de X et Y est approximée par convolution et où $O(.)$ étant le reste négligeable de l'approximation :

$$X = I * \begin{bmatrix} -1 & 0 & 1 \end{bmatrix} \approx \frac{\partial I}{\partial x}(x)$$
$$Y = I * \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \approx \frac{\partial I}{\partial y}(y)$$

La taille et la forme de la fenêtre w utilisée influence les résultats. Harris a par exemple noté que l'utilisation d'une fenêtre carrée et binaire rend l'opération sujette au bruit. C'est pourquoi, elle est remplacée par une fenêtre gaussienne qui a pour effet de lisser les pixels et donc de minimiser cette sensibilité.

E est intimement lié à la fonction d'auto-corrélation avec pour tenseur de structure la matrice M, capable de décrire la structure locale du pixel considéré. De ce tenseur M, on note λ_1 et λ_2 ses valeurs propres, descriptive et proportionnelle de la courbure de la fonction d'auto-corrélation. Pour chaque pixel $P_{(x,y)} \in I$, un couple (λ_1, λ_2) décrivant sa structure, de son niveau d'appartenance à une région uniforme, contour ou coin.

5.1.2 Classification des attributs structurels

Grâce à la méthode présentée précédemment, chaque pixel peut être décrit par son couple (λ_1, λ_2). En fonction des valeurs conjointes que prennent chacun des λ, il est possible d'associer chaque pixel à une grande classe structurelle. La Figure 5.2 permet d'illustrer les 4 grandes classes relatives aux valeurs de couples (λ_1, λ_2) sur un repère orthonormé, que nous nommerons pour la suite $diagramme_\lambda$. On visualise donc, que pour deux faibles valeurs de λ, le pixel considéré doit appartenir à une région uniforme. Dans le cas où seulement un

des λ a une forte valeur, le pixel appartiendra forcément à un contour (horizontal ou vertical). Enfin, dans le cas particulier où les deux λ sont importants, le pixel doit appartenir à un coin.

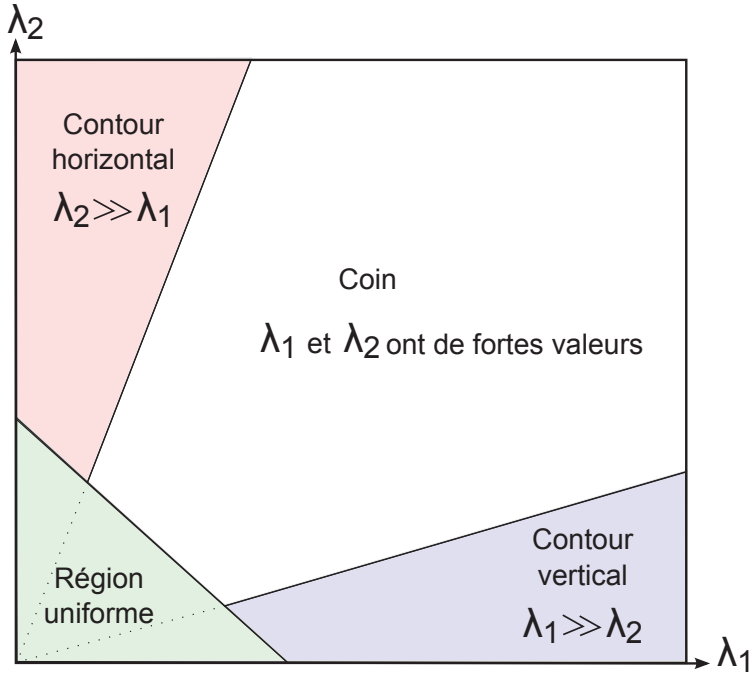


Figure 5.2 : Classification des configurations de couple (λ_1, λ_2)

La Figure 5.3 permet d'illustrer sur un cas concret une classification des pixels en 4 classes, où les pixels noirs appartiennent à une région uniforme, les verts aux contours horizontaux, les bleus aux contours verticaux et les rouges aux coins.

Par cette représentation, on distingue quatre grandes catégories. Dans la démarche proposée, nous souhaitons quantifier les pixels ayant changé de classe ou en d'autres termes ayant migré. Le nombre et la répartition des différentes classes a une influence sur la sensibilité des mesures de migration. En effet, plus le nombre de classes est important, et partitionne donc le $diagramme_\lambda$ de manière fine, plus la quantité de migrations observables augmente. Il est question ici d'observations calculatoires, mais il serait également intéressant de faire le parallèle avec les changements observables par l'humain. L'idée sous-jacente serait de réussir à trouver une manière optimale de subdivision du $diagramme_\lambda$ avec comme critère la maximisation de la corrélation entre les changements observables par l'humain et les migrations observables d'une façon calculatoire.

Afin de décrire au mieux des subdivisions de l'espace et dans un souci de

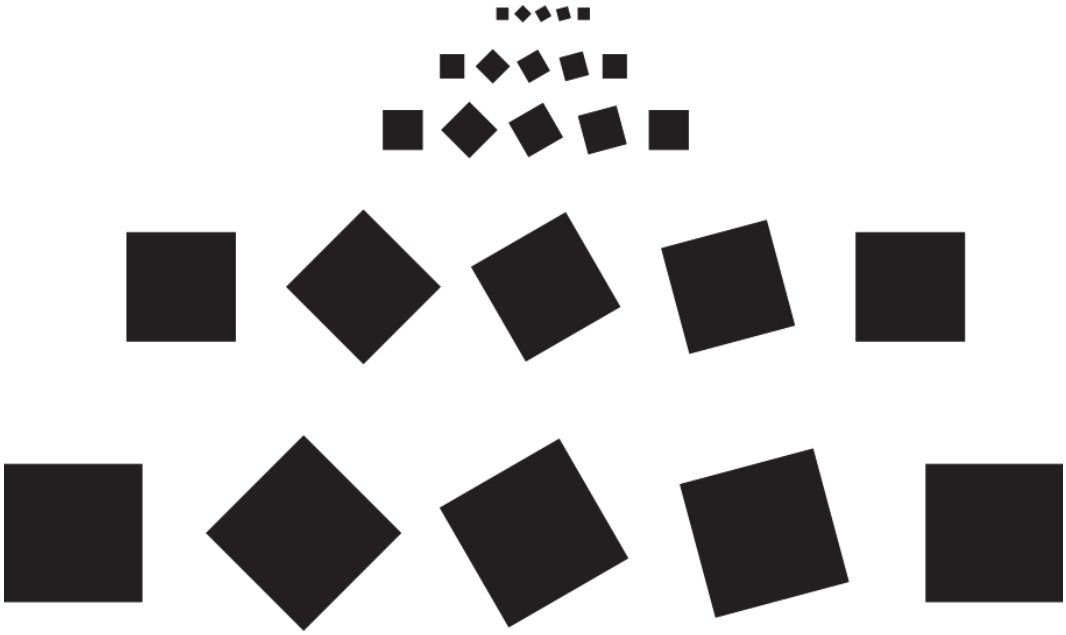
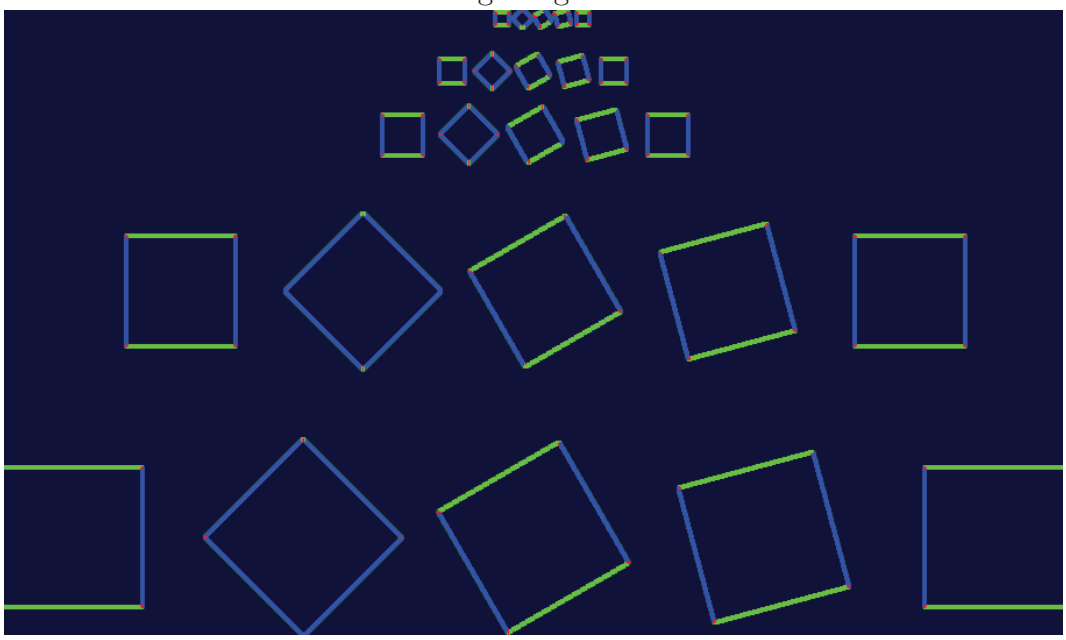


Image originale.



Résultats de classification.

Figure 5.3 : Résultats de classification avec 4 classes : zone uniformes, contours verticaux et horizontaux et coins.

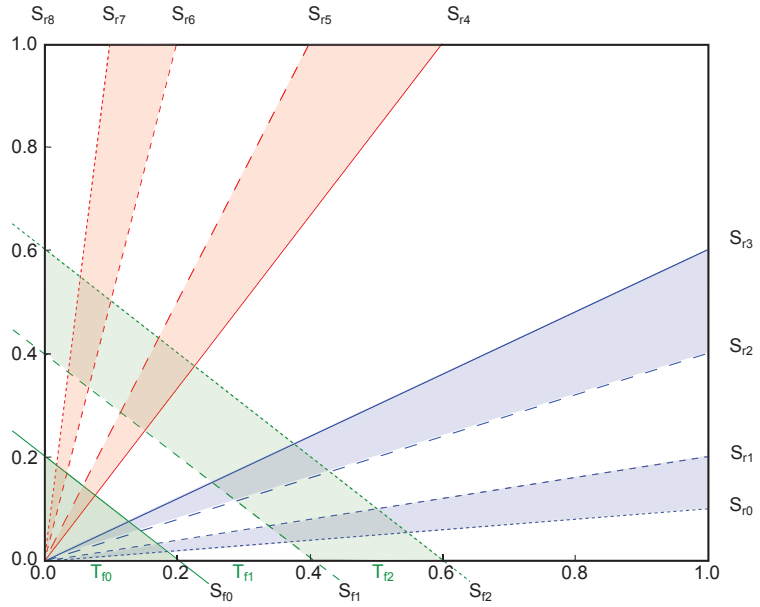


Figure 5.4 : Division du $diagramme_\lambda$ par des droites.

généricité nous noterons :

- Nsr : le nombre de subdivisions radiales. Ce nombre de subdivisions permet de régler la sensibilité à des changements de structures allant de contour horizontal à vertical (en passant par la classe coin).
- Nsf : le nombre de subdivisions de force. Ce nombre de subdivisions permet de régler la sensibilité à des changements de force (d'intensité) de structure, allant de région uniforme à contour très marqué.
- Sr_i : L'équation de fonction d'indice i permettant la subdivision radiale du $diagramme_\lambda$, avec $\{i \in \mathbb{R}; 0 \leq i < Nsr\ \}$
- Sf_j : L'équation de fonction d'indice j permettant la subdivision par force du $diagramme_\lambda$, avec $\{j \in \mathbb{R}; 0 \leq i < Nsf\ \}$
- Ar_i : L'aire radiale $_i$. Avec $Ar_i = \int_0^1 Sr_i \, \mathrm{d}\lambda_1 - \int_0^1 Sr_{i-1} \, \mathrm{d}\lambda_1$
- Af_j : L'aire de force $_j$. Avec $Af_j = \int_0^1 Sf_j \, \mathrm{d}\lambda_1 - \int_0^1 Sf_{j-1} \, \mathrm{d}\lambda_1$
- Ar : l'ensemble des Ar_i
- Af : l'ensemble des Af_i

La Figure 5.4 illustre un partitionnement de l'espace pour $Nsr = 9$ et $Nsf = 4$ avec :

$$\begin{aligned} Sr_0 &= 0.1 * \lambda_1 & Sr_4 &= 10 * \lambda_1 \\ Sr_1 &= 0.2 * \lambda_1 & Sr_5 &= 5 * \lambda_1 \\ Sr_2 &= 0.4 * \lambda_1 & Sr_6 &= 2.5 * \lambda_1 \\ Sr_3 &= 0.6 * \lambda_1 & Sr_7 &= (1/6) * \lambda_1 \end{aligned}$$

et

$$\begin{aligned} Sf_0 &= -\lambda_1 + 0.2 \\ Sf_1 &= -\lambda_1 + 0.4 \\ Sf_2 &= -\lambda_1 + 0.6 \end{aligned}$$

A noter que ce partitionnement est entièrement défini par des équations de droites linéaires. Cependant, on peut remarquer que le découpage par "force" de manière linéaire rend ce dernier anisotrope. La Figure 5.5 illustre donc un partitionnement par arcs de cercles, afin de garantir l'isotropie à la valeur radiale.

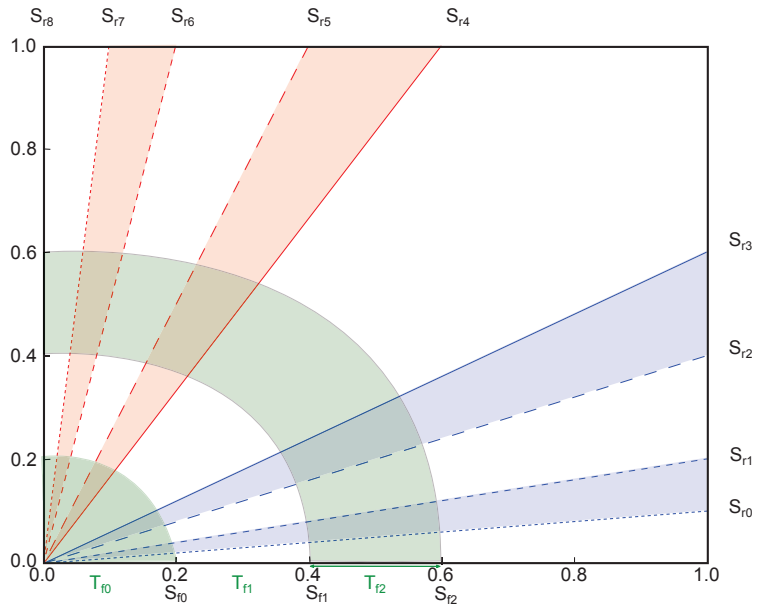


Figure 5.5 : Division du $diagramme_\lambda$ par équation de courbe.

L'espace étant complètement partitionné, il est possible d'affecter pour tout pixel une seule et unique classe d'appartenance à une aire Ar_i et une aire Af_j. A titre d'exemple, un pixel décrit par $\lambda_1 = 0.8$ et $\lambda_2 = 0.8$ est inscrit dans l'aire Ar_4 et l'aire Af_3. Soit $\forall\ P_{I(x,y)} \in I\ \exists! Ar_i \in Ar, P_{I(x,y)} \in Ar_i$ et de la même manière $\forall\ P_{I(x,y)} \in I\ \exists!\ Af_j \in Af, P_{I(x,y)} \in Af_j$

Critères de partitionnement

Bien que la sensibilité d'observation des migrations puisse être en partie fixée par le nombre de partitions Nsf et Nsr, les valeurs des attributs des équations Sf_j et Sr_i sont extrêmement influentes.

Il existe de nombreuses méthodes et d'algorithmes capables d'affecter des valeurs à ces attributs. Car si nous reformulons ce problème, il s'agit de résoudre une problématique de partitionnement d'un espace bidimensionnel en le subdivisant afin de regrouper les éléments de composition identique ou similaire. Ces problématiques sont courantes dans de nombreux domaines et bien au-delà de l'analyse d'images. Cependant, on peut catégoriser ces partitionnements en 2 grandes familles : les fixes et les adaptatifs.

La méthode de partitionnement fixe, consiste à utiliser les mêmes équations de partitionnement de manière indépendante et invariante à l'image considérée. Ces équations sont généralement fixées *a priori*, par l'utilisation d'observations expertes, ou bien par le biais d'algorithmes d'optimisation suite à des phases d'apprentissage. L'idée sous-jacente est de déterminer le partitionnement qui maximise l'invariance à l'image traitée. Dans le cadre de notre démarche, il peut être envisageable que l'œil et le cerveau humain ont des seuils fixes de sensibilité perceptible de changement de structures et ce peu importe l'image visualisée. Cette sensibilité invariante aux images peut sembler surprenante et en contradiction avec un effet connu et largement utilisé en perception de la qualité qui est l'effet de masquage. Ce phénomène est basé sur le principe qu'une distorsion dans l'image est plus ou moins visible en fonction de sa localisation ou plutôt des stimuli qui l'entourent. Par exemple un bruit gaussien localisé dans une région uniforme est nettement plus visible que s'il est au milieu d'une texture. De ce constat, la sensibilité aux dégradations ne peut être invariante à l'image car dépendante de la quantité initiale de régions uniformes et texturées. Cependant, un article récent [QaG11] a démontré le très faible apport de la prise en compte de l'effet de masquage dans la prédiction de la qualité avec référence complète et référence réduite dans le cadre de distorsions de type JPEG. De ce fait, l'hypothèse d'un partitionnement fixe pourrait être une solution acceptable pour certains types de dégradations.

La seconde hypothèse consiste à adapter dynamiquement la localisation et la taille de chaque partition en fonction des propriétés intrinsèques de chaque image considérée. A titre d'exemple, une structure de données adaptative couramment utilisée pour du partitionnement d'espace, est le quadtree, autant utilisée en indexation dans les bases de données [M+03], que pour de la compression d'images [SF94] ou sa version 3D nommée octree utilisée en optimisation de détection de collisions pour du rendu 3D [JT80]. L'intérêt de ce type de structure est sa capacité à s'adapter aux données, en ayant un partitionnement très fin aux endroits où l'information est dense et un partitionnement grossier où l'information est éparse. La construction d'un quadtree est liée à une recherche d'homogénéité (variance minimale, attributs semblables, etc.). Le but étant donc de maximiser ce critère d'homogénéité.

Dans le cadre de notre étude, nous proposons la maximisation du critère d'équi-proportion du nombre de pixels total $NPsf_j$ inclus dans chaque classe Sf_j afin d'adapter le partitionnement aux propriétés de chaque image. Pour NP, le nombre total pixel de I :

$$NP = \sum_{j=0}^{Nsf} (NPsf_j) \tag{5.3}$$

avec

$$NPsf_0 \approx NPsf_1 \approx NP_T f_T \tag{5.4}$$

Par cette méthode, des partitions de petites tailles se concentrent où les couples de λ sont les plus denses et de grandes partitions pour les couples les moins représentés. Cette approche maximise l'observabilité de faibles variations structurelles pour les populations les plus représentées et réduit la visibilité des variations des populations atypiques. L'idée sous-jacente est de considérer que la sensibilité humaine a des seuils adaptatifs, variables, et très influencés par la quantité de régions uniformes et texturées dans l'image pour les tâches de perception d'artéfacts.

Pour généraliser, les diverses méthodes de partitionnement ainsi que les quantités et tailles des partitions, sont les paramètres qui tendent à maximiser ou minimiser l'observabilité des changements structurels.

5.1.3 Mesure et modélisation des migrations par graphe multi-étiqueté

Nous venons de décrire une méthode permettant de catégoriser dans une classe structurelle Ar_i et Af_j tous les pixels $P_{I(x,y)}$ d'une image quelconque I.

Mais notre démarche vise à mesurer et modéliser les changements structurels d'une image après divers traitements (dégradations), tels qu'introduit par les procédés de compression ou de transmission. Considérons donc maintenant, non pas une image mais un couple $(Io, Id_{typ,mag})$ où Io est une image d'origine, sans aucune dégradation, tandis que $Id_{typ,mag}$ est cette même image mais ayant subi une dégradation de type typ et de magnitude mag.

L'idée développée consiste à classer tous les pixels de Io et $Id_{type,magnitude}$ dans différentes classes, puis de mesurer la quantité de pixels ayant migré après dégradation. On imagine que plus une image est dégradée, plus la quantité de pixel ayant changé de structure (migré de classe) est importante. Concrètement il s'agit de mesurer et modéliser la migration des populations de pixels. Cette migration peut être vue comme un flux entre sites. Ce qui n'est pas sans rappeler les problématiques de mesure de trafic de réseaux routiers, de charge de réseaux téléphoniques ou d'étude de migrations de population... Dans toutes ces problématiques l'utilisation de modèles à base de graphe est très courante, efficace et repose sur de solides bases théoriques [Deo74] [FF56]. Pour cette raison, nous choisissons d'utiliser le formalisme des graphes pour notre modélisation.

Dans ce chapitre, pour tout couple d'images $(Io, Id_{typ,mag})$ nous associons un graphe orienté multi-étiqueté [1] $G_{(Io,Id)} = \langle S, A, \phi_S, \phi_A, \psi_S, \psi_A \rangle$. Tel que :

- S est l'ensemble fini des sommets,
- $A = S \times S$ est l'ensemble des arcs orientés,
- ϕ_S est l'ensemble fini des étiquettes de sommet,
- ϕ_A est l'ensemble fini des étiquettes d'arc,
- $\psi_S : S \longrightarrow \wp(\phi_S) - \{\emptyset\}$ est la fonction qui associe à chaque sommet un ensemble non vide de caractéristiques,
- $\psi_A : A \longrightarrow \wp(\phi_A) - \{\emptyset\}$ est la fonction qui associe à chaque arc un ensemble non vide de caractéristiques,

Sommets et étiquettes

L'ensemble des sommets S est défini par : $S = Ar \times Af$, à savoir le produit cartésien des ensembles de classe Ar et Af définie en section 5.1.2.

Pour rappel, tout pixel de I est décrit par un couple (λ_1, λ_2), et pour chaque valeur de couple (λ_1, λ_2) on peut associer un couple de classes structurelles (Ar, Af).

1. un graphe orienté multi-étiqueté est un graphe dont les sommets sont reliés par des arrêtes orientées (des arcs) et dont les sommets et arcs possèdent plusieurs valuations [CS03].

Puisque S est un ensemble incluant toutes les combinaisons de (Ar, Af), il existe donc pour tout pixel d'une image I, un sommet du graphe capable de représenter sa structure locale. $\forall\ P_{I(x,y)} \in I\ \exists!\ S_{(Sr,Sf)} \in S, P_{I(x,y)} \in S_{(Sr,Sf)}$

Nous associons à chaque sommet plusieurs étiquettes ϕ_S définies par un uplet constitué de :

- Nom : une étiquette de nom, issue du couple (Ar_i, Af_j) qu'il représente.
- NP_o : le nombre de pixels associés à ce sommet dans l'image d'origine, avec $\{NP_o \in \mathbb{N}; 0 \leq NP_o < NP\ \}$
- NP_d : le nombre de pixels associés à ce sommet dans l'image dégradée, avec $\{NP_d \in \mathbb{N}; 0 \leq NP_d < NP\ \}$

Arcs et étiquettes

L'ensemble des arcs E est défini par : $E = S \times S$. Pour chaque couple $(Io, Id_{typ,mag})$, les arcs symbolisent le potentiel de migration des pixels d'une classe à une autre entre l'image originale et sa version dégradée.

Nous associons à chaque arc plusieurs étiquettes ϕ_A par un uplet constitué de :

- Nom : une étiquette de nom, issue du nom du sommet d'origine et du nom du sommet de destination
- NP_od : le nombre de pixels ayant migré du sommet d'origine vers le sommet destination, avec $\{NP_od \in \mathbb{N}; 0 \geq NP_od < NP_o\ \}$
- $NP_{od}L$: le pourcentage que représente NP_od par rapport au nombre NP_o du sommet d'origine. C'est un pourcentage local car lié à l'information de son sommet d'origine, avec $\{NP_{od}L \in \mathbb{R}; 0 \geq NP_od_L < 1.0; NP_od_L = \frac{NP_od}{NP_o}\ \}$. Si l'on considère simplement les étiquettes NP_od_L, le graphe peut être appelé graphe probabiliste. [2]
- $NP_{od}G$: le pourcentage que représente NP_od par rapport au nombre de pixel total de l'image NP . C'est un pourcentage global car lié à une information globale de l'image. $NP_{od}G = \frac{NP_od}{NP}$

L'intérêt de $NP_{od}L$ et $NP_{od}G$ est de rendre intelligible et interprétable les valeurs de NP_{od}. En effet, il est plus compréhensible de dire : 98% des pixels de contours horizontaux sont devenus des pixels de régions uniformes, mais cela ne représente que 4% de tous les pixels de l'image. Contrairement à une

2. Dans ce type de graphe, la somme des pondérations des arêtes sortantes est égale à un.

valeur de $NP_{od} = 10485$ pixels migrants...

Analyse par graphe multi-étiqueté

La Figure 5.6 permet d'illustrer la modélisation par un graphe multi-étiqueté associé à un partitionnement $Nsf = 2$ et $Nsr = 3$ en y faisant figurer toutes les étiquettes des sommets ψ_S et l'étiquetage de deux arcs $f1r0\_f1r1$ et $f1r2\_f1r2$. Il est intéressant de noter que l'arc $f1r2\_f1r2$, a un sommet source et destination identique. Il s'agit d'une migration de pixel sur soi-même, ce qui est finalement un non changement de classe structurelle. Le cas où le $NP_{od}L = 100\%$ de ce type d'arc informe que tous les pixels de cette classe sont restés identiques et qu'il n'y a donc pas de dégradation observable.

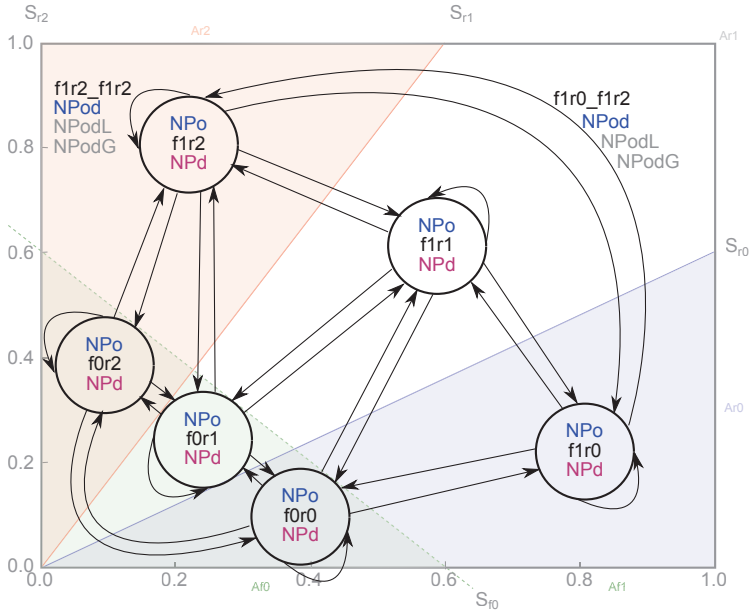


Figure 5.6 : Graphe valué associé à un partitionnement $Nsf = 2$ et $Nsr = 3$.

En résumé, la modélisation basée sur les multiples étiquettes d'arcs et de sommets permet d'observer les changements structurels des pixels à deux niveaux de détail. Le premier niveau d'observation est effectué en étudiant uniquement les étiquettes des sommets. Il est possible de mesurer en chaque sommet l'évolution du nombre de pixels appartenant à ce sommet. Cette évolution, ou similarité est quantifiée par $sim\_S = NP_o(S) - NP_d(S)$. A titre d'exemple, une image fortement dégradée par un flou gaussien doit avoir une augmentation des NP_d dans les sommets représentant les structures uniformes ($sim\_S < 0$) et une réduction de NP_d dans les autres sommets ($sim\_S > 0$).

Cependant, avec ce niveau d'analyse, il n'est pas possible d'identifier les classes ayant le plus contribuées à l'augmentation d'un NP_d particulier. C'est pour ce type de problématique qu'interviennent les arcs et leurs étiquettes. De plus, l'observation des étiquettes des arcs peut être un élément clef pour la reconnaissance et la caractérisation des différents types de dégradation. L'observation des arcs entrants aux sommets de régions uniformes permet de détecter l'introduction de flou dans l'image. Tandis que l'observation des arcs sortants de ces sommets permet de détecter l'introduction de bruit ou de faux contours.

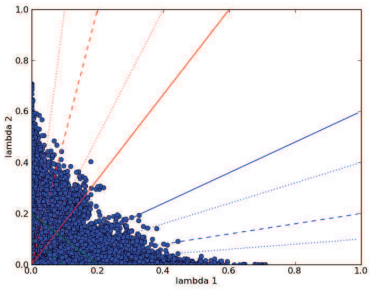
Afin de donner du crédit à ces hypothèses, nous avons mis en place de nombreuses expérimentions et proposé des extensions aux modèles pour des objectifs précis, tels que la prédiction de la qualité perçue et la reconnaissance du paramètre de compression q utilisé lors de compressions JPEG.

5.2 Sensibilité des statistiques aux dégradations

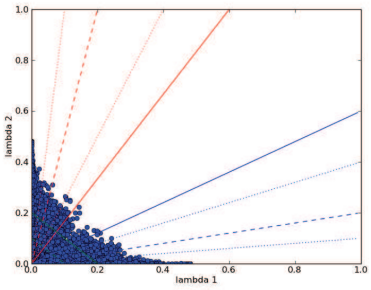
Cette section a pour objectif de valider les hypothèses énoncées dans la section précédente. Tout d'abord, il est intéressant d'observer visuellement dans quelle mesure les valeurs conjointes (λ_1, λ_2) de chaque pixel de l'image sont sensibles aux différentes dégradations.

Une dégradation très classique est l'introduction de flou par l'application d'un noyau Gaussien paramétré par son écart-type σ. Nous proposons sur la Figure 5.7 de visualiser l'effet de ce flou sur l'image et sur la répartition des λ. Sur cette figure, chaque pixel de l'image de gauche trouve son équivalent, sa caractérisation par son couple (λ_1, λ_2) matérialisée par un cercle plein sur l'image de droite. On remarque que plus la dégradation augmente, plus la répartition des λ est différente de la répartition d'origine. Nous observons donc que ces λ sont sensibles à cette dégradation. Nous remarquons également que cette répartition a tendance à se concentrer vers des valeurs proches de 0 pour les fortes valeurs de σ. Pour rappel, quand ces deux valeurs sont proches de zéro, le pixel considéré est censé appartenir à une région uniforme. De ce fait nous pouvons conclure que tous les pixels de l'image, aussi bien, les contours marqués, que les coins ou les textures se transforment de plus en plus en pixels de région uniforme. Ce qui est complètement en accord avec les effets connus de ce filtre, aussi bien d'un point de vue analyse du signal que d'un point de vue perceptuel.

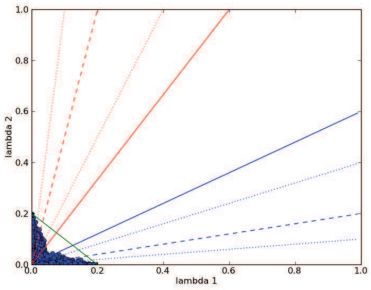
Dans le cadre de ces observations préliminaires, il est également intéressant d'observer la différence de répartition entre deux images sans aucune dégradation mais dont le contenu est différent. Sur la Figure 5.8 nous pouvons com-



Sans dégradation



σ 0.533821



σ 1.164031

Figure 5.7 : Évolution des valeurs conjointes (λ_1, λ_2) par introduction de flou.

parer les répartitions (sans dégradation) entre l'image "Building2" et l'image "Parrots". Sans grande surprise, les répartitions sont très différentes, ceci s'explique simplement par le fait que les images ont des contenus très différents, où l'image "Building2" contient de nombreux contours et coins très marqués du fait de la présence d'un bâtiment, tandis que l'image "Parrots" est principalement constituée d'un fond complètement flou et de textures légères sur le plumage des perroquets. Du fait de cette diversité, la répartition semble avoir un bon pouvoir descriptif, ce qui peut être mis en pratique pour des problématiques de reconnaissance ou classification d'images. Dans ces problématiques, les analyses par regroupement de structures similaires semblent performants [MT09].

Sur cette même figure, il est également possible de visualiser l'effet de diverses dégradations marquées. Les images utilisées pour produire ces répartitions sont extraites de la base d'images LIVE [SWCBa]. Nous pouvons remarquer que l'introduction de flou avec un σ élevé affecte très fortement les deux images et produit des répartitions finales très proches, bien que les images d'origines soient très différentes. Ce constat est également identique pour l'introduction d'un bruit blanc de forte intensité. Cependant, au lieu de concentrer les répartitions vers les faibles λ, c'est l'effet inverse qui se produit, où de nouveaux contours marqués et coins semblent apparaître. Dans les deux cas, ce sont les dégradations qui sont en cours de caractérisation. En effet, quand les dégradations sont extrêmement importantes, ce sont elles qui remplacent le contenu original de l'image. De ce fait, peu importe l'image de référence, les répartitions finales sont très proches. Sur cette figure, il est également possible d'observer l'effet de la dégradation JPEG 2000, qui affiche une certaine ressemblance avec le flou en compactant la répartition vers les faibles λ tout en semblant préserver les fortes structures. Enfin, c'est peut être la compression JPEG qui semble le moins affecter cette répartition, tout en ayant quelques petites ressemblances avec l'effet du bruit, en faisant apparaître de nouvelles structures fortes dans les régions coins et contours. Ce faible impact de la forte compression JPEG sur la répartition peut paraître surprenant, car les effets de cette compression font pourtant disparaître de nombreux détails, textures et coins. Mais dans le même temps, cette dégradation et son effet de bloc fait également apparaître de nouvelles structures, des contours horizontaux/verticaux (plus ou moins marqués) ainsi que des coins aux bords de ces fameux blocs.

La conclusion de cette première observation est que toutes les dégradations affectent les pixels et leurs caractéristiques structurelles définies par leurs λ. Cependant, il apparaît que certaines dégradations soient plus facilement observables. En ce sens, les dégradations de type flou dont les changements structurels sont unilatéraux, peuvent être observés en ne s'intéressant qu'à l'augmentation des NP_d des nœuds de régions uniformes. Au contraire, cette observation

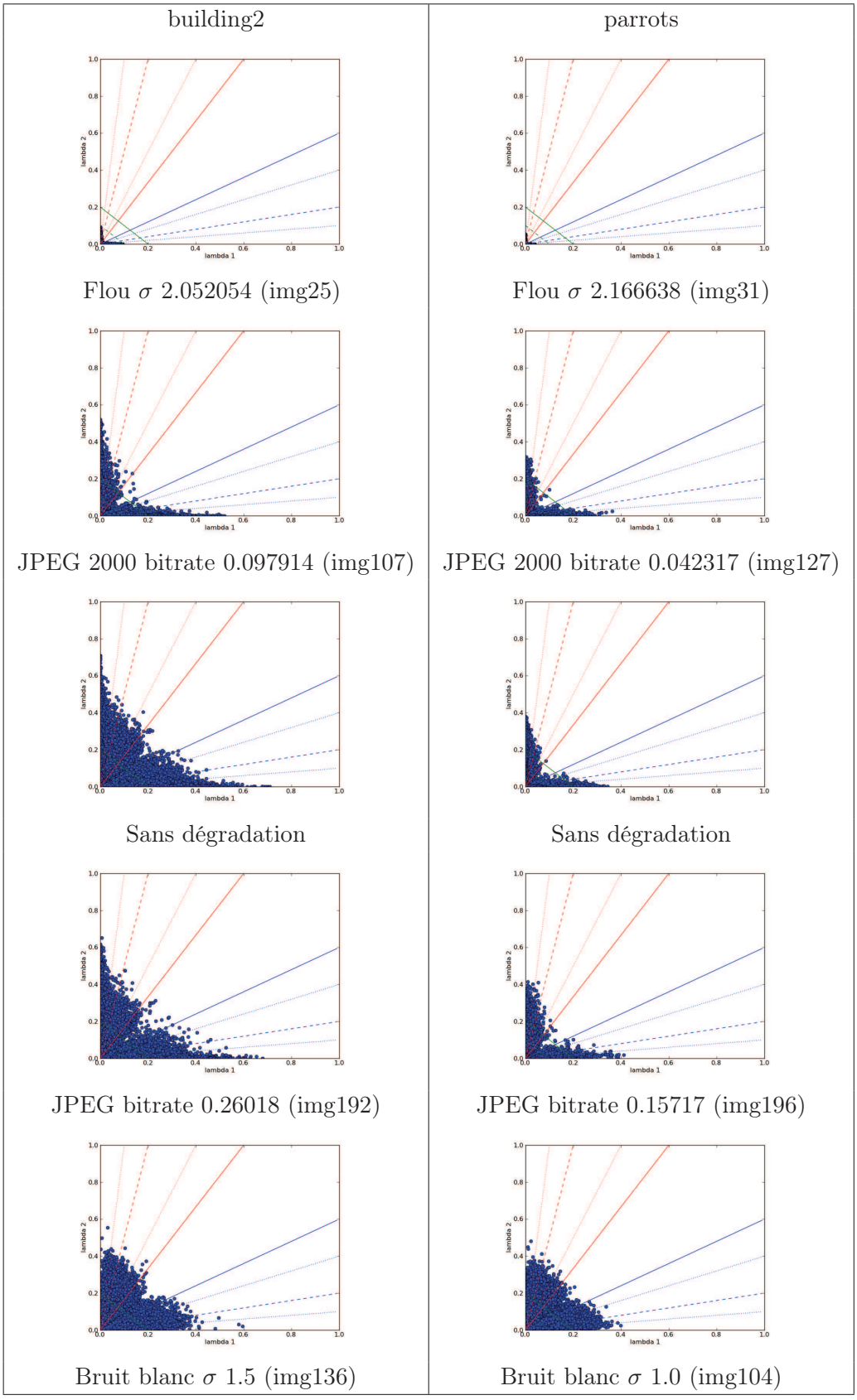
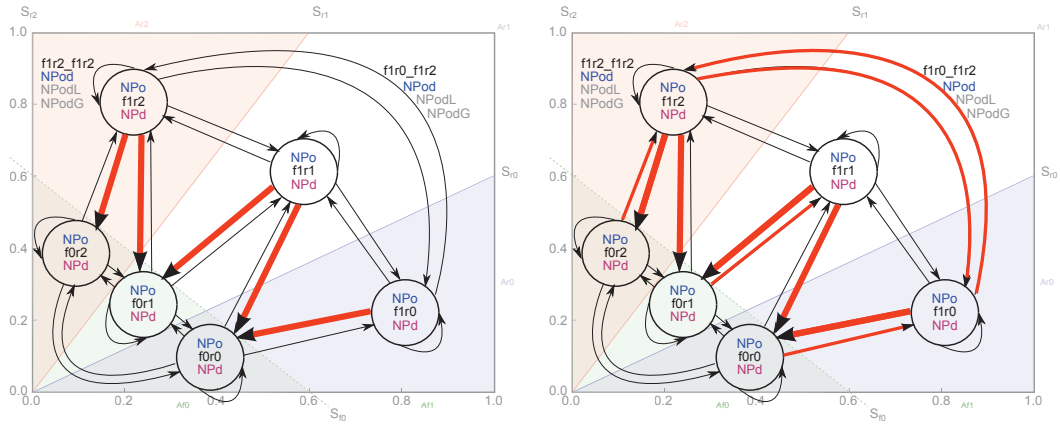


Figure 5.8 : Évolution des (λ_1, λ_2) par introduction de fortes dégradations.

s'avère inutile pour des dégradations plus complexes qui font à la fois apparaître et disparaître des structures, pouvant ainsi garantir des NP_d constants malgré d'importantes migrations. Afin de mieux appréhender le phénomène d'augmentation ou de constance des NP_d pour d'importantes dégradations, nous proposons d'accentuer les arcs les plus exploités sur la Figure 5.9 pour les dégradations de type flou et JPEG.



Dégradations par flou Dégradation JPEG

Figure 5.9 : Visualisation des migrations les plus importantes sur le modèle proposé.

Ayant mis en avant la puissance d'expressivité de l'observation des migrations, nous proposons dans la section 5.2.1 de tenter de prédire la qualité perçue (le MOS) à travers l'exploitation de ce modèle.

5.2.1 Prédiction de la qualité perçue

Nous estimons que le modèle proposé est à même de quantifier les changements structurels de tous les pixels d'une image pour tout type de distorsions, sachant que l'humain est capable d'observer ce type de perturbations du signal. C'est pourquoi nous pensons qu'il est possible d'approximer la qualité perçue d'une image en trouvant une formulation basée sur la mesure des migrations structurelles.

En ce sens, nous proposons d'effectuer une expérimentation en utilisant les images dégradées de la base LIVE et ses notes subjectives. En posant l'hypothèse que l'humain considère qu'une image est de plus en plus dégradée si de plus en plus de pixels ont changé, il semble envisageable de prédire le score moyen des observateurs (MOS : Mean Opinion Score) en exploitant les $NP_{od}L$

ou $NP_{od}G$ de notre modèle. D'où :

$$MOS_{NP_{od}L} = \sum(NPsf_{NP_{od}L}), \tag{5.5}$$

$$MOS_{NP_{od}G} = \sum(NPsf_{NP_{od}G}). \tag{5.6}$$

Dans notre modèle, l'observation des migrations est en grande partie contrôlée par le partitionnement du $diagramme_\lambda$. Nous proposons, dans un premier temps, un découpage $SplitA$ par équation de droite et relativement large en fixant $N_{sr} = 9$ et $N_{sf} = 4$ avec les équations N_{sr} tels que décrit sur la Figure 5.4 et une modification sur les valeurs des équations N_{sf} :

$$\begin{aligned} Sf_0 &= -\lambda_1 + 0.05, \\ Sf_1 &= -\lambda_1 + 0.1, \\ Sf_2 &= -\lambda_1 + 0.2. \end{aligned}$$

Le choix de ces valeurs est issu des observations des répartitions sur des images naturelles avant et après dégradations, et il apparaît que les répartitions sont principalement proches des faibles valeurs de λ et quasiment inexistantes au dessus de 0.4.

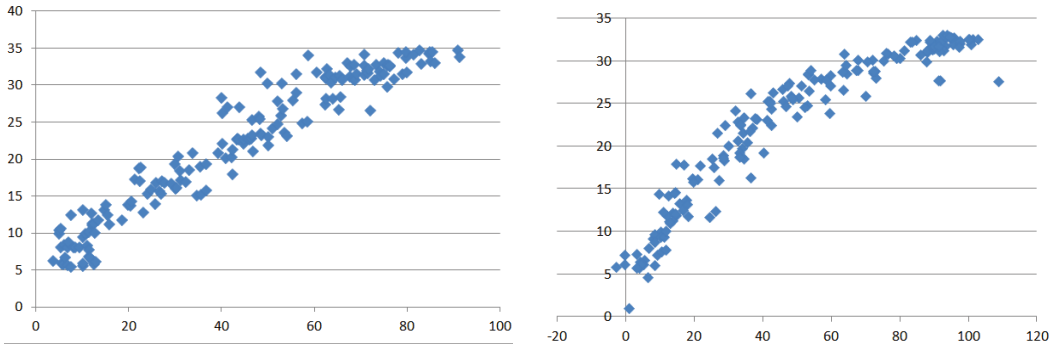
Nous fournissons dans le Tableau 5.1 des scores de corrélation de Pearson pour les dégradations JPEG et JPEG 2000 qui sont les dégradations classiques introduites par les standards de compression. Les résultats fournis sont obtenus sans utilisation de méthode d'ajustement de valeurs afin d'éviter l'introduction de tout biais de mesure dans cette démarche exploratrice.

	$MOS_{NP_{od}L}$	$MOS_{NP_{od}G}$
JPEG	**0,934**	0,869
JPEG2000	**0,962**	0,809

Tableau 5.1 : Performance d'estimation de la qualité perçue par le modèle de découpage $SplitA$ (sans utilisation de fitting).

En complément, nous fournissons également sur la Figure 5.10 le nuage de points des prédictions pour ces deux types de dégradations, où l'axe horizontal

présente les MOS et l'axe vertical présente les valeurs prédites par le modèle proposé.



JPEG 2000 $MOS_{NP_{od}L}$ JPEG $MOS_{NP_{od}L}$

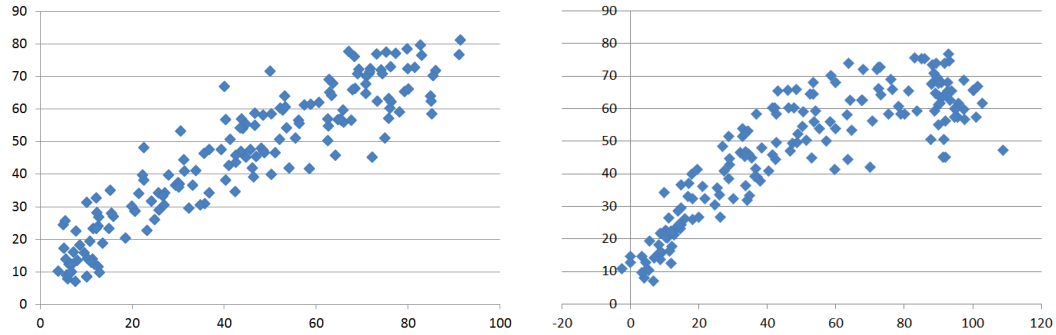
Figure 5.10 : Visualisation des prédictions par le modèle par découpage $SplitA$.

Les résultats obtenus lors de cette expérience sont très encourageants et valident l'hypothèse selon laquelle le modèle est en mesure d'estimer la qualité perçue. De plus, il apparait que la prédiction $MOS_{NP_{od}L}$ est plus pertinente que la version $MOS_{NP_{od}G}$. Néanmoins, les scores obtenus par $MOS_{NP_{od}G}$ sont tout de même relativement élevé. De ce fait, cette mesure est également porteuse d'informations. De plus, en observant ces valeurs, il apparaît que les migrations globales observables sont relativement faibles. Comme expliqué précédemment, la sensibilité de notre mesure est intimement liée à la finesse du découpage. Nous proposons donc un second type de découpage $SplitE$ afin d'être capable de mesurer plus finement les migrations. Le découpage radial étant déjà relativement fin et puisque les deux types de dégradations considérées ont tendance à créer des régions de plus en plus uniformes, nous affinons le découpage en fixant $N_{sf} = 13$ avec pour valeurs d'équations :

$$
\begin{aligned}
Sf_0 &= -\lambda_1 + 0.01 & Sf_6 &= -\lambda_1 + 0.06 \\
Sf_1 &= -\lambda_1 + 0.015 & Sf_7 &= -\lambda_1 + 0.1 \\
Sf_2 &= -\lambda_1 + 0.025 & Sf_8 &= -\lambda_1 + 0.2 \\
Sf_3 &= -\lambda_1 + 0.035 & Sf_9 &= -\lambda_1 + 0.3 \\
Sf_4 &= -\lambda_1 + 0.055 & Sf_{10} &= -\lambda_1 + 0.4 \\
Sf_5 &= -\lambda_1 + 0.05 & Sf_{11} &= -\lambda_1 + 0.5 \\
Sf_{12} &= -\lambda_1 + 0.6
\end{aligned}
$$

	$MOS_{NP_{od}L}$	$MOS_{NP_{od}G}$
JPEG	0,848	**0,892**
JPEG2000	**0,922**	0,861

Tableau 5.2 : Performance d'estimation de la qualité perçue par le modèle de découpage $SplitE$ (sans utilisation de fitting)



JPEG2000 $MOS_{NP_{od}L}$ JPEG $MOS_{NP_{od}L}$

Figure 5.11 : Visualisation des prédictions par le modèle de découpage $SplitE$.

Par ce découpage, les résultats donnés sur la Figure 5.2 sont globalement moins élevés en comparaison avec le partitionnement précédent, avec tout de même des augmentations notables de performances pour les $MOS_{NP_{od}G}$. En effet, avec ce découpage plus fin, plus de migrations sont observables et donnent plus de sens à l'utilisation d'observation de pourcentages globaux. De plus, bien que la corrélation ait baissée pour les $MOS_{NP_{od}L}$, ce découpage offre tout de même des prédictions intéressantes en exploitant une plage de valeurs plus large. Il y a donc bien une augmentation de la sensibilité de détection des migrations. Cette sensibilité est peut être même trop importante en comparaison avec la perception humaine. En effet, au delà de certains seuils, l'œil et le cerveau humain ne sont plus aptes à différencier certaines variations, ce qui est encore plus vrai si l'on y ajoute l'effet de masquage. Une perspective intéressante serait de déterminer et d'appliquer ces seuils de perception afin de réduire la sursensibilité de ce type de découpage.

En complément, et puisque $MOS_{NP_{od}G}$ semble fournir des informations de plus en plus corrélées avec l'humain, une idée serait de combiner ces deux informations. D'autant plus si l'on observe le nuage JPEG $MOS_{NP_{od}G}$ de la Figure 5.12 possédant des valeurs compactes pour les MOS élevés, là où JPEG $MOS_{NP_{od}L}$, visible sur la Figure 5.11, a une grande dispersion de son nuage. Nous proposons trois variantes de combinaison de ces informations locales et globales :

$$MOS_{L.G} = MOS_{NP_{od}G} \times MOS_{NP_{od}L} \tag{5.7}$$

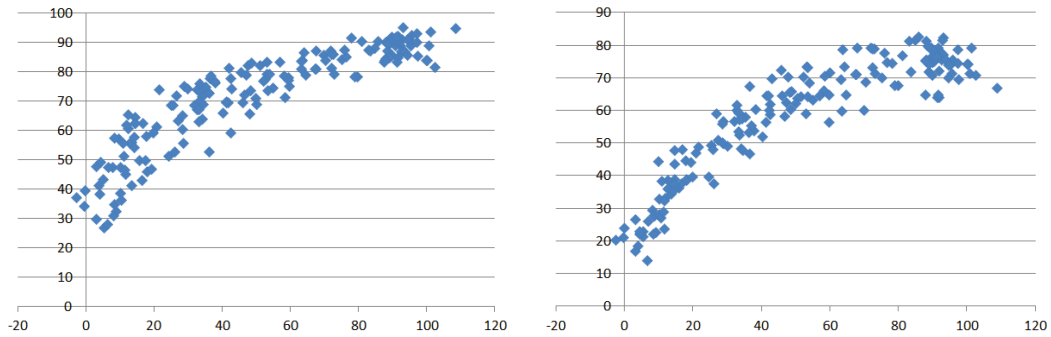


Figure 5.12 : Visualisation des prédictions par le modèle de découpage $SplitE$ sur compression JPEG.

$$MOS_{mean(LG)} = \frac{MOS_{NP_{od}G} + MOS_{NP_{od}L}}{2} \tag{5.8}$$

$$MOS_{sqrt(L.G)} = \sqrt{MOS_{NP_{od}G} \times MOS_{NP_{od}L}} \tag{5.9}$$

$MOS_{NP_{od}L}$	$MOS_{NP_{od}G}$	$MOS_{L.G}$	$MOS_{mean(LG)}$	$MOS_{sqrt(L.G)}$
0,848	0,892	0,913	0,963	0,977

Tableau 5.3 : Performance par combinaison des $MOS_{NP_{od}G}$ et $MOS_{NP_{od}L}$ pour un partitionnement $SplitE$ sur compression JPEG.

Le Tableau 5.3 permet de visualiser les performances en terme de gains de corrélation obtenus avec les différentes formulations de combinaison des $MOS_{NP_{od}G}$ et $MOS_{NP_{od}L}$. Il apparaît clairement qu'utiliser conjointement ces deux informations permet d'augmenter la corrélation avec le jugement humain. Le but de ces formulations n'est pas de proposer une combinaison optimale, mais plutôt de prouver le concept de complémentarité et de combinaison de ces informations. Nous estimons que bien d'autres formulations et poids sur ces mesures peuvent augmenter encore cette corrélation. En ce qui concerne l'utilisation de pondérations sur les mesures, nous proposons, à titre d'illustration, une nouvelle mesure consistant à filtrer les NP_{od} dans la formule de $MOS_{NP_{od}L}$ en ne conservant que les migrations des directions privilégiées par type de dégradation. Par exemple, la compression JPEG 2000 a tendance à favoriser les mêmes arcs que la dégradation par introduction de flou. Nous proposons donc une mesure MOS_{Lf} comme étant la version filtrée de $MOS_{NP_{od}L}$ où seules les migrations des arcs privilégiés sont considérées. En effet, nous pensons que les arcs non privilégiés ont des comportements chaotiques ou aléatoires, ce qui

peut rendre les mesures moins stables. Un comparatif des corrélations obtenues par cette méthode sur les dégradations JPEG 2000 est présenté par le tableau 5.4.

$MOS_{NP_{od}L}$	MOS_{Lf}	$MOS_{sqrt(L.G)}$	$MOS_{sqrt(Lf.G)}$
0,922	0,942	0,929	0,957

Tableau 5.4 : Performance par combinaison des $MOS_{NP_{od}G}$ et $MOS_{NP_{od}L}$ pour un partitionnement $SplitE$ sur compression JPEG 2000 avec filtrage des $MOS_{NP_{od}L}$.

Par ce filtrage, nous ne conservons que les migrations les plus importantes et par conséquent nous augmentons la corrélation. Encore une fois, ces propositions ne sont pas là pour prouver une formulation optimale, mais plutôt pour encourager des études plus approfondies. En effet des pondérations sur les mesures de migrations peuvent avoir un effet positif dans la recherche de corrélation maximale avec l'humain. Ce sont donc des pistes et perspectives encourageantes pour de futurs travaux.

Par ces différentes expérimentations, nous venons de démontrer dans quelle mesure le modèle proposé est à même de prédire la qualité perçue. Dans cette expérimentation, différentes ouvertures ont été proposées, telles que l'optimisation d'un choix de partitionnement, et le filtrage des migrations observées. Ces ouvertures ne sont pas approfondies car elles nécessitent des processus d'optimisation, toujours délicats et dépendants des bases d'images de référence utilisées et des applications réelles visées. Nous préférons orienter notre étude en proposant une caractérisation plus précise de l'influence de chaque type de dégradations sur les migrations. En effet, l'idée serait d'être capable de caractériser les migrations en fonction de valeurs objectives de dégradation, tel que peut l'être le paramètre σ pour le flou ou le facteur q pour la compression JPEG. Ceci permettra ensuite d'effectuer une mise en correspondance entre ce type de paramètre et la qualité subjective.

Dans cette direction, nous proposons dans la section 5.2.2, une extension à notre modèle afin de caractériser la dégradation JPEG en fonction de son facteur q.

5.2.2 Estimation du facteur q de la compression JPEG

Dans cette section, nous proposons de caractériser et de modéliser finement la compression JPEG par notre modèle basé sur les migrations structurelles. La compression JPEG est contrôlée par le facteur de qualité q, compris entre

1 et 100%. Ce facteur sert à diviser la table des pas de quantification utilisés après la DCT, il ne permet donc pas réellement d'atteindre un débit binaire fixé mais plutôt de fixer un facteur de qualité.

Pour notre caractérisation, nous décidons d'utiliser les 29 images de référence de la base d'images LIVE reconnue pour son contenu varié. Cette base de données contient également des images déjà compressées par JPEG, mais ne rend accessible que des informations de qualité subjective et de débit binaire, sans fournir les valeurs du facteur q ayant permis d'obtenir ces images résultats. Pour cette raison, nous décidons de créer nous même les différentes versions compressées de ces images pour des paramètres q connus et fixés. Afin d'avoir une caractérisation relativement fine, nous générons les images pour les valeurs $q = [2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100]$. Chaque image peut être identifiée par Ir_iq_j avec $i \in [1, 29]$ et j une valeur comprise dans l'intervalle q énoncé précédemment.

Pour chaque image Ir_iq_j nous évaluons les migrations de structure $NP_{od}L$ en utilisant un partitionnement $SplitA$ tels que décrit dans la section précédente. Nous stockons toutes les migrations sous la forme d'une matrice de transition Tr_iq_j. L'utilisation de ce genre de matrice est assez répandue lors de la manipulation de graphes, mais dans un souci de clarté, nous fournissons par la Figure 5.13 une illustration de cette formulation afin d'éviter toute confusion et ainsi faciliter la lecture des matrices présentées par la suite. Nous pouvons également introduire une ouverture pour de futurs travaux, car dans cette utilisation de notre modèle, nous n'utilisons que les $NP_{od}L$. De ce fait, ce graphe de migration peut être appelé graphe probabiliste et sa matrice de transition peut être nommée matrice stochastique. Par cette observation, ce sont tous les travaux utilisant les processus de Markov [Par07] qui semblent accessibles.

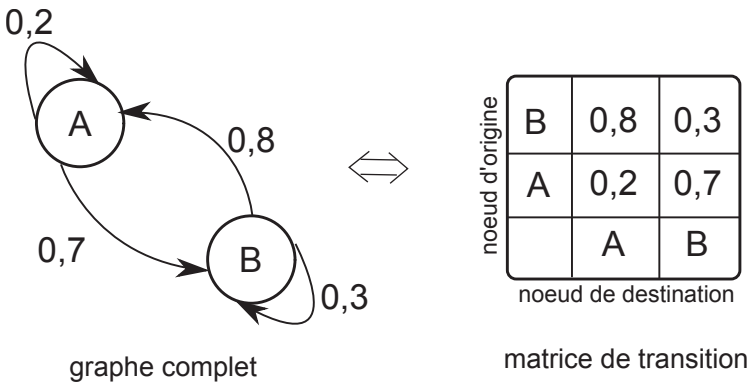


Figure 5.13 : Graphe complet et sa matrice de transition associée.

Afin d'aider à la compréhension de notre modélisation, nous fournissons

sur la figure 5.14 l'architecture de construction de cette dernière. Sur cette Figure, il est possible de visualiser les images Ir_1q_{100} et $Ir_{29}q_{100}$ ainsi que leurs versions les plus dégradées notées Ir_1q_2 et $Ir_{29}q_2$.

Une fois les Tr_iq_j calculées pour chaque image, nous produisons une matrice $Tr_{moy}q_j$ comme étant la moyenne des Tr_iq_j afin de modéliser un comportement moyen de migration pour chaque facteur q utilisé. De ces $Tr_{moy}q_j$ il est possible d'observer l'évolution moyenne des migrations de chaque arc en fonction du facteur q.

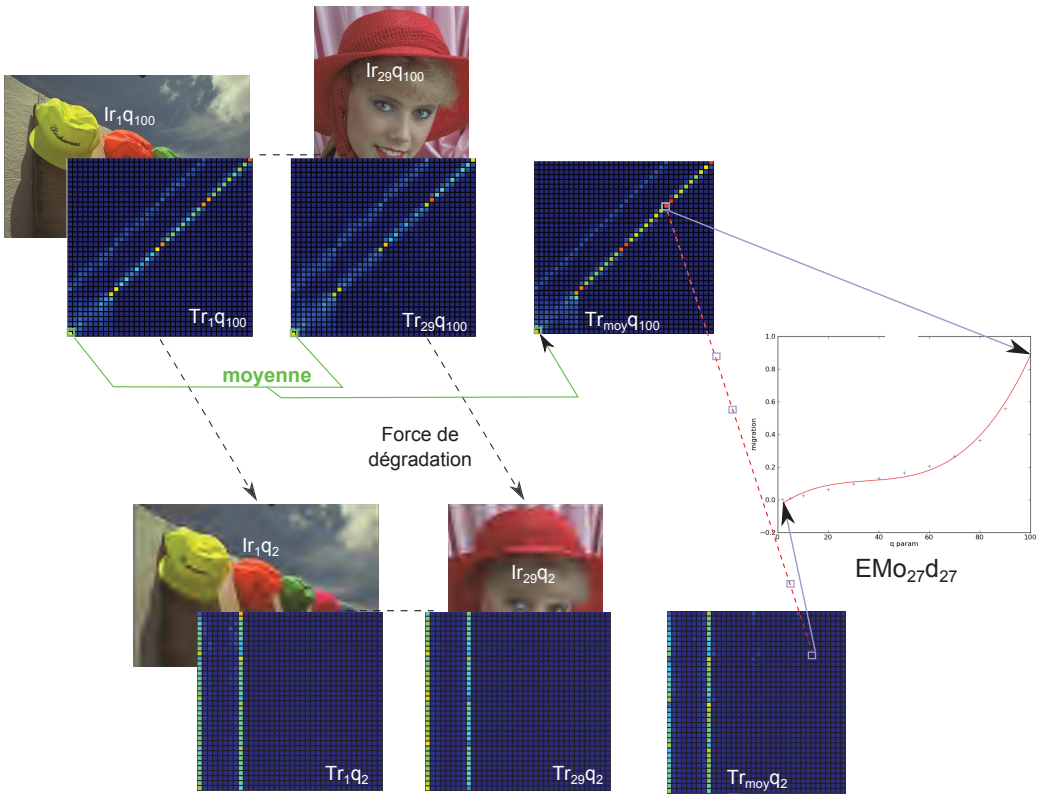


Figure 5.14 : Architecture de construction du modèle d'évolution des migrations pour un type de dégradation.

Cette évolution de migrations entre chaque nœud d'origine et de destination peut être modélisé par une équation EMo_kd_l dont les valeurs prises par k et l sont dépendantes de la finesse du partitionnement. A titre d'exemple et pour faire référence à la figure fournie, illustrant un partitionnement $SplitA$, les valeurs prises par $k \in [1;36]$ et celles de $l \in [1;36]$. Le résultat final obtenu par cette modélisation est une matrice, contenant une équation EMo_kd_l pour

chaque arc, modélisant ainsi l'évolution de toutes les migrations au cours de l'augmentation des dégradations.

Afin de rentrer un peu plus en détails dans les résultats obtenus, nous présentons sur la Figure 5.15 les matrices de transitions $Tr_{moy}q_j$ pour différentes valeurs de j.

En ce qui concerne l'interprétation nous pouvons tout d'abord constater que pour $j = 100$ toutes les migrations sont proches de 100% sur la diagonale, très proche de la matrice identité. Cette diagonale représente les boucles [3]. Si les migrations sur une boucle indique 100% cela indique qu'il n'y a aucun changement de structure observé. Dans le cas présent, nous pouvons conclure que même avec le paramètre de qualité maximum $q = 100$, notre modèle est capable d'observer quelques faibles changements de structure. Il peut être utile de rappeler que la compression JPEG minimum (obtenue par le paramètre $q = 100$) est une compression avec pertes, introduisant donc quelques modifications des pixels. Nous pouvons conclure que notre étude de migrations est suffisamment sensible pour observer ces infimes pertes d'information.

Pour $j = 80$, nous pouvons observer que la diagonale des boucles a évolué, il y a donc plus de pixels ayant changé de structure que pour $j = 100$.

Le nouvelles migrations les plus observables créent une seconde diagonale dans la partie haute gauche de la matrice, partant de la valeur 9 à 26. Cela indique que le type des structures sont préservés mais que leurs intensités ont diminuées. A titre d'exemple, si nous nous intéressons aux pixels de type "coins marqués", nous pouvons constater qu'ils restent de type coins après dégradation, mais qu'ils sont moins marqués. Ces constats sont similaires et accentués pour $j = 40$.

Pour $j = 20$, nous pouvons constater une forme de dispersion des migrations autours des deux diagonales citées précédemment. Les dispersions autour de la diagonale indiquent des changements de types de structures. A titre d'exemple, si l'on considère toujours les pixels de coins, ils ont tendance à évoluer un peu plus en pixels de contours horizontaux ou verticaux. Avec ce niveau de compressions, il est donc possible d'observer à la fois des changements de type de structures et également des évolutions de leurs intensités. Ces phénomènes s'accentuent pour $j = 5$.

Enfin, pour $j = 2$ nous observons des lignes verticales aux valeurs 0, 8, 9, 17, 18, 26. Ce comportement indique que des pixels de plusieurs types et intensités de structures évoluent en des pixels de contours horizontaux et verticaux

3. une boucle est un arc partant d'un sommet et arrivant sur lui-même

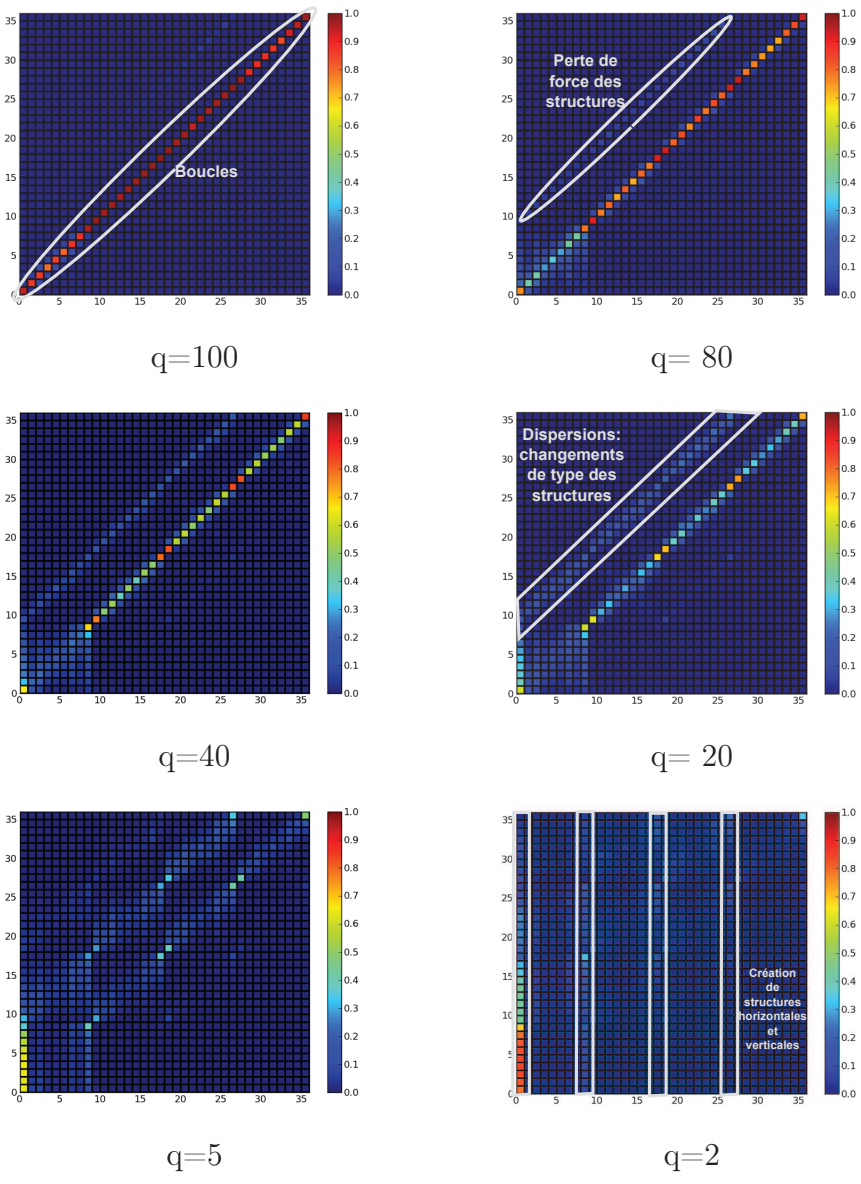


Figure 5.15 : Visualisation des $Tr_{moy}q_j$ pour différents niveaux de compression JPEG.

marqués. Ce qui confirme l'effet de bloc connu de ce type de compression, avec également une concentration au sommet f0r0 indiquant la création de nombreuses régions uniformes.

Par ces premières interprétations, nous pouvons constater que la quantité et le type des migrations évoluent en fonction du facteur q. Afin d'observer plus en détails ces variations, nous proposons sur les Figures 5.16-(b) et (d) d'observer l'évolution des migrations sur deux arcs de migration différents.

Tout d'abord sur la Figure 5.16-(b) nous nous intéressons aux migrations d'une boucle (arc $f2r2\_f2r2$). Pour q proche de 0, il y peu de migrations observables, tandis que pour q proche de 100, il y a quasiment 100% de migration. Puisque nous considérons une boucle, cela indique que plus le facteur q est élevé, moins il y a de changements de structure observables. Ce qui est également intéressant, c'est la monotonie et la linéarité de cette migrations.

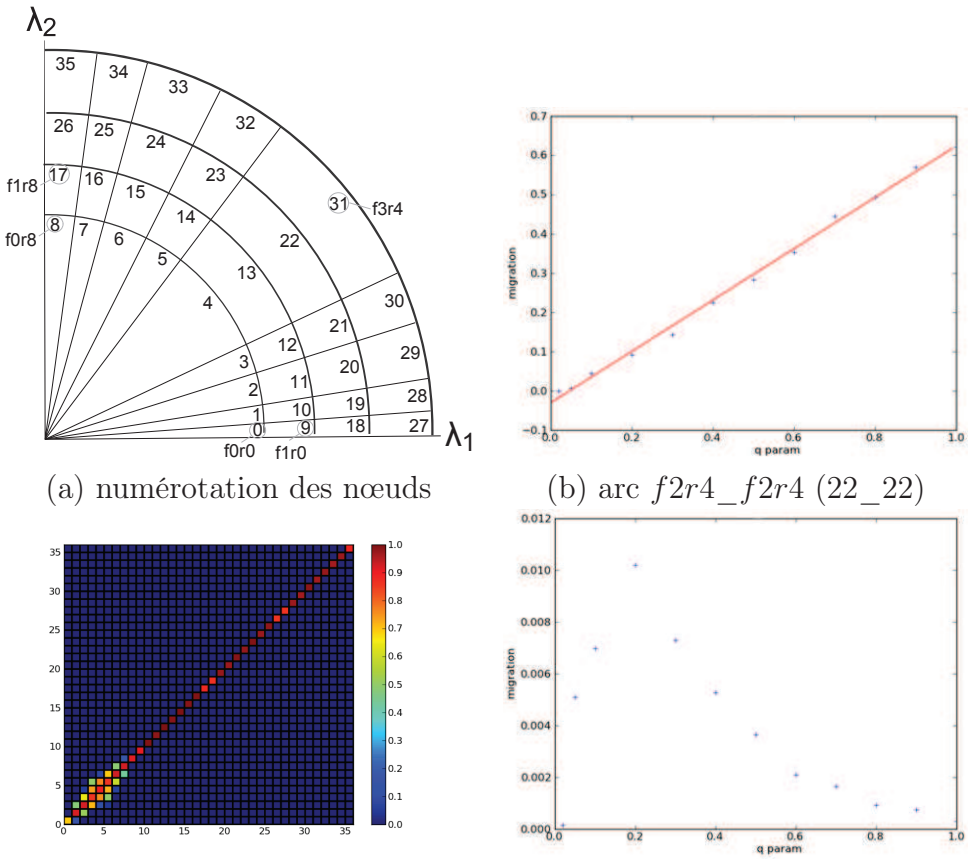
En ce qui concerne la Figure 5.16-(d), cette linéarité n'est plus apparente. Cependant, cet arc n'est pas une boucle, le comportement donc effectivement différent. Nous pouvons observer que pour les facteurs q extrêmes (proche de 0 et 100), quasiment aucune migration n'est observable. Cependant, un pic unique apparaît pour $q = 20$. Après l'observation de tous les arcs qui ne sont pas des boucles, nous avons constaté ce même type de profils, avec cependant des différences concernant, l'amplitude et la localisation du pic.

Afin d'étendre l'observation de linéarité des migrations de l'arc $f2r2\_f2r2$, nous proposons de visualiser sur la Figure 5.16-(c) la corrélation d'évolution des migrations en fonction de celle du facteur q. L'observation est sans appel ; il y a des corrélations très proches de 99% sur toute la diagonale de la matrice de transition. Du fait de cette très forte corrélation et de la monotonie de l'évolution, il semble tout à fait possible de caractériser tous les $EMo_k d_l$ (avec $k = l$) par des fonctions linéaires de la forme $y = ax + b$ permettant ainsi d'approximer et de prédire le facteur q d'une image en fonction des migrations observées.

Mesure de performances de la modélisation

En ce sens, nous décidons de prédire le facteur q en utilisant les fonctions inverses $EMo_k d_l$ ayant une corrélation avec le facteurs q supérieure à 99% :

$$q_p = EMo_k d_l;\ correlation(EMo_k d_l) \geq 0,99 \tag{5.10}$$



(a) numérotation des nœuds (b) arc $f2r4\_f2r4$ (22_22)

(c) corrélation migration$NP_{od}L/q$ (d) arc $f2r4\_f2r1$ (22_19)

Figure 5.16 : Étude des migrations par arc de la compression JPEG.

Nous proposons de visualiser sur la Figure 5.16 les prédictions q_p par rapport au facteur q réel sur la base d'images LIVE (ayant permis de créer le modèle) et sur la base CSIQ ayant un contenu de 30 images de référence complètement différentes des images de la base LIVE.

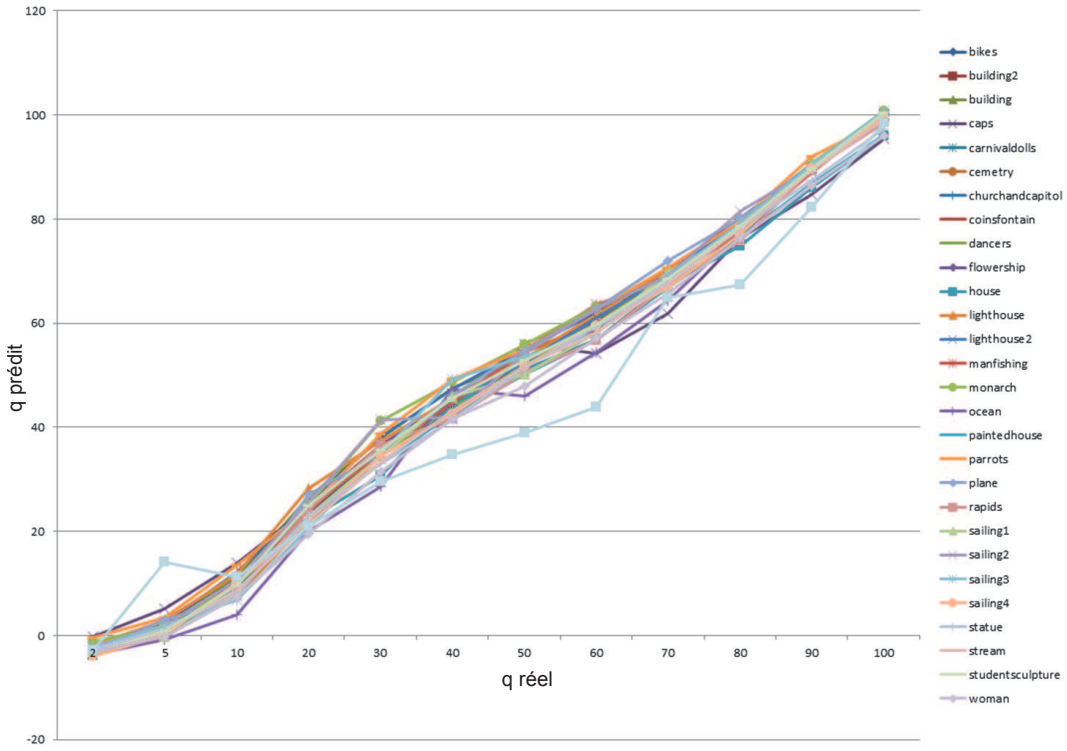
Sur ces graphiques, nous pouvons remarquer à quel point la prédiction est précise et cohérente sur ce grand ensemble d'images différentes. Afin d'étudier plus en détails notre approche, nous avons considéré deux cas de figures de prédictions. Le premier cas se base sur la construction de notre modèle en utilisant les images de référence de la base LIVE, tandis que le deuxième se base sur la construction du modèle en utilisant les images de référence de la base CSIQ. Afin d'évaluer de manière quantitative l'influence de ce paramètre, nous proposons de mesurer sur le tableau 5.5 la corrélation entre la prédiction du facteur q de chaque type de modèle et le facteur réel. A partir de ce tableau nous pouvons constater que quelque soit le modèle utilisé, il est possible de prédire le facteur q avec fiabilité sur des images complètement différentes du modèle. Cependant, nous pouvons tout de même noter que le modèle basé sur LIVE offre des performances légèrement meilleures. Les images ayant servi de référence ont donc une influence, mais elle est tout de même très faible.

	modèle basé sur LIVE	modèle basé CSIQ
base testée LIVE	**0,993**	0,991
base testée CSIQ	**0,989**	0,985

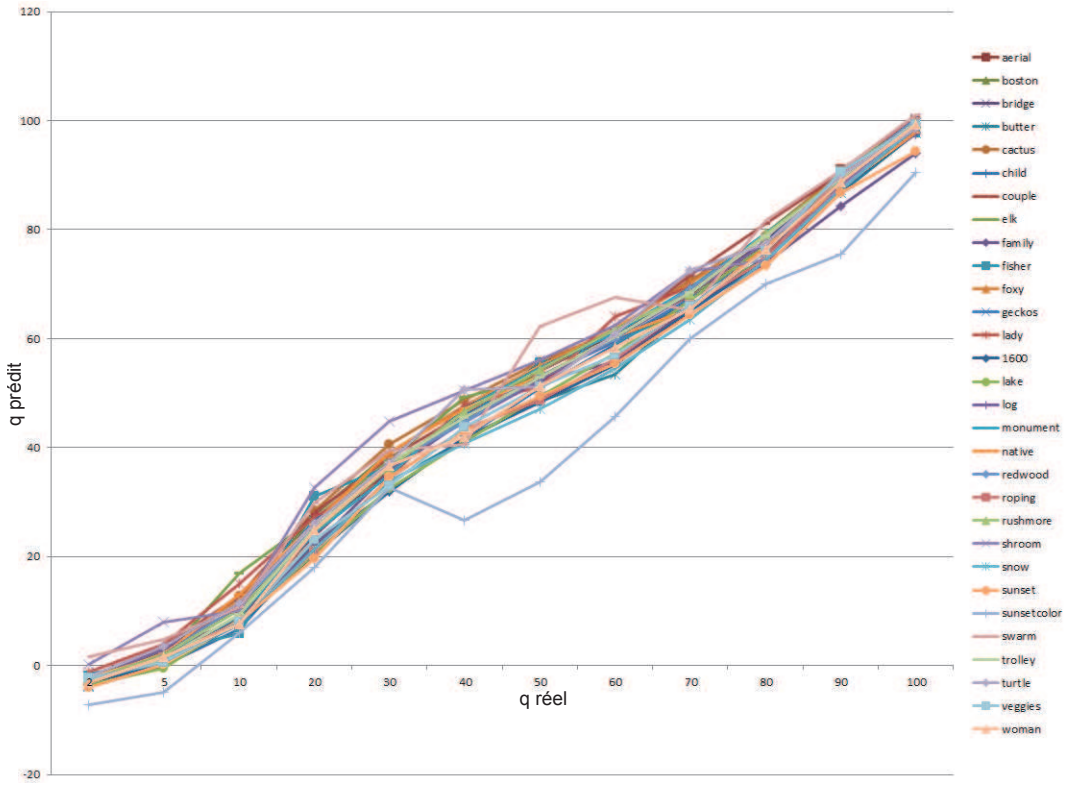
Tableau 5.5 : Niveaux de corrélation entre les modèles utilisés et les bases de test.

Confrontation du facteur q à la qualité perçue

Pour compléter cette étude, nous proposons d'évaluer l'éventuel rapport existant entre le facteur objectif q de la compression JPEG et la qualité perçue par le biais des MOS subjectifs. Dans cette démarche, nous décidons de considérer toutes les bases de données subjectives de la littérature contenant les MOS pour les dégradations de type JPEG, à savoir les bases LIVE [SWCBa, SWCBb], Toyama [HKS], IVC [CA05], TID2008 [PLE+08] et CSIQ [LC10] (décrites en Section 3.4.4). Il est à noter que toutes ces bases exceptée Toyama, ne fournissent pas directement la valeur du facteur q, mais simplement le débit binaire ou un indice de force de compression allant de 1 à 5. L'idée est d'utiliser notre prédicteur de facteur q pour obtenir cette information manquante pour les autres bases. En ce qui concerne la fiabilité des prédictions sur ces différentes bases, nous tenons à rappeler que nous venons de prouver que



prédiction du facteur q sur la base d'images LIVE



prédiction du facteur q sur la base d'images CSIQ

Figure 5.17 : Prédiction du facteur q par le modèle construit sur la base LIVE.

notre modèle basé sur LIVE a donné de très bon résultats de prédiction sur des images complètement différentes, issues de la base CSIQ. La Figure 5.18 illustre le rapport entre la qualité subjective (sur l'axe vertical) et le facteur q (sur l'axe horizontal) pour les différentes bases.

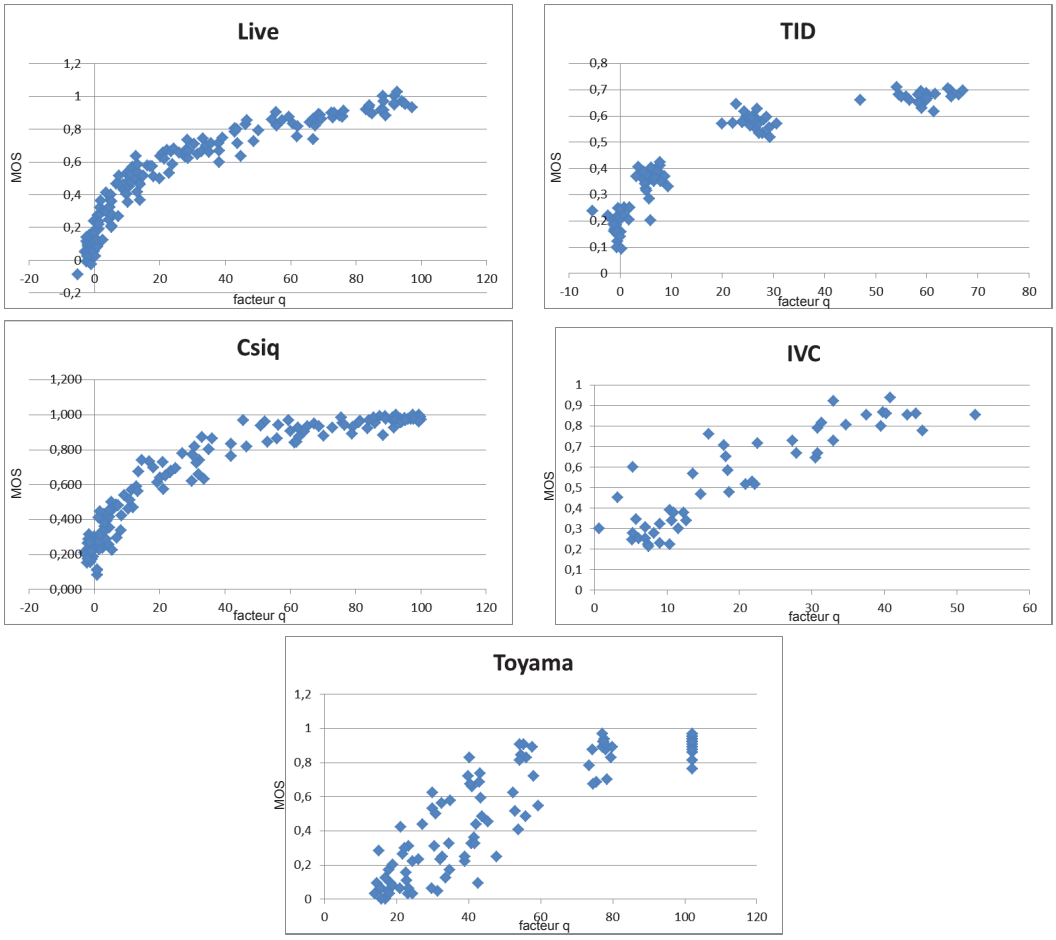


Figure 5.18 : Rapport entre le facteur q et les MOS sur les bases d'images

Par cette méthode, nous avons la possibilité de fusionner, unifier et comparer les différents résultats subjectifs de la littérature tels que présentés par la Figure 5.21. Le premier constat est que certaines bases ne relatent pas le comportement subjectif sur toute l'échelle possible du facteur q, avec par exemple un $q_{max} \approx 50$ pour la base IVC alors que pour la base TID, nous observons des concentrations pour quelques intervalles réduits de facteurs q et de grandes zones sans mesure. Nous pouvons noter que les valeurs subjectives fournies sont également dépendantes de cette plage de configurations testée. En effet, il semble que l'humain, juge et donne des scores dépendants des dégradations minimales et maximales présentées lors de l'évaluation (ou du trainning avant l'évaluation). Nous pouvons également noter que la base Toyama fournit un

nuage de point très dispersé contrairement aux base LIVE et CSIQ qui offre des nuages très compacts et très proches les uns des autres. De ces observations, il est également possible de mieux comprendre la difficulté de produire une métrique de qualité efficace sur toutes les bases avec une même configuration.

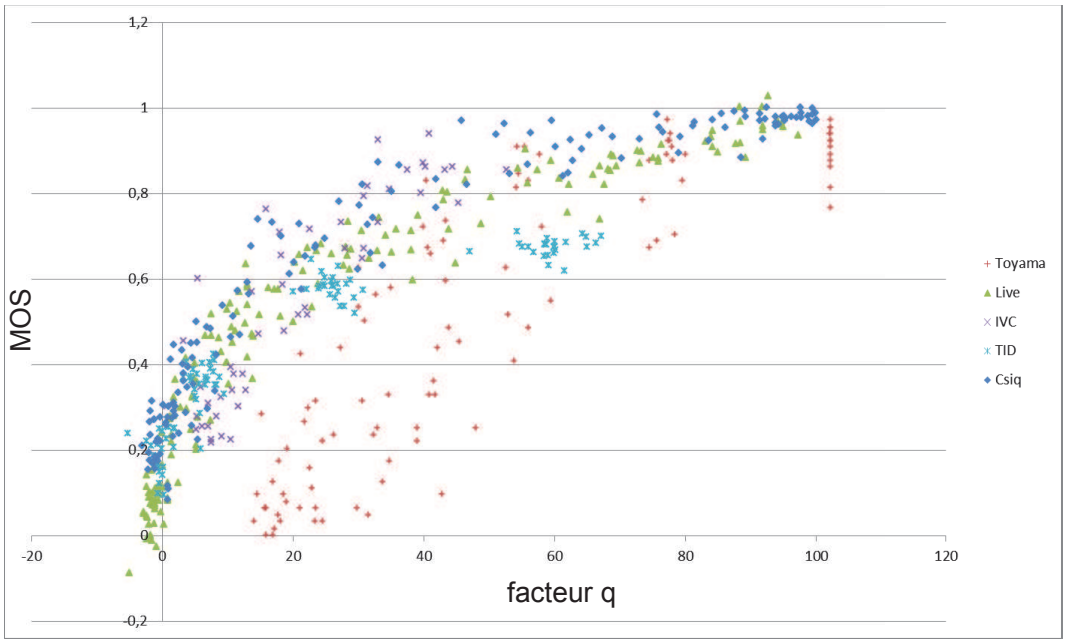


Figure 5.19 : Unification des bases LIVE, CSIQ, TID, IVC et Toyama pour la compression JPEG

Nous constatons que les bases LIVE, CSIQ, TID et IVC fournissent des données cohérentes et proches les unes des autres, contrairement à la base Toyama qui fournit un profil nettement différent. Pour cette raison, nous décidons de retirer la base Toyama de l'étude pour éviter une modélisation erronée. Ayant noté un rapport visible entre l'évolution du facteur q et du MOS sur ces bases unifiées, nous proposons de modéliser ce rapport à travers une régression non linéaire par une fonction logistique :

$$y = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp\left(\beta_2 (x - \beta_3)\right)} \right) + \beta_4 x + \beta_5 \tag{5.11}$$

avec :

$$\beta_1 = 1,10237192657757$$
$$\beta_2 = 0,109711711034500$$
$$\beta_3 = -5,58082290446593$$
$$\beta_4 = 0,00409415320762215$$
$$\beta_5 = 0,0253142443469712$$

La Figure 5.21 permet de visualiser le modèle obtenu par régression non linéaire.

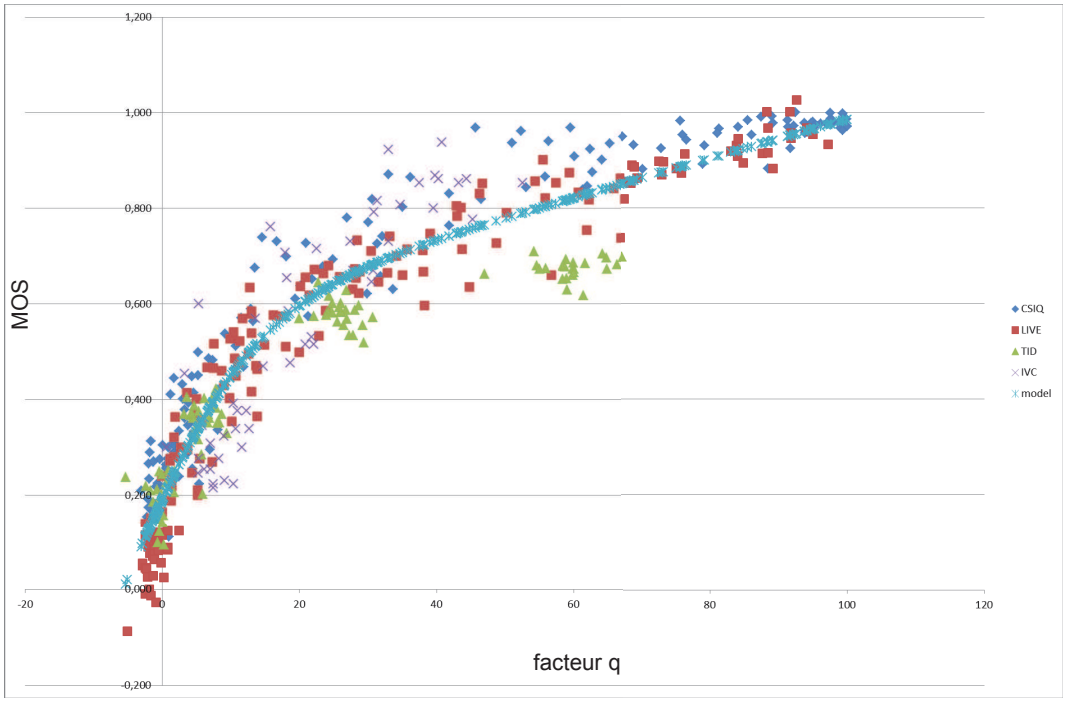


Figure 5.20 : Modélisation de la relation entre le facteur q et le MOS par unification des bases LIVE2, CSIQ, TID, IVC pour la dégradation JPEG

Ayant déterminé un modèle associant le facteur q objectif au paramètre subjectif MOS sur l'ensemble des bases d'images utilisées, nous pouvons proposer une nouvelle métrique Quality by prediction of Q factor ($QpQf$).

La Figure 5.21 permet de visualiser sur l'axe horizontal le MOS subjectif des différentes bases associées aux MOS prédits par notre approche sur l'axe vertical. Nous pouvons constater l'exploitation de toute la dynamique de valeur, la linéarité et la faible dispersion malgré la variété des bases traitées.

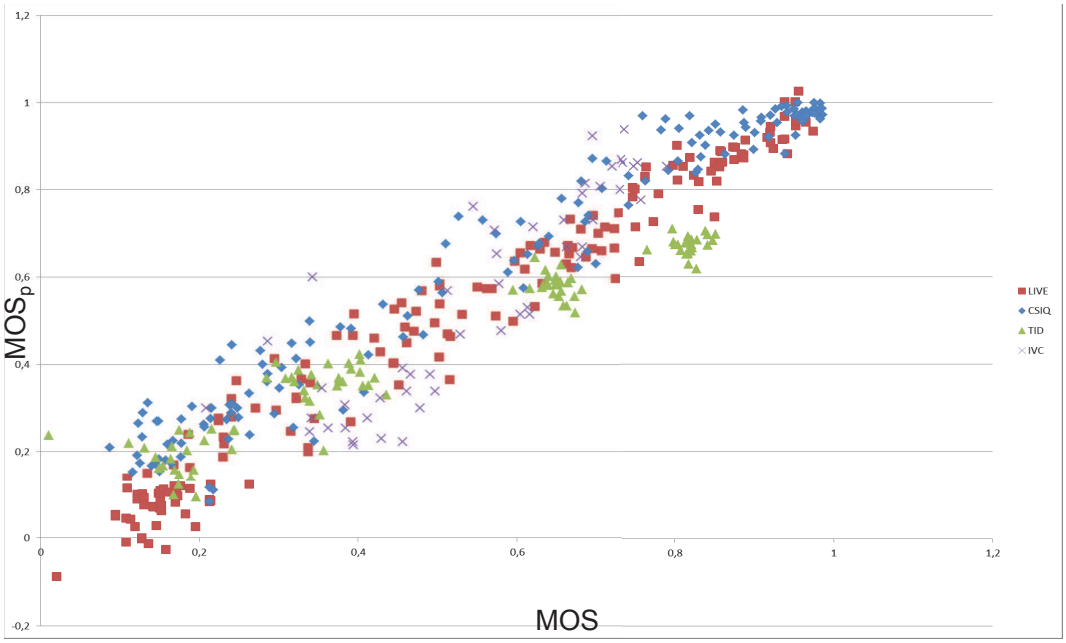


Figure 5.21 : MOS versus MOS prédit par notre modèle sur les base LIVE2, CSIQ, TID, IVC pour des dégradations JPEG

D'un point de vue performance objective et comparaison avec l'état de l'art, nous fournissons sur les Tableaux 5.6, 5.7, 5.8 et 5.9, les mesures de Pearson, Spearman et RMSE pour les bases LIVE2, CSIQ, TID2008 et IVC.

Les nuages de points de l'approche proposée ainsi que ceux des métriques VIF et SSIM sont quant à eux représentés sur les Figures 5.22, 5.23, 5.24 et 5.25. Les résultats sont fournis sans aucun ajustement des valeurs, afin de s'affranchir de l'introduction d'un quelconque biais de mesure.

Tableau 5.6 : Performance sur la base LIVE

D\M	VSNR	PSNR	VIF	Pdiff
Corr. Pears.	0.945	0.888	0.958	0.706
Corr. Spear.	0.965	0.901	**0.983**	0.708
RMSE	48.2	**43.9**	59.5	48.5
D\M	PSNR HVS	SSIM	IFC	QpQf
Corr. Pears.	0.925	0.927	0.859	**0.984**
Corr. Spear.	0.934	0.976	0.942	0.978
RMSE	46.4	59.1	58.2	59.5

En observant la Figure 5.22 et les données du Tableau 5.6 nous pouvons conclure que notre proposition offre la meilleure corrélation avec le jugement

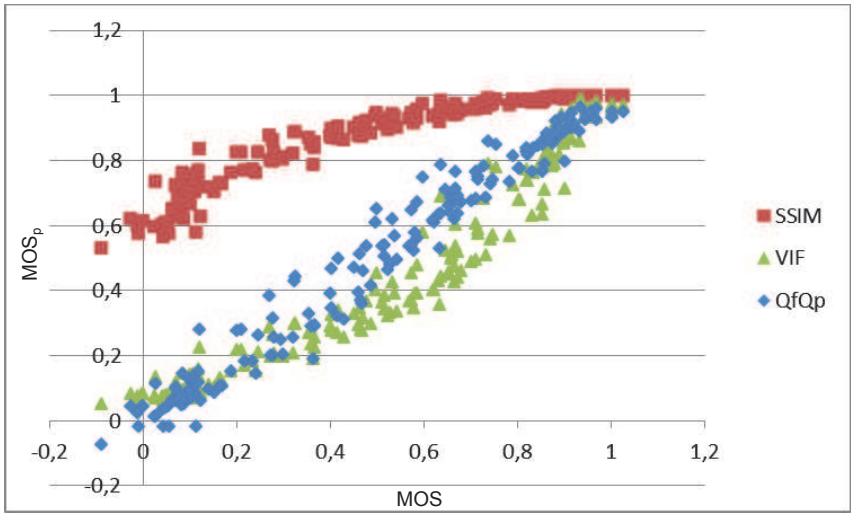


Figure 5.22 : Comparaison avec SSIM et VIF sur la base LIVE

humain tout en exploitant complètement la plage des valeurs disponibles, contrairement à la faible dynamique de valeurs de la métrique SSIM.

Tableau 5.7 : Performance sur la base CSIQ

D\M	VSNR	PSNR	VIF	Pdiff
Corr. Pears.	0.724	0.893	0.958	0.655
Corr. Spear.	0.903	0.900	**0.970**	0.665
RMSE	35.2	28.7	**0.652**	59.6
D\M	PSNR HVS	SSIM	IFC	QpQf
Corr. Pears.	0.911	0.916	0.794	**0.974**
Corr. Spear.	0.922	0.954	0.940	0.956
RMSE	26.6	0.668	8.807	0.665

Pour la base CSIQ dont les résultats sont fournis en Figure 5.23 et Tableau 5.7, nous pouvons avoir un constat similaire à celui de la base LIVE2, avec cependant une légère dispersion de notre métrique pour les faibles facteurs de qualité en comparaison avec la métrique VIF. Mais cette dispersion est compensée pour les facteurs qualité importants.

Encore une fois, c'est notre modèle qui fournit la meilleure corrélation pour la base TID visualisable sur la Figure 5.24 et le Tableau 5.8 avec encore une fois une légère dispersion pour les faibles valeurs de facteur qualité.

En ce qui concerne la base IVC, disponible sur la Figure 5.25 et le Tableau 5.9, nous pouvons noter le profil relativement plat et la faible dynamique de

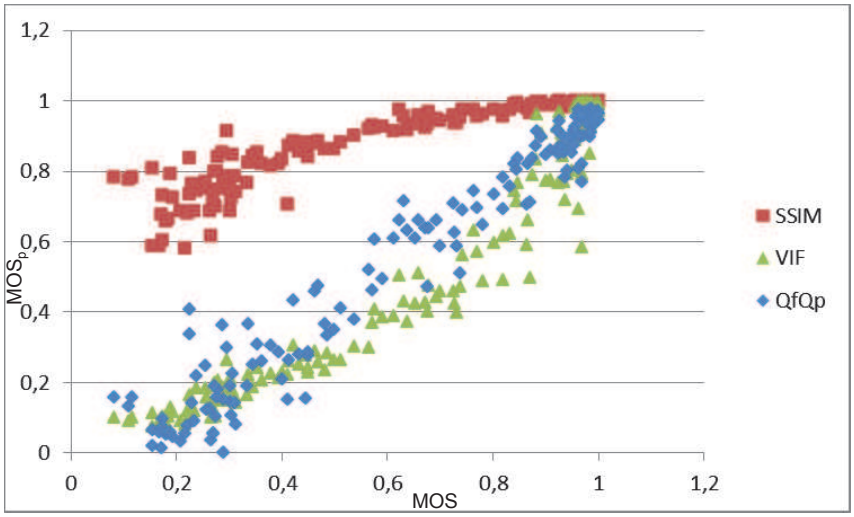


Figure 5.23 : Comparaison avec SSIM et VIF sur base CSIQ

Tableau 5.8 : Performance sur la base TID

D\M	VSNR	PSNR	VIF	Pdiff
Corr. Pears.	0.900	0.889	0.931	0.643
Corr. Spear.	0.913	0.901	0.918	0.673
RMSE	26.4	26.8	**0.100**	58.6
D\M	PSNR HVS	SSIM	IFC	QpQf
Corr. Pears.	0.943	0.931	0.779	**0.965**
Corr. Spear.	**0.942**	0.925	0.816	0.928
RMSE	24.4	0.442	3.49	0.117

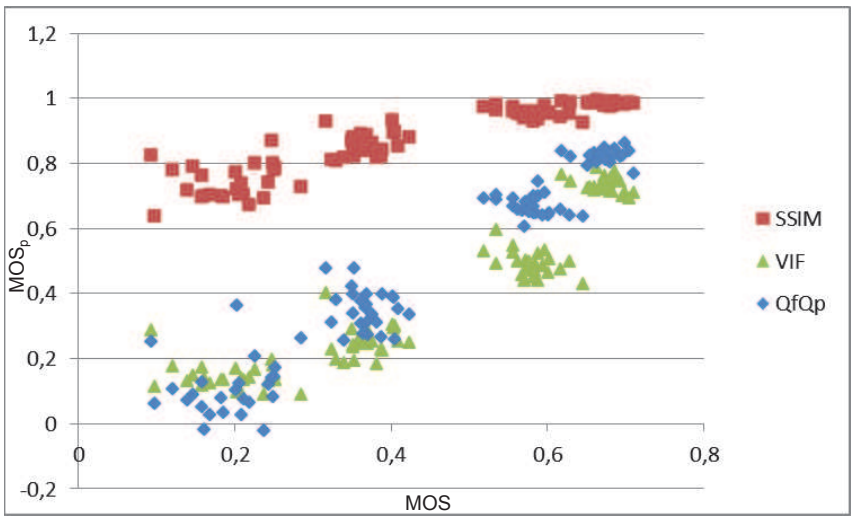


Figure 5.24 : Comparaison avec SSIM et VIF sur la base TID

Tableau 5.9 : Performance sur la base IVC

D\M	VSNR	PSNR	VIF	Pdiff
Corr. Pears.	0.653	0.590	**0.922**	0.181
Corr. Spear.	0.650	0.664	0.922	0.219
RMSE	22.1	26.2	0.220	37.4
D\M	PSNR HVS	SSIM	IFC	QpQf
Corr. Pears.	0.661	0.833	0.921	0.866
Corr. Spear.	0.654	0.922	**0.953**	0.899
RMSE	22.0	0.409	2.35	**0.123**

valeur des toutes les métriques. Cela s'explique par la faible dynamique des facteurs q utilisés lors de l'expérimentation subjective. Ceci démontre l'impertinence de cette base d'images qui est d'ailleurs construite avec des images très anciennes (Lena, Barbara) qui n'ont plus de lien avec les possibilités actuelles de capture d'images.

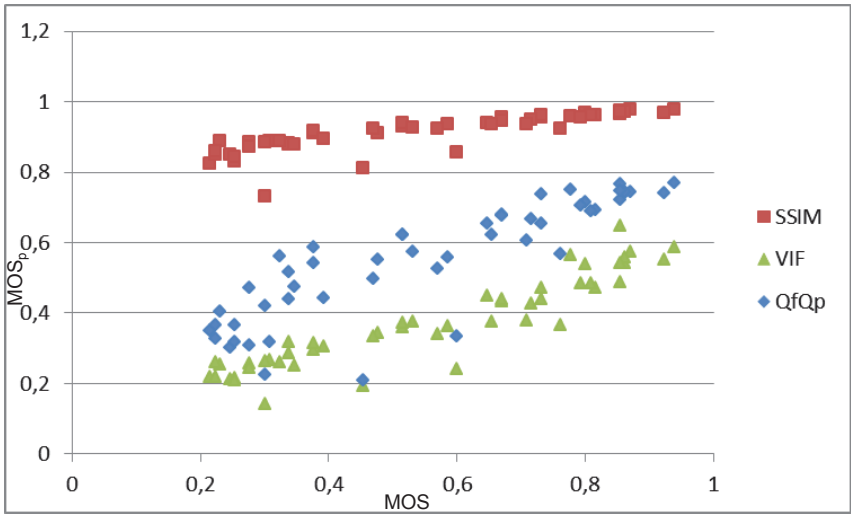


Figure 5.25 : Comparaison avec SSIM et VIF sur la base IVC

Nous pouvons constater que les performances objectives de notre approche confirment les observations visuelles, en fournissant des prédictions fiables et robustes, malgré la variété des bases traitées. Ayant validé les atouts de ce modèle en tant que métrique, nous pouvons également ajouter qu'il peut être intégré dans un schéma de compression en tant qu'outil d'optimisation débit/distorsion (RDO :Rate-distortion optimization).

Avec la possibilité d'associer le facteur q de la compression JPEG et le MOS pour n'importe quelle image, il est maintenant possible de déterminer le facteur

q optimal garantissant le meilleur compromis entre niveau de compression et l'aboutissement à une qualité subjective minimale. Ce type d'optimisation peut être très utile dans les problématiques de compression d'images sur les réseaux sans fil, tel qu'introduit dans le chapitre 4.

5.3 Conclusion

Nous avons proposé dans ce chapitre d'étudier de manière détaillée l'impact de plusieurs dégradations sur les changements structurels de l'image. Dans cette objectif, nous proposons une méthodologie et un modèle à base de graphe multi-étiqueté afin de caractériser et quantifier ces changements.

De ce modèle, nous avons proposé une première méthode de prédiction de la qualité perçue en utilisant une quantification de toutes les migrations structurelles de l'image pour les dégradations JPEG et JPEG 2000.

Nous avons ensuite focalisé notre étude sur les distorsions de type JPEG, car il est encore aujourd'hui l'outil de référence pour la compression des images dans le domaine grand public. Dans cette étude, nous avons modélisé finement le rapport existant entre le facteur q de la compression et les migrations structurelles associées. Nous avons démontré à quelle mesure le modèle proposé est capable de quantifier et prédire les déformations structurelles en fonction du facteur q.

Ce travail a également permis d'unifier les bases de données subjectives pour les dégradations JPEG. Il en est sorti qu'il existe un lien fort entre le facteur q objectif et le MOS subjectif. Une mise en équation de ce rapport a permis le développement d'une nouvelle métrique de qualité, spécialement adaptée à la compression JPEG, offrant de très bonnes performances de prédiction. Ce modèle reliant le q et le MOS pourrait permettre également le développement d'un outil d'optimisation débit/distorsion, pouvant potentiellement être exploité dans de nombreuses applications, dont la compression pour les réseaux sans fil est un exemple.

Notre travail a permis de montrer le potentiel de notre modélisation, en l'illustrant sur des cas pratiques, orientés sur les dégradations dues à la compression. Les pistes restant à explorer sont de modéliser finement d'autres types de dégradations et leur lien avec la perception. Il est également prometteur d'exploiter la puissance de nos graphes en tant que classifieur de types de dégradations à travers l'utilisation de techniques de *graphe matching*.

Conclusion

Notre objectif était d'étudier et d'intégrer des facteurs humains et psychovisuels dans les problématiques de compression et la transmission des images numériques. Ce travail étant effectué dans le but d'augmenter la qualité de l'expérience lors de la manipulation des images dans les contextes et usages actuels et futurs. Il a été mené dans le cadre de l'élaboration du nouveau standard de compression AIC par l'intermédiaire du projet ANR CAIMAN.

1 Rappel des contributions

Pour atteindre cet objectif, nous nous sommes particulièrement intéressés aux détecteurs de points d'intérêt en tant qu'extracteurs de caractéristiques structurelles de l'image.

Nos travaux [NLF12] ont, dans un premier temps, permis de révéler puis de quantifier le fait que les détecteurs de points d'intérêt (Harris, SIFT et SURF), s'ils sont paramétrés et configurés de manière particulière, partagent des propriétés communes avec les mécanismes impliqués dans la saillance visuelle. Par ces résultats, nous avons pu développer un nouveau type de prédicteur de saillance visuelle ascendante, compétitif au regard des méthodes de l'état de l'art (Itti et Achanta) sur plus de 1200 images évaluées. Par l'utilisation des

détecteurs de points d'intérêt et leurs nombreuses optimisations, notre proposition garantit une simplicité d'implémentation tout en étant peu coûteuse en temps d'exécution. Ainsi, notre prédiction de saillance peut être intégrée très facilement et rapidement dans bon nombre d'applications existantes, assurant ainsi des gains de performance notables en intégrant des paramètres psychovisuels dans des chaînes de traitements qui usuellement, n'intègrent pas ce genre de considérations.

Dans un second temps, nous avons exploité le pouvoir descriptif des structures de l'image dont bénéficient les détecteurs de points d'intérêt ainsi que leur sensibilité aux dégradations, afin de produire une métrique de qualité à référence réduite, nommée QIP [NLF10b]. Cette métrique n'a besoin que de 22 octets pour estimer la qualité perçue, tout en tirant profit du faible coût de calcul du détecteur de coins de Harris. Afin d'intégrer plus de paramètres psychovisuels, une extension de cette métrique, QIP-HSM [NLF11a] a également été proposée en exploitant une hiérarchisation spatiale du contenu de l'image par la saillance [NLF10a] dans l'optique de pondérer les artéfacts détectés. Grâce aux faibles temps de calcul et à la référence réduite, notre métrique QIP a pu être intégrée en tant qu'organe décisionnel dans une chaîne de transmission d'images sur canal wifi MIMO [ANP$^+$11a, ANP$^+$11b]. Elle a ainsi permis de garantir le décodage optimal des images en réception en estimant la qualité perçue. Par ce biais, l'utilisateur de ce système a la garantie de disposer de la meilleure image possible dans des conditions de faible couverture du réseau et malgré d'importantes erreurs de transmission. Ces gains de qualité visuelle ont été quantifiés objectivement par les métriques PSNR et SSIM. En complément, une campagne d'évaluation subjective a été menée afin de vérifier les gains en terme de qualité de l'expérience, sur dix-sept observateurs. Elle a par conséquent révélé que notre approche augmente la qualité de l'expérience dans plus de 90% des cas.

Par la suite, nous avons mené une étude détaillée de l'impact de plusieurs dégradations sur les changements structurels de l'image en exploitant les variations des valeurs propres des tenseurs de structure. Grâce au fort pouvoir descriptif de ces statistiques de l'image, nous avons proposé une méthodologie et un modèle à base de graphes multi-étiquetés à même de caractériser et quantifier divers changements et artéfacts introduits dans les images. De ce modèle, nous avons proposé une première méthode de prédiction de la qualité perçue en utilisant une quantification des migrations structurelles pour les dégradations JPEG et JPEG 2000. Nous avons ensuite focalisé notre étude sur les distorsions de type JPEG, car il est encore aujourd'hui l'outil de référence pour la compression. Dans cette étude, nous avons modélisé finement le rapport existant entre le facteur de qualité q de la compression et les migrations structurelles observables objectivement. Nous avons démontré dans quelle mesure le modèle

proposé est capable de quantifier et de prédire les déformations structurelles en fonction du facteur q. Ce travail a également permis d'unifier les bases de données subjectives de la littérature pour les dégradations JPEG. Il en est apparu un lien fort entre le facteur q objectif et le MOS subjectif. Une mise en équation de ce rapport a permis le développement d'une nouvelle métrique de qualité, spécialement adaptée à la compression JPEG, offrant d'excellentes performances de prédiction.

Pour finir, nous avons entamé le développement d'un service-web, ImQual [NLF11b], spécialement dédié à la comparaison et l'utilisation des métriques de qualité. Ce service a pour but de référencer de manière continue l'état de l'art des métriques de qualité, en fournissant des mesures de performance détaillées sur chacune d'elles et sur diverses bases d'images, afin de pouvoir les comparer facilement. Ce service permet également d'utiliser en ligne l'intégralité des métriques disponibles, afin de démocratiser d'autres métriques que le PSNR et ainsi d'étendre l'usage des métriques à des domaines scientifiques qui n'en sont pas spécialistes. Bien que non encore finalisé, ce projet est actuellement encouragé et soutenu par le comité JPEG et la CIE, pour à terme, en faire l'outil de référence pour les mesures de performance.

2 Perspectives

Nos travaux ont démontré la possibilité d'exploiter les détecteurs de points d'intérêt pour des tâches de prédiction de saillance visuelle et de qualité perçue des images. Notre démarche a permis d'obtenir d'ores et déjà des résultats significatifs au regard de l'état de l'art. Néanmoins, il semble que des gains en termes de performance peuvent être obtenus par l'utilisation de processus d'optimisation de paramètres comme les réseaux de neurones ou les SVM.

A partir de ces travaux, plusieurs perspectives s'ouvrent à nous. Il serait donc intéressant d'étendre la démarche suivie en considérant l'aspect temporel et ce en traitant les problématiques de saillance et de qualité sur les vidéos. Cette extension semble assez naturelle et séduisante, car n'oublions pas que les détecteurs de points d'intérêt ont été initialement conçus pour les problématiques de suivi d'objets en mouvement sur des séquences d'images et que la saillance visuelle y est très sensible.

Il est également envisageable d'ajouter une quatrième dimension à ces travaux en considérant les vidéos 3D. Encore une fois, grâce à la puissance d'appariement de ces détecteurs, il semble prometteur de les utiliser pour étudier les différentes profondeurs des objets de la scène et identifier les régions d'oc-

clusions. Il est donc envisageable de proposer une nouvelle génération de métriques de qualité spatio-temporelle 3D, exploitant la puissance et la flexibilité de paramétrage des détecteurs de points d'intérêt afin de gérer de manière conjointe la saillance et la visibilité d'artéfacts tout en assurant un faible coût calculatoire grâce aux évolutions continues de ce type d'outils.

Enfin, c'est notre modèle de caractérisation des migrations à base de graphes multi-étiquetés, exploitant les évolutions des structures de l'image, qui semble le plus prometteur et le plus ouvert à de futurs travaux. A court terme, il pourrait être utilisé pour décrire et caractériser finement d'autres dégradations, telles que le flou, le ringing, le bruit de moustique ou de nombreux types de défauts introduits par des erreurs de transmission, et toujours en y associant la perception humaine. Par cette caractérisation, il semblerait assez naturel de s'en servir comme outil d'optimisation débit/distorsion et ce dans l'optique de garantir la meilleure qualité de service et d'expérience.

A moyen terme, en plus des nouvelles distorsions, ce modèle pourrait être étendu en y ajoutant les dimensions temporelles et 3D au même titre que pour nos métriques QIP. Au vu des graphes au cœur de ce modèle, ce sont tous les travaux sur les chaînes de Markov qui semblent exploitables pour ces ajouts de dimensionnalités. De plus, par les caractérisations de divers artéfacts, il semble tout à fait possible, en utilisant des techniques de graph matching, de développer des méthodes de reconnaissance et de prédiction de types de défauts. Ce type d'outils pourrait en pratique servir comme classifieur exploitable par des métriques de qualité devant connaître le type de défaut présent dans l'image avant d'en estimer la qualité.

A plus long terme, il serait intéressant d'étendre les travaux de caractérisation des artéfacts par statistiques structurelles dans le domaine compressé. Ceci permettra d'agir directement sur le flux binaire des images et vidéos et évitera la décompression avant la phase d'estimation de la qualité. Une des applications cible serait la télévision sur IP où des flux importants sont partagés sur plusieurs canaux.

Bibliographie

[AHES09a] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1597–1604, 2009.

[AHES09b] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Salient region detector. http://ivrg.epfl.ch/, 2009.

[AHR05] M. Al-Hames and G. Rigoll. A multi-modal graphical model for robust recognition of group actions in meetings from disturbed videos. In *IEEE International Conference on Image Processing (ICIP)*, volume 3, pages 421–424. IEEE, 2005.

[ALRH] H. Alers, H. Liu, J. Redi, and I. Heynderickx. TUD image quality database: Eye-tracking release 2. http://mmi.tudelft.nl/iqlab/eye tracking 2.html.

[ALRH10] H. Alers, H. Liu, J. Redi, and I. Heynderickx. Studying the risks of optimizing the image quality in saliency regions at the expense of background content. In *IS&T SPIE Electronic Imaging, Image Quality and System Performance VII*, Jan 2010.

[ANP+11a] J. Abot, M. Nauge, C. Perrine, M.-C. Larabi, C. Bergeron, C. Olivier, and Y. Pousset. A robust content-based JPWL

transmission over a realistic mimo channel under perceptual constraints. In *IEEE International Conference on Image Processing (ICIP)*, 2011.

[ANP+11b] J. Abot, M. Nauge, C. Perrine, M.C. Larabi, C. Bergeron, C. Olivier, Y. Pousset, et al. Maximisation perceptuelle de la qualité de transmission JPWL via un canal MIMO réaliste. *Dans Groupement de Recherche en Traitement du Signal et des Images (GRESTI)*, 2011.

[Bea76] P.R. Beaudet. Context dependent interpolation. In *Image Science Mathematics Symposium. November*, 1976.

[Bea78] P.R. Beaudet. Rotationally invariant image operators. In *International Joint Conference on Pattern Recognition*, volume 579, pages 579–583, 1978.

[Bec73] P. Beckmann. *Orthogonal Polynomials for Engineering and Physicists*. The Golem Press, 1973.

[Bio] Bioinformaticssite. http://www.bioinformatics.org/oeil-couleur/dossier/index.html.

[BTVG06] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *European Conference on Computer Vision (ECCV)*, pages 404–417, 2006.

[BZCM08] N. J. Butko, L. Zhang, G. W. Cottrell, and J. R. Movellan. Visual saliency model for robot cameras. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2398–2403, may 2008.

[CA05] P. Le Callet and F. Autrusseau. Subjective quality assessment IRCCyN/IVC database, 2005. http://www.irccyn.ec-nantes.fr/ivcdb/.

[CGK+11] R. Calinescu, L. Grunske, M. Kwiatkowska, R. Mirandola, and G. Tamburrelli. Dynamic qos management and optimization in service-based systems. *IEEE Transactions on Software Engineering*, 37(3):387–409, 2011.

[CH07] D.M. Chandler and S.S. Hemami. Vsnr: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing (ICIP)*, 16(9):2284–2298, 2007.

[CO03] J. Caviedes and F. Oberti. No-reference quality metric for degraded and enhanced video. In *SPIE Visual Communications and Image Processing*, 2003.

[com86] JPEG committee, 1986. http://www.jpeg.org/jpeg/index.html.

[Con04] ICC International Color Consortium. specification 1: 2004-10. *Image technology colour management-Architecture, profile format, and data structure*, 2004.

[CPF95] J.-P. Cocquerez and Sylvie P.-F. *Analyse d'images : filtrage et segmentation*. Masson, 1995.

[CPV05] Y. Chartois, Y. Pousset, and R. Vauzelle. A SISO and MIMO radio channel characterization with 3d ray tracing propagation model in urban environment. *The European conference on propagation and systems (ECPS)*, 2005.

[CR68] F.W. Campbell and J.G. Robson. Application of fourier analysis to the visibility of gratings. *The Journal of Physiology*, 197(3):551, 1968.

[Cro84] F.C. Crow. Summed-area tables for texture mapping. *Computer Graphics*, 18(3):207–212, 1984.

[CRT01] N. Chiurtu, B. Rimoldi, and E. Telatar. On the capacity of multi-antenna gaussian channels. *IEEE International Symposium on Information Theory*, page 53, 2001.

[CS03] P.-A. Champin and C. Solnon. Measuring the similarity of labeled graphs. In *International Conference on Case-Based Reasoning*, pages 80–95. Springer, 2003.

[CWB+04] Y.C. Chung, J.M. Wang, R.R. Bailey, S.W. Chen, and S.L. Chang. A non-parametric blur measure based on edge analysis for image processing applications. In *IEEE Conference on Cybernetics and Intelligent Systems*, volume 1, pages 356–360. IEEE, 2004.

[Dal93] S. Daly. The visible differences predictor: an algorithm for the assessment of image fidelity. *Digital images and human vision*, 4:124–125, 1993.

[Dal94] S. Daly. A visual model for optimizing the design of image processing algorithms. In *International Conference on Image Processing (ICIP)*, volume 2, pages 16–20. IEEE, 1994.

[Dan51] G. B. Dantzig. Application of the simplexmethod to a transportation problem. In *Activity Analysis of Production and Allocation, JohnWiley and Sons*, pages 359–373, 1951.

[DBN+04] O. Déforges, M. Babel, N. Normand, B. Parrein, J. Ronsin, J.P. Guédon, L. Bédat, et al. Le lar aux mojettes. pages 165–168, 2004.

[Deo74] N. Deo. *Graph Theory with Applications to Engineering and Computer Science (Prentice Hall Series in Automatic Computation)*. Prentice-Hall, Inc., 1974.

[DKL84] A.M. Derrington, J. Krauskopf, and P. Lennie. Chromatic mechanisms in lateral geniculate nucleus of macaque. *The Journal of Physiology*, 357(1):241–265, 1984.

[DN82] L. Dreschler and H.H. Nagel. Volumetric model and 3d trajectory of a moving car derived from monocular tv frame sequences of a street scene. *Computer Graphics and Image Processing*, 20(3):199–228, 1982.

[Dub] B. Dubuc. Le cerveau à tous les niveaux. http://lecerveau.mcgill.ca/.

[DVKG+00] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, and A.C. Bovik. Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing (ICIP)*, 9(4):636–650, 2000.

[EAP+06] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli. Two new full-reference quality metrics based on hvs. In *Second International Workshop on Video Processing and Quality Metrics*, Scottsdale USA, 2006.

[ECR66] C. Enroth-Cugell and J.G. Robson. The contrast sensitivity of retinal ganglion cells of the cat. *The Journal of Physiology*, 187(3):517–552, December 1966.

[ELZ+10] U. Engelke, H. Liu, H.J. Zepernick, I. Heynderickx, and A. Maeder. Comparing two eye-tracking databases: The effect of experimental setup and image presentation time on the creation of saliency maps. In *Picture Coding Symposium (PCS)*, pages 282–285. IEEE, 2010.

[EMZ09] U. Engelke, A. Maeder, and H.J. Zepernick. Visual attention modelling for subjective image quality databases. In *IEEE In-*

ternational Workshop on Multimedia Signal Processing, pages 1–6. IEEE, 2009.

[Eng08] U. Engelke. *Perceptual quality metric design for wireless image and video communication*. Department of Signal Processing, School of Engineering, Blekinge Institute of Technology, 2008.

[EZ11] U. Engelke and H.n Zepernick. Quality of experience of multimedia services: Past, present, and future. In *QoE for Multimedia Content Sharing at European Interactive TV (EuroITV)*, 2011.

[FC75] J. Feldman and J.D. Cowan. Large-scale activity in neuralnets i : Theory with application to motoneuron pool responses. *Biological Cybernetics*, 1(17):29–38, 1975.

[FF56] L.R. Ford and D. R. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics 8*, 11(20):399–404, 1956.

[FMLBR05] C. Fernandez-Maloigne, M.C. Larabi, B. Bringier, and N. Richard. Spatio temporal characteristics of the human color perception for digital quality assessment. 1:203–206, 2005.

[FYZ+11] Z. Fang, D. Yang, W. Zhang, H. Chen, and B. Zang. A comprehensive analysis and parallelization of an image retrieval algorithm. In *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 154–164, april 2011.

[GHT11] S. Gauglitz, T. Höllerer, and M. Turk. Evaluation of interest point detectors and feature descriptors for visual tracking. *International journal of computer vision*, 94(3):335–360, 2011.

[GPRZ07] P. Gastaldo, G. Parodi, J. Redi, and R. Zunino. No-reference quality assessment of jpeg images by using cbp neural networks. *Artificial Neural Networks (ICANN)*, pages 564–572, 2007.

[GW89] C.D. Gilbert and T.N. Wiesel. Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *Journal of Neuroscience*, 7(9):2432–2442, 1989.

[Har11] J. Harel. A saliency implementation in matlab. urlhttp://www.klab.caltech.edu/ harel/share/gbvs.php, 2011.

[HKS] Y. Horita, Y. Kawayoke, and Z. M. Parvez Sazzad. Image quality evaluation database. http://mict.eng.u-toyama.ac.jp/mict/index2.html.

[HM06] J.A. Hirsch and L.M. Martinez. Circuits that build visual cortical receptive fields. *Trends in Neurosciences*, 1(29):30–39, 2006.

[HS88a] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Manchester, UK, 1988.

[HS88b] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.

[HS04] A. Hakeem and M. Shah. Ontology and taxonomy collaborated framework for meeting classification, 2004.

[HW62] D.H. Hubel and T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, (160):106–154, 1962.

[HW68] D.H. Hubel and T.N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 1:215–243, 1968.

[IEA00] IEA international ergonomics association, 2000. http://www.iea.cc/01_what/What is Ergonomics.html.

[IKN98] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(11):1254–1259, 1998.

[ISO98] ISO. Ergonomic requirements for office work with visual display terminals (VDTs)–part 11: Guidance on usability, 1998.

[ITU84] ITU-R. Recommendation bt.601 : Encoding parameters of digital television for studios, 1984.

[ITU02] ITU-R. Recommendation BT.500-11 : Methodology for the subjective assessment of the quality of television pictures, 2002.

[ITU05] ITU-R. Document 6Q/131-E : Comparison of DSCQS and ACR, 2005.

[ITU08] ITU-T. Recommendation P.910 : Subjective video quality assessment methods for multimedia applications, 2008.

[ITU09] ITU-R. Recommendation BT.500-12 : Methodology for the subjective assessment of the quality of television pictures, 2009.

[JBZ95] M. Karpinski R. Karp M. Luby J. Blomer, M. Kalfane and D. Zuckerman. Technical report tr-95-048. Technical report, International Computer Science Institute, 1995.

[JEDT09] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[JPE91] JPEG. ITU-T recommendation T.81 | iso/iec is 10918-1, 1991. http://www.jpeg.org/jpeg/.

[JPE00] JPEG2000. Iso/cei 15444-1, 2000. http://www.jpeg.org/jpeg2000/.

[JPE09] JPEG XR. Iso/iec 29199, 2009.

[JPW91] JPWL. Iso/iec 15444-11:2007; jpeg2000 image coding system - part 11: Wireless jpeg2000, 1991. http://www.jpeg.org/jpeg/.

[JT80] C. L. Jackins and S. L. Tanimoto. Oct-trees and their use in representing three-dimensional objects. *Comput. Graphics and Image Processing*, 14(3):249–270, 1980.

[KOD09] C. Keimel, T. Oelbaum, and K. Diepold. Improving the verification process of video quality metrics. In *Quality of Multimedia Experience (QoMEx)*, Jul 2009.

[Koe84] J.J. Koenderink. The structure of images. *Biological cybernetics*, 50(5):363–370, 1984.

[KR82] L. Kitchen and A. Rosenfeld. Gray-level corner detection. *Pattern Recognition Letters*, 1(2):95–102, 1982.

[KSJ00] E.R. Kandel, J.H. Schwartz, and T.M. Jessell. Mcgraw-hill medical. *Principles of Neural Science*, 2000.

[KU85] C. Koch and S. Ullman. hifts in selective visual attention : towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[Kuf53] S.W. Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, 1(16):37–68, 1953.

[Kus05] T. M. Kusuma. *A perceptual-based objective quality metric for wireless imaging*. Curtin University of Technology, Perth, Australia, 2005.

[KZ03] T. M. Kusuma and H.-J. Zepernick. On perceptual objective quality metrics for in-service picture quality monitoring. In *3rd ATcrc Telecommunications and Networking Conference and Workshop*, 2003.

[LC10] E.C. Larson and D.M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, 2010.

[Lin90] T. Lindeberg. Scale-space for discrete signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 12(3):234–254, 1990.

[Lin93] T. Lindeberg. Discrete derivative approximations with scale-space properties: A basis for low-level feature extraction. *Journal of Mathematical Imaging and Vision*, 3(4):349–376, 1993.

[LMLCBT06] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(5):802–817, 2006.

[Low99] D.G. Lowe. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[Low04] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[Lub95] J. Lubin. A visual discrimination model for imaging system design and evaluation. *Vision models for target detection and recognition*, 2:245–357, 1995.

[M+01] R.H. Masland et al. The fundamental plan of the retina. *Nature neuroscience*, 4:877–886, 2001.

[M+03] C. Murray et al. Oracle® spatial user's guide and reference 10g release 1 (10.1). *Redwood City, Oracle Corporation*, page 602, 2003.

[Mar74] D. Marr. The computation of lightness by the primate retina. *Vision Research*, 14(12):1377–1388, 1974.

[Mas01] R.H. Masland. Neuronal diversity in the retina. *Current opinion in neurobiology*, 11(4):431–436, 2001.

[MB09] A.K. Moorthy and A.C. Bovik. Visual importance pooling for image quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):193–201, 2009.

[MDWE02] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi. A no-reference perceptual blur metric. In *International Conference on Image Processing (ICIP)*, volume 3, pages III–57. IEEE, 2002.

[Mey90] Y. Meyer. *Ondelettes et opérateurs*, volume I. 1990.

[MF77] J.W. Modestino and R.W. Fries. Edge detection in noisy images using recursive digital filtering. *Computer graphics and image processing*, 6(5):409–433, 1977.

[MGPB+05] I. McCowan, Da. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 27(3):305–317, 2005.

[MM07] R.H. Masland and P.R. Martin. The unsolved mystery of vision. *Current biology*, 17(15):R577–R582, 2007.

[MMS04] R. Mantiuk, K. Myszkowski, and H.P. Seidel. Visible difference predicator for high dynamic range images. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 3, pages 2763–2769. IEEE, 2004.

[Mon81] G. Monge. Mémoire sur la théorie des déblais et de remblais. In *Histoire de l'Académie Royale des Sciences de Paris*, pages 666–704, 1781.

[Mor77] H. Moravec. Towards automatic visual obstacle avoidance. In *International Joint Conferences on Artificial Intelligence*, page 584, Cambridge, 1977.

[Mor80] H. Moravec. *Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover*. Stanford University, 1980.

[MPD04] M.J. McMahon, O.S. Packer, and D.M. Dacey. The classical receptive field surround of primate parasol ganglion cells is mediated primarily by a non-gabaergic pathway. *The Journal of neuroscience*, 24(15):3736–3745, 2004.

[MS01] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 525–531. IEEE, 2001.

[MS04] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International journal of computer vision*, 60(1):63–86, 2004.

[MS05] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(10):1615–1630, 2005.

[MT09] R. Melnyk and R. Tushnytskyy. Image classification by pattern and structure features clustering. In *International Journal of Computing*, pages 52–60, 2009.

[MVF12] M.G. Martini, B. Villarini, and F. Fiorucci. A reduced-reference perceptual image and video quality metric based on edge preservation. *Journal on Advances in Signal Processing (Eurasip)*, 66, 2012.

[MYL+11] P. Mainali, Q. Yang, G. Lafruit, L. Van Gool, and R. Lauwereins. Robust low complexity corner detector. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(4):435–445, 2011.

[Nad00] M. Nadenau. *Integration of human color vision models intp high quality image compression*. Ecole polytechnique federale de Lausanne, 2000.

[Nag83] H.H. Nagel. Displacement vectors derived from second-order intensity variations in image sequences. *Computer Vision, Graphics, and Image Processing (CGVIP)*, 21(1):85–117, 1983.

[NLF10a] M. Nauge, M.-C. Larabi, and C. Fernandez. A hierarchical saliency map generation based on the human visual system properties. In *Workshop on Picture Coding and Image Processing (WPCIP)*, 2010.

[NLF10b] M. Nauge, M.C. Larabi, and C. Fernandez. A reduced-reference metric based on the interest points in color images. In *Picture Coding Symposium (PCS)*, pages 610–613. IEEE, 2010.

[NLF11a] M. Nauge, M. Larabi, and C. Fernandez. Quality estimation based on interest points through hierarchical saliency maps. In *European Workshop on Visual Information Processing (EUVIP)*, pages 186–191. IEEE, 2011.

[NLF11b] M. Nauge, M.-C. Larabi, and C. Fernandez. Imqual: A web-service dedicated to image quality evaluation and metrics bench-

mark. In *IS&TSPIE Electronic Imaging, Image Quality and System Performance VIII*, Etats-Unis, 2011.

[NLF12] M. Nauge, M.-C. Larabi, and C. Fernandez. A statistical study of the correlation between interest points and gaze points. In *IS&T/SPIE Electronic Imaging, Human Vision and Electronic Imaging XVII (HVEI)*, 2012.

[NLMLCB07] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barbba. Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric. 2:169–172, 2007.

[Nob88] J.A. Noble. Finding corners. *Image and Vision Computing*, 6(2):121–128, 1988.

[N.R09] N.Ramin. *Vers une métrique sans reference de la qualité patiale d'un signal vidéo dans un contexte multimedia*. 2009.

[NS97] I. Novikov and E.M. Semenov. *Haar series and linear operators*. Mathematics and its applications. Kluwer Academic, 1997.

[oQoEiMSS] QUALINET European Network on Quality of Experience in Multimedia Systems and Services. Cost action ic 1003. http://www.qualinet.eu/.

[Par07] E. Pardoux. *Processus de Markov et applications*. Dunod, 2007.

[PBT09] J. Petit, R. Brémond, and J.P. Tarel. Saliency maps of high dynamic range images. In *Proceedings of the 6th Symposium on Applied Perception in Graphics and Visualization*, pages 134–134. ACM, 2009.

[PH09] M. Pedersen and J.Y. Hardeberg. Survey of full-reference image quality metrics. In *Global Congress on Intelligent Systems (GCIS)*, Gjovik, Norway, June 2009.

[Pic00] R.W. Picard. Toward computers that recognize and respond to user emotion. *IBM Systems Journal*, 39(3.4):705–719, 2000.

[Pic01] R.W. Picard. Computers that sense, recognize, and respond to human emotion. 2001.

[PLE+08] N. Ponomarenko, V. Lukin, K. Egiazarian, J. Astola, M. Carli, and F. Battisti. Color image database for evaluation of image quality metrics, 2008. http://www.ponomarenko.info/tid2008.htm.

[QaG11] M.T. Qadri and K.T. Tan andM. Ghanbari. The impact of spatial masking in image quality meters. *Global Journal of Computer Science and Technology*, 11(20), 2011.

[RCFM09] N. Richard, A.S. Capelle, and C. Fernandez-Maloigne. A complete scheme for colour morphology with perceptual integration. In *SPIE Optical Engineering + Applications*, pages 744319–744319. International Society for Optics and Photonics, 2009.

[RLFM08] V. Rosselli, M.-C. Larabi, and C. Fernandez-Maloigne. Modelling the anisotropic contrast sensitivity by the estimation of the perception threshold. *AIC Interim Meeting - Colour Effects & Affects*, 2008.

[RRFM07] E. Rollo, N. Richard, and C. Fernandez-Maloigne. perceptual colour image watershed. In *26th session of CIE, Beijing, China*, 2007.

[RTG98] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *IEEE International Conference on Computer Vision (ICCV)*, pages 59–66, Jan 1998.

[RW98] P. Robertson and T. Worz. Bandwidth-efficient turbo trellis-coded modulation using punctured component codes. *IEEE Journal on Selected Areas in Communications*, 16(2):206–218, 1998.

[SB06] H.R. Sheikh and A.C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing (ICIP)*, 15(2):430–444, 2006.

[SBC10] M.A. Saad, A.C. Bovik, and C. Charrier. A dct statistics-based blind image quality index. *IEEE Signal Processing Letters*, 17(6):583–586, 2010.

[SDG79] K.S. Shanmugam, F.M. Dickey, and J.A. Green. An optimal frequency domain filter for edge detection in digital pictures. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, (1):37–49, 1979.

[SF94] E. Shusterman and M. Feder. Image compression via improved quadtree decomposition algorithms. *IEEE Transactions on Image Processing (ICIP)*, 3(2):207–215, 1994.

[SH92] L.G. Shapiro and R. Haralick. Computer and robot vision. *Addison-Wesley*, 1992.

[She04] H. R. Sheikh. *Image Quality Assessment Using Natural Scene Statistics*. University of Texas at Austin, 2004.

[SHJB05] M.F. Sabir, R.W. Heath Jr, and A.C. Bovik. Unequal power allocation for JPEG transmission over MIMO systems. In *Conference on Signals, Systems and Computers*, pages 1608–1612. Citeseer, 2005.

[SKC+05] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. Technical report, DTIC Document, 2005.

[ST94] J. Shi and C. Tomasi. Good features to track. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600. IEEE, 1994.

[SV00] S. Saha and R. Vemuri. An analysis on the effect of image activity on lossy coding performance. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, volume 3, pages 295–298. IEEE, 2000.

[SWCBa] H.R. Sheikh, Z. Wang, L. Cormack, and A.C. Bovik. LIVE image quality assessment database release 1. http://live.ece.utexas.edu/research/quality.

[SWCBb] H.R. Sheikh, Z. Wang, L. Cormack, and A.C. Bovik. LIVE image quality assessment database release 2. http://live.ece.utexas.edu/research/quality.

[Tat07] B.W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14), 2007.

[TG80] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 113:97–316, 1980.

[TKCT10] Y. Tong, H. Konik, F. Cheikh, and A. Tremeau. Full reference image quality assessment based on saliency map analysis. *Journal of Imaging Science*, 54(3):30503–30503, 2010.

[TP86] V. Torre and T.A. Poggio. On edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, (2):147–163, 1986.

[Tre85] A. Treisman. Preattentive processing in vision. *Computer Vision,Graphics and Image Processing*, (31):156–177, 1985.

[Tre91] A. Treisman. Search, similarity, and integration of features between and within dimensions. *Journal of Experimental Psychology : Human Perception and Performance*, 3(17):652–676, 1991.

[TS85] A. Treisman and J. Souther. Search asymmetry : a diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology-General*, (114):285–310, 1985.

[TS02] J.B. Troy and T. Shou. The receptive fields of cat retinal ganglion cells in physiological and pathological states: where we are after half a century of research. *Progress in retinal and eye research*, 21(3):263, 2002.

[Urb] M. Urban. Harris interest operator. http://cmp.felk.cvut.cz/cmp/courses/dzo/resources/.

[VJ01] P. Viola and M. Jones. Robust real-time object detection. *IJCV*, 2001.

[VJ04] P. Viola and M.J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[VQE08] VQEG. Phase i : Final report from the video quality experts group on the validation of objective models of multimedia quality assessement, 2008.

[Wat87] A.B. Watson. The cortex transform: rapid computation of simulated neural images. volume 39, pages 311–327. Elsevier, 1987.

[WBE00] Z. Wang, A.C. Bovik, and BL Evan. Blind measurement of blocking artifacts in images. In *International Conference on Image Processing (ICIP)*, volume 3, pages 981–984. IEEE, 2000.

[WBSS04] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing (ICIP)*, 13(4):600–612, 2004.

[WCF89] J. M. Wolfe, K. R. Cave, and S. L. Franzel. Guided search : an alternative to the feature integration model for visual search. *Journal of Experimental Psychology : Human Perception and Performance*, 15(3):419–433, 1989.

[Wid84] H. Widdel. Theoretical and applied aspects of eye movement research, chapter operational problems in analysing eye movements. *Elsevier*, pages 21–29, 1984.

[Win09] S. Winkler. On the properties of subjective ratings in video quality experiments. In *Quality of Multimedia Experience (QoMEx)*, pages 139–144. IEEE, 2009.

[WJY09] S. Wan, P. Jin, and L. Yue. An approach for image retrieval based on visual saliency. In *Image Analysis and Signal Procession (IASP)*, pages 172–175, april 2009.

[WS05] Z. Wang and E.P. Simoncelli. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In *SPIE Human Vision and Electronic Imaging*, volume 5666, pages 149–159, 2005.

[Yar67] A. L. Yarbus. Eye movements and vision. In *Plenum Press*, 1967.

[Yee04] H. Yee. Perceptual metric for production testing. *Journal of Graphics Tools*, 9(4):33–40, 2004.

[You87] R.A. Young. The Gaussian derivative model for spatial vision: I. Retinal mechanisms. *Spatial Vision*, pages 273–293, 1987.

[ZH83] O. Zuniga and R. Haralick. Corner detection using the facet model. *IEEE Conference on Computer Vision Pattern Recognition*, pages 30–37, 1983.

Bibliographie de l'auteur

1 Revues internationales avec comité de lecture

- **M. Nauge**, M.-C. Larabi, C. Fernandez. Can we build speed up saliency map using interest points detectors ? *Journal of Electronic Imaging (JEI)*, en soumission (2012).
- **M. Nauge**, M.-C. Larabi, C. Fernandez. Maximizing QoE on realistic MIMO channel. *Journal of Visual Communications and Image Processing (VCIP)*, en préparation (2013).
- **M. Nauge**, M.-C. Larabi, C. Fernandez. Caractérisation d'une image par migrations structurelles. *IEEE Pattern Analysis and Machine Intelligence (PAMI)*, en préparation (2013).

2 Conférences internationales avec comité de lecture

- **M. Nauge**, M.-C. Larabi, C. Fernandez. A Reduced-Reference Metric Based on the Interest Points in Color Images. *Picture Coding Symposium (PCS)*, Décembre 2010. Pages 610-613. Nagoya, Japon.
- **M. Nauge**, M.-C. Larabi, C. Fernandez. A Hierarchical Saliency Map Generation Based on the Human Visual System Properties. *Workshop on Picture Coding and Image Processing (WPCIP)*, Décembre 2010, Nagoya, Japon.

- **M. Nauge**, M.-C. Larabi, C. Fernandez. A Web-Service Dedicated to Image Quality Evaluation and Metrics Benchmark. *IS&T/SPIE Electronic Imaging, Image Quality and System Performance VIII*, 2011, San Francisco, Etats-Unis.
- J. Abot, **M. Nauge**, C. Perrine, M.-C. Larabi, C. Bergeron, C. Olivier, Y. Pousset. A Robust Content-Based JPWL Transmission Over a Realistic MIMO Channel Under Perceptual Constraints. *18th IEEE International Conference on Image Processing (ICIP)*, 2011, Bruxelles, Belgique.
- **M. Nauge**, M.-C. Larabi, C. Fernandez. Quality Estimation Based on Interest Points Through Hierarchical Saliency Maps, *European Workshop on Visual Information Processing (EUVIP)*, Juillet 2011. Pages 186-191, Paris, France.
- **M. Nauge**, M.-C. Larabi, C. Fernandez. A statistical study of the correlation between interest points and gaze points. *IS&T/SPIE Electronic Imaging, Human Vision and Electronic Imaging XVII (HVEI)*, Janvier 2012, San Francisco, Etats-Unis.

3 Conférences nationales avec comité de lecture

- **M. Nauge**, M.-C. Larabi, C. Fernandez. Benchmark de métriques de qualité sur bases de données d'images compressées. *Compression et Représentation des Signaux Audiovisuels (CORESA)*, Octobre,2010, France.
- J. Abot, **M. Nauge**, C. Perrine, M.-C. Larabi, C. Bergeron, C. Olivier, Y. Pousset. Maximisation Perceptuelle de la Qualité de Transmission JPWL via un Canal MIMO Réaliste. *Groupement de Recherche en Traitement du Signal et des Images (GRETSI)*, 2011, France.

4 Conférences sans acte

- **M. Nauge**, M.-C. Larabi, C. Fernandez. Caractérisation des distorsions par l'exploitation des points d'intérêt. *GDR 720 Information, Signal, Images et ViSion, (ISIS)*, Septembre 2012, Télécom Paristech Paris.
- **M. Nauge**, M.-C. Larabi. Subjective evaluation of potential coding technologies, *WG1N5682, ISO/IEC JTC1 SC29/WG1 Tokyo meeting*, Février 2011.