RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Khider – BISKRA
Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie
Département d'informatique

N° d'ordre :.....

T H È S E

Présentée en vue de l'obtention du diplôme de
Doctorat LMD en Informatique

Option : Sciences et techniques de l'image

Titre

L'apport de la perception/l'attention visuelle à l'amélioration de la fusion d'images multi-focales

Par

BABAHENINI Sarra

Devant le jury composé de :

DJEROU Leila	**Professeur**	**Université de Biskra**	**Président**
CHERIF Foudil	**Professeur**	**Université de Biskra**	**Rapporteur**
DJEFFAL Abdelhamid	**Professeur**	**Université de Biskra**	**Examinateur**
MELKEMI Kamal Eddine	**Professeur**	**Université de Batna 2**	**Examinateur**

Année universitaire : **2020-2021**

ملخص الأطروحة

حتى يومنا هذا، تم تطوير العديد من طرق دمج الصور متعددة البؤر. يعتبر حساب متوسط الصور المصدر بكسلًا ببكسل أبسط طريقة مزج الصور ببعضها البعض، ولكن هذه الطريقة تؤدي عمومًا إلى تأثيرات غير مرغوب فيها مثل تقليل تباين الصورة المدمجة.

بشكل عام ، يمكن تصنيف طريقة اندماج الصور متعددة البؤر في المجال المكاني ومجال التحويل. عادةً ما تكون تقنية دمج الصور متعددة البؤر التي تعطي صورة مدمجة عالية الدقة معقدة وتستغرق وقتًا طويلاً.

في هذه الأطروحة ، نقوم بتطوير تقنيات دمج الصور متعددة البؤر وهي غير مكلفة ولا تستغرق الكثير من الوقت ، ولكنها تؤدي إلى صورة مدمجة عالية الجودة. لهذا نستخدم مناهج تستند إلى الإدراك البصري، وحساب البروز البصري بعدة طرق ودمجها في حساب دمج الصورة.

الكلمات المفتاحية: نظام الرؤية البشرية، كشف البروز البصري، خريطة البروز، اندماج صورة متعدد البؤر، خريطة الوزن، تحلل المصفوفة.

Résumé de la thèse

A ce jour, de nombreuses méthodes de fusion d'images multi-focales ont été développées. La méthode de fusion la plus simple consiste à faire la moyenne des images sources pixel par pixel mais cette méthode conduit généralement à des effets indésirables tels que la réduction du contraste de l'image fusionnée. Généralement, la méthode de fusion d'images multi-focales peut être classée en domaine spatial et domaine de transformation. La technique de fusion d'images multi-focales qui donne une image fusionnée de haute précision est généralement compliquée et prend beaucoup de temps.
Dans cette thèse, nous développons des techniques de fusion d'images multi-focales qui sont peu coûteuses et ne prennent pas beaucoup de temps, mais il en résulte une image fusionnée de haute qualité. Pour cela nous utilisons des approches basées sur la perception visuelle, en calculant la saillance visuelle de plusieurs façons et en l'intégrant dans le calcul de fusion d'images.

Mots clés: système visuel humain, détection de la saillance visuelle, carte de saillance, fusion d'images multi-focales, cartes de poids, transformations en contourlet, matrice de décomposition.

Abstract of the thesis

To date, many multi-focus image fusion methods have been developed. The simplest fusion method is to average the source images pixel by pixel, but this method generally leads to unwanted effects such as reducing the contrast of the fused image. Generally, the multi-focus image fusion method can be classified into spatial domain and transformation domain. The multi-focus image fusion technique which gives a high precision fused image is usually complicated and time consuming.
In this thesis, we develop multi-focus image fusion techniques which are inexpensive and do not take much time, but results in a high quality fused image. For this we use approaches based on visual perception, calculating visual salience in several ways and integrating it into the image fusion calculation.

Keywords: human vision system; visual saliency detection; saliency map; multi-focus image fusion; weight map; contourlet transform; matrix decomposition.

Articles et conférences

- Sarra Babahenini, Foudil Cherif, and Fella Charif. Comparative Study of Noise Robustness in Visual Attention Models. In International Conference on Electrical Engineering and Control Applications, pages 1259–1269. Springer, 2019
- Sarra Babahenini, Foudil Cherif, Fella Charif, Abdelmalik Taleb-Ahmed, and Yassine Ruichek. Using saliency detection to improve Multi-Focus Image Fusion. International Journal of Signal and Imaging Systems Engineering, 12(3): 81-92, 2021.

Remerciements

A l'issue de la rédaction de cette recherche, je suis convaincue que la thèse est loin d'être un travail solitaire. En effet, je n'aurais jamais pu réaliser ce travail doctoral sans le soutien d'un grand nombre de personnes dont la générosité, la bonne humeur, et l'intérêt manifesté à l'égard de ma recherche m'ont permis de progresser dans cette phase délicate de "chercheur". Je souhaiterais tout d'abord, remercier, et surtout dédier cette thèse à mes parents, mon mari et mes enfants Mohamed Ilyes et Mohamed Racim. Sans tous les sacrifices qu'ils ont consentis, leur soutien et leur amour, rien de tout cela n'aurait été possible.

Je tiens à remercier mon directeur de thèse, Pr CHERIF Foudil pour la confiance qu'il m'a accordé en acceptant d'encadrer ce travail de doctorat, pour ses multiples conseils, sa patience et son soutien.

Je voudrais également remercier : Dr CHARIF Fella et Pr TALEB-AHMED Abdelmalik pour leur engagement et leur aide. Ils m'ont transmis leur savoir pour mener à bien cette thèse.

Je remercie également les membres de jury Pr DJEROU Leila, Pr DJEFFAL Abdelhamid et Pr MELKEMI Kamal Eddine pour l'honneur qu'ils m'ont fait en acceptant de juger notre travail.

Au passage, je remercie mes sœurs Hadjer et Cherifa, ma chère tante Farah et mes beaux parents qui m'ont soutenu et m'ont appuyé moralement.

Enfin, merci à mes collègues et mes amis pour les moments inoubliables qu'on a passé ensemble.

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1

Introduction

1.1 Contexte de la thèse

La perception visuelle, en tant qu'outil de gestion de l'information ; nous permet de la modélisation des parties attirantes selon la vision humaine. La détection de saillance de l'image est la clé de l'extraction des informations d'image. L'extraction des régions de saillance d'image est un modèle de perception visuelle, nécessaire dans la plupart des méthodes de traitement d'image basées sur le contenu de l'image, car les composants d'image importants fournissent l'information la plus complète sur une image entière. Par conséquent, l'extraction des régions saillantes des images facilite efficacement de nombreuses applications d'image telles que la récupération d'image, la compression d'image adaptative, la reconnaissance d'objet, le redimensionnement d'image, « l'amélioration des images ».
Les humains peuvent facilement se concentrer sur les parties saillantes des images selon l'expérience et le jugement, mais les machines sont incapables de reproduire précisément une telle capacité. De nombreux chercheurs ont étudié cette question sur la base de la biologie, la physiologie et la neurobiologie. Dans ces études, certaines caractéristiques que les régions saillantes devraient avoir, y compris l'unicité, le caractère aléatoire, et des caractéristiques surprenantes, sont acquises.
La fusion d'images, une branche importante de la fusion de données, est le processus de combinaison d'informations pertinentes de deux images ou plus en une seule image où l'image résultante sera plus informative que n'importe laquelle des images d'entrée. L'image résultante devrait être plus adaptée à la perception visuelle et à la perception machine ou au traitement informatique. L'objectif de la fusion d'images est de réduire l'incertitude et de minimiser la redondance dans la sortie ainsi que de maximiser les informations pertinentes particulières à une application ou à une tâche.

La clé de la fusion d'images réside dans le choix d'une méthode de fusion fiable et efficace pour déterminer le coefficient de fusion. De nos jours, avec les progrès rapides de la technologie, il est désormais possible d'obtenir des informations à partir d'images multi-sources pour produire une information de haute qualité à partir d'un ensemble d'images. Cependant, en raison de la profondeur de champ limitée des lentilles optiques dans les appareils photo, il n'est souvent pas possible d'obtenir une image contenant tous les objets pertinents « au point » afin qu'une scène puisse être prise dans un ensemble d'images avec différents réglages de mise au point de chaque image. Outre les solutions utilisant l'optique spécialisée, et l'imagerie computationnelle, la façon de résoudre ce problème est la fusion d'images multi-focales.
La fusion d'images multi-focales est une branche de la fusion d'images qui intègre plusieurs images sources avec différents réglages de mise au point sur la même scène dans une image composite contenant tous les objets mis au point. L'objectif de la fusion d'images multi-focales est de produire une image qui contient tous les objets pertinents mis au point en extrayant et en synthétisant les objets focalisés des images sources. L'hypothèse de base de la fusion d'images multi-focales est que l'objet focalisé est plus net que l'objet non focalisé, et la netteté est liée à certaines mesures d'informations calculées. Au cours de la dernière décennie, un certain nombre de mesures de netteté pour la fusion d'images multi-focales ont été proposées.
L'objectif de notre travail est de développer des techniques de fusion d'images multi-focales qui donnent une image fusionnée de haute précision qui convient mieux à la perception humaine ou machine et à d'autres tâches de traitement d'images.

1.2 Enoncé du problème

A ce jour, de nombreuses méthodes de fusion d'images multi-focales ont été développées. La méthode de fusion la plus simple consiste à faire la moyenne des images sources pixel par pixel mais cette méthode conduit généralement à des effets indésirables tels que la réduction du contraste de l'image fusionnée. Généralement, la méthode de fusion d'images multi-focales peut être classée en domaine spatial et domaine de transformation. La technique de fusion d'images multi-focales qui donne une image fusionnée de haute précision est généralement compliquée et prend beaucoup de temps, ce qui est d'une importance vitale pour la qualité de la fusion. Dans cette thèse, nous développons des techniques de fusion d'images multi-focales qui sont peu coûteuses et ne prennent pas beaucoup de temps, mais il en résulte une image fusionnée de haute qualité. Pour cela nous utilisons des approches

basées sur la perception visuelle, en calculant la saillance visuelle de plusieurs façons et en l'intégrant dans le calcul de fusion d'images.

1.3 Contributions

Dans cette thèse, nous proposons un nouveau cadre de la fusion d'images multi-focales par l'intégration de la saillance dans le calcul des algorithmes de fusion, ceci s'insère dans la technologie de fusion d'images au niveau des pixels. Afin d'atteindre l'objectif de la thèse, les contributions de ce travail sont les suivantes :

- Étude comparative de la robustesse au bruit dans les modèles d'attention visuelle, pour cela nous avons de corrompu une image par du bruit de différentes variances et étudié ensuite ses effets avec cinq métriques.
- Nous utilisons trois algorithmes de détection de saillance visuelle (VSD) pour extraire les informations de saillance à partir d'images visuellement différentes. Les résultats de ces algorithmes ont été utilisés comme une carte de poids, permettant de ne conserver dans l'image fusionnée que les régions focalisées et ensuite intégrés dans la technique de fusion d'images
- Proposition d'un algorithme de fusion d'images multi-focales basé sur un filtre d'image guidé rapide, qui utilise un échantillonnage des images d'entrées et effectue une reconstruction à la fin, permettant ainsi de réduire considérablement le temps de calcul sans perte de la qualité de fusion.

1.4 Organisation de la thèse

Après l'introduction dans le présent chapitre, le reste de la thèse est présenté comme suit :
Dans le chapitre 01 : propose une étude bibliographique sur les sujets de la perception visuelle, où nous allons présenter quelques propriétés de l'œil et du système visuelle humain, ainsi que les techniques de calcul et modélisation de l'attention visuelle.
Dans le chapitre 02 : nous allons exposer un état de l'art sur les différentes méthodes et les modèles de fusion d'image, ainsi qu'une étude théorique, suivi d'un bilan montrant la force et les faiblesses de chaque méthode.
Dans le chapitre 03 : nous allons examiner, si les effets du bruit ont une influence sur leur taux de reconnaissance, pour cela nous sélectionnons trois techniques de calcul récentes de la saillance visuelle : la détection de la saillance à l'aide de la technique de la carte de

saillance humaine, la détection de la saillance à l'aide de la technique de transformation de contour (CT) et la détection de la saillance à l'aide de la technique de décomposition matricielle de faible rang et structurée (LSMD), et nous évaluons les effets du bruit et leur influence sur le taux de reconnaissance en utilisant les métriques, basées sur les courbes AUC, F-mesure et MAE.

Dans, le chapitre 04 nous avons intégré trois techniques de détection de saillance dans notre méthode, le choix est justifié par le fait qu'elles sont structurellement différentes, ce qui nous permettra de comparer l'influence de l'attention visuelle avec notre objectif principal: la fusion d'images multi-focus.

Dans le chapitre 05, nous avons proposé, une nouvelle méthode de fusion d'images par filtrage guidé, la contribution clé est de présenter une méthode de fusion de filtre guidée multi-rapide pour réduire le temps.

Enfin, une conclusion générale et quelques perspectives clôturerons cette thèse.

Chapitre 2

Perception visuelle

2.1 Introduction

Le système humain traite l'information de son environnement au travers d'organes sensoriels dédiés. Mais la quantité d'information qui parvient à nos yeux à chaque instant est très élevée, cette information ne peut donc pas être traite en totalité. C'est pourquoi nous focalisons notre attention seulement sur une partie de l'information visuelle présente. Cette focalisation attentionnelle vers une région particulière du champ visuel va entraîner un mouvement de nos yeux vers cette région, nous permettant ainsi de la percevoir plus en détail c'est ce que l'on appelle l'attention "ouvert". Ce déploiement de l'attention visuelle implique deux mécanismes : un premier dépendant du stimulus, dit "ascendant" ou "Bottom-Up", et un second singulier à l'observateur, "descendant" ou "Top-Down".

2.2 Le système visuel humain

L'information visuelle est reçue par l'œil puis transmise au cerveau après différents traitements préliminaires réalisés par la rétine. Les traitements de plus haut niveau sont réalisés dans le cerveau au niveau des aires visuels que l'on trouve dans la partie arrière des deux hémisphères cérébraux figure 2.1. D'après les études physiologiques, les zones corticales dédiées au traitement de l'information visuelle occupent une place importante. De plus, les connexions avec les autres aires du cerveau font du système visuel un ensemble très complexe. La figure 2.1 montre les principaux traitements qui vont faire l'objet de notre étude :
Les traitements rétiniens puis après transmission de l'information via le nerf optique, les traitements effectués par la première aire du cortex visuel primaire (aire V1).

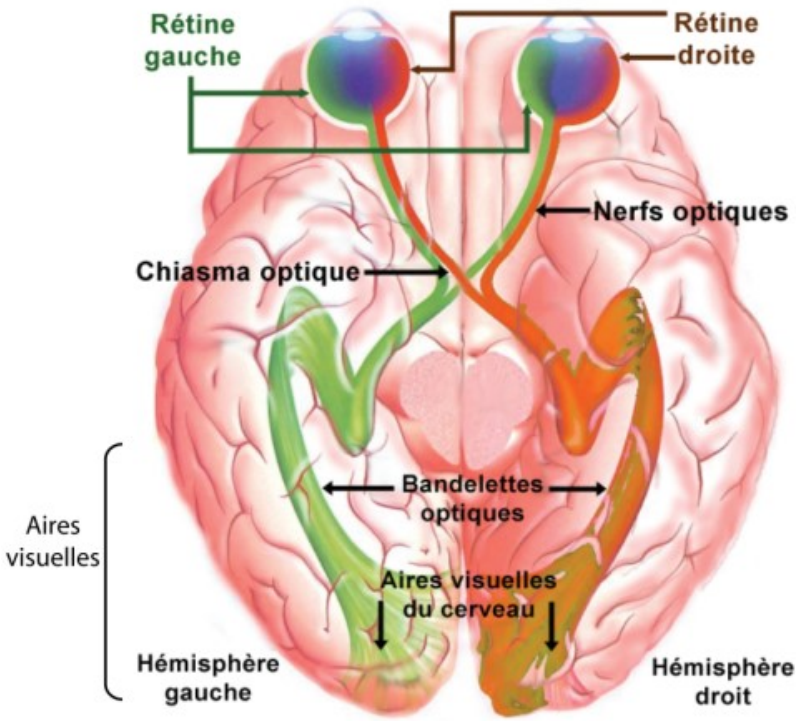


Figure 2.1: Schéma global du système visuel humain

L'œil est le capteur du système visuel. Il est composé d'un système optique adaptatif constitué des éléments suivants figure 2.2 :

- La cornée, barrière protectrice transparente entre le milieu extérieur et l'œil à proprement parler. Elle constitue l'un des premiers systèmes de focalisation.
- L'iris et le cristallin qui permettent respectivement la modulation de la quantité de lumière entrante et l'ajustement de la focale du système optique.
- La rétine est une fine membrane tapissant le globe oculaire. Par analogie avec un appareil photo, elle est en quelque sorte la « pellicule » de l'œil, chargée de capter les rayons lumineux pour les transmettre au système nerveux central..

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

La rétine est située sur le fond de l'œil. Elle est constituée d'un assemblage de cellules de différents types, organisées en couches et ayant des propriétés particulières. La figure

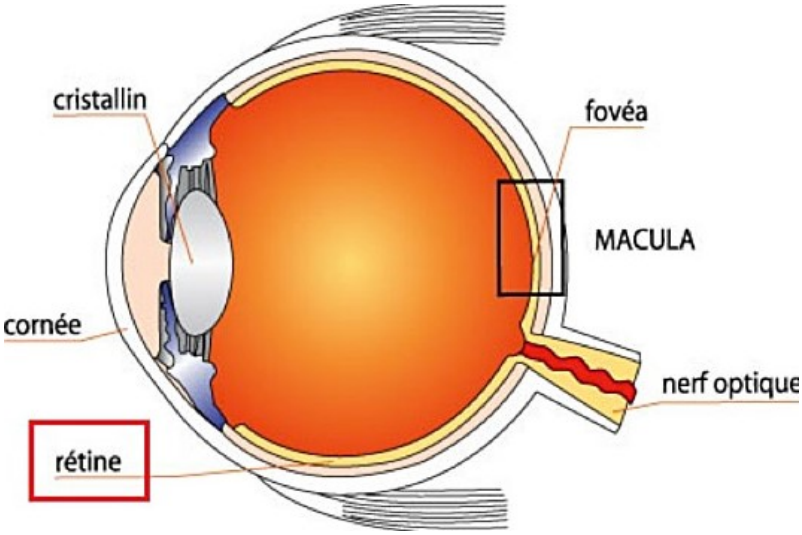


Figure 2.2: Schéma général de l'œil

2.2 montre l'organisation de ces couches cellulaires. On y trouve : la couche des photo-récepteurs (cônes et bâtonnets) qui captent l'information lumineuse, la couche des cellules horizontales, des cellules bipolaires et des cellules amacrines et enfin la couche des cellules ganglionnaires. Ces couches sont séparées par des zones de connexions synaptiques : la couche plexiforme externe (PLE) et la couche plexiforme interne (PLI)

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

- Les bâtonnets :
  - Ils sont responsables de la vision scotopique : c'est à dire une vision où l'intensité lumineuse est faible (vision de nuit) ;
  - Ils sont très sensibles à la lumière ;
  - Ils ont une réponse lente aux variations d'illumination.
- Les cônes :
  - Ils sont responsables de la vision photopique : vision dans les hautes intensités lumineuses (vision de jour) ;
  - Ils sont responsables de la vision de précision (haute résolution) ;
  - Ils déterminent la luminosité des scènes visualisées ;
  - Ils permettent l'extraction des couleurs grâce à 3 types de cônes (figure 2.3) :

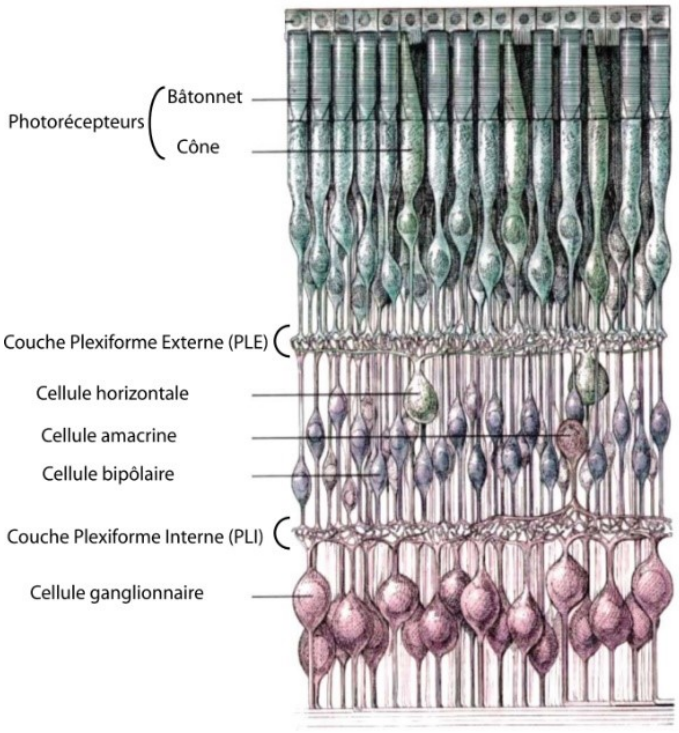


Figure 2.3: Organisation des couches cellulaire de la rétine

* Type "L" ("Long", rouge) sensible aux grandes longueurs d'onde visibles ;
* Type "M" ("Medium", vert) sensible aux longueurs d'onde visibles moyennes;
* Type "S" ("Short", bleu) sensible aux longueurs d'onde visibles courtes.

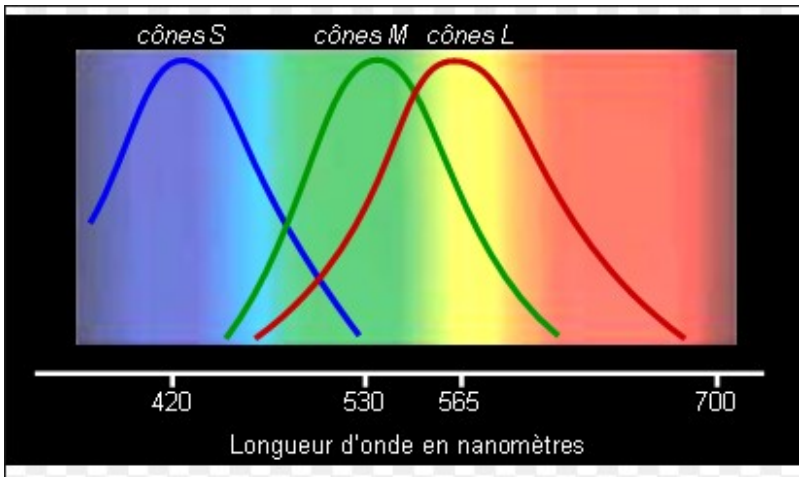


Figure 2.4: Sensibilité des différents types de photo-récepteur a la longueur d'onde de la lumière

2.3 Processus attentionnels

L'attention peut être définie selon de plusieurs façons, celles-ci reposent sur des classifications de l'attention, qui se fait en fonction de l'approche choisie, ainsi les classifications utilisées recouvrent des notions plus ou moins différentes et permettent de répondre à des questions variées et renvoient la représentation de l'attention visuelle à des manières différentes. En effet, il semble qu'en fonction des problématiques de recherche, les questions ont évolués. Nous nous attarderons dans notre étude sur trois questions : Les processus de l'attention peuvent se différencier en termes de :

- comportement moteur (vont-ils être associés à une action ?),
- traitement (l'information sera-t-elle traitée différemment en fonction de ceux-ci ?),
- sélection de l'information (quelle information vont-ils sélectionner ?).

La focalisation attentionnelle vers une région particulière du champ visuel est donc influencée par deux types de processus : l'un dit « Bottom-Up » et l'autre dit « Top-Down », qu'on décrira par la suite.

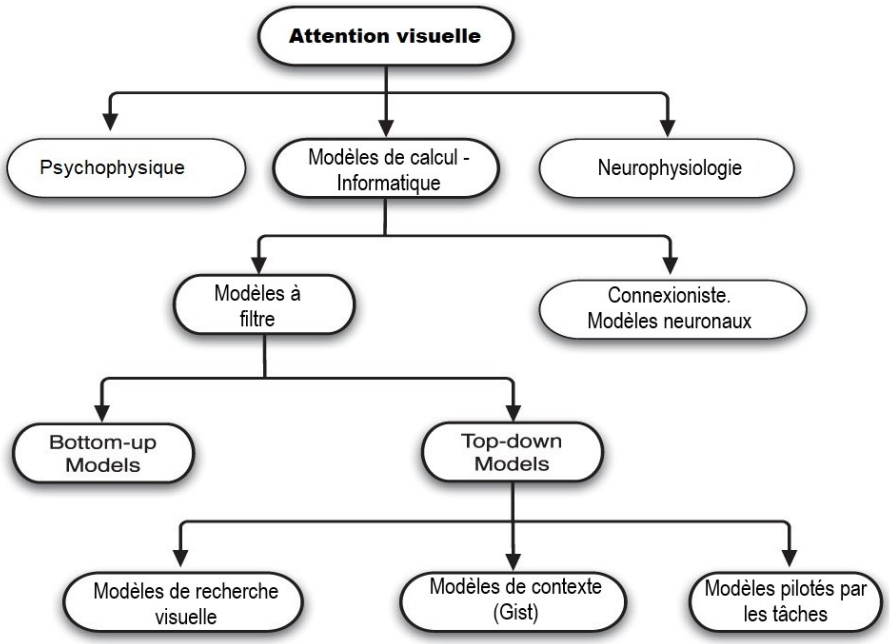


Figure 2.5: Classification des modèles de l'attention visuelle (43) et (7)

L'attention ou la discrimination visuelle caractérise intuitivement les parties d'une scène qui se détachent par rapport aux parties voisines et attirent notre attention. Examiner

les mécanismes du système visuel humain dans la sélection des régions d'intérêt est un problème de recherche important et fondamental en neurosciences et en psychologie.
Pour le système visuel humain, il est impossible de percevoir simultanément la scène entière dans le même acte sensoriel. Seule une petite région de la scène est analysée en détail à chaque instant. La région actuellement visitée ne sera pas toujours la même que celle fixée par les yeux et le regard passera à la prochaine région intéressante. Ce processus se poursuit jusqu'à ce que la scène entière soit perçue. L'ordre dans lequel la scène est analysée est déterminé par le mécanisme de l'attention sélective.
Le schéma précédent figure 2.5, montre une classification des différents modèles de l'attention visuelle. Nous allons expliquer dans ce qui suit ces modèles, nous nous intéressons dans notre cas à la modélisation informatique de l'attention visuelle.
Un modèle d'attention visuelle (43), (7) approche le comportement observé et/ou prédit de l'attention visuelle des humains et non humains. Les modèles peuvent être descriptifs, mathématiques, algorithmiques ou informatiques et tenter d'imiter, d'expliquer et/ou de prédire tout ou partie du comportement visuel attentif. Un modèle informatique de l'attention visuelle comprend non seulement une description du processus de calcul de l'attention, mais peut également être testé en fournissant des entrées d'images, similaires à celles qu'un expérimentateur pourrait présenter à un sujet, puis en voyant comment le modèle fonctionne par comparaison.
Les modèles informatique d'attention, dans la mesure où ils prédisent relativement bien l'attractivité visuelle de différents emplacements d'une scène, peuvent offrir une solution plus simple et plus rentable au problème de l'élaboration de critères d'évaluation rapides, objectifs et impartiaux. L'hypothèse est que les lieux marqués comme très saillants par le modèle devraient attirer le regard d'une majorité de clients potentiels.

2.3.1 Modèles neuronaux

Le modèle informatique (computational model) (42) est un modèle mathématique utilisant le calcul pour étudier des systèmes complexes. Typiquement, on met en place une simulation avec les paramètres souhaités et on laisse tourner l'ordinateur. On regarde alors la sortie pour interpréter le comportement du modèle.
Les modèles informatiques de la cognition sont axés sur l'utilisation des réseaux de neurones. Ces architectures ont été inspirées par des recherches sur le fonctionnement du calcul dans le cerveau et les travaux ultérieurs ont abouti à des modèles cognitifs distincts. Actuellement, la modélisation informatique semble être l'approche la plus prometteuse à

bien des égards et offre plus de flexibilité et de puissance d'expression que les autres approches.
Parmi les différents modèles informatiques cognitifs, les réseaux de neurones sont le modèle connexionniste le plus couramment utilisé. L'approche connexionniste qui prévaut aujourd'hui était à l'origine connue sous le nom de traitement distribué parallèle (PDP). Il s'agit d'une approche de réseau de neurones artificiels qui a utilisé la nature parallèle du traitement neuronal et la nature distribuée de la représentation neuronale.
Les réseaux de neurones ont été inspirés par la structure du cerveau humain. Une variante particulière du réseau de neurones est appelée le réseau de neurones récurrent. Il incarne la philosophie selon laquelle l'apprentissage nécessite de se souvenir.

2.3.2 Modèles à filtres

Lorsque nous observons notre environnement visuel, nous ne percevons pas toutes ses composantes comme étant également intéressantes. Certains objets « sortent » automatiquement et sans effort de leur environnement, c'est-à-dire qu'ils attirent notre attention visuelle, de manière ascendante - "bottom up", vers eux.
Les processus bottom-up sont donc des mécanismes exogènes guidée uniquement par les stimuli présent de le champ visuel sans aucune volonté de la part de l'observateur. La plupart des modèles de contrôle ascendant de l'attention sont basés sur le concept de carte de saillance, c'est-à-dire une carte bidimensionnelle explicite qui code la visibilité des objets dans l'environnement visuel. A partir de cette idée intuitive, plusieurs définitions peuvent être proposées, selon le critère qui fait que l'objet va être remarqué.La saillance s'appuient uniquement sur le contenu de l'image et les régions saillante d'une image sont donc les même quel que soit l'observateur (15)
Les Processus top-down, appelé encore descendant , sont des mécanisme endogène faisant intervenir la volonté du sujet ,ces mécanisme sont lies a la tache mais aussi à la sémantique du stimulus et au vécu propre, a l'idiosyncrasie, de chaque observateur.
Une distinction importante entre les modèles est de savoir s'ils s'appuient sur des influences ascendantes (Buttom-up), des influences descendantes (Top-down) ou une combinaison des deux (40). Les indices de Buttom-up sont principalement basés sur les caractéristiques d'une scène visuelle (déterminée par un stimulus), alors que les indices de Top-down (déterminés par un objectif) sont déterminés par des phénomènes cognitifs tels que les connaissances, les attentes, les récompenses et les objectifs actuels.
Le dénominateur commun de ces définitions et de pratiquement toutes les autres définitions des processus ascendants et descendants peut être résumé comme suit (40):

1. Le traitement de l'information est organisé hiérarchiquement.

2. Les niveaux inférieurs de la hiérarchie représentent des informations détaillées sur le stimulus, tandis que les niveaux supérieurs représentent des informations plus intégrées.

3. L'échange d'informations entre les niveaux est bidirectionnel.

2.3.2.1 Modèles Top-down

L'attention descendante (Top-down) est lente, axée sur les tâches, volontaire et en boucle fermée. L'un des exemples les plus célèbres de guidage par le haut de l'attention est celui de Yarbus en 1967, qui a montré que les mouvements oculaires dépendaient de la tâche actuelle.
Les mouvements oculaires différaient considérablement dans chacun de ces cas. Les modèles ont exploré trois sources principales d'influences top-down en réponse à cette question comment décider où regarder ? Certains modèles traitent de la recherche visuelle dans laquelle l'attention est attirée sur les caractéristiques d'un objet cible que nous recherchons. D'autres modèles étudient le rôle du contexte de la scène ou de l'essentiel pour contraindre les emplacements que nous examinons. Dans certains cas, il est difficile de dire précisément où et ce que nous examinons car une tâche complexe régit la fixation des yeux, par exemple, la conduite. Tandis que, en principe, les demandes de tâches sur l'attention englobent les deux autres facteurs, dans la pratique, les modèles se sont concentrés sur chacun d'eux séparément. La disposition des scènes a également été proposée comme source d'attention descendante et est considérée ici avec le contexte de la scène.
a- Modèles de recherche visuelle (visual search models)
La recherche visuelle (28) est un paradigme clé dans la recherche sur l'attention sélective. Le point de départ de la plupart des théories actuelles de la recherche visuelle a été la « théorie de l'intégration des caractéristiques » de l'attention visuelle de Treisman en 1980. Un certain nombre de questions clés qui ont été soulevées dans les tentatives de tester cette théorie sont encore des questions de recherche pertinentes aujourd'hui :

1. Le rôle et (le mode de) fonction des mécanismes ascendants et descendants dans le contrôle ou le « guidage » visuel chercher;

2. en particulier, le rôle et la fonction des mécanismes de mémoire implicites et explicites ;

3. la mise en œuvre de ces mécanismes dans le cerveau ; et

4. la simulation de processus de recherche visuelle dans des modèles informatiques ou, respectivement, neurocomputationnels (de réseau).

Les modèles attentionnels ont généralement été validés contre les mouvements oculaires d'observateurs humains. Les mouvements oculaires transmettent des informations importantes sur les processus cognitifs tels que la lecture, la recherche visuelle et la perception de la scène. En tant que tels, ils sont souvent traités comme des substituts des changements d'attention. Par exemple, dans la perception des scènes et la recherche visuelle, lorsque le stimulus est plus encombré, les fixations deviennent plus longues et les saccades plus courtes. La difficulté de la tâche (lire, par exemple, pour comprendre, plutôt que pour lire l'essence, ou rechercher une personne dans une scène, ou regarder la scène pour un test de mémoire) influence évidemment le comportement des mouvements oculaires. Bien que les modèles de prévision de l'attention et des mouvements oculaires soient souvent validés par rapport à des données oculaires, il existe de légères différences dans la portée, les approches, les stimuli et le niveau de détail. Les modèles de prédiction des mouvements oculaires (programmation de saccades) tentent de comprendre les fondements mathématiques et théoriques de l'attention. Certains exemples incluent les processus de recherche (par exemple, la théorie de la recherche optimale), les modèles de maximisation de l'information.
Il faut noté que les mouvements oculaires ne racontent pas toujours toute l'histoire et qu'il existe d'autres métriques pouvant être utilisées pour l'évaluation du modèle. Par exemple, la précision dans le signalement correct d'un changement dans une image (c.-à-d. Aveuglement par la recherche) ou dans la prédiction des éléments qui retiennent l'attention dont vous vous souviendrez montre des aspects importants de l'attention qui ne sont pas pris en compte par la seule analyse des mouvements oculaires.

b- Modèles de Contexte (context (Gist) models)

Des études sur la perception des scènes ont montré que les observateurs reconnaissent une scène du monde réel d'un seul coup d'œil. Au cours de ce processus rapide de vision, le système visuel forme une représentation spatiale du monde extérieur suffisamment riche pour saisir le sens de la scène, en reconnaissant quelques objets et autres informations saillantes dans l'image, pour faciliter la détection d'objets et le déploiement de attention. Cette représentation fait référence à l'essentiel (gist) d'une scène, qui comprend tous les niveaux de traitement, des caractéristiques de bas niveau (par exemple, la couleur, les fréquences spatiales) aux propriétés d'image intermédiaires (par exemple, la surface, le volume) et

les informations de haut niveau (par exemple, les objets , activation des connaissances sémantiques). Par conséquent, l'essentiel peut être étudié à la fois au niveau perceptif et conceptuel.

Il est important de noter que l'essentiel ne révèle pas nécessairement la catégorie sémantique d'une scène. Chun et Jiang ont montré que les cibles apparaissant dans des configurations répétées par rapport à certains objets d'arrière-plan étaient détectées plus rapidement. Les associations sémantiques entre les objets d'une scène (par exemple, un ordinateur est souvent placé sur un bureau) ou des indices contextuels jouent également un rôle important dans la direction des mouvements oculaires.

Les modèles informatiques actuels de l'attention visuelle se concentrent sur les informations ascendantes (Bottom-Up) et ignorent le contexte de la scène. Cependant, des études en cognition visuelle montrent que les humains utilisent le contexte pour faciliter la détection d'objets dans des scènes naturelles en dirigeant leur attention vers des régions de diagnostic. Dans (32) un modèle de guidage de l'attention basé sur la configuration globale de la scène. Nous montrons que les statistiques des caractéristiques de bas niveau (Top-Down) à travers l'image de la scène déterminent où un objet spécifique (par exemple une personne) doit être situé. Les mouvements oculaires humains montrent que les régions choisies par le modèle descendant concordent avec les régions examinées par des observateurs humains effectuant une tâche de recherche visuelle pour les personnes.

La saillance des régions de l'image lors de la tâche de détection d'objets. Les informations contextuelles fournissent un raccourci vers des systèmes de détection d'objets efficaces.

c- Modèles dirigé par les tâches (Task-driven models)

La tâche a une forte influence sur le déploiement de l'attention. Nous pouvons confirmer que les scènes visuelles sont interprétées selon les besoins pour répondre aux demandes des tâches. Ainsi Hayhoe et Ballard ont montré qu'il existait une relation étroite entre la cognition visuelle et les mouvements oculaires lorsqu'il s'agissait de tâches complexes.

Il ont constaté que les sujets effectuant une tâche guidée visible dirigeaient la majorité des ajustements vers des emplacements pertinents pour la tâche.

Il est souvent possible de déduire l'algorithme qu'un sujet a en tête à partir du schéma de ses mouvements oculaires. Par exemple, dans une tâche de "copie de bloc" dans laquelle les sujets devaient répliquer un assemblage de blocs de construction élémentaires, l'algorithme des observateurs pour effectuer la tâche était révélé par des schémas de mouvements oculaires.

Les sujets ont d'abord sélectionné un bloc cible dans le modèle pour vérifier la position du bloc, puis ont fixé l'espace de travail pour placer le nouveau bloc à l'emplacement

correspondant. D'autres recherches ont étudié des comptes rendus de haut niveau sur le comportement du regard dans des environnements naturels pour des tâches telles que la fabrication de sandwich, la conduite, le cricket et la marche Sodhi et ont étudié l'impact des distractions au volant, comme le réglage de la radio ou la réponse au téléphone, sur les mouvements des yeux (7).

2.3.2.2 Modèles de Buttom-up

Les régions d'intérêt qui attirent notre attention de manière ascendante (Buttom-up) doivent être suffisamment distinctives par rapport aux entités environnantes. Ce mécanisme attentionnel est également appelé exogène, automatique, réflexif ou périphérique.
L'attention ascendante (Buttom-up) est rapide, involontaire et très probablement rétroactive, Un exemple type d'attention ascendante consiste à regarder une scène avec une seule barre horizontale parmi plusieurs barres verticales est immédiatement attiré par la barre horizontale.
Elle est liée à la notion de saillance qui dénit une région saillante d'une image comme celle qui se distingue de son voisinage par certains attributs visuels, Cette région saillante attire l'attention. La saillance s'appuie donc uniquement sur le contenu de l'image et les régions saillantes d'une image sont donc les mêmes quel que soit l'observateur.

Carte de Saillance

la notion de carte de saillance est considérée comme la façon la plus courante de conceptualiser les mécanismes liés à la saillance, qui peut être définie l'information perceptuelle permettant à certains objets ou régions de la scène de ressortir fortement de leur voisinage et ainsi d'attirer l'attention visuelle de l'observateur.
Elle a été proposée par Koch et Ullman en 1985 (19), et se fond sur la description de l'architecture du système visuel en régions cérébrales sous-tendant des "cartes cognitives" relativement spécialisées dans le traitement de caractéristiques perceptives spécifiques (couleur, orientation, mouvement, etc.).
La carte de saillance est donc une carte topographique arrangée qui représente la saillance visuelle, elle est combine les différents éléments visuels qui contribuent à la sélection attentive d'un stimulus (couleur, orientation, mouvement, ...) et intègre l'information normalisée à partir des cartes de caractéristiques individuels vers une mesure globale de visibilité.
La carte de saillance comprend les éléments suivants:

- Une représentation anticipée composée d'un ensemble des cartes de caractéristiques calculées en parallèle.

- Une carte de saillance topographique ou chaque emplacement encode la combinaison des propriétés sur toutes les cartes de caractéristiques.
- Une cartographie sélective des propriétés d'un seul emplacement visuel représentée d'une manière non-topographique centrale, à partir de la carte de saillance.
- Un réseau WTA (Winner-take-all) implémente le processus de sélection.
- Inhibition de l'emplacement choisi, qui provoque un décalage automatique vers l'emplacement prochain le plus visible.

Les cartes de caractéristiques encodent la visibilité avec une dimension particulière des caractéristiques, et la carte de saillance combine les informations de chaque carte de caractéristiques. La saillance à un emplacement donné est déterminée par le degré de différence entre cet emplacement et ces voisins.

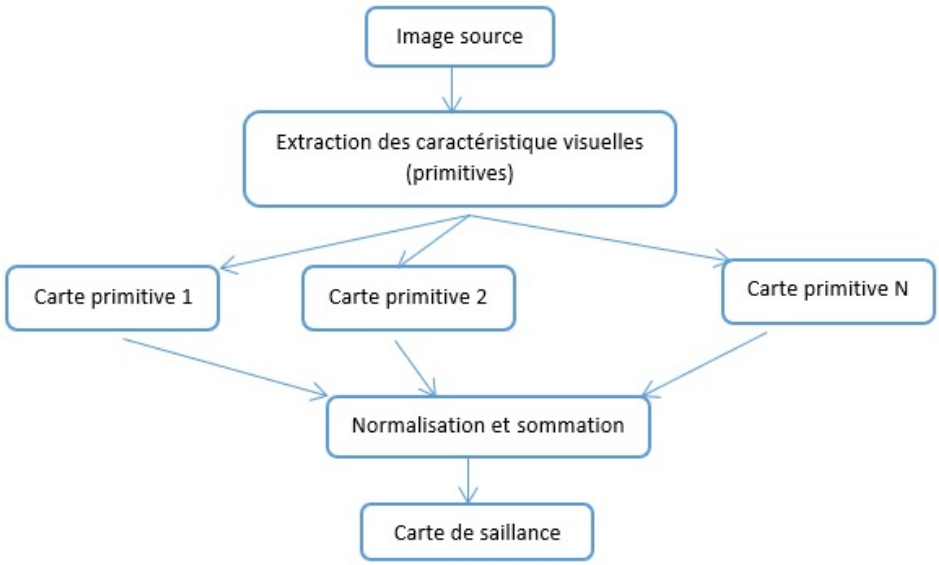
a- Carte de saillance 2D

Nous allons présenter ici les principales cartes de saillance 2D d'attention visuelle qui s'inspirent de la biologie de système visuel humaine « HVS ». Ces modèles d'attention regroupent les caractéristique base sur les orientations, l'intensité et les couleurs, ces derniers donnent naissance à des cartes qui mettent en évidence de région de la scène différentes de leur voisinage selon la caractéristique considérer, les différentes cartes sont ensuite fusionnées en une carte de saillance Ce modèle, propose par Koch et Ullman (19), est inspiré par les études de Treisman et Gelade en 1980 sur la théorie de l'intégration des caractéristiques pour l'attention visuelle. Selon cette théorie, l'attention visuelle est guidée par la combinaison des caractéristiques de bas niveau comme l'intensité de luminosité, la couleur et l'orientation. Ainsi, dans ce premier modèle de Koch and Ullman (figure 2.6), une image d'entrée est décomposée en plusieurs cartes, une carte par caractéristique de bas niveau. Cette architecture devenue par la suite une référence.

Nous présentons ci dessous un exemple (figure 2.7) d'obtention d'une carte de saillance sur une image donnée (a); les cartes de saillances couleur, intensité et orientation sont données respectivement en (b), (c) et (d); la carte de saillance finale est donnée en (e) alors que (f) et (g) représentent respectivement les deux et les cinq points de fixation les plus saillants.

b- Carte de saillance 3D

La carte de saillance est un modèle d'attention visuelle sélectif qui utilise des caractéristiques purement ascendantes d'une image comme la couleur, l'intensité et l'orientation.



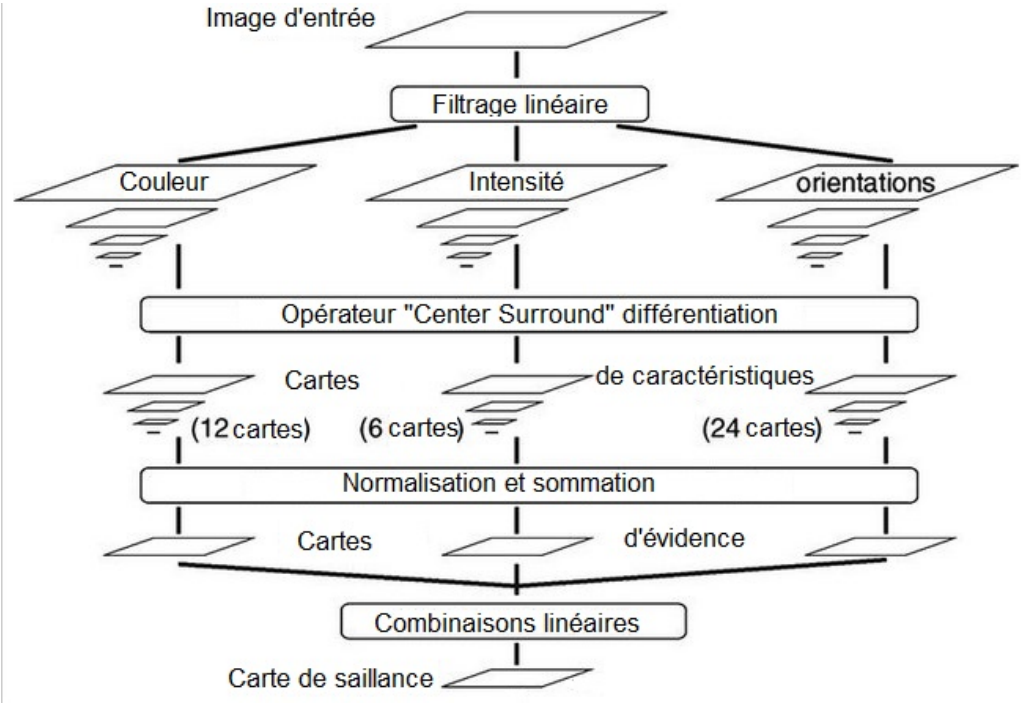


Figure 2.6: Architecture biologiquement plausible d'un modèle d'attention visuelle proposée par Koch and Ullman.

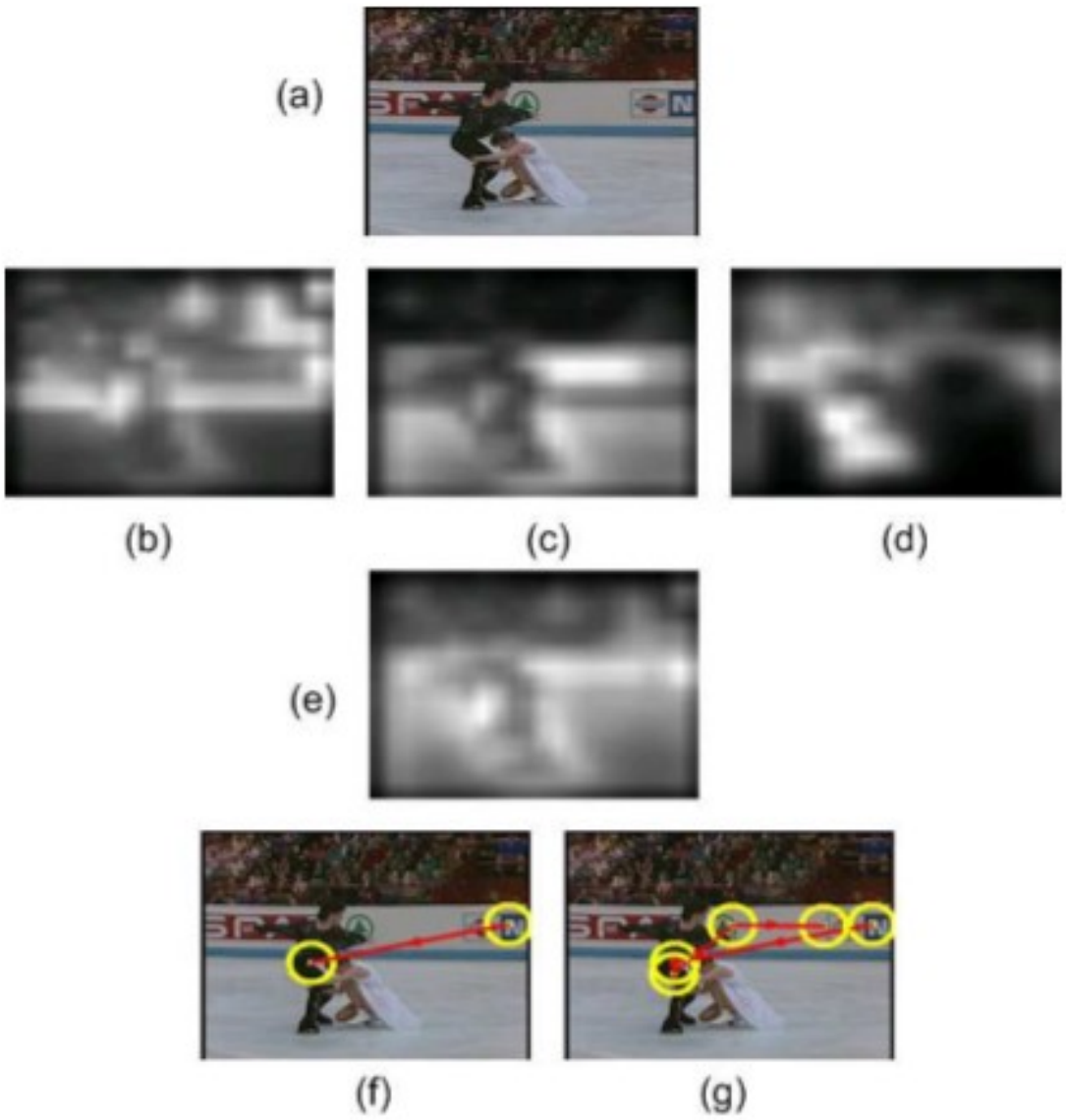


Figure 2.7: Régions saillantes d'une image (cercles jaunes) correspondants aux zones claires de la carte de saillance

Une autre caractéristique ascendante de l'entrée visuelle est la profondeur, la distance entre l'œil (ou le capteur) et les objets dans le champ visuel.Un exemple de calcul de la saillance 3D est présenté dans la figure (figure 2.8)
Les modèles de saillance visuelle 3D existants peuvent être classés en trois catégories : la pondération en profondeur, la saillance en profondeur et la stéréovision (39), (44).

- **Pondération en profondeur**
  L'approche de pondération en profondeur utilise les informations de profondeur comme facteur de pondération dans l'étape de fusion des caractéristiques visuelles 2D. La saillance 3D finale de chaque emplacement de la scène est directement liée à sa profondeur.
  Makis et al. ont proposé un modèle basé sur la disparité du flux et du mouvement de l'image. Zhang et al. ont conçu un algorithme d'attention visuelle stéréoscopique pour la vidéo 3D basé sur de multiples stimuli perceptifs. Chamaret et al. ont construit une méthode d'extraction de région d'intérêt (ROI) pour le rendu 3D adaptatif.

- **Profondeur-saillance**
  Les approches profondeur-saillance prennent les informations de profondeur comme facteur de saillance supplémentaire dans l'étape de fusion des caractéristiques visuelles 2D. Les caractéristiques de profondeur sont combinées avec des caractéristiques visuelles 2D en utilisant une stratégie de mise en commun des cartes de saillance pour obtenir une carte de saillance 3D finale.

- **Stéréo-vision**
  Les approches de pondération en profondeur et de saillance en profondeur contiennent toutes deux une étape dans laquelle les caractéristiques visuelles 2D sont extraites et combinées avec des informations de profondeur pour calculer des cartes de saillance 3D. L'approche par vision stéréo n'utilise que l'image stéréoscopique des deux vues et n'a pas besoin d'une carte de profondeur pour détecter la saillance visuelle 3D. Le modèle de détection de saillance 3D est basé sur les mécanismes de la perception stéréoscopique dans le HVS.

2.4 Modèles d'attention

Dans cette partie, les modèles sont expliqués en fonction de leur mécanisme pour obtenir la saillance. Certains modèles appartiennent à plusieurs catégories. Nous nous sommes plus intéressés par les modèles de saillance que par les approches qui détectent et segmentent la

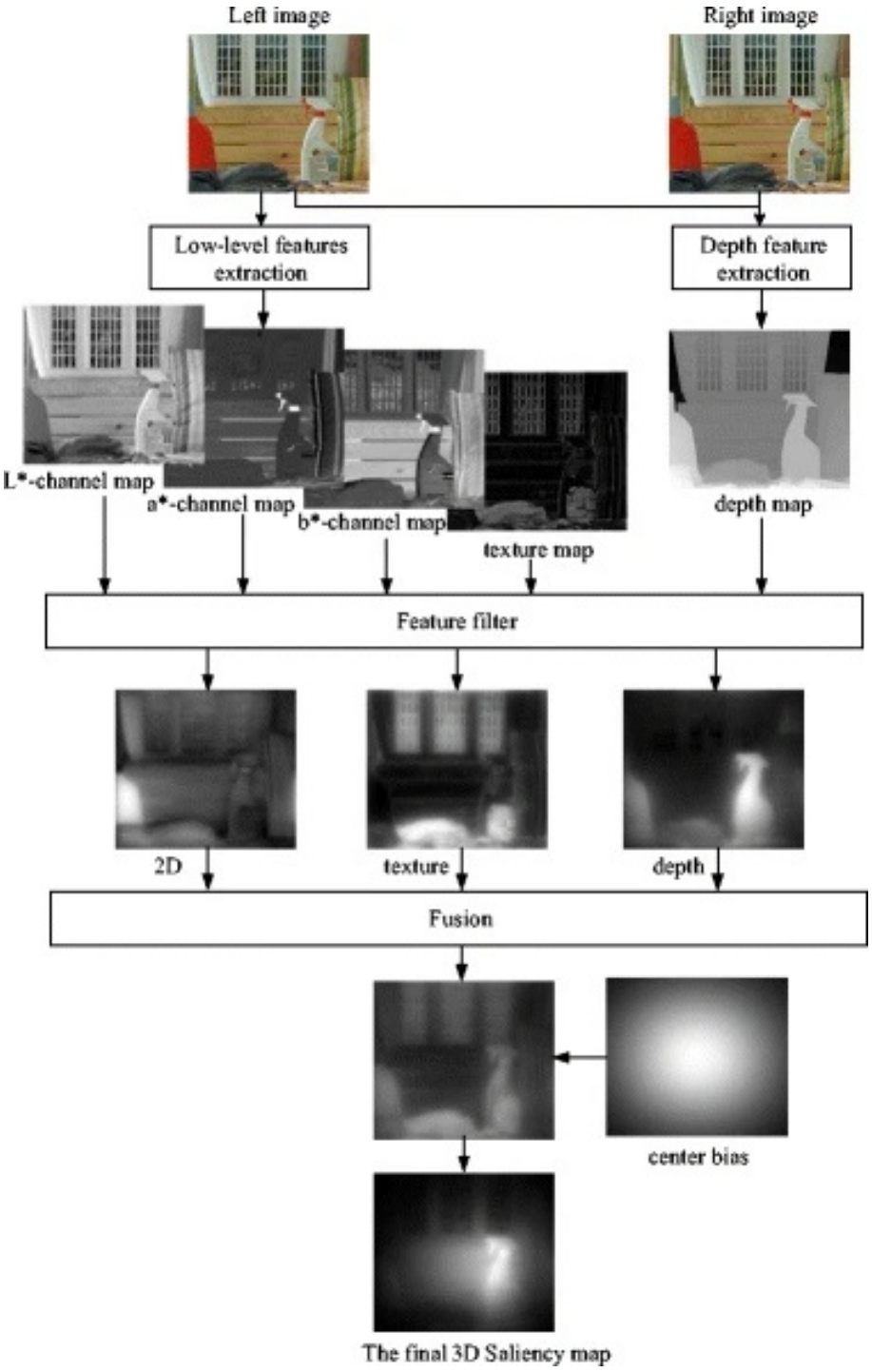


Figure 2.8: Exemple de carte de saillance 3D

région ou l'objet le plus saillant d'une scène. Bien que ces modèles utilisent un opérateur de saillance au stade initial, leur objectif principal n'est pas d'expliquer le comportement attentionnel. Cependant, certaines méthodes ont inspiré des modèles de saillance ultérieurs.

2.4.1 Modèles Cognitifs

La plupart des modèles attentionnels sont directement ou indirectement inspirés des concepts cognitifs. Le modèle de base d'Itti et al. (15) utilise trois canaux de caractéristiques : couleur, intensité et orientation. Ce modèle a été la base des modèles ultérieurs et la référence standard pour les comparaisons. Il a été démontré qu'il est en corrélation avec les mouvements oculaires dans les tâches de visualisation humaine. Le modèle fonctionne globalement comme suit : L'image d'entrée est sous-échantillonnée en une pyramide gaussienne et chaque niveau de pyramide est décomposé en canaux pour le rouge (R), le vert (V), le bleu (B), le jaune (Y), l'intensité (I) et les orientations locales (O). A partir de ces canaux, des « cartes de caractéristiques » centre-surround f_l sont construites et normalisé pour les différentes caractéristiques l. Dans chaque canal, les cartes sont additionnées sur échelle et normalisé à nouveau :

$$f_l = N(\sum_{c=2}^{4} \sum_{s=c+3}^{c+4} f_{l,c,s}), \forall l \in L_I \cup L_C \cup L_O \tag{2.1}$$

$$L_I = \{I\}, L_C = \{RG, BY\}, L_O = \{0^o, 45^o, 90^o, 135^o\} \tag{2.2}$$

Ces cartes sont additionnées linéairement et normalisées une fois de plus pour produire les « cartes de visibilité » :

$$C_I = f_I, C_C = N(\sum_{l \in L_C} f_l), C_O = N(\sum_{l \in L_O} f_t) \tag{2.3}$$

Enfin, les cartes de visibilité sont à nouveau combinées linéairement pour générer la carte de saillance : $S = \frac{1}{3} \sum_{k \in \{I,C,O\}} C_k$.
Plusieurs améliorations ont été introduites par plusieurs auteurs au modèle initial d'Itti, notamment :

- Amélioration des fonctions de sensibilité au contraste, la position de décomposition perceptive, le masquage visuel et les interactions centre-surround.
- Extension du modèle au domaine spatio-temporel en fusionnant des informations achromatiques, chromatiques et temporelles.

- Modélisation de la recherche visuelle comme un problème d'optimisation de gain descendant en maximisant le rapport signal sur bruit (SNR) de la cible par rapport aux distracteurs au lieu d'apprendre des fonctions de fusion explicites.

- Introduction d'opérateurs de symétrie isotrope et de symétrie radiale et de la symétrie des couleurs.

- Introduction d'un modèle basé sur un système de vision de bas niveau en trois étapes : traitement des stimuli visuels, simulation des mécanismes d'inhibition présents dans les cellules du cortex visuel normalisation de leur réponse au contraste du stimulus et enfin intégration des informations intégrées à plusieurs échelles en effectuant une transformation en ondelettes inverse directement sur les poids calculés à partir de la non-linéarisation des sorties corticales.

Les modèles cognitifs ont l'avantage d'élargir notre vision des fondements biologiques de l'attention visuelle. Cela aide en outre à comprendre les principes de calcul ou les mécanismes neuronaux de ce processus ainsi que d'autres processus dépendants complexes tels que la reconnaissance d'objets.

2.4.2 Modèles Bayésiens

Plusieurs modèles statistiques de saillance visuelle ont également été développés. Dans ces modèles, un ensemble de statistiques ou de distributions de probabilité (PD) des variables visuelles (16) sont calculées à partir de la scène que le sujet regarde ou d'un ensemble de scènes naturelles, et une variété de mesures de saillance visuelle sont définies sur ces statistiques ou PD, y compris l'auto-information, pouvoir discriminant, surprise bayésienne et inverse de vraisemblance.
La modélisation bayésienne (46) est ainsi utilisée pour combiner des preuves sensorielles avec des contraintes préalables. Dans ces modèles, les connaissances antérieures (par exemple, le contexte ou l'essentiel (gist) de la scène) et les informations sensorielles (par exemple, les caractéristiques de la cible) sont combinées de manière probabiliste selon la règle de Bayes (par exemple, pour détecter un objet d'intérêt).
La connaissance de l'objet cible comprend la caractéristique d'apparence O , le contexte F l'environnement (dans lequel la cible apparaît probablement) et l'emplacement X. Supposons que la distribution d'une caractéristique ne change pas avec l'emplacement :

$$P(F, C, X) = P(F, C)P(X)$$

Compte tenu de la caractéristique et du contexte à l'emplacement, la probabilité de l'objet cible peut être calculé comme suit :

$$P(O|F,C,X) = \frac{P(O,F,C,X)}{P(F,C,X)} = \frac{1}{P(F,C)} P(F|O,C)P(C|O)P(O|X) \tag{2.4}$$

Cela implique l'hypothèse que la distribution d'une caractéristique en un point sur la cible ne change pas avec l'emplacement:

- Le premier terme du côté droit de l'équation 2.4, $\frac{1}{P(F,C)}$, ne dépend que des caractéristiques visuelles observées au point et de son contexte, qui est indépendant de toute connaissance que nous avons sur la classe cible, et est donc un facteur ascendant. Il fournit une mesure de la probabilité de trouver un ensemble de mesures locales dans des scènes naturelles. Ce terme correspond à la définition de saillance, et est la mesure de saillance employé.

- Le deuxième terme, $P(F|O,C)$, représente la connaissance descendante de l'apparence cible et de sa contribution à la recherche. Les régions de l'image avec des caractéristiques peu susceptibles d'appartenir à l'objet cible font l'objet d'un veto et les régions avec des caractéristiques assistées sont améliorées.

- Le troisième terme, $P(C|O)$ représente la connaissance descendante du contexte cible. Les régions de l'image avec des caractéristiques susceptibles d'appartenir au contexte cible seront prises en compte. Par exemple, lorsque nous recherchons un avion dans une image, nous allons observer le ciel et ignorer la forêt.

- Le quatrième terme, $P(X|O)$ est indépendant des caractéristiques visuelles et reflète toute connaissance préalable de l'endroit où la cible est susceptible d'apparaître.

2.4.3 Modèles de la théorie de décision

Une formulation de la saillance visuelle à l'aide de la théorie décisionnelle (9), a été initialement proposée pour le traitement descendant (reconnaissance d'objets) par Gao et Vasconcelos en 2005, elle a été ensuite étendue au problème de la saillance ascendante. Ainsi la théorie de la décision indique que les systèmes perceptifs évoluent pour produire des décisions sur les états de l'environnement avoisinant qui sont optimales au sens de la théorie de la décision (par exemple, une probabilité d'erreur minimale).
Sous cette formulation, l'optimalité est définie dans la probabilité minimale de sens d'erreur, sous une contrainte de calcul. La saillance des caractéristiques visuelles à un emplacement donné du champ visuel est définie comme le pouvoir de ces caractéristiques de discriminer

entre le stimulus à l'emplacement et une hypothèse nulle.
Pour la saillance ascendante, il s'agit de l'ensemble des caractéristiques visuelles qui entourent l'emplacement considéré. La discrimination est définie au sens de la théorie de l'information et le détecteur de saillance optimal est dérivé pour une classe de stimuli conforme aux propriétés statistiques connues des images naturelles. Il a été montré que sous l'hypothèse que la saillance est déterminée par le filtrage linéaire, le détecteur optimal se compose de ce que l'on appelle généralement l'architecture standard de V1 : La figure 2.9) montre que le champ visuel est projeté dans des cartes de caractéristiques qui tiennent compte de la couleur, de l'intensité, de l'orientation, de l'échelle, etc. Les fenêtres centrales et périphériques sont ensuite analysées à chaque emplacement pour déduire le pouvoir discriminant de chaque caractéristique à cet emplacement.

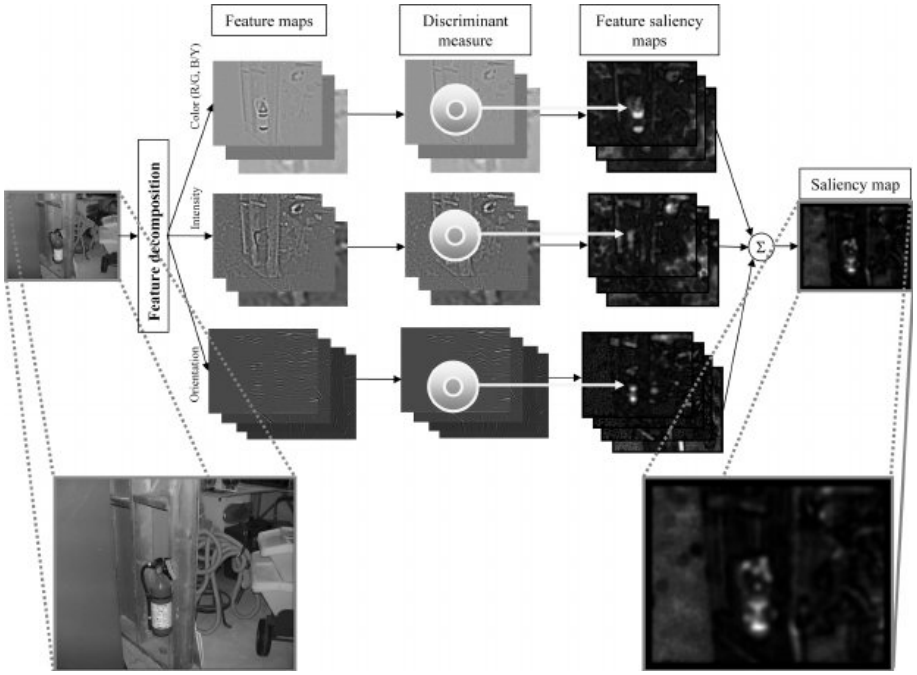


Figure 2.9: Utilisation de la théorie de décision pour le calcul de la saillance

La saillance caractéristique est définie comme le pouvoir de discriminer entre le centre et l'entourage. La saillance globale est définie comme le pouvoir discriminant de l'ensemble des caractéristiques et (pour les scènes naturelles) peut être approximée par la somme de toutes les saillances des caractéristiques.

2.4.4 Modèles se basant sur la théorie de l'information

Ces modèles sont basés sur le fait que le calcul de saillance localisée sert à maximiser les informations échantillonnées dans son environnement. Ils traitent de la sélection des parties les plus informatives d'une scène et de l'élimination du reste.(7).
Rosenholtz a conçu un modèle de recherche visuelle qui pourrait également être utilisé pour la prédiction de saillance sur une image en visionnage libre. Premièrement, les caractéristiques de chaque point, p_i, sont dérivées dans un espace de caractéristiques uniforme approprié (par exemple, un espace de couleur uniforme). Ensuite, à partir de la distribution des caractéristiques, la moyenne μ , et la covariance, $\sum$, des caractéristiques du distracteur sont calculées. Le modèle définit ensuite la saillance cible comme la distance de Mahalanobis, , entre le vecteur de caractéristiques cible, T, et la moyenne de la distribution du distracteur, où $\Delta^2 = (T-\mu)' \sum^{-} 1(T-\mu)$.
Bruce et Tsotsos ont proposé en 2009 un modèle d'attention ascendante basé sur le principe de maximiser les informations échantillonnées à partir d'une scène. L'opération proposée est basée sur la mesure d'auto-information de Shannon et est réalisée dans un circuit neuronal, ayant des liens étroits avec les circuits existant dans le cortex visuel des primates.
Hou et Zhang ont introduit l'approche de la longueur de codage incrémentiel (ICL) pour mesurer le gain d'entropie respectif de chaque caractéristique. L'objectif était de maximiser l'entropie des caractéristiques visuelles de l'échantillon. Li et al. ont proposé un modèle de saillance visuelle basé sur l'entropie conditionnelle pour l'image et la vidéo. La saillance a été définie comme l'incertitude minimale d'une région locale compte tenu de sa zone environnante (à savoir, l'entropie conditionnelle minimale) lorsque la distorsion de perception est prise en compte. Enfin,
Wang et al. ont introduit un modèle pour la saccade humaine qui est considérée comme un processus dynamique de recherche d'informations. Ils proposent ainsi un modèle basée sur le principe de maximisation de l'information, pour simuler les trajets de balayage saccadés « scanpaths » humains sur des images naturelles. Le modèle intègre trois facteurs connexes guidant les mouvements oculaires de manière séquentielle :

1. les réponses sensorielles de référence,
2. l'écart de résolution fovéa-périphérie
3. la mémoire de travail visuelle

Ils calculent trois cartes de réponse de filtre multibande pour chaque mouvement oculaire qui sont ensuite combinées en cartes de réponse de filtre résiduel multibande. Enfin, ils calculent des informations perceptives résiduelles (RPI) à chaque emplacement.

2.4.5 Modèles graphiques

Un modèle graphique est un modèle probabiliste dans lequel un graphique désigne la structure d'indépendance conditionnelle entre les variables aléatoires. Les modèles Attention traitent alors les mouvements oculaires comme une série chronologique. Comme il existe des variables cachées influençant la génération des mouvements oculaires, des approches telles que les modèles de Markov cachés (HMM), les réseaux bayésiens dynamiques (DBN) et les champs aléatoires conditionnels (CRF) ont été incorporées. Nous citons ci-après quelques travaux, listés dans le papier de l'état de l'art de (7).

Salah et al. ont proposé une approche de l'attention et l'ont appliquée à la reconnaissance des chiffres manuscrits et des visages. Dans la première étape, une carte de saillance ascendante est construite à l'aide de fonctionnalités simples. Au niveau intermédiaire, les informations « quoi » et « où » sont extraites en divisant l'espace image en régions uniformes et en entraînant un perceptron à couche unique sur chaque région de manière supervisée. Finalement, ces informations sont combinées au niveau associatif avec un modèle de Markov observable discret (OMM). Les régions visitées par une fovéa sont traitées comme des états de l'OMM. Une inhibition de retour permet à la fovéa de se focaliser sur les autres positions de l'image.

Harel et al. introduisent Graph-Based Visual Saliency (GBVS), qui se compose de deux étapes(figure 2.10 : d'abord former des cartes d'activation sur certains canaux de caractéristiques (similaire à Itti et al.), il en résulte alors un graphique entièrement connecté sur tous les emplacements de grille de chaque carte de caractéristiques. Les pondérations entre deux nœuds sont attribuées proportionnellement à la similarité des valeurs des caractéristiques et à leur distance spatiale. La deuxième étape consiste à les normaliser d'une manière qui met en évidence la visibilité et admet la combinaison avec d'autres cartes. Le modèle est simple, et biologiquement plausible dans la mesure où il est naturellement parallélisable.

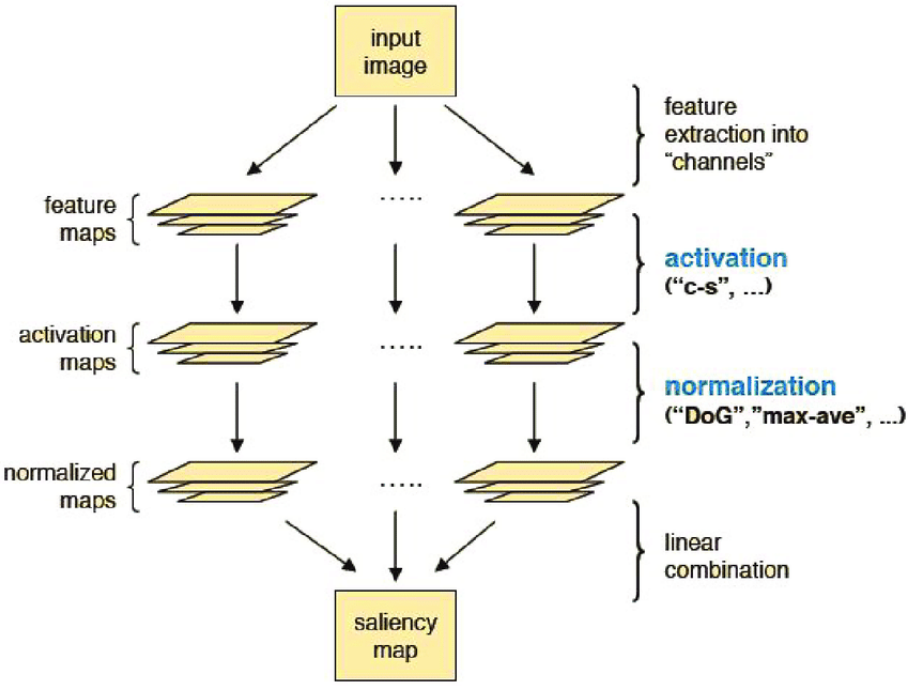


Figure 2.10: Modèle graphique pour le calcul de la saillance

Avraham et Lindenbaum, ont introduit le modèle E-saliency (Extended saliency) en utilisant une approximation de modèle graphique pour étendre leur modèle de saillance statique basé sur des auto-similitudes. L'algorithme est essentiellement une méthode pour estimer la probabilité qu'un candidat soit une cible. Pang et al. ont présenté un modèle stochastique de l'attention visuelle basé sur la théorie de la détection de signaux pour la recherche visuelle et l'attention. Ils ont proposé un réseau bayésien dynamique pour prédire où les humains se concentrent généralement dans une scène vidéo.

Hikkerur et al. ont proposé un modèle basé sur des hypothèses selon lesquelles le but du système visuel est de savoir où se trouve et que le traitement visuel se déroule de manière séquentielle. Dans ce modèle, l'attention apparaît comme l'inférence dans un modèle graphique bayésien qui met en œuvre des interactions entre les zones ventrale et dorsale.

Les modèles graphiques pourraient être considérés comme une version généralisée des modèles bayésiens. Cela leur permet de modéliser des mécanismes d'attention plus complexes dans l'espace et le temps, ce qui se traduit par un bon pouvoir de prédiction. Les inconvénients résident dans la complexité du modèle, notamment en ce qui concerne la formation et la lisibilité.

2.4.6 Modèles d'analyse Spectrale

Au lieu de traiter une image dans le domaine spatial, les modèles de cette catégorie dérivent la saillance dans le domaine fréquentiel. Hou et Zhang ont développé le modèle de saillance résiduelle spectrale basé sur l'idée que les similitudes impliquent des redondances. Ils proposent que des singularités statistiques dans le spectre puissent être responsables de régions anormales dans l'image, où les objets proto deviennent visibles. Étant donné une image d'entrée $I(x)$, l'amplitude $\mathcal{A}(f)$ et la phase $\mathcal{P}(f)$ sont dérivées. Ensuite, le log-spectre $\mathcal{L}(f)$ est calculé à partir de l'image sous-échantillonnée. A partir de $\mathcal{L}(f)$, le résidu spectral $\mathcal{R}(f)$ peut être obtenu en multipliant $\mathcal{L}(f)$ par $h_n(f)$, qui est un filtre de moyenne locale nxn, et en soustrayant le résultat à lui-même. En utilisant la transformée de Fourier inverse, ils construisent la carte de saillance dans le domaine spatial. La valeur de chaque point de la carte de saillance est ensuite mise au carré pour indiquer l'erreur d'estimation. Enfin, ils lissent la carte de saillance avec un filtre gaussien $g(x)$ pour un meilleur effet visuel. L'ensemble du processus est résumé ci-dessous 2.5:

$$\begin{aligned}
\mathcal{A}(f) &= \mathcal{R}(\mathcal{F}[I(x)]),\\
\mathcal{P}(f) &= \varphi(\mathcal{F}[I(x)]),\\
\mathcal{L}(f) &= log(\mathcal{A}(f)),\\
\mathcal{R}(f) &= \mathcal{L}(f) - h_n(f) * \mathcal{L}(f),\\
S(x) &= g(x) * \mathcal{F}^{-}1[exp(\mathcal{R}(f) + \mathcal{P}(f))]^2
\end{aligned} \tag{2.5}$$

Achanta et al. ont mis en œuvre une approche à réglage de fréquence pour la détection des régions saillantes en utilisant des caractéristiques de faible niveau de couleur et de luminance. Tout d'abord, l'image RVB d'entrée I est transformée en espace colorimétrique CIE Lab. Ensuite, la carte de saillance scalaire S pour l'image I est calculée comme $S(x, y) = \|I_\mu - I_{\omega hc}\|$, où I_μ est la moyenne arithmétique du vecteur caractéristique de l'image, $I_{\omega hc}$ est une version floue gaussienne de l'image originale utilisant un noyau binomial séparable $5x5$, $\|.\|$ est la norme $L2$ (distance euclidienne), et x, y sont les coordonnées des pixels.

Bian et Zhang ont proposé le modèle de blanchiment spectral « Spectral Whitening (SW) » basé sur l'idée que le système visuel contourne les caractéristiques redondantes (fréquemment présentes, non informatives) tout en répondant aux caractéristiques rares (informatives).

Les modèles d'analyse spectrale sont simples à expliquer et à mettre en œuvre. Bien que toujours très réussi, la plausibilité biologique de ces modèles n'est pas très claire

2.4.7 Modèles de Classification

Des approches d'apprentissage automatique ont été utilisées pour modéliser l'attention visuelle en apprenant des modèles à partir de fixations oculaires enregistrées ou de régions saillantes étiquetées. En règle générale, le contrôle de l'attention fonctionne comme une fonction de « stimuli-saillance » pour sélectionner, repondérer et intégrer les stimuli visuels d'entrée. Notez que ces modèles peuvent ne pas être purement ascendants car ils utilisent des fonctionnalités qui guident l'attention de haut en bas (par exemple, des visages ou du texte).

Kienzle et al. ont introduit une approche ascendante non paramétrique pour apprendre l'attention directement à partir des données de suivi oculaire. Le modèle consiste en un mappage non linéaire d'un patch d'image à une valeur réelle, entraîné pour produire des sorties positives sur les fixations et des sorties négatives sur des patchs d'image sélectionnés au hasard. La fonction de saillance est déterminée par sa maximisation des performances de prédiction sur les données observées. Un système à base de machines à vecteurs de support (SVM) a été entraînée pour déterminer la saillance à l'aide des intensités locales.

Peters et Itti (36) ont entraîné un simple classificateur de régression pour capturer l'association dépendante de la tâche entre une scène donnée et les emplacements préférés à regarder pendant que des sujets humains jouaient à des jeux vidéo. La figure 2.11 illustre leur modèle d'apprentissage des influences descendantes et dépendantes de la tâche sur la position des yeux.

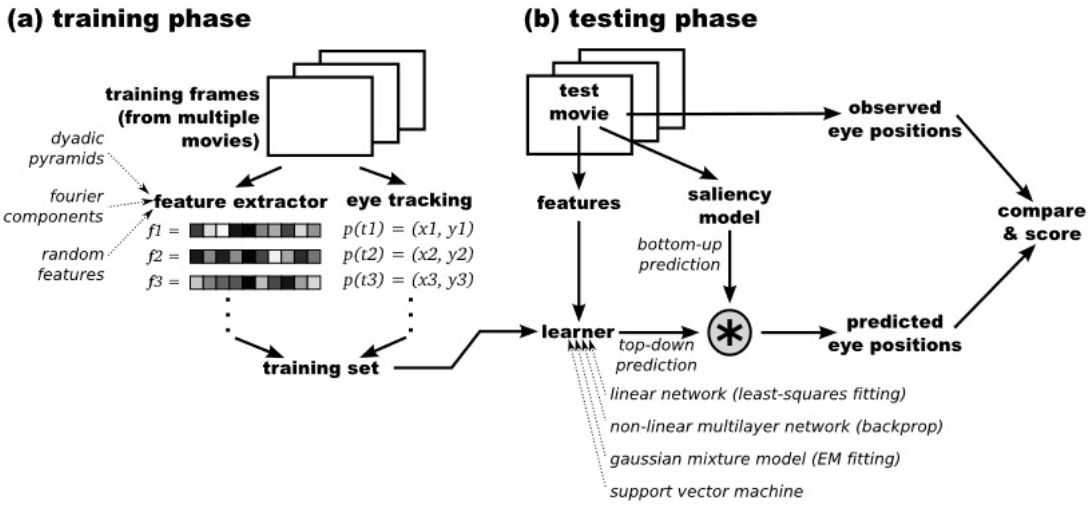


Figure 2.11: Modèle basée machine learning (36) pour le calcul de la saillance

Tout d'abord, dans (a) la phase d'apprentissage, ils compilent un ensemble d'apprentissage

contenant des vecteurs de caractéristiques et des positions oculaires correspondant à des images individuelles de plusieurs clips de jeux vidéo qui ont été enregistrés pendant que les observateurs jouaient aux jeux de manière interactive.
Ensuite, dans (b) la phase de test, ils utilisent un clip de jeu vidéo différent pour tester le modèle. Les images du clip de test sont transmises en parallèle à un modèle de saillance ascendant, ainsi qu'à l'extracteur de caractéristiques descendant, qui génère un vecteur de caractéristiques utilisé pour générer une carte de prédiction de la position des yeux de haut en bas. Enfin, les cartes de prédiction ascendantes et descendantes peuvent être combinées via une multiplication ponctuelle, et les cartes individuelles et combinées peuvent être comparées à la position réelle observée de l'œil.
Enfin Judd et al.(18), ont entraîné une SVM linéaire à partir de données de fixation humaine en utilisant un ensemble de caractéristiques d'image de bas, moyen et haut niveau pour définir les emplacements saillants. Les vecteurs de caractéristiques provenant d'emplacements fixes et d'emplacements aléatoires se sont vu attribuer des étiquettes de classe +1 et -1, respectivement. Leurs résultats sur un ensemble de données de 1 003 images observées par 15 sujets (rassemblés par les mêmes auteurs) montrent que la combinaison de toutes les caractéristiques susmentionnées et de la distance par rapport au centre de l'image produit les meilleures performances de prédiction de fixation oculaire.
Les approches de calcul de la saillance basées sur l'apprentissage automatique sont dépendantes des données disponibles, qui augmentent au fur et à mesure que sur les mouvements oculaires et avec une plus large diffusion de dispositifs de suivi oculaire prenant en charge la collecte de données de masse. Cependant, cela rend les modèles dépendants des données, influençant ainsi la comparaison équitable des modèles.

2.5 Modèles à base d'apprentissage

À partir de 2015, l'émergence des technologies d'apprentissage profond en vision par ordinateur leur a permis de devenir progressivement la direction dominante dans la détection d'objets en saillie. En effet, les modèles de recherche de saillance basés sur l'apprentissage profond permettent l'extraction de caractéristiques sémantiques de haut niveau à différentes échelles au lieu de créer les fonctionnalités artisanales conventionnelles.
Un modèle basé sur CNN utilise généralement un regroupement de couches convolutives, de mise en commun et entièrement connectées. Les régions les plus saillantes d'une image sont fournies par les neurones avec de grands champs récepteurs, tandis que les neurones

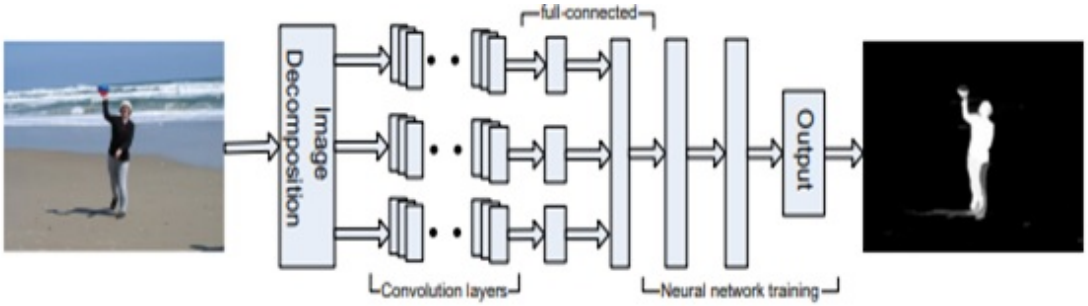


Figure 2.12: L'architecture globale de l'extraction de saillance visuelle à l'aide de CNN.

avec de petits champs récepteurs génèrent des informations locales qui peuvent être exploitées pour affiner les cartes de saillance produites par les couches supérieures.
Les modèles de détection d'objets en saillie basés sur l'apprentissage en profondeur peuvent être divisés en deux catégories principales. La première catégorie comprend des modèles qui utilisent des perceptrons multicouches (MLP) pour prédire le score de saillance des caractéristiques profondes extraites de chaque unité d'imagerie. La deuxième catégorie comprend des modèles basés sur des réseaux entièrement convolutifs (basés sur FCN). Les modèles utilisent en général un algorithme d'encodeur automatique basé sur un réseau de neurones convolutifs (CNN) pré-entraîné sur une tâche de classification d'images à grande échelle, et combinent les représentations résultantes avec des informations de scène globales. (13), (21)

2.6 Conclusion

Face à une scène visuelle complexe, l'homme peut localiser efficacement la région d'intérêt et analyser la scène en traitant sélectivement des sous-ensembles d'entrées visuelles.
L'attention a été employée pour affiner la recherche et accélérer le processus. L'attention visuelle est un sujet actif dans les domaines de la vision par ordinateur, des neurosciences et de l'apprentissage en profondeur. Il est largement utilisé dans la segmentation d'objets, la reconnaissance d'objets, la génération de légendes d'images et la réponse visuelle aux questions.
Dans ce chapitre, nous avons défini la perception visuelle puis nous avons discuté des avancées récentes dans la modélisation de l'attention visuelle en mettant l'accent sur les modèles de saillance ascendants.
Les progrès dans ce domaine pourraient grandement aider à résoudre d'autres problèmes de vision difficiles tels que l'interprétation de scènes encombrées et la reconnaissance d'objets.

De plus, de nombreuses applications technologiques peuvent en bénéficier. Plusieurs facteurs influençant l'attention visuelle ascendante ont été découverts par des chercheurs en comportement et ont encore inspiré la communauté de la modélisation.
La plupart des recherches de modélisation précédentes se sont concentrées sur la composante ascendante de l'attention visuelle. Bien que les efforts antérieurs dans la recherche soient appréciés, plusieurs pistes restent encore à explorer, comme celle du manque de principes de calcul pour une attention axée sur les tâches, et celle de la prise en compte des exigences des tâches variant dans le temps, en particulier dans les environnements interactifs, complexes et dynamiques.
Au cours des dernières années, l'apprentissage en profondeur a connu une croissance rapide. De nombreux réseaux de neurones convolutifs et réseaux de neurones récurrents ont obtenu de bien meilleures performances dans diverses tâches de vision par ordinateur et de traitement du langage naturel, par rapport aux méthodes traditionnelles précédentes. Ainsi depuis 2016, le deep learning a été introduit dans le domaine de l'attention visuelle et diverses méthodes ont été proposées.

Chapitre 3

Fusion d'images

3.1 Introduction

La fusion consiste à produire une nouvelle image qui conserve une partie de l'information contenue dans chacune des images originales. L'objectif est de créer une synergie, c'est-à-dire d'obtenir une image fusionnée géométriquement et/ou sémantiquement plus riche qu'une image initiale.
De nombreuses méthodes sont capables de réaliser une fusion d'images. Elles diffèrent par la manière selon laquelle elles favorisent telle ou telle caractéristique des images originales. Le choix d'une méthode est donc conditionné par l'application. La mise en œuvre d'une fusion d'images nécessite plusieurs manipulations préalables qui interfèrent directement sur la qualité du produit fusionné.

Figure 3.1: Principe de la fusion d'images

Les contraintes principales seront donc liées à l'obtention d'une image aisément interprétable par une majorité d'utilisateurs. L'image devra présenter des qualités esthétiques, en particulier vis-à-vis des couleurs restituées, et offrir une lisibilité maximale. Enfin, dans la mesure du possible, l'obtention de couleurs naturelles sera recherchée.

La fusion peut présenter plusieurs intérêts qui, le cas échéant, peuvent se cumuler, par exemple : l'amélioration de la résolution spatiale, la combinaison d'informations diachroniques ou la réduction du nombre d'images à traiter.
La fusion d'images fait référence à la technique consistant à réunir deux ou plusieurs images ou images en une seule image qui contient les structures et topographies significatives de chacune des images constituantes d'origine. En tant que tel, il s'agit d'une procédure qui permet de rassembler toutes les informations vitales à partir de plusieurs images et d'inclure ces informations dans moins d'images, de préférence une seule image. L'image résultante devient plus informative et précise par rapport à l'une des images constitutives.
La fusion d'images peut être obtenue à différents niveaux, le plus simple étant la fusion d'images au niveau du pixel. Cette technique consiste à faire une moyenne des images sources pixel par pixel. Cependant, cette méthode est associée à des lacunes qui augmentent la dépendance à l'égard d'autres méthodes fiables.

3.2 Classification des méthodes de fusion d'images

Ces dernières années, il y a eu une recrudescence des études explorant le domaine des techniques de fusion d'images multi-focales, qui a été associée à son application croissante dans des domaines autres que l'armée. Ainsi les travaux de recherche axés sur l'utilisation de la fusion d'images multi-focales dans la télédétection, la surveillance ainsi que le diagnostic médical, ont augmenté considérablement. Des applications ont également été réalisées dans le domaine de la photographie.
Les méthodes de fusion d'images multi-focales peuvent être classées en trois catégories (27), qui sont le domaine spatial, le domaine de transformation et les techniques basées sur l'apprentissage profond.

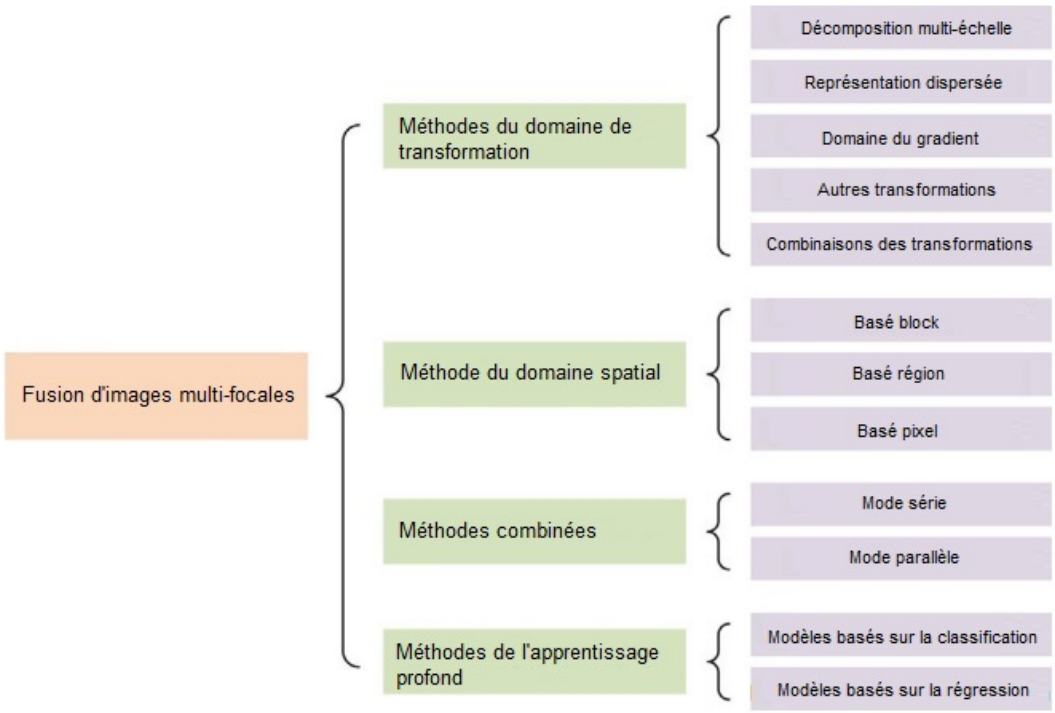


Figure 3.2: Classification des techniques de fusion d'images

3.2.1 Domaine spatial

La technique de fusion du domaine spatial utilise des caractéristiques spatiales locales telles que le gradient, la fréquence spatiale et l'écart type local. Les valeurs des pixels de deux images ou plus sont rassemblées et manipulées pour obtenir les résultats souhaités par la technique de fusion de domaine spatial.

Dans cette catégorie de méthodes, les images sources sont fusionnées dans le domaine spatial, c'est-à-dire en utilisant certaines caractéristiques spatiales des images. Par rapport aux méthodes de domaine de transformation, la caractéristique la plus importante des méthodes de domaine spatial est qu'elles ne contiennent pas l'étape de transformation inverse pour reconstruire l'image fusionnée.

Certaines méthodes de domaine spatial peuvent appliquer des techniques de transformation d'image telles que la transformation en ondelettes et la représentation éparse pour la mesure du niveau d'activité, mais elles n'ont pas besoin d'effectuer la transformation inverse. Dans les méthodes de domaine spatial, un objectif général consistant à générer une carte de poids pour chaque image source, et l'image fusionnée est calculée comme la moyenne pondérée de toutes les images sources.

Selon la manière de traitement de pixels adoptée, les méthodes de domaine spatial peuvent être regroupées en méthodes basées sur des blocs, des méthodes basées sur des régions et des méthodes basées sur des pixels.
Les techniques du domaine spatial sont sensibles au bruit et s'accompagnent de certaines limitations comme le flou d'image, la distorsion spatiale. Elles souffrent en plus d'artefacts dans l'image fusionnée en raison de mauvaises décisions dans les sous-régions (27). Ces problèmes peuvent être résolus en utilisant des méthodes de domaine de transformation.

3.2.2 Domaine de transformation

Les méthodes du domaine de transformation se déroulent en trois étapes principales comme illustré sur la figure 3.3. Dans la première étape les images sources sont converties en un

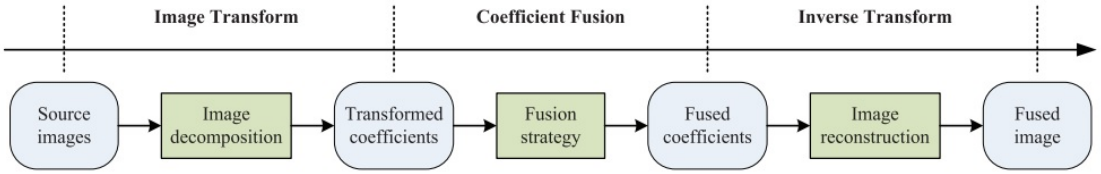


Figure 3.3: Schéma général des méthodes de domaine de transformation

domaine de transformation en appliquant une approche de décomposition/représentation d'images. Ensuite, les coefficients transformés sont fusionnés par une stratégie de fusion préconçue contenant des étapes tenant compte de la mesure du niveau d'activité, de la règle de fusion et de la vérification de cohérence. Enfin, l'image fusionnée est reconstruite en effectuant la transformée inverse correspondante sur les coefficients fusionnés.
Selon la transformation d'image appliquée, les méthodes principales de transformation peuvent être divisées en méthodes basées sur la décomposition multi-échelle (MSD), les méthodes basées sur la représentation clairsemée (SR), les méthodes basées sur le domaine de gradient (GD), les méthodes basées sur d'autres transformations et des méthodes basées sur la combinaison de différentes transformations. (27)

3.2.3 A base de l'apprentissage profond

Les techniques basées sur l'apprentissage profond considèrent la génération de carte de mise au point dans la fusion d'images comme un problème de classification. Plus précisément, la mesure du niveau d'activité est connue sous le nom d'extraction de caractéristiques, tandis que le rôle de la règle de fusion est similaire à celui d'un classificateur utilisé dans

les tâches de classification générale.
Ainsi, il est théoriquement possible d'utiliser des CNN pour la fusion d'images. L'architecture CNN pour la classification visuelle est un cadre de bout en bout, dans lequel L'entrée est une image tandis que la sortie est un vecteur d'étiquette qui indique la probabilité pour chaque catégorie. Entre ces deux extrémités, le réseau se compose de plusieurs couches convolutives (une couche non linéaire comme ReLU suit toujours une couche convolutive), des couches de pool max et des couches entièrement connectées. Les couches convolutives et de mise en commun maximale sont généralement considérées comme une partie d'extraction de caractéristiques dans le système, tandis que les couches entièrement connectées existant à l'extrémité de sortie sont considérées comme la partie de classification.
Les avantages de la méthode de fusion basée sur les CNN par rapport aux méthodes classiques sont doubles. Premièrement, il surmonte la difficulté de concevoir manuellement des règles compliquées de mesure du niveau d'activité et de fusion. Deuxièmement, la mesure du niveau d'activité et la règle de fusion peuvent être générées conjointement via l'apprentissage d'un modèle CNN. Le résultat appris peut-être considérer comme une solution « optimale » dans une certaine mesure, et est donc susceptible d'être plus efficace que les solutions conçues manuellement.

3.3 Description des méthodes de fusion d'images multi -focales

Dans cette section, nous allons décrire quelques méthodes récentes de fusion d'images multi-focales qui existant dans la littérature, selon la classification établie dans la section précédente.

3.3.1 Méthodes du domaine spatial

3.3.1.1 Méthodes basées pixel

Les méthodes du domaine spatial qui fonctionne au niveau des pixels traite directement de la position des pixels de l'image d'entrée. Ce sont les valeurs des pixels qui sont manipulées directement pour obtenir le résultat souhaité.
Pour cela, ils utilisent le principe de la somme pondérée linéaire, l'image fusionnée est donc calculée comme la somme pondérée de toutes les images sources et le problème central est d'obtenir la carte de poids pour chaque image source.
Dans ces méthodes, une mesure du niveau d'activité, également connue sous le nom de mesure de mise au point dans la fusion d'images multi-focales, est d'abord appliquée pour évaluer la saillance des pixels dans les images sources. Ensuite, les mesures de mise au

point obtenues à partir de différentes images sources sont comparées pour générer une carte de poids au niveau des pixels.

Dans la plupart des cas, la carte de poids appelée aussi carte de décision car la fusion d'images multi-focales peut être considérée comme un problème de classification dans lequel la propriété de mise au point (c'est-à-dire focalisée ou défocalisée) de chaque pixel est déterminée.

Pour certaines méthodes de fusion, la carte de poids est utilisée directement pour obtenir l'image fusionnée. Cependant, afin d'obtenir des poids ou des résultats de classification plus précis, davantage de méthodes tentent d'affiner la carte de poids ou de décision obtenue en ajoutant une étape de vérification de cohérence, dans laquelle diverses techniques de filtrage d'images sont fréquemment utilisées. Dans cette situation, les cartes poids/décision avant et après le raffinement sont généralement désignées par carte poids/décision initiale et carte poids/décision finale.

Par rapport à la règle de fusion précédente utilisée pour générer la carte de poids/décision initiale, la règle de fusion finale peut soit rester la même (par exemple, les deux sont une sélection maximale ou une moyenne pondérée) ou apporter un changement (par exemple, de la sélection maximale à la moyenne pondérée).

En outre, il existe également des méthodes basées sur les pixels qui visent à promouvoir les performances de fusion en concevant des règles de fusion plus compliquées. Par exemple, certaines méthodes divisent les images sources en régions avec des attributs différents (par exemple, focalisé/défocalisé/limite, texture/lisse) et adoptent différentes règles de fusion en fonction de leurs caractéristiques respectives.

Les méthodes de fusion de domaine spatial basées sur les pixels peuvent être classées sous trois aspects : la mesure du niveau d'activité, la règle de fusion et le raffinement de la carte poids/décision :

Les mesures conventionnelles du niveau d'activité telles que la variance, SF (spatial frequency), EOG (energy of gradient), EOL(energy of Laplacian), SML (sum-modified-laplacian) sont également fréquemment utilisées dans les méthodes de domaine spatial basées sur les pixels.

En outre, il existe également un certain nombre de méthodes basées sur les pixels qui adoptent les approches de décomposition d'image utilisées dans les méthodes de domaine de transformation comme mesure de mise au point.

Certaines méthodes basées sur les pixels appliquent des modèles d'extraction de caractéristiques plus avancés pour obtenir des mesures de mise au point robustes. Des exemples

typiques incluent la méthode basée sur LBP (51), la méthode basée sur la fréquence locale orientable, la méthode basée sur SIFT dense (25), la méthode basée sur la surface, la méthode basée sur le modèle de bord, la méthode basée sur la structure saillante (24), la méthode basée sur la matrice Hessienne, etc. Enfin, de nombreuses méthodes basées sur les pixels tentent en se basant sur des approches de filtrage d'images, d'obtenir une mesure de mise au point en calculant la différence d'image entre l'image source d'origine et son image lissée.

Pour la règle de fusion, la sélection maximale et la moyenne pondérée restent les règles les plus largement utilisées dans les méthodes de fusion basées sur les pixels.

L'hypothèse de base de ces méthodes est que les régions avec des attributs différents doivent être fusionnées par des règles différentes. Une méthode courante consiste à diviser les images sources en régions focalisées, en régions défocalisées et en régions limites, et la fusion des régions limites nécessite généralement des schémas plus complexes pour améliorer la qualité visuelle des images fusionnées.

Il existe également des méthodes basées sur les pixels qui convertissent la tâche d'estimation de la carte de poids en résolution d'un problème d'optimisation, telles que la méthode basée sur un modèle variationnel, les méthodes basées sur les marches aléatoires (RW), le la méthode basée sur le champ aléatoire conditionnel (CRF), la méthode basée sur le modèle multi-matting, etc.

Les approches de raffinement de la carte poids/décision sont basées sur des techniques de filtrage d'images, qui incluent le filtrage morphologique et le filtrage statistique. L'objectif étant d'éliminer les régions isolées susceptibles d'être mal classées dans la carte de décision initiale, tandis que les filtres préservant les contours tels que le filtre guidé et le filtre bilatéral sont principalement conçus pour rendre les poids en régions frontalières plus lisses et naturelles. Il existe d'autres approches de raffinement de carte de poids/décision qui incluent celles basées sur le champ aléatoire de Markov (MRF), celles basées sur RW, celles basées sur la coupe normalisée, celles basées sur l'appariement de caractéristiques locales, basé sur une coupe de graphe, basé sur un modèle de contour actif, etc.

3.3.1.2 Méthodes basées blocs

En 2001, Li et al. (22) ont introduit une méthode de fusion d'images multi-focales dans le domaine spatial basée sur un schéma de division de blocs, dans lequel chaque image source est divisée en un certain nombre de blocs de taille fixe. La fréquence spatiale est utilisée comme mesure du niveau d'activité de chaque bloc et une règle de fusion adaptative basée sur un seuil est utilisée pour obtenir le bloc fusionné. L'image fusionnée est finalement

construite après l'application d'une approche de vérification de cohérence basée sur un filtrage majeur.

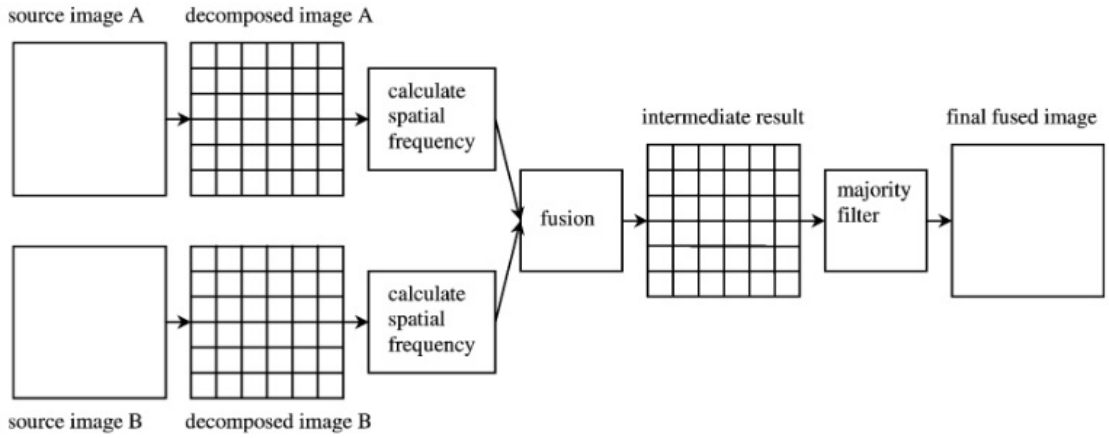


Figure 3.4: Schéma de principe pour la fusion d'images multi-focales: Méthode de Li et al basée blocs (22)

Depuis lors, les méthodes basées sur les blocs ont émergé comme une direction active dans la fusion d'images multi-focales et diverses améliorations ont été apportées à la mesure du niveau d'activité, à la règle de fusion, à la stratégie de division des blocs, etc.
Nous allons, dans ce qui suit donné un aperçu de quelques techniques, nous allons principalement insisté sur les différences et les améliorations proposées en s'appuyant sur l'état de l'art présenté dans la référence (27).
Huang et Jing (14) ont présenté une évaluation d'un ensemble de mesures de mise au point fréquemment utilisées dans la fusion d'images multi-focales, notamment la variance, l'énergie de gradient (EOG), l'énergie du laplacien (EOL), le laplacien à somme modifiée (SML), la fréquence spatiale. (SF), etc. Ils ont également proposé une mesure du niveau d'activité combiné EOL et PCNN pour la fusion basée sur des blocs, conduisant à un schéma populaire (c'est-à-dire combinant une mesure de mise au point traditionnelle avec un modèle PCNN) dans la fusion d'images multi-focales. Ils ont aussi présenté un schéma générique pour les techniques de fusion d'images multi-focales basée sur la sélection de blocs figure 3.5.

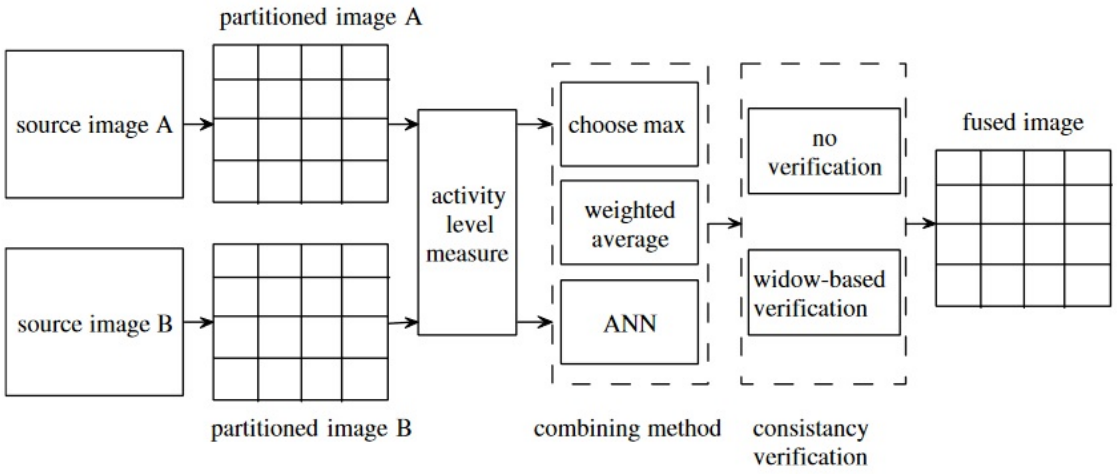


Figure 3.5: Schéma générique pour la fusion d'images multi-focales basée sur la sélection de blocs d'images à partir d'images sources. (14).

Zhan et al. ont proposé une mesure de mise au point basée sur la congruence de phase (PC) pour fusionner des images multi-focales. Les mesures du niveau d'activité basées sur la DCT ont également été appliquées à la fusion d'images multi-focales par blocs. En plus de développer des mesures d'activité plus efficaces, certains chercheurs ont tenté d'appliquer des mesures d'activité multiples pour remplacer la manière ci-dessus basée sur une mesure unique et concevoir des modèles de classification correspondants comme règle de fusion pour combiner les blocs d'images sources.
Li et al. ont proposé une règle basée sur le réseau de neurones pour déterminer la propriété de mise au point des blocs d'image source via les mesures d'activité, y compris la SF, la visibilité et la caractéristique de bord.
Kausar et Majid ont introduit la forêt aléatoire comme classificateur pour la détermination de la propriété de mise au point sur la base de neuf caractéristiques locales couramment utilisées telles que la visibilité, la SF, la variance, l'EOG, les caractéristiques basées sur DWT et les caractéristiques basées sur DCT. Un schéma de classification basé sur le vote majoritaire a également été adopté dans cette catégorie de méthodes de fusion. Toutes les méthodes de fusion par blocs mentionnées ci-dessus sont basées sur une taille de bloc fixe qui est définie empiriquement. De toute évidence, la taille du bloc a un impact crucial sur les résultats de fusion finaux. La manière basée sur une taille fixe est très susceptible d'introduire des effets de blocage indésirables dans l'image fusionnée. Pour résoudre ce problème, des stratégies améliorées de division de blocs ont été proposées par les chercheurs. Zhang et al. ont appliqué l'algorithme génétique pour obtenir une taille de bloc optimale

pour la fusion d'images multifocales. L'algorithme génétique définit la taille du bloc en tant que variable 2D (hauteur et largeur) à optimiser. En adoptant une fonction de fitness appropriée comme objectif, la solution optimale peut être obtenue à partir de valeurs aléatoirement initialisées avec un certain nombre de générations, qui contiennent des opérations telles que le croisement, la mutation et la sélection.

Aslantas et Kurban ont proposé une approche similaire d'optimisation de la taille des blocs basée sur un algorithme d'évolution différentielle pour la fusion d'images multi-focales. D'autres algorithmes de calcul évolutifs tels que l'optimisation des essaims de particules, l'optimisation basée sur la biogéographie et la colonie d'abeilles artificielles ont également été utilisés dans la fusion d'images multi-foyers par blocs . Cependant, dans ces méthodes basées sur l'optimisation, la taille de bloc obtenue par l'algorithme d'optimisation est toujours fixe pour une image donnée, ce qui peut encore provoquer des effets de blocage.

Pour résoudre ce problème, De et Chanda ont appliqué la structure quad-tree pour réaliser une division de blocs adaptative pour la fusion d'images multi-focales. Les tailles des différents blocs d'une image donnée sont différentes et déterminées par le contenu spécifique qu'ils contiennent.

Bai et al. ont également proposé une méthode de fusion d'images multi-foyers basée sur un arbre quadruple avec une stratégie de décomposition en arbre quadruple améliorée et une mesure de mise au point basée sur SML. Dans les méthodes basées sur les blocs introduites ci-dessus, la fusion de chaque bloc est effectuée indépendamment des autres. Il existe également des méthodes basées sur les blocs qui se concentrent sur la relation entre les différents blocs pendant la fusion.

Dans leur papier, Wu et al. a proposé une méthode de fusion d'images multi-foyers basée sur la division de blocs superposés et a adopté le modèle de Markov caché (HMM) pour prendre en compte à la fois la clarté du bloc actuel et la compatibilité avec ses blocs voisins. Guo et al. ont proposé une méthode de fusion d'images multi-foyers basée sur des blocs dans laquelle une région adaptative pour chaque bloc est construite en fonction de la similitude entre deux patchs. Pour chaque région adaptative, les informations SML et de profondeur sont utilisées conjointement pour définir la mesure d'activité pour la fusion. Zhang et Levine ont proposé un modèle de représentation éparse robuste et multitâche pour obtenir une mesure d'activité pour la fusion d'images multi-foyers. Dans leur méthode, un bloc image et ses 8 voisins connectés sont décomposés conjointement pour obtenir les coefficients clairsemés et les erreurs de reconstruction pour le calcul de la mesure d'activité.

3.3.2 Méthodes du domaine de transformation

3.3.2.1 Méthodes basées sur la décomposition multi-échelle

a- Méthodes basées sur la transformée en Ondelette

La transformée en ondelettes discrète (DWT) est une autre approche pour la fusion d'images multi-focales qui a été explorée de manière approfondie par Pajares et Cruz (33). Cette approche implique de définir différentes règles de fusion pour la combinaison séparée des coefficients de sous-bande basse fréquence et haute fréquence. Ce processus est ensuite suivi d'une vérification de cohérence. Les coefficients DWT sont obtenus par la combinaison séparée des sous-bandes basse fréquence et haute fréquence. Ces coefficients sont ensuite combinés convenablement afin d'obtenir la meilleure qualité dans l'image suivante. Enfin, un test de cohérence est effectué pour vérifier la validité du résultat. L'image fusionnée finale est obtenue en effectuant la transformée en ondelettes discrète inverse (IDWT), suite à la combinaison des coefficients. L'inconvénient majeur associé à cette méthode est qu'elle donne une faible résolution spatiale de l'image fusionnée finale. Par conséquent, la complexité de l'algorithme de fusion détermine la qualité des résultats. Malheureusement, les approches de fusion d'images multi-focales basées sur le DWT nécessitent l'application de plusieurs opérations de convolution. En conséquence, il y a beaucoup de consommation d'énergie et de temps passé pendant le traitement. Ce résultat rend ces méthodes moins applicables dans des scénarios en temps réel. De plus, le processus de transformation en ondelettes conduit à des bords manquants de l'image ; ainsi, certains artefacts importants de l'image peuvent être perdus et la qualité de l'image réduite. Ces problèmes associés à l'imagerie multi-focale DWT ont conduit les travaux de recherche à s'orienter vers le domaine de la transformée en cosinus discrète (DCT).

b- Méthodes basées les pyramides

La pyramide laplacienne est une approche courante pour réaliser la fusion d'images multi-focales. Wang and Chang (45) ont proposé une méthode d'utilisation de cette approche en trois étapes :

- Dans la première étape, les pyramides laplaciennes des images sources constitutives sont créées en trois sous-étapes. Cette étape commence au niveau inférieur, auquel cas un filtre passe-bas est appliqué. La réduction de la taille de l'image est ensuite réalisée par sous-échantillonnage, après quoi l'interpolation et la différenciation sont effectuées.

- La deuxième étape de l'approche de la pyramide laplacienne implique l'utilisation des règles de fusion pour fusionner chacun des niveaux de la nouvelle pyramide laplacienne. La règle d'information de région maximale est appliquée au niveau supérieur, tandis qu'aux niveaux restants, elle adopte la règle d'énergie de région maximale.
- Dans la dernière étape, il devient possible d'obtenir l'image fusionnée par l'inverse de la transformée pyramidale laplacienne.

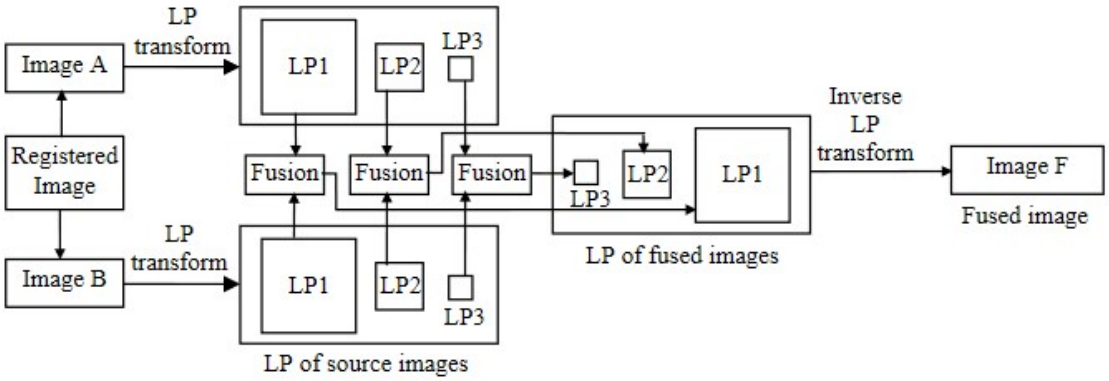


Figure 3.6: Principe de la technique à base de pyramides. (45).

Cette approche de fusion d'images possède l'avantage de produire une performance passionnante car elle préserve mieux les informations de bord dans une image. Cependant, cette approche ne s'adapte pas aux changements d'intensité soudains dans la résolution spatiale des images constitutives2

c- Méthodes basées sur l'analyse en composantes principales (ACP)

La méthode d'analyse en composantes principales (ACP) est une autre approche standard de fusion d'images qui relève des techniques du domaine spatial(29). Il s'agit d'une méthode d'analyse de données qui implique la transformation de variables corrélées en de nouvelles variables non corrélées les unes aux autres. Cette caractéristique est fondamentale pour éviter la redondance des informations contenues dans une image en raison de la réduction du nombre de variables. La fusion des images est réalisée en effectuant une moyenne pondérée des images individuelles à fusionner.

Pour chaque source d'image, les poids sont obtenus à partir du vecteur propre qui correspond à la plus grande valeur propre des matrices de covariance. Cette méthode est associée à l'avantage de favoriser l'atteinte de bonnes performances. Cependant, cela peut également produire une dégradation spectrale, ce qui est un résultat indésirable.

La possibilité de combiner les techniques de la pyramide laplacienne et de l'ACP a été mise en évidence dans les travaux de Verma et al. en 2016. La méthode de la pyramide laplacienne est utilisée pour créer différents niveaux d'une image d'entrée. En utilisant cette approche, les niveaux supérieurs sont fusionnés à l'aide de l'algorithme PCA, tandis que la méthode traditionnelle de fusion d'images DWT est utilisée pour fusionner les autres niveaux. La combinaison des deux approches est associée à des avantages tels que la préservation des bords et le lissage de l'image. Malheureusement, l'image résultante peut également avoir un contraste réduit. Certaines des lacunes constatées dans la combinaison des méthodes laplaciennes et ACP ont incité des travaux de recherche à améliorer l'efficacité.

Une solution fiable a été proposée par Yin et al. (50), grâce à la technique proposée basée sur la détection compressive. Les images sources initient un processus de décomposition à l'aide d'une transformée de contourlet non sous-échantillonnée. Dans cette technique, un modèle de réseau de neurones à couplage d'impulsions à double couche est appliqué dans la combinaison des sous-bandes du passe-bas. En outre, la règle de fusion basée sur la rétention des bords est utilisée pour combiner les sous-bandes passe-haut. La matrice aléatoire gaussienne est ensuite utilisée pour fusionner les coefficients creux.

Enfin, l'image fusionnée est reconstruite à l'aide de l'algorithme Compressive Sampling Matched Pursuit. La nécessité de réduire l'erreur de reconstruction nécessite de calculer les coefficients fusionnés à l'aide d'une matrice de mesure.

Bavirisetti et al. ont proposé l'utilisation d'un filtre guidé pour réaliser la fusion d'images. Cette méthode commence par la décomposition multi-échelle des images sources à l'aide d'un filtre GF. Ensuite, une carte de saillance de deuxième phase est générée avec des cartes de poids de calcul qui correspondent aux couches de détail. Enfin, l'image fusionnée finale est générée lorsque les couches de détails sont combinées à l'aide de cartes de poids. L'amélioration majeure de la méthode proposée est associée à la détection de la saillance visuelle basée sur le filtre d'image guidée (GF) (6), une fonctionnalité qui a permis d'extraire les aspects visuellement importants à partir d'images différentes de la même scène.

Par conséquent, la nécessité de surmonter les problèmes susmentionnés a incité notre proposition à utiliser trois modèles de carte de saillance de la scène focalisée pour la fusion d'images multifocales.

3.3.2.2 Méthodes basées sur le domaine du gradient

L'idée de base de la méthode de fusion d'images basée sur le domaine du gradient est de fusionner les représentations de gradient des images sources, puis de reconstruire l'image fusionnée en limitant son gradient au gradient fusionné. Dans (38), Piella a proposé une approche variationnelle pour la fusion d'images dans le domaine du gradient basée sur le tenseur de structure. La technique fonctionne comme suit :

- Les images sources sont empilées dans une image à valeurs multiples et son tenseur de structure est calculé sur la base des cartes de gradient de chaque image source.
- Le tenseur de structure contient donc les informations de gradient combinées de toutes les images source, et le gradient cible peut être représenté par les valeurs propres et les vecteurs propres du tenseur de structure.
- Une version pondérée du tenseur de structure est adoptée dans cette méthode pour de meilleures performances et la carte de poids de chaque image source est définie comme sa magnitude de gradient normalisée.
- Enfin, l'image fusionnée est reconstruite sous une forme optimisée en minimisant la différence entre le gradient de l'image fusionnée et le gradient cible obtenu à partir du tenseur de structure pondéré.

Socolinsky et Wolff ont proposé une approche de fusion d'images qui intègre les informations d'un ensemble de données d'images multispectrales. Ils généralisent le contraste d'image, lié aux gradients d'image, en le définissant pour les images multispectrales en termes de géométrie différentielle. Ils utilisent ces informations de contraste pour reconstruire le champ vectoriel de gradient optimal, afin de produire l'image fusionnée.
Plus tard, Wang et al. ont fusionné les images dans le domaine de gradient en utilisant des poids dépendant des variations locales d'intensité des images d'entrée. A chaque position de pixel, ils construisent une matrice de contraste pondérée par l'importance. La racine carrée de la plus grande valeur propre de cette matrice donne la magnitude du gradient fusionné, et le vecteur propre correspondant donne la direction du gradient fusionné.
Récemment, Hara et al. ont utilisé un schéma de pondération inter-image pour optimiser la somme pondérée de l'amplitude du gradient, puis reconstruire les gradients fusionnés pour produire l'image fusionnée. L'étape d'optimisation a tendance à ralentir cette technique. De plus, leur technique comprend une carte de saillance intra-image seuillée manuellement, nécessitant l'intervention de l'utilisateur.

Enfin Paul et al en 2016(34), ont proposé dans « Multi-Exposure and Multi-Focus Image Fusion in Gradient Domain », un nouvel algorithme de fusion d'images. L'algorithme peut être appliqué pour fusionner une séquence d'images en couleur ou en niveaux de gris. voir Figure 3.7.

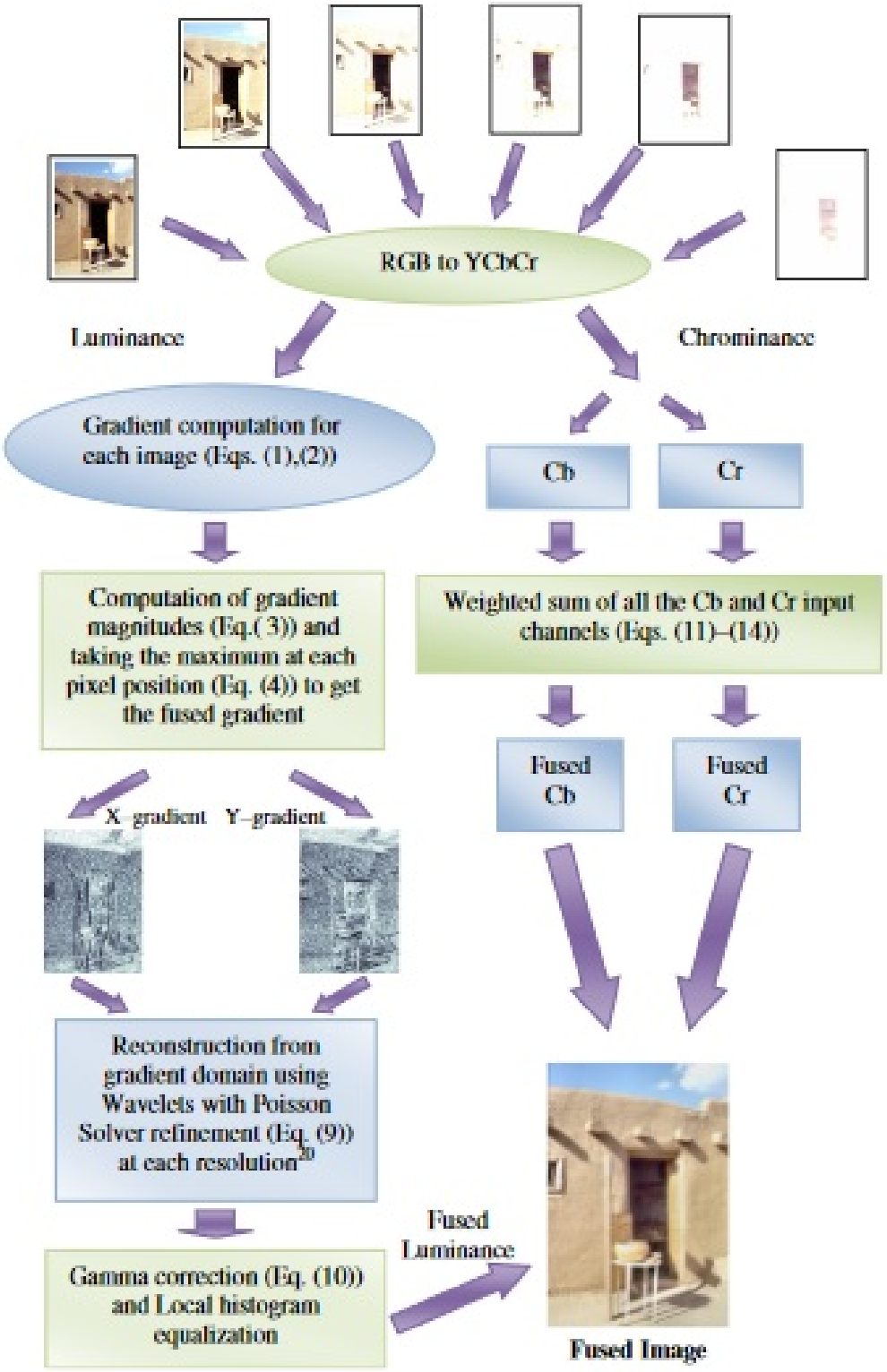


Figure 3.7: Organigramme de l'algorithme de fusion d'images proposé par Paul et al. (34).

L'algorithme proposé fonctionne dans l'espace colorimétrique $YCbCr$.a Le canal de luminance (Y) représente les informations sur la luminosité de l'image et c'est dans ce canal que les variations et les détails sont les plus visibles, car le système visuel humain est plus sensible à la luminance (Y) que à la chrominance (C_b, C_r). Cette observation importante a deux conséquences principales pour l'algorithme de fusion proposé. Tout d'abord, il indique que la fusion des canaux de luminance et de chrominance doit se faire de manière différente, et que c'est dans le canal de luminance que la partie la plus avancée de la fusion doit être effectuée. Deuxièmement, il révèle que la même procédure utilisée pour la fusion des canaux de luminance peut être utilisée pour fusionner des images à canal unique (c'est-à-dire des images en représentation en niveaux de gris). Dans ce qui suit, la technique de fusion de luminance proposée est décrite, suivie de la fusion de chrominance.

3.3.2.3 Méthodes basées sur la DCT (Discrete Cosine Transform)

Les travaux de Haghighat (10) ont été une approche efficace pour la fusion d'images multi-focus basées sur la variance calculée dans le domaine DCT. L'algorithme de fusion a été appliqué à trois étapes. Lors de la première étape, les images sources ont été divisées en partitions de 8 x 8 blocs, après quoi les coefficients DCT pour chaque bloc ont été calculés. Cette étape est suivie du calcul des variances des blocs correspondants à partir des images sources comme mesure d'activité. Enfin, le bloc avec un niveau d'activité élevé est considéré comme le bloc approprié à partir de l'image source. Cette approche a été associée à des inconvénients tels que l'impossibilité de générer une image focalisée souhaitable, en plus d'être coûteuse en temps de calcul. Un besoin a été identifié dans la recherche pour résoudre les limites de l'approche de fusion conventionnelle basée sur la détection compressive. Ces limitations découlent de facteurs tels que le caractère aléatoire de la matrice de mesure et les différentes propriétés de divers coefficients clairsemés décomposés.

3.3.3 Méthodes basées sur l'apprentissage profond

Depuis 2017, le deep learning a été introduit dans le domaine du MFIF et diverses méthodes ont été proposées. Les méthodes MFIF basées sur l'apprentissage profond peuvent être classées de différentes manières.(53)
La plupart des algorithmes MFIF basés sur l'apprentissage profond existants sont supervisés, tandis que plus de 10 approches non supervisées ont été proposées depuis 2018. Les méthodes MFIF basées sur l'apprentissage profond peuvent également être regroupées en méthodes basées sur une carte de décision et de bout en bout.

Dans les approches basées sur la carte de décision, une carte de décision qui indique le niveau de concentration (ou le niveau d'activité) est générée en premier. La fusion d'images est alors effectuée selon cette carte de décision. Dans ces méthodes, l'apprentissage en profondeur est normalement appliqué à la génération de carte de décision, suivie d'éventuelles étapes de post-traitement pour obtenir une meilleure carte de décision.
En revanche, les algorithmes MFIF de bout en bout produisent directement l'image fusionnée en alimentant le réseau en images sources. Il convient de mentionner que la plupart des méthodes basées sur des cartes de décision sont similaires aux algorithmes basés sur le domaine spatial, tandis que les méthodes de bout en bout sont similaires à celles basées sur le domaine de transformation.

3.3.3.1 Méthode supervisée basée sur l'apprentissage profond: Méthodes basées sur CNN

Liu et al. (26) ont proposé la première méthode MFIF basée sur CNN, qui utilise un CNN pour apprendre un mappage des images sources vers la carte de mise au point, comme le montre la figure 3.8. De cette façon, la mesure du niveau d'activité et la règle de fusion qui ont été traitées séparément dans les méthodes conventionnelles sont apprises conjointement. Après un ensemble d'étapes de post-traitement, la carte de décision finale est obtenue, qui est ensuite utilisée pour produire une image fusionnée via une approche moyenne pondérée au niveau des pixels. Depuis lors, diverses méthodes basées sur CNN ont été proposées, notamment des méthodes basées sur des cartes de décision et des méthodes de bout en bout.

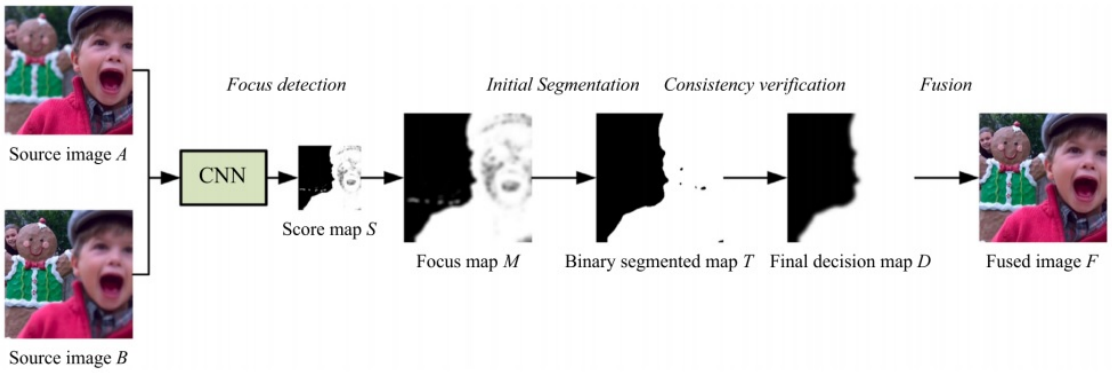


Figure 3.8: Le schéma de principe de la méthode MFIF basée sur CNN. Le CNN produit la carte de mise au point, qui est ensuite traitée par des étapes de post-traitement. (26).

3.3.3.2 Méthodes basées sur l'apprentissage profond non supervisé

Les méthodes supervisées nécessitent une grande quantité de données d'apprentissage étiquetées. Cependant, la vérité terrain n'est normalement pas disponible dans MFIF, donc presque toutes les méthodes supervisées utilisent des ensembles de données synthétiques qui ne sont en fait pas réalistes. Par conséquent, il est souhaitable d'avoir des méthodes MFIF non supervisées. Les méthodes sont entièrement convolutives et la plupart d'entre elles sont des algorithmes de bout en bout. voir figure 3.9

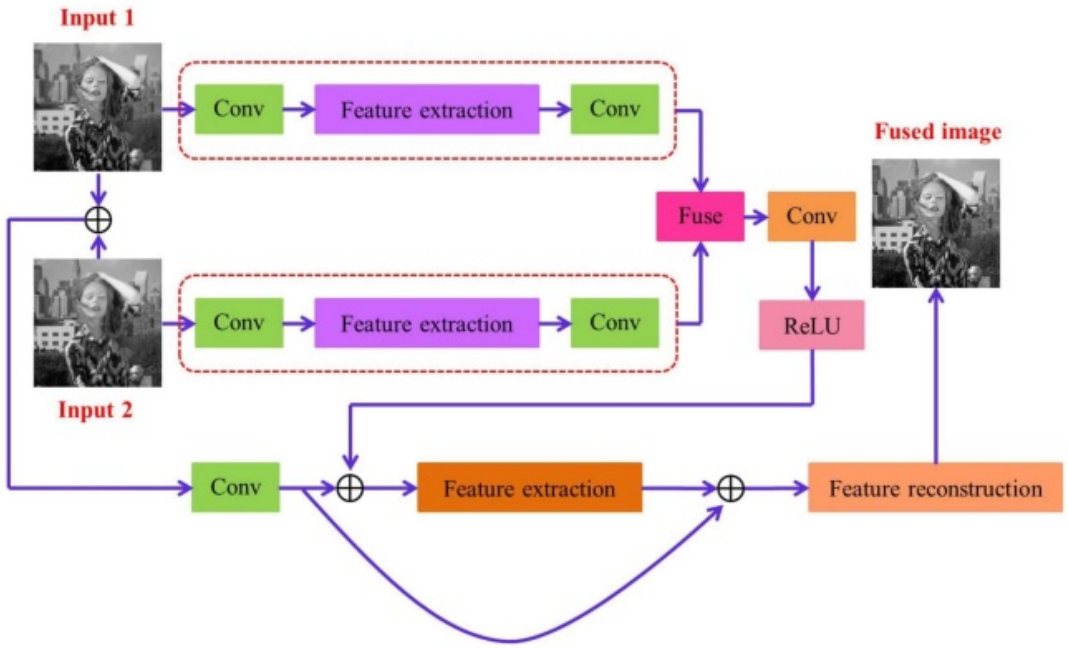


Figure 3.9: L'architecture réseau du MFNet (47).

Comme montré sur la figure 3.9 Yan et al. (47) ont proposé le premier algorithme MFIF non supervisé basé sur CNN (MFNet). La clé pour réaliser un entraînement non supervisée dans MFNet est l'utilisation d'une fonction de perte basée sur le SSIM de non-référence, qui est une métrique d'évaluation de fusion d'images largement utilisée mesurant la similarité structurelle entre les images source et l'image fusionnée. La fonction de perte est :

$$Loss(x_1, x_2, \hat{y}) = 1 - \sum_{\omega=1}^{N} S(x_1, x_2, \hat{y}|\omega) \tag{3.1}$$

où x_1, x_2, $\hat{y}$ sont les images sources et l'image fusionnée, ω est une fenêtre locale, et S est définie comme la valeur SSIM entre le patch image x et$\hat{y}$ sur une fenêtre locale ω.
En entraînant le modèle d'apprentissage en profondeur, les paramètres sont optimisés pour

réduire la valeur de SSIM, améliorant ainsi la similitude structurelle entre les images source et l'image fusionnée. De cette façon, l'étiquette de vérité terrain n'est pas nécessaire. MFNet prend une paire d'images multi-focus en entrée et peut directement sortir l'image tout-focus.
MFNet se compose de quatre sous-réseaux principaux : trois sous-réseaux d'extraction de caractéristiques et un sous-réseau de reconstruction de caractéristiques.
Les sous-réseaux d'extraction de caractéristiques sont chargés d'extraire les caractéristiques des images source et de la moyenne de deux images d'entrée. Le sous-réseau de reconstruction est utilisé pour produire l'image fusionnée en fonction des caractéristiques.

3.4 Conclusion

Dans ce chapitre, nous avons présenté un aperçu complet des méthodes de fusion d'images multi-focus existantes. Nous avons pour cela utiliser plusieurs références mais la principale source était l'état de l'art établi en 2020 dans le papier « Multi-focus image fusion: A Survey of the state of the art » par Yu Liua , Lei Wanga , Juan Chenga , Chang Li, Xun Chenb. (27).
Pour suivre les derniers développements dans ce domaine, une nouvelle taxonomie est introduite pour regrouper les méthodes de fusion d'images multi-focus existantes en quatre catégories principales : méthodes de domaine de transformation, méthodes de domaine spatial, méthodes combinant domaine de transformation et domaine spatial, et méthodes d'apprentissage en profondeur.
Chaque catégorie est ensuite divisée en plusieurs sous-catégories pour une classification claire. Nous nous sommes intéressés à la description d'un certain nombre de méthodes représentatives dans chaque sous-catégorie. Nous avons également présenté brièvement au début du chapitre les définitions et les motivations des techniques de fusion ainsi que les principales applications de la fusion d'images multi-focales.
Malgré les grands progrès réalisés ces dernières années, nous pensons qu'il reste encore plusieurs défis majeurs dans le domaine de la fusion d'images multi-focales.
Premièrement, la stratégie de fusion pour les régions frontalières doit encore être améliorée. Les régions limites indiquent les régions entre les régions focalisées et les régions défocalisées dans les images source et elles se situent généralement dans les zones où la profondeur a un changement brusque.
Deuxièmement, le problème d'enregistrement erroné causé par des objets en mouvement ou le bougé de l'appareil photo est rarement résolu par les méthodes de fusion existantes.

Dans la fusion d'images multi-focales, étant donné que les images sources sont capturées à des moments différents, des objets en mouvement ou un bougé de l'appareil photo peuvent se produire pendant le processus de capture, ce qui entraînera un défaut de repérage partiel ou global entre les différentes images sources.
Troisièmement, peu de travaux se concentrent sur le développement d'algorithmes de fusion d'images multi-focales pour des applications spécifiques dans les domaines de la biologie, de la médecine, de l'industrie, etc. La plupart des publications existantes sur la fusion d'images multi-focales ont adopté des images naturelles pour vérifier l'efficacité d'une nouvelle méthode proposée.

Chapitre 4

Étude comparative de la robustesse au bruit dans les modèles d'attention visuelle

4.1 Introduction et motivations

Les humains sont des experts pour identifier rapidement et avec précision l'objet du premier plan le plus visible de la scène. Plusieurs modèles de détection de saillance sont maintenant disponibles et ont montré des applications réussies dans divers domaines. En utilisant des techniques d'attention visuelle, nous examinons dans ce chapitre, si les effets du bruit ont une influence sur leur taux de reconnaissance.

Pour répondre à cette question, nous sélectionnons trois techniques récentes de calcul de la saillance visuelle : la détection de la saillance à l'aide de la technique de la carte de saillance humaine, la détection de la saillance à l'aide de la technique de transformation en contourlet (CT) et la détection de la saillance à l'aide de la technique de décomposition matricielle de faible rang et structurée (LSMD), et nous évaluons les effets du bruit et leur influence sur le taux de reconnaissance en utilisant les métriques, basées sur les courbes AUC, F-mesure et MAE.

Notre principale contribution est de corrompre une image par du bruit de différentes variances et d'étudier ses effets.

4.2 Étude comparative de la robustesse au bruit dans les modèles d'attention visuelle

Cette section est consacrée à une étude comparative des trois modèles de l'attention visuelle que nous allons décrire ci-dessous, la détection de saillance à l'aide de la technique de suivi oculaire humain, la détection de saillance à l'aide de la technique de transformation en contourlet (CT) et la détection de saillance à l'aide de la décomposition matricielle de faible rang et structurée (LSMD).

4.2.1 Description des techniques de l'attention visuelle

4.2.1.1 Détection de saillance à l'aide de suivi oculaire humain

Cette technique (18) utilise une approche d'apprentissage pour former un classificateur directement à partir des données de suivi oculaire humain, qui peuvent être utilisées pour prédire les emplacements de fixation. Les caractéristiques sont divisées en trois catégories : les caractéristiques de bas niveau qui incluent les caractéristiques d'un modèle de saillance simple décrit par Torralba (31) et Rosenholtz (41) et Itti et Koch (17) basé sur des pyramides de sous-bandes qui considèrent l'intensité, l'orientation et le contraste des couleurs comme des caractéristiques importantes pour la saillance ascendante et inclut les trois canaux. Les caractéristiques de niveau intermédiaire de deuxième catégorie, qui utilisent le détecteur de ligne d'horizon comme un aspect naturel d'un objet saillant, enfin les auteurs ajoutent la détection de visage au modèle et plusieurs autres fonctionnalités de niveau supérieur à la troisième catégorie.
Ces fonctionnalités sont donc utilisées pour entraîner et tester un modèle, l'ensemble d'images est divisé en images d'entraînement et de test, et SVM avec noyau linéaire est utilisé pour entraîner le modèle.

4.2.1.2 Détection de la saillance en utilisant la transformée en contourlet (CT)

Abouelaziz et al, dans (1) utilisent l'image couleur comme entrée et génèrent deux cartes de saillance (locale et globale) pour les combiner en une carte de saillance finale, donc la Transformée en Contourlet génère les cartes de caractéristiques. La carte de saillance locale est le résultat de la combinaison des cartes de caractéristiques à chaque niveau.

$$S_L(x,y) = \sum [argmax(f_s^L(x,y), f_s^a(x,y), f_s^b(x,y)] * I_{KxK} \tag{4.1}$$

Où f_s^L, f_s^a, f_s^b sont les cartes de caractéristiques à l'échelle s pour les canaux L,a,b. La carte de saillance globale calcule la distribution globale des caractéristiques locales. De fcs(x,y), un emplacement (x,y) peut être représenté comme un vecteur de caractéristiques avec une taille de (3 × N) de toutes les cartes de caractéristiques (24 caractéristiques pour chaque emplacement). La vraisemblance des caractéristiques à un endroit donné peut être définie par la fonction de densité de probabilité (PDF) avec une distribution normale.

$$S_G(x,y) = (\log(p(f(x,y))^{-1})^{1/2} * I_{KxK} \tag{4.2}$$

La carte de saillance finale est le résultat de la combinaison des deux cartes $S_G(x,y)$ et $S_L(x,y)$. La combinaison est définie comme :

$$S(x,y) = M(S_L(x,y) * e^{(S_G(x,y))}) * I_{KxK} \tag{4.3}$$

4.2.1.3 Détection de saillance à l'aide d'un modèle de décomposition matricielle creuse et structurée (LSMD)

Dans cette technique (35), la carte de saillance visuelle est calculée en quatre étapes : tout d'abord, nous commençons par partitionner l'image d'entrée afin d'extraire les caractéristiques de bas niveau comme la couleur RVB, les pyramides orientables et Filtre de Gabor, pour construire un espace de caractéristiques de dimension D, puis nous effectuons un clustering à décalage moyen en N patchs de base. La deuxième étape consiste à construire un arbre d'index pour représenter la structure de l'image en utilisant un algorithme de clustering de kmeans hiérarchique. L'arbre d'index de résultat stocke les informations de structure sous-jacentes d'une image originale et les impose à S en tant que contrainte structurée. La troisième étape consiste à décomposer la matrice F en :

$$min_{L,S} || L ||_* + \lambda * || S ||_l \tag{4.4}$$

s.t. F=L+S Où F est la matrice de caractéristiques d'une image d'entrée, qui peut être décomposée en une matrice de faible rang L plus une matrice creuse S correspondant respectivement à l'arrière-plan non saillant et à l'objet saillant (20). La dernière étape consiste à transformer la représentation de l'image du domaine des caractéristiques au domaine spatial.

4.2.2 Data sets utilisés et métriques d'évaluation

4.2.2.1 Data sets

Dans cette étude comparative, nous avons utilisé le jeu de données DUT-OMRON (48). Cet ensemble de données va nous servir pour évaluer les algorithmes de détection d'objets

saillants précédemment décrits sur des images avec plus d'un seul objet saillant et un arrière-plan relativement complexe. Il contient 5168 images naturelles de haute qualité, où chaque image est redimensionnée pour avoir une longueur de côté maximale de 400 pixels. Les annotations sont disponibles sous la forme de cadres de délimitation et de masques d'objets binaires au niveau des pixels. En outre, des annotations de fixation oculaire sont également fournies, ce qui rend cet ensemble de données approprié pour évaluer simultanément les modèles de localisation et de détection d'objets saillants ainsi que les modèles de prédiction de fixation.

4.2.2.2 Métriques d'évaluation

Nous évaluons les performances à l'aide des mesures utilisées dans (8) sur la base de la zone de chevauchement entre l'annotation de l'image terrain vérité (Ground Truth) et la prédiction de saillance, y compris la courbe PR (précision-rappel), la courbe ROC (caractéristique de fonctionnement du récepteur) et l'AUC (Area Under courbe ROC). Ils sont définis comme suit :

$$FPR = \frac{FP}{FP + TN} \tag{4.5}$$

$$TPR = \frac{TP}{TP + FN} \tag{4.6}$$

$$F_\xi = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{(\xi^2 + 1) \times precision \times recall}{(\xi^2 \times precision + recall)} \right) \tag{4.7}$$

$$MAE = \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{|S(i,j) - G(i,j)|}{H \times W} \tag{4.8}$$

La précision correspond au taux de pixels saillants correctement attribués, et le rappel est la fraction de pixels saillants détectés appartenant à l'objet saillant dans l'image terrain de vérité. Pour une carte de saillance en niveaux de gris, dont les valeurs de pixels sont dans la plage [0,255], nous faisons varier le seuil de 0 à 255 pour obtenir une série de segmentations d'objets saillants. La courbe PR est créé, en calculant la précision et la valeur de rappel sat à chaque seuil.

La courbe ROC peut également être générée sur la base des taux de vrais positifs et des taux de faux positifs obtenus lors du calcul de la courbe PR.

4.3 Expérimentation, évaluation et bilan

4.3.1 Comparaison avec la vérité terrain

1) Comparaison qualitative : Les résultats de la comparaison qualitative par la vérité terrain et les trois méthodes sont illustrés à la figure 4.1. La figure 4.1 montre que la plupart des méthodes de détection de saillance peuvent gérer des images bien simples avec un arrière-plan relativement homogène. Ils peuvent générer une carte de saillance de haute qualité, mais la meilleure comme nous le voyons sur la figure 4.1 et presque similaire à la vérité terrain est la méthode « LSMD ».

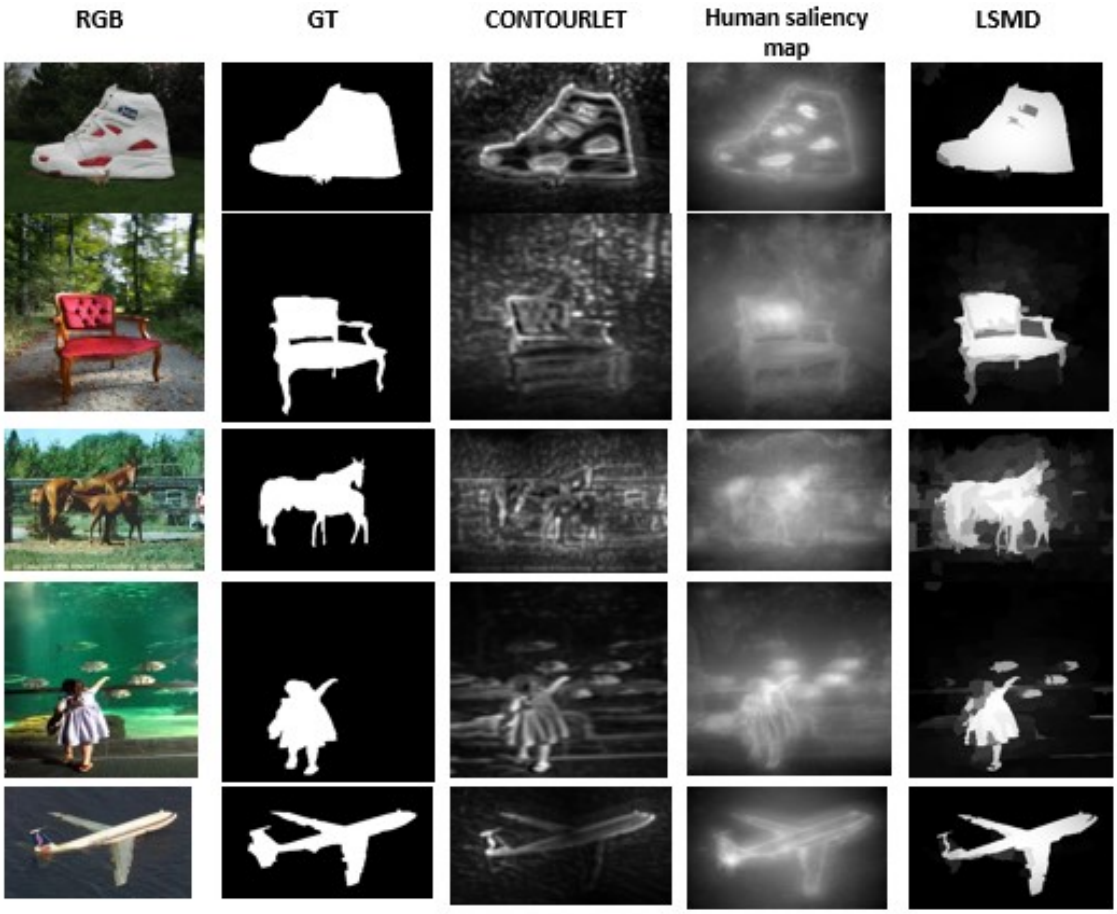


Figure 4.1: Comparaison qualitative : Colonne 1 : images RVB, colonne 2 : Image terrain de vérité, colonne 3 : carte contourlet, colonne 4 : carte de saillance suivi oculaire, colonne 5 : carte LSMD

2) Comparaison quantitative : La figure 4.2 montre les résultats quantitatifs des trois méthodes. On sait que la méthode "LSMD" est compétitive et qu'elle est meilleure que les autres méthodes en termes de courbe P-R, ROC et F-Measure. Surtout, la précision peut rester au-dessus de 80% dans une large plage de seuil. **A. Analyse de la carte de**

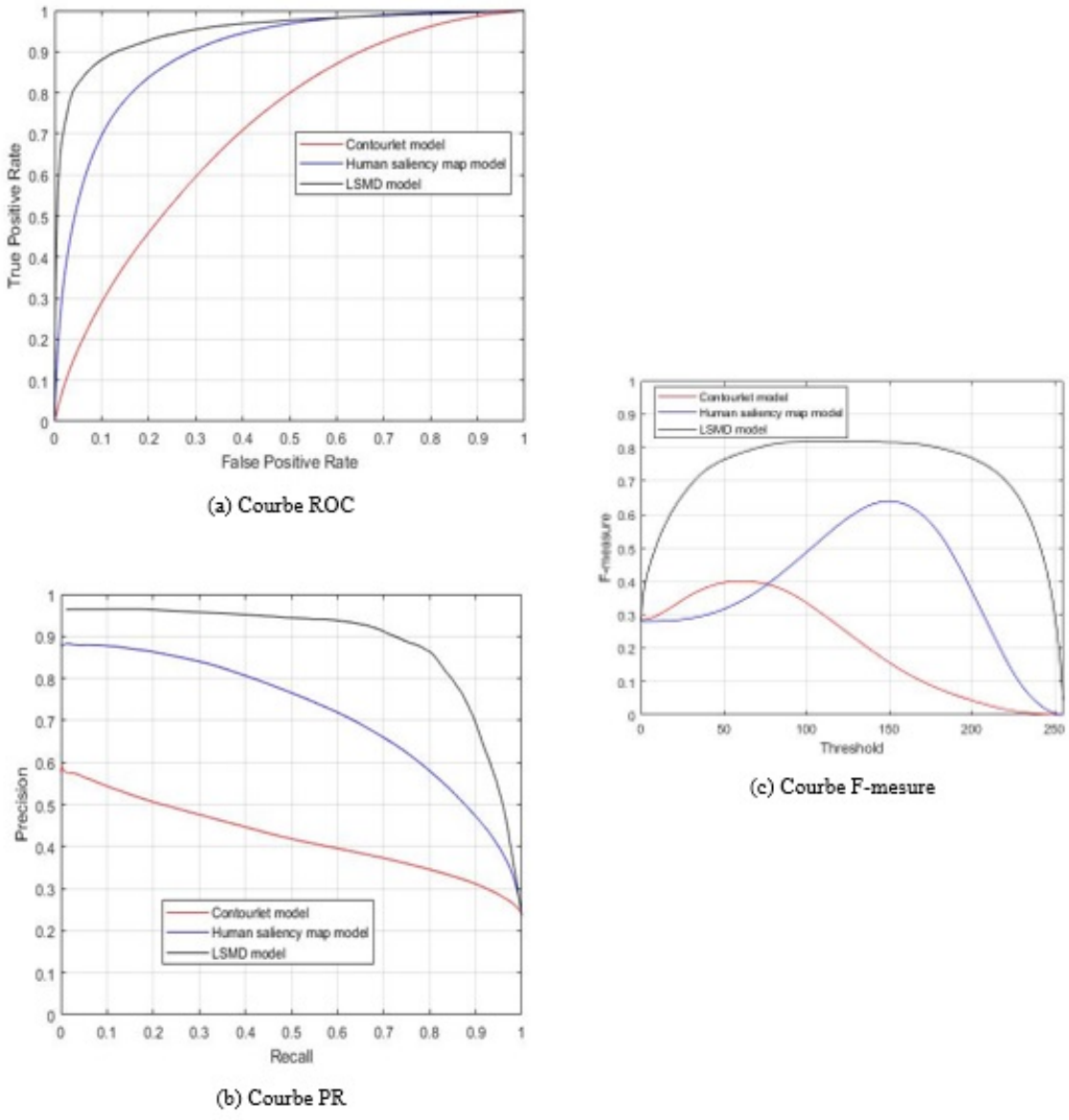
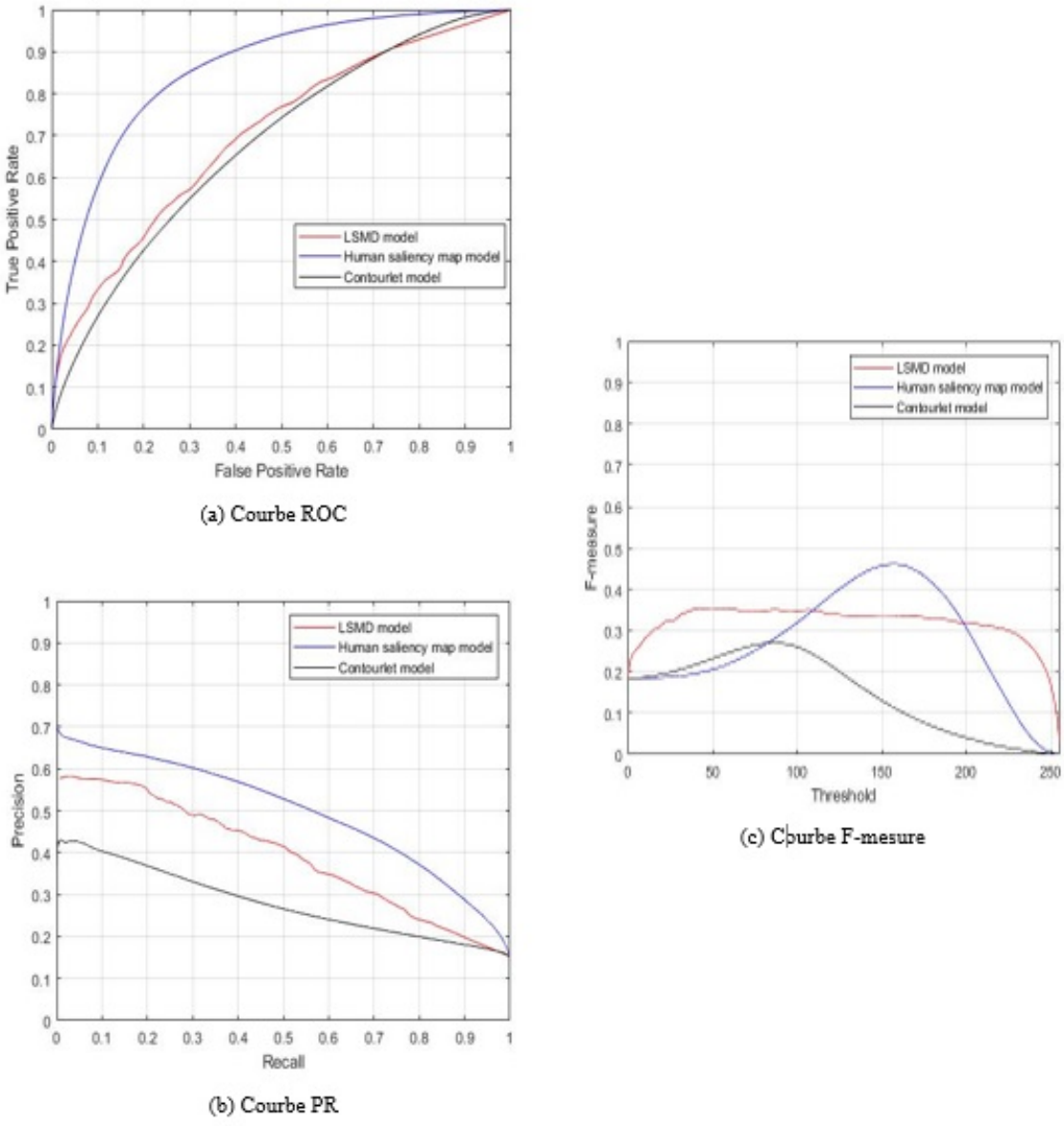


Figure 4.2: Comparaisons qualitatives (sans bruit) des cartes de saillance produites par différentes approches avec l'ensemble de données « DUT-OMRON » en termes de ROC, de courbes PR et de F-mesure c.

saillance par rapport au bruit.

Les images étudiées étaient corrompues par un bruit gaussien avec différentes variances,

dont 35/255, 25/255, 15/255 et 5/255. D'après le tableau 4.1, les figures 4.2, 4.3 et 4.4



(a) Courbe ROC

(b) Courbe PR

(c) Courbe F-mesure

Figure 4.3: Comparaisons quantitatives (ajout de bruit) des cartes de saillance produites par différentes approches avec l'ensemble de données « DUT-OMRON » en termes de ROC, de courbes PR et de F-mesure c.

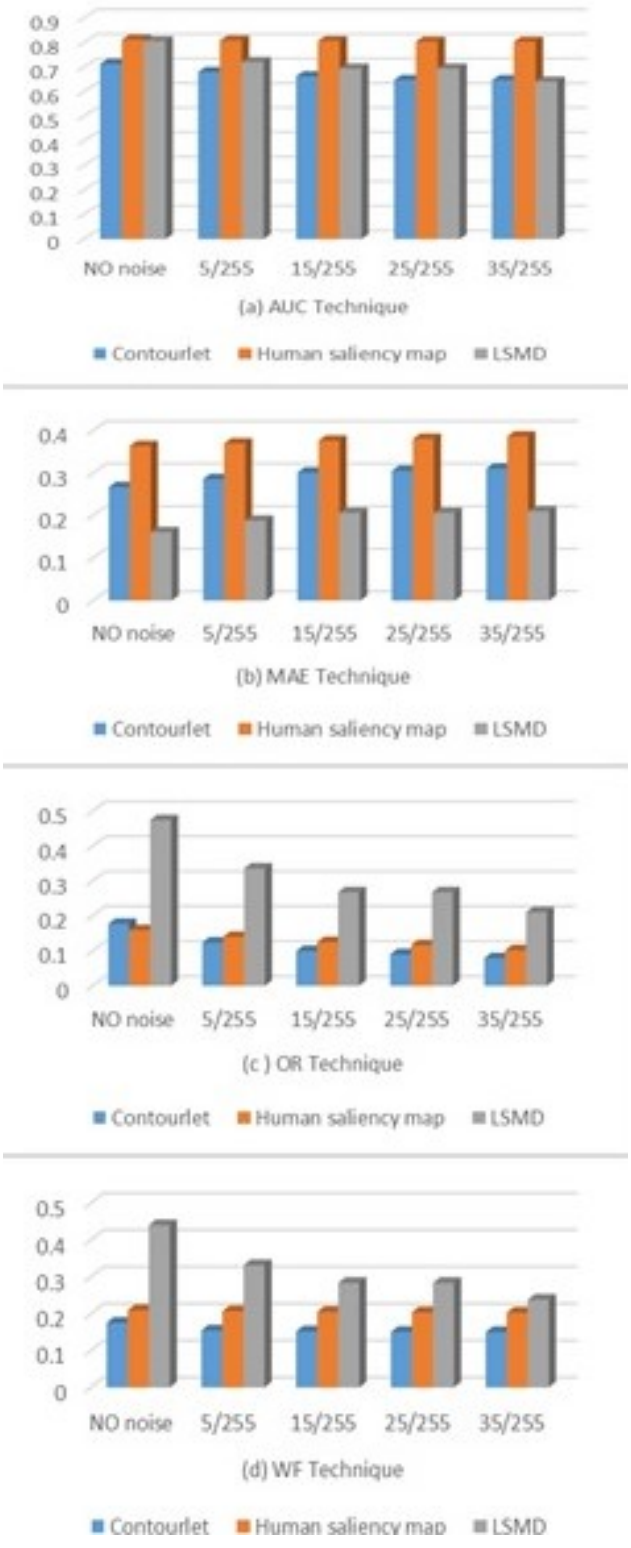


Figure 4.4: Importance des fonctionnalités dans trois modèles différents

(les cinq métriques d'évaluation quantitatives), nous pouvons voir que lorsque la variance augmente progressivement, le modèle de la carte de saillance humaine peut résister au bruit, mais le modèle LSMD est affecté par le bruit. Le modèle de carte de saillance humaine est robuste au bruit.

Variance	INDEX	Contourlet	Human saliency map	LSMD
NO noise	MAE↓	0,26712	0,36353	0,16115
	AUC↑	0,71403	0,81245	0,80639
	WF↑	0,17529	0,2091	0,4403
	OR↑	0,17742	0,15898	0,47398
5/255	MAE↓	0,28596	0,36948	0,18796
	AUC↑	0,68093	0,80893	0,72159
	WF↑	0,15485	0,20696	0,33167
	OR↑	0,12371	0,1402	0,33594
15/255	MAE↓	0,30148	0,37608	0,20632
	AUC↑	0,66339	0,80692	0,69476
	WF↑	0,15217	0,20511	0,28371
	OR↑	0,099769	0,12515	0,26781
25/255	MAE↓	0,30559	0,38041	0,20632
	AUC↑	0,64823	0,80447	0,69476
	WF↑	0,15078	0,20403	0,28371
	OR↑	0,090345	0,11631	0,26781
35/255	MAE↓	0,3104	0,38605	0,2103
	AUC↑	0,6478	0,80412	0,64204
	WF↑	0,15017	0,203	0,23827
	OR↑	0,079005	0,10195	0,21051

Table 4.1: Résultats expérimentaux avec différents bruits

4.4 Conclusion

Afin d'améliorer les résultats obtenus à partir de plusieurs modèles de détection de saillance, nous avons implémenté deux modèles informatiques de saillance visuelle.
Le premier est la détection de saillance à l'aide de la transformée Contourlet, ce modèle combinait deux cartes de saillance : locale et globale, celles-ci sont combinées pour produire une seule carte de saillance.
Le second que nous avons appelé modèle LSMD (Low-rank and Structured sparse Matrix Decomposition) qui a une excellente efficacité et atteint les performances supérieures sur l'ensemble de données de référence public et enfin le modèle de classification de la saillance visuelle « carte de la saillance humaine » basée sur le bottom-up calcul et sémantique d'image descendante pour correspondre aux mouvements oculaires réels.
Nous avons évalué les modèles de performances basés sur les courbes ROC, PR et F-mesure c, les performances globales du modèle LSMD sont bonnes et donnent de meilleurs résultats par rapport aux autres modèles, cependant lorsque l'image est polluée par le bruit, nous voyons que le modèle de Le LSMD est très sensible par rapport au reste des modèles. Sur la base de la discussion précédente et de l'étude comparative des techniques présentées, les travaux présenté dans ce chapitre nécessitent une étude plus approfondie à l'avenir et peuvent être étendus dans les directions suivantes :

- faire des recherches sur de nouvelles techniques basées sur des algorithmes d'apprentissage automatique, en particulier sur des architectures CNN.
- ajouter plus d'ensembles de données dans les tests.
- proposer une approche combinant deux techniques LSMD et CT puis tenter d'améliorer les performances, comme par exemple tirer parti des mérites des deux techniques.

Chapitre 5

Utilisation de la détection de saillance pour améliorer la fusion d'images multi-focales

5.1 Introduction

Nous nous intéressons dans ce chapitre à la fusion d'images multi-focales, en utilisant la saillance visuelle. Les valeurs de la carte correspondante sont utilisées comme des poids dans le processus de fusion des images d'entrée. En effet, le système de vision humaine (HVS) a une grande capacité à reconnaître et à focaliser les objets et les régions des scènes, qui sont visuellement plus distincts et plus visibles. Par conséquent, le calcul de la saillance visuelle représente une information essentielle dans le processus de fusion des images d'entrée. Nous avons intégré trois techniques de détection de saillance dans notre méthode, le choix est justifié par le fait qu'elles sont structurellement différentes, ce qui nous permettra de comparer l'influence de l'attention visuelle avec notre objectif principal : la fusion d'images multi-focales. Les principales contributions de notre proposition sont résumées comme suit :

- Nous utilisons trois algorithmes de détection de saillance visuelle (VSD) pour extraire les informations de saillance à partir d'images visuellement différentes.
- Un algorithme de fusion d'images basé sur une carte de poids, calculé à partir de trois méthodes récentes de saillance visuelle, permettant de ne conserver dans l'image fusionnée que les régions focalisées.

- Des simulations approfondies sont effectuées sur un ensemble de données de fusion d'images multi-focus (X.Zuzhang, 2019) et Lytro Multi-focus Dataset (30).

5.2 Description de la technique proposée

La méthode de fusion proposée est présentée plus en détail dans cette section. La recherche dans le domaine de la fusion d'images, a établi que les régions focalisées d'images fusionnées fournissent plus d'informations que les régions défocalisées lors de la comparaison d'images multi-focales (37). Petrovic et Xydeas (2005) expliquent cet événement en réitérant que les régions focalisées semblent être plus saillantes que les régions défocalisées. L'algorithme proposé a été développé sur la base d'une carte de poids dans le but d'intégrer des informations de région focalisée dans l'image fusionnée. L'un des trois modèles existants a été utilisé dans le calcul de la carte de saillance visuelle dans la première procédure. Cette étape a été suivie du calcul des cartes de poids des images sources individuelles, après quoi les résultats ont été normalisés dans l'intervalle [0,1]. Les chercheurs ont utilisé ces cartes de poids pour illustrer des informations complémentaires. La dernière étape impliquait la génération d'images fusionnées en obtenant le produit des cartes de poids et des images sources. L'algorithme 1 représente toutes les étapes nécessaires appliquées sur l'algorithme proposé, tandis que la figure 5.1 illustre la méthodologie proposée à travers un schéma fonctionnel. Un résumé des étapes importantes de l'algorithme proposé est également présenté dans la figure 5.1.

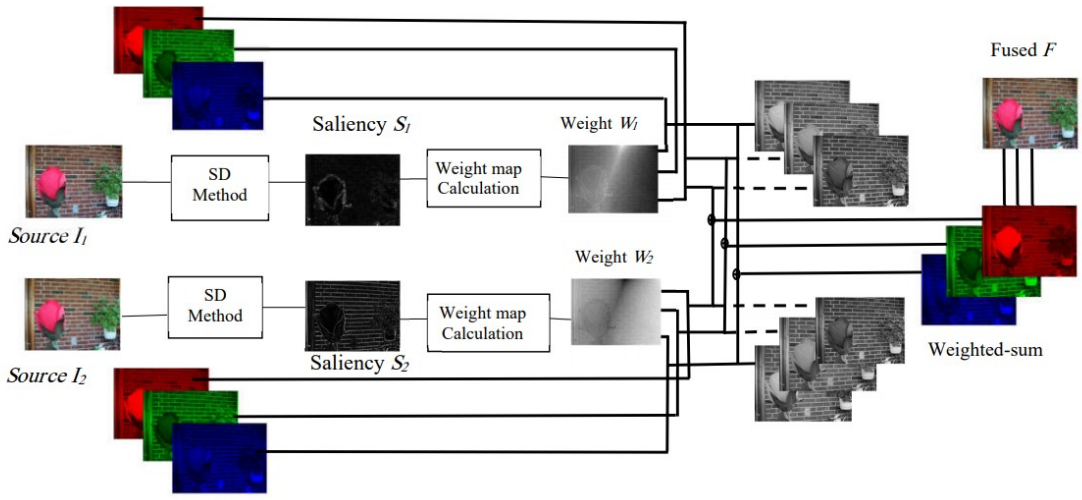


Figure 5.1: Principe de la méthode proposée

Première étape. Décomposer l'image d'entrée I en images rouge, vert et bleue $I_{R,G,B}$. Deuxième étape. Calculer les saillances visuelles de l'image d'entrée multi-focus en utilisant les trois algorithmes de détection de la saillance :

- Saliency Detection Using human saliency map
- Saliency Detection Using Contourlet Transform (CT)
- Saliency Detection Using low-rank and Structured sparse. Matrix Decomposition (LSMD) model

Troisième étape. Dans cette étape, nous devons combiner chaque région focalisée des images d'entrée en une seule image. Ce qui peut être accompli en attribuant des cartes de poids appropriées aux images d'entrée. Nous utilisons la normalisation des cartes de saillance comme poids pour identifier les régions focalisées et défocalisées des images d'entrée. Ensuite, les cartes de poids W_i sont déterminées à partir des cartes de saillance extraites en les normalisant comme :

$$W_i = \frac{S_i}{\sum_{k=1}^{2} S_k}, i = 1, 2 \tag{5.1}$$

Quatrième étape. Multipliez les images RVB avec les cartes de poids et fusionnez les images RVB pondérées pour obtenir une image finale.

$$F_{R,G,B} = \sum_{i=1}^{2} W_{i.} * I_{iR,G,B} \tag{5.2}$$

Algorithm 1: Algorithme Image fusion

Input: image $I_{1R,G,B}$, $I_{2R,G,B}$
Output: image $F_{R,G,B}$
for $k \leftarrow 1$ **to** 2 **do**
  $S_k = VSD(I_{kR,G,B})$
end
for $k \leftarrow 1$ **to** 2 **do**
  $W_k = \frac{S_k}{\sum_{i=1}^{2} S_i}$
end
$F_{R,G,B} = W_{1.} * I_{1R,G,B} + W_{2.} * I_{2R,G,B}$

5.3 Résultats et discussions

5.3.1 Evaluation des performances de la méthode proposée

5.3.1.1 Métriques utilisées

Divers critères objectifs ont été appliqués dans le processus d'évaluation des performances de la méthode de fusion proposée pour l'analyse comparative. La mesure de la quantité d'informations de bord transférées des images source à l'image fusionnée a été donnée en rémanence des bords $Q^{AB/F}$ (37). La similitude structurelle entre l'image source et l'image fusionnée est calculée à l'aide de la métrique basée sur la structure de l'image (SSIM) proposé par Yang et al. (2008) (49). Alors que la perte de fusion $L^{AB/F}$ est utilisée dans l'évaluation de la quantité d'informations perdues pendant le processus de fusion, les artefacts de fusion $N^{AB/F}$ sont applicables pour évaluer le nombre d'artefacts introduits au cours du processus de fusion (37).
Une meilleure performance de fusion est observée lorsque les valeurs de $L^{AB/F}$ and $N^{AB/F}$ sont relativement petites. Selon Haghighat et al. (2011) (11), la quantité d'informations sur les caractéristiques transférées de l'image source à l'image fusionnée peut être calculée à l'aide de la métrique d'informations mutuelles sur les caractéristiques (FMI). La meilleure qualité de fusion est enregistrée lorsque les valeurs de $Q^{AB/F}$, FMI et SSIM sont relativement élevées.
L'évaluation de la performance de la méthode proposée a été réalisée en comparant différents algorithmes en combinaison avec différentes métriques d'évaluation. Cependant, il est important de noter qu'une telle analyse de la performance ne pourrait être réalisée isolément étant donné que les différentes méthodes fonctionnent différemment selon les critères spécifiques utilisés. Tout en considérant toutes les métriques, cinq méthodes ont été comparées à l'aide d'un mécanisme de notation appelé score de classement. Ces mesures sont effectuées en supposant que toutes les métriques se présentent avec la même importance dès le début. La classification des cinq méthodes a été calculée selon les différentes métriques d'évaluation dans chaque paire d'images avec une plage de classement de 1 à 5 enregistrée selon l'ordre des cinq algorithmes. L'algorithme est représenté ci-dessous :

$$\begin{aligned} Rank = {} & Rank(FMI) + Rank(Q^{AB/F}) + Rank(SSIM) \\ & + Rank(N^{AB/F}) + Rank(L^{AB/F}) \end{aligned} \tag{5.3}$$

Il est important de noter que la meilleure performance globale de l'algorithme évalué est celle avec les plus petites valeurs du score de classement. Un exemple est indiqué dans le tableau 1 où l'image 1 sur le modèle d'Abouelaziz a un score de classement de 1 en FMI,

un score de 2 en $Q^{AB/F}$, un score de 4 en $L^{AB/F}$, un score de 1 en $N^{AB/F}$, et 1 dans QSSIM, ce qui fait que le score de classement global est de 9 à partir de la somme des scores, c'est-à-dire (1+2+4+1+1). De même, le classement final de MGF est de 17, ce qui indique que la performance globale du modèle d'Abouelaziz est bien meilleure que celle de MGF (6) lorsque l'on considère les métriques d'évaluation globales. Des comparaisons de la méthode proposée sont faites avec des méthodes populaires existantes telles que la fusion d'images guidées multi-échelles et la fusion d'images multi-focus basée sur la détection de saillance (SDMF) (4).

5.3.1.2 Datasets

Pour vérifier les méthodes proposées, nous utilisons des bases de données bicolores ; un ensemble de données de fusion d'images multi-focus (X.Zuzhang, 2019), où l'image de référence est disponible et Lytro Multi-focus Dataset (30) où l'image de référence n'est pas disponible. Les figures 5.2 et 5.3 présentent quelques images de test.

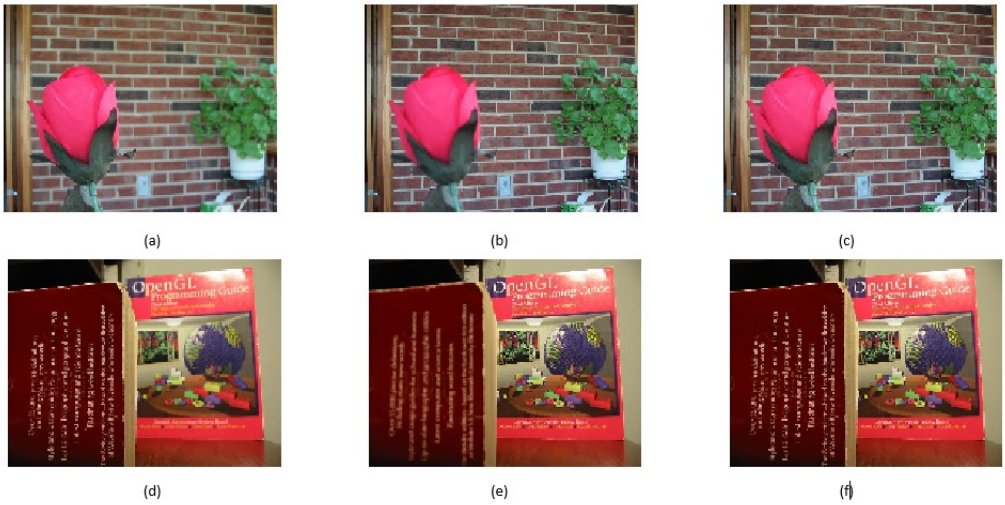


Figure 5.2: Images de référence et Images sources de la fleur et du livre (fusion d'images multi-focales) : (a ,d) image source 1, (b,e) image source 2 et (c,f) image de référence.

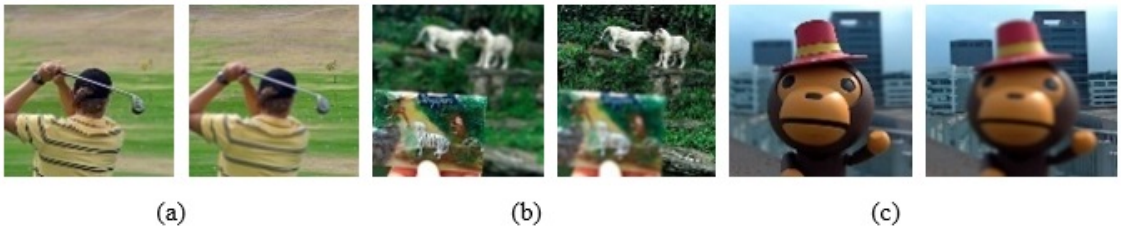


Figure 5.3: Images sources (a) Golf , (b) Zoo et (c) Toy (Lytro dataset) .

Les résultats expérimentaux sont présentés dans cette section. Nous décrivons les résultats et l'analyse des deux ensembles de données. Les performances de notre méthode peuvent être vérifiées qualitativement par inspection visuelle et quantitativement à l'aide de métriques de fusion.

5.3.2 Expérimentations

Les expériences sont réalisées dans l'environnement MATLAB 2018b à l'aide d'un processeur Intel Core i7-6700HQ avec une vitesse d'horloge de 2,60 GHz.
Une analyse qualitative des trois ensembles de données d'images multi-focus, y compris le golf, le zoo et les jouets, a été présentée à la figure 8. Les figures 5.4 à 5.6 représentent la qualité visuelle des ensembles de données. Une analyse qualitative approfondie du jouet, du zoo et du golf a été obtenue en zoomant sur des parties de régions particulières des images fusionnées.
Un affichage visuel du modèle proposé et de trois cartes de saillance (MGF et SDMF) a été illustré sur les figures 5.4 à 5.6 et les sous-figures (a) à (e). De plus, les parties agrandies des sous-figures (a) à (e) ont été illustrées dans les sous-figures (f) à (j). La figure 4 affiche avec force les images fusionnées de l'ensemble de données de jouets dans des rectangles rouges et verts. Cependant, une combinaison du modèle proposé et du modèle d'Abouelaziz peut créer des régions plus ciblées dans les images. Même si le reste des images était visuellement bon, ils n'ont pas atteint la qualité souhaitée pour l'inclusion dans les résultats.
Notamment, la partie agrandie du modèle de Howen et Judd illustrée sur les figures 5.4 (b) et (c) contient certaines distorsions visuelles, bien que des améliorations distinctes avec le modèle proposé puissent être observées sur la figure 5.4 (a). Les distorsions d'image du modèle de cabine de Howen et Judd sont également observées sur la figure 5.5 des sous-figures (f) à (j), en particulier lorsqu'elles sont comparées au modèle proposé.

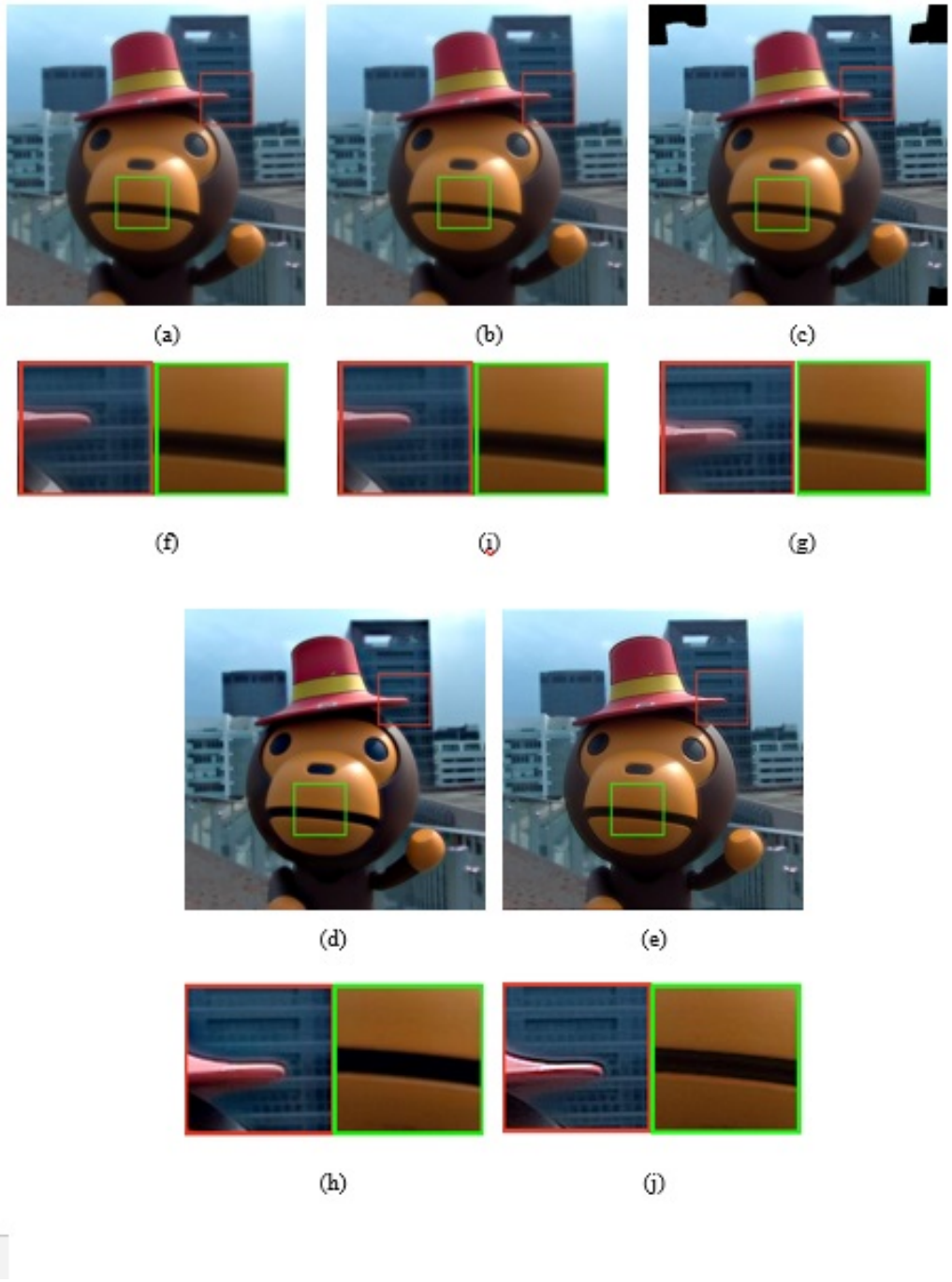


Figure 5.4: Comparison de la qualité visuelle pour toy dataset (a) méthode proposée + modèle de Abouelaziz, (b) méthode proposée + modèle de Judd, (c) méthode proposée+ modèle de Peng,(d) modèle MGF,(e) modèle SDMF. sous figures(f)-(j) montre la version zoomée des zones focalisées de (a)-(e) respectivement.

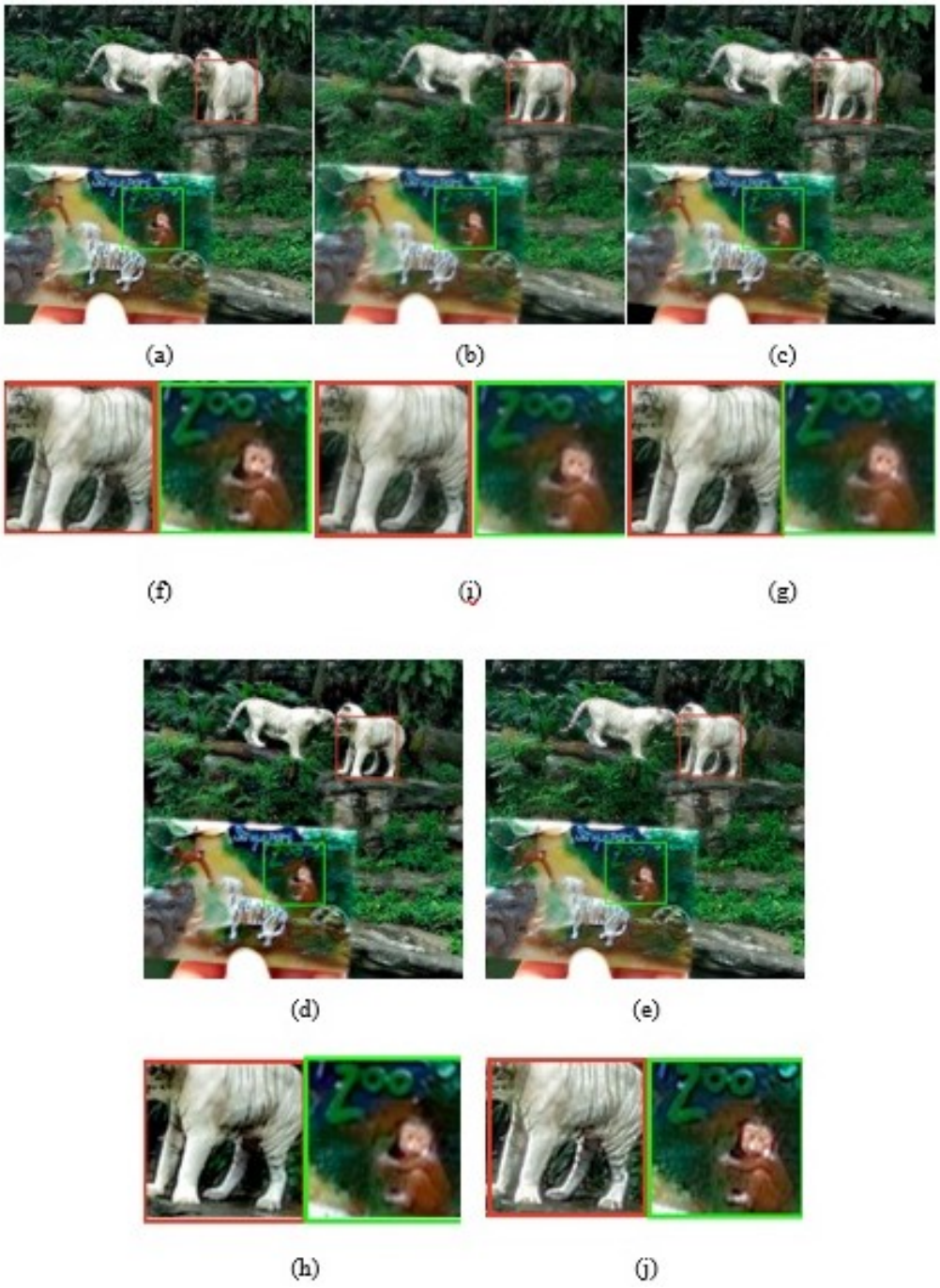


Figure 5.5: Comparison de la qualité visuelle pour zoo dataset (a) méthode proposée + modèle de Abouelaziz, (b) méthode proposée + modèle de Judd, (c) méthode proposée+ modèle de Peng,(d) modèle MGF,(e) modèle SDMF. sous figures(f)-(j) montre la version zoomée des zones focalisées de (a)-(e) respectivement.

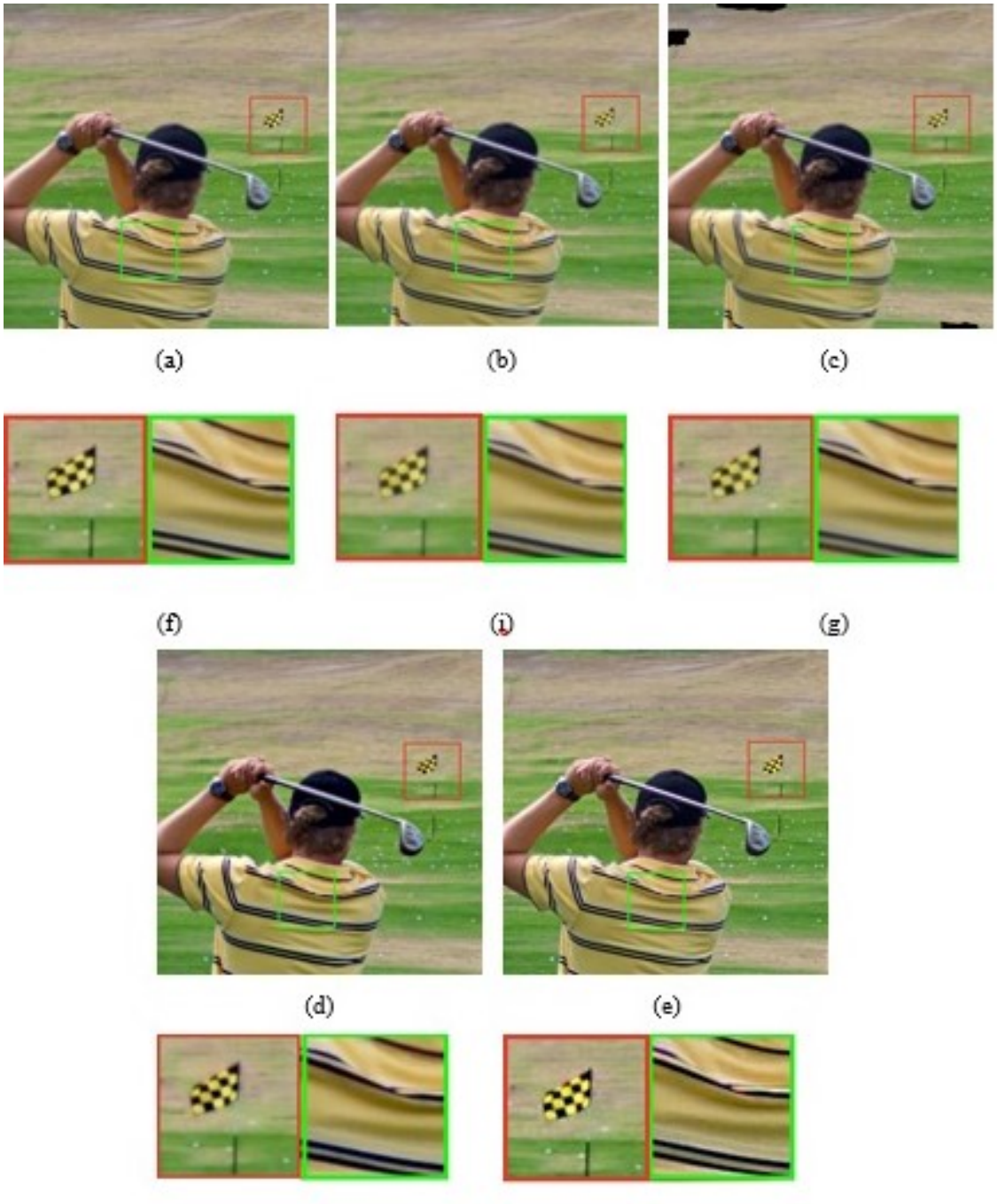


Figure 5.6: Comparison de la qualité visuelle pour golf dataset (a) méthode proposée + modèle de Abouelaziz, (b) méthode proposée + modèle de Judd, (c) méthode proposée+ modèle de Peng,(d) modèle MGF,(e) modèle SDMF. sous figures(f)-(j) montre la version zoomée des zones focalisées de (a)-(e) respectivement.

Des résultats similaires ont été observés avec l'ensemble de données de golf, selon lequel la méthode proposée a entraîné une combinaison de régions plus ciblées des images sources, qui présentaient peu d'artefacts par rapport aux méthodes existantes, comme illustré à la figure 5.6 . De plus, il ressort clairement de la figure 5.6 (sous-figures g et f) que le zoom sur les images produites par la méthode proposée a produit plus de détails sur l'arrière-plan ainsi que le dossier dans l'image fusionnée par rapport à d'autres méthodes.
Nous avons également comparé les résultats de la méthode proposée aux techniques de fusion multi-échelles récentes. Les résultats ont révélé que les performances de la méthode proposée liée au modèle d'Abouelaziz étaient soit comparables soit supérieures aux méthodes existantes. Étant donné que la métrique $Q^{AB/F}$ a produit les valeurs les plus élevées ; on peut conclure que l'utilisation du modèle d'Abouelaziz conduit à la production d'images de meilleure qualité, qui sont considérées comme les plus importantes dans le processus de fusion. De plus, il est évident que le modèle proposé produit la valeur la plus élevée d'informations mutuelles de caractéristiques (FMI), ce qui implique que le transfert d'informations des images sources vers les images fusionnées est maximal lorsque la méthode proposée est appliquée. La méthode proposée était également associée aux meilleures performances de SSIM, indiquant une réduction du niveau de distorsion structurelle entre les images sources et les images fusionnées par rapport aux méthodes existantes. En comparant le $L^{AB/F}$ et le $N^{AB/F}$, il a été constaté que les valeurs des méthodes proposées étaient inférieures à celles des méthodes existantes. En fin de compte, les tableaux 5.1 et 5.2 indiquent clairement qu'une combinaison de la méthode proposée et du modèle d'Abouelaziz produit les résultats les plus satisfaisants concernant la qualité visuelle et l'évaluation objective par rapport à d'autres méthodes de fusion.

Images	Methods	$Q^{AB/F}$	$L^{AB/F}$	$N^{AB/F}$	FMI	SSIM	Rank
flower	OMAM	0.9003	0.0997	0	0.8726	0.9981	**9**
	OMJM	0.9014	0.0986	0	0.8712	0.9967	9
	OMPM	0.8444	0.1463	0.0093	0.8614	0.9643	20
	MGF	0.8817	0.0926	0.0257	0.8552	0.9902	17
	SDMF	0.8677	0.0640	0.0683	0.8598	0.9959	16
book	OMAM	0.8885	0.1115	0	0.8863	0.9874	**6**
	OMJM	0.8129	0.1871	0	0.8802	0.9832	14
	OMPM	0.7340	0.2646	0.0014	0.8759	0.9528	22
	MGF	0.8740	0.1173	0.0087	0.8773	0.9469	16
	SDMF	0.8678	0.0588	0.0735	0.8760	0.9843	13

Table 5.1: Comparaison des mesures de performance des méthodes proposées avec les différentes méthodes de fusion d'images multi-focales, en utilisant les datasets

Images	Methods	$Q^{AB/F}$	$L^{AB/F}$	$N^{AB/F}$	FMI	Rank
Toy	OMAM	0.9224	0.0775	0	0.9312	7
	OMJM	0.7683	0.2280	*0.0037*	0.9359	15
	OMPM	0.7832	0.2148	0.0020	0.9189	15
	MGF	0.9062	0.0754	0.0185	0.9221	9
	SDMF	*0.8880*	0.0577	0.0543	0.9166	14
Zoo	OMAM	0.9296	0.0704	0	0.8585	**4**
	OMJM	0.7090	0.2890	0.0019	*0.8577*	13
	OMPM	0.8126	0.1842	0.0032	0.8292	14
	MGF	0.9165	0.0704	0.0132	0.8454	10
	SDMF	0.8684	0.0904	0.0413	0.8087	15
Golf	OMAM	0.8971	0.1029	0	0.9412	7
	OMJM	0.8310	0.1665	*0.0025*	0.9442	13
	OMPM	*0.8716*	0.1251	0.0033	0.9323	14
	MGF	0.8713	0.0638	0.0649	0.9283	15
	SDMF	0.8895	*0.1015*	0.0090	0.9349	11

Table 5.2: Comparaison des mesures de performance des méthodes proposées avec les différentes méthodes de fusion d'images multi-focales, en utilisant Lytro datasets

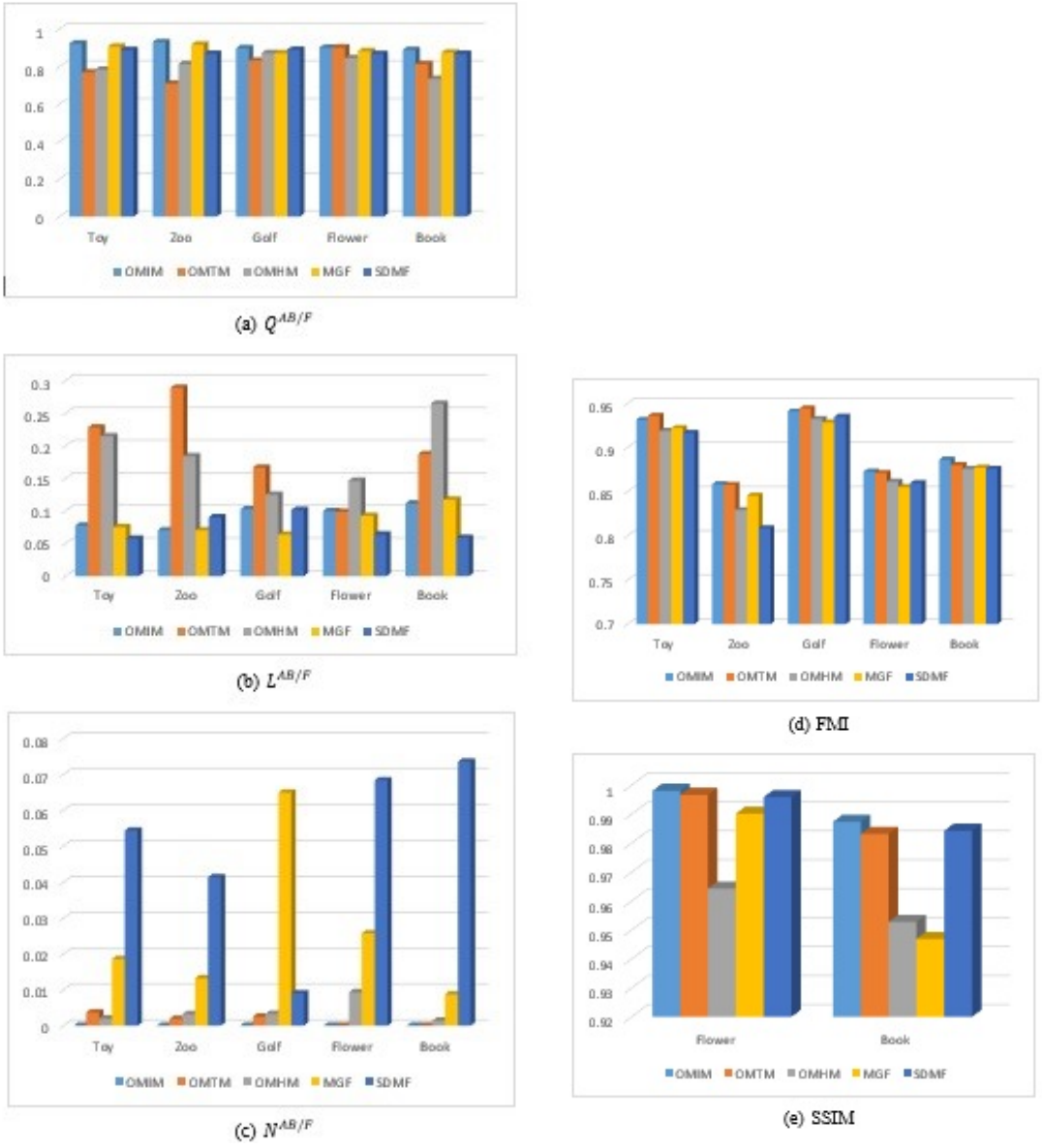


Figure 5.7: Analyse quantitative des méthodes proposées avec les différentes méthodes de fusion

5.4 Conclusion

Nous nous sommes appuyés sur des protocoles de détection de saillance visuelle pour identifier une nouvelle méthode applicable à la fusion d'images multi-focus en couleur. Dans l'étude, l'extraction des caractéristiques saillantes a été réalisée à l'aide de trois méthodes récemment proposées pour le calcul de la carte de saillance utilisée dans le processus de

fusion. La nouvelle méthode proposée a été testée sur différentes paires d'images. Alors qu'une évaluation qualitative a été menée pour évaluer les améliorations par inspection visuelle, des mesures de fusion objectives ont été utilisées pour évaluer les améliorations quantitatives. Les résultats de ces évaluations ont été comparés avec SDMF et MGF, qui forment une combinaison de techniques de fusion multi-échelle de pointe. Les résultats de l'évaluation quantitative ont indiqué que la méthode proposée avait une prévalence plus élevée par rapport à différentes stratégies et à d'autres résultats d'analyse visuelle avec des niveaux de fiabilité élevés. En tant que telle, la méthode proposée peut obtenir des améliorations de bord supérieures pour les images par rapport à d'autres techniques. Il existe des preuves cohérentes montrant que la méthode proposée surpasse les méthodes existantes, y compris MGF et SDMF. Une surperformance similaire a été observée lorsque la capacité de la méthodologie proposée à fournir des informations à partir de la source a été comparée à SSIM, $Q^{AB/F}$, $L^{AB/F}$ $N^{AB/F}$, ainsi comme FMI. Cependant, l'algorithme proposé présente une limitation majeure dans sa capacité à être appliqué en temps réel. En conséquence, les performances de la nouvelle méthode proposée peuvent être considérablement améliorées grâce à l'intégration d'images multi-saillantes en tant qu'entrées. Les recherches futures devraient donc se concentrer sur la parallélisation de la méthode proposée en plus d'intégrer l'extension multi-échelle via des algorithmes pull-push (52). Nous prévoyons également d'intégrer la détection de saillance basée sur des méthodes d'apprentissage profond dans notre étude, afin de voir leur influence dans la fusion d'images multi-focus.

Chapitre 6

Algorithme de fusion d'images multi-focales basé sur un filtre d'image guidé rapide

6.1 Introduction

Une nouvelle méthode de fusion d'images par filtrage guidé rapide est proposée dans ce chapitre, ce qui constitue notre troisième contribution dans cette thèse. Les résultats expérimentaux montrent que la méthode proposée donne une performance comparable aux approches de fusion de pointe. Plusieurs avantages de l'approche de fusion d'images proposée sont mis en évidence dans ce qui suit.

- Un algorithme de fusion d'images à usage général basé sur FGF est développé pour répondre à diverses applications de fusion d'images.
- La contribution clé consiste à présenter une méthode de fusion de filtre guidée rapide pour réduire le temps.

6.2 Description de la technique proposée

6.2.1 Définition du filtre guidée et des travaux en relation

Le filtre guidé (23) est une technique dérivée d'un modèle linéaire local, il calcule la sortie de filtrage en considérant le contenu d'une image de guidage, qui peut être l'image d'entrée elle-même ou une autre image différente.
Le filtre guidé peut être utilisé comme un opérateur de lissage préservant les bords comme le

filtre bilatéral, mais possède de meilleurs comportements près des bords: il peut transférer les structures de l'image de guidage vers la sortie de filtrage.
Il est basé sur un modèle linéaire local, ce qui le rend utilisable pour d'autres applications telles que le matage d'images, le sur-échantillonnage et la colorisation. En théorie, le filtre guidé suppose que la sortie de filtrage O est un modèle linéaire dans un voisinage local centré au pixel k.

$$O_i = a_k.I_i + b_k, \forall i \in w_k \tag{6.1}$$

où w_k désigne une fenêtre de taille $(2.r + 1) \times (2.r + 1)$ et les coefficients linéaires a_k et b_k sont constants dans w_k et peuvent être calculés en minimisant la différence au carré entre l'image de sortie O et l'image d'entrée P.

$$\begin{aligned} a_k &= \frac{\frac{1}{|w|}\sum_{i\in\Theta_k} I_i p_i - \mu_k \bar{p_k}}{\sigma_k^2 + \varepsilon}, \\ b_k &= \bar{p_k} - a_k u_k \end{aligned} \tag{6.2}$$

où μ_k et σ_k sont la moyenne et la variance de I dans la fenêtre w_k , $|w|$ est le nombre de pixels dans w_k, ε est un paramètre de régularisation défini par l'utilisateur pour contrôler le degré de lissage. $\bar{p_k}$ est la moyenne de p dans w_k . La sortie de filtrage peut être estimée par:

$$q_i = \bar{a}_i.I_i + \bar{b}_i, \tag{6.3}$$

Initialement, Li et al. (23) ont proposé un algorithme de fusion basé sur la GF que nous appelons la fusion par filtre guidé (GFF). Dans la méthode GFF, les images sources sont décomposées en couches de base et de détail sur la base d'un filtre moyen.
Ensuite, les cartes de saillance et les cartes de poids initial correspondantes sont calculées à l'aide d'opérateurs laplacien et gaussien.
Dans l'étape suivante, les cartes de poids initiales sont affinées à l'aide de GF pour obtenir les cartes de poids finales correspondant à chaque couche de base et de détail.
Enfin, les couches de base et de détail sont combinées avec ces poids.
Cette approche a présenté des performances supérieures par rapport aux méthodes de fusion basées sur la décomposition multi-échelles existantes pour de nombreuses applications de fusion d'images. Cependant, il a quelques limitations qui sont abordées par divers nouveaux algorithmes de fusion.
Bavirisetti et al. (6) ont proposé une méthode de fusion (GFS) basée sur la GF et les statistiques d'images. Cette méthode est conçue pour s'appliquer à la fois aux images mono et multi-capteurs. Même si les résultats de GFS sont raisonnables, le temps d'exécution est

très élevé puisque la règle de fusion employée dans cet algorithme dépend des propriétés statistiques du voisinage. À l'exception de GFF et GFS, les autres méthodes dépendent de l'application. Toutes les méthodes de fusion basées sur GF dépendent d'autres outils et techniques d'extraction de carte de saillance ou de processus de construction de carte de poids. La majorité d'entre eux sont coûteux en temps de calcul et la comparaison d'exécution n'est pas prioritaire. De plus, aucun d'entre eux n'est testé sur des jeux de données vidéo.
Le MGF, peut très bien combiner des informations utiles sur l'image source dans l'image fusionnée, en exploitant les avantages de la décomposition d'image à plusieurs échelles et de la propriété de transfert de structure du GF, développant ainsi un nouveau processus d'extraction de saillance visuelle et de construction de carte de poids. La décomposition d'images à plusieurs échelles est appropriée pour représenter et manipuler des caractéristiques d'images à différentes échelles. La propriété de transfert de structure du GF activé par notre algorithme peut transférer les structures d'une image source dans l'autre. Le processus de détection de saillance visuelle basé sur des informations de couche de détail avec un GF multi-échelle développé dans notre algorithme peut identifier des informations d'image source importantes. Le processus de construction de la carte de poids basé sur la saillance visuelle peut intégrer des informations complémentaires pixel par pixel. Par conséquent, la méthode proposée est capable de transférer des informations d'image source.

6.2.2 Description de la méthode proposée

6.2.2.1 Définition du l'algorithme : Fast Guided Filter

Pour accélérer le temps de GF Kaiming He (12) propose un nouvel algorithme FGF: Fast Guided Filter, où on remarque presque aucune dégradation visible dans l'image, le principe de l'algorithme consiste à sous-échantillonner l'image d'entrée et l'image de guidage en tant que rapport de sous-échantillonnage s, FGF peut diminuer la complexité temporelle de la forme $O(N)$ à $O(N/s^2)$.
Le filtre proposé peut être résumé par l'algorithme ci dessous, où nous remarquons que consiste à échantillonner l'image avec un rapport s, , que les étapes de 2 à 5 représentent le fonctionnement du filtre et que l'étape 6 concerne la reconstruction de l'image (sur-échantillonnage):

Algorithm 2: Algorithme FGF

$1 : I' = f_{subsample}(I, s)$
$\quad p' = f_{subsample}(p, s)$
$\quad r' = r/s$
$2 : mean_I = f_{mean}(I', r')$
$\quad mean_p = f_{mean}(p', r')$
$\quad corr_I = f_{mean}(I'_. * I', r')$
$\quad corr_{I_p} = f_{mean}(I'_. * p', r')$
$3 : var_I = corr_I - mean_{I_.} * mean_I$
$\quad cov_{I_p} = corr_{I_p} - mean_{I_.} * mean_p$
$4 : a = cov_{I_{p.}}/(var_t + \varepsilon)$
$\quad b = mean_p - a_. * mean_f$
$5 : mean_a = f_{mean}(a, r')$
$\quad mean_b = f_{mean}(b, r')$
$6 : mean_a = f_{upsample}(mean_a, s)$
$\quad mean_b = f_{upsample}(mean_b, s)$
$7 : q = mean_{a_.} * I + mean_b$

6.2.2.2 Algorithme de fusion d'images multi-focales basé sur un filtre d'image guidé rapide

L'algorithme de fusion d'images multi-focales basé sur un filtre d'image guidé rapide peut être décrit par les étapes suivantes, où nous avons gardé globalement l'architecture de l'algorithme de fusion en utilisant un filtre guidé et intégrer le filtre guidé rapide où l'image est sous-échantillonnée en entrée, traitée et enfin reconstruite en sortie.

1. Considérons les images d'entrée $I_1(x, y)$ et $I_2(x, y)$ de la même taille, nous avons effectué une décomposition multi-échelle de I_1 et I_2 en utilisant FGF pour obtenir les couches de base B_{11},B_{12} :

   $$B_{11} = FGF(I_1, I_2, r_1, 31)$$
   $$B_{12} = FGF(I_2, I_1, r_1, 31)$$

   I_1, I_2 est utilisé comme image de guidage pour le filtrage. Par conséquent, les informations structurelles de I_2 sont utilisées pour lisser I_1.

   Un fonctionnement similaire peut être observé sur I_2 en prenant I1 comme image de guidage. Les couches de base consécutives sont générées comme :

   $$B_{k1} = FGF(B_{k-11}, B_{k-12}, r_k, 3k)$$
   $$B_{k2} = FGF(B_{k-12}, B_{k-11}, r_k, 3k)$$
   $$\forall k = 1, ..., n$$

2. Les couches de détail D_{1k}, D_{2k} sont obtenues en trouvant la différence entre les images sources et les couches de base :

$$D_{1k} = B_{k-11} - B_{k1}$$
$$D_{2k} = B_{k-12} - B_{k2}$$

3. Calculer les saillances visuelles des images d'entrée multi-focales, en mesurant les informations saillantes S_1, S_2:

$$S_1 = |D_1|$$
$$S_2 = |D_2|$$

4. Les cartes de poids w_i sont déterminées à partir des cartes de saillance extraites en les normalisant comme suit:

$$\begin{aligned} w_1^k &= \frac{S_1^k}{\sum_{i=1}^{2} S_i^k}, \forall k = 1, 2, ..., n \\ w_2^k &= \frac{S_2^k}{\sum_{i=1}^{2} S_i^k}, \forall k = 1, 2, ..., n \end{aligned} \tag{6.4}$$

où k désigne l'échelle. On peut noter que nous calculons des poids pour chaque échelle.

5. Pour les informations de la couche de détail, nous devons intégrer à chaque échelle K à l'aide des cartes de poids w_{k1} et w_k2 en utilisant une combinaison linéaire:

$$D_{kF} = w_1^k \times D_{1k} + w_2^k \times D_{2k} \tag{6.5}$$

Enfin, nous devons combiner les couches de détails fusionnées obtenues à chaque échelle :

6. La couche de base fusionnée B est générée en faisant converger les couches de base à l'échelle finale n comme suit :

$$B_F = \frac{1}{2}(B_1^n + B_2^n) \tag{6.6}$$

7. L'image fusionnée est générée, en combinant les couches de base et de détail finales comme suit :

$$F = B + D \tag{6.7}$$

6.3 Résultats et bilan

6.3.1 Métriques utilisées dans l'évaluation

Afin d'évaluer les performances des méthodes de fusion proposées, nous utilisons les métriques suivantes: l'information mutuelle (MI) qui mesure la quantité d'informations transférées à l'image fusionnée à partir des images sources, l'écart type (SD) , la fréquence spatiale (SF), $Q^{AB/F}$: les informations totales transférées des images sources vers les images fusionnées, $L^{AB/F}$: la perte totale d'information , et $N^{AB/F}$: dénote le bruit ou les artefacts ajoutés dans l'image fusionnée en raison du processus de fusion.
En général, les valeurs MI et $Q^{AB/F}$ plus élevées indiquent le meilleur résultat fusionné. $L^{AB/F}$ est introduit pour évaluer les informations perdues au cours du processus de fusion. Les informations perdues sont disponibles dans les images sources mais pas dans l'image fusionnée. $N^{AB/F}$ représente les artefacts de fusion qui ont été introduits dans l'image fusionnée. Il est clair que plus $L^{AB/F}$ et $N^{AB/F}$ sont petits, meilleure est l'image fusionnée. Il convient de noter que les compléments $Q^{AB/F}$, $L^{AB/F}$ et $N^{AB/F}$ indiquent que la somme de tous ces éléments devrait donner l'unité.
Des valeurs plus élevées de $Q^{AB/F}$, et SD indiquent une meilleure qualité de fusion de l'image fusionnée. En revanche, lorsque la valeur de $N^{AB/F}$ est faible, les performances de fusion sont meilleures.
Les performances de fusion de l'image fusionnée sont meilleures avec l'augmentation de l'indice numérique de $Q^{AB/F}$, SF et SSIM. Au contraire, les performances de fusion sont meilleures lorsque la valeur de $N^{AB/F}$ est faible.

6.3.2 Résultats et discussion

Le Whole Brain Atlas est une base de données de référence pour évaluer les performances des méthodes de fusion d'images médicales multimodales établies par Keith A. Johnson et J. Alex Becker à la Harvard Medical School. La base de données Whole Brain Atlas se compose de quatre types d'imagerie : CT, IRM, PET et SPECT avec la description de la structure cérébrale normale et anormale, toutes les images de la base de données sont co-alignées. Pour valider nos propositions nous avons donc utilisés des images de cette base, et en particulier des images de niveaux de gris du dataset CT et IRM.

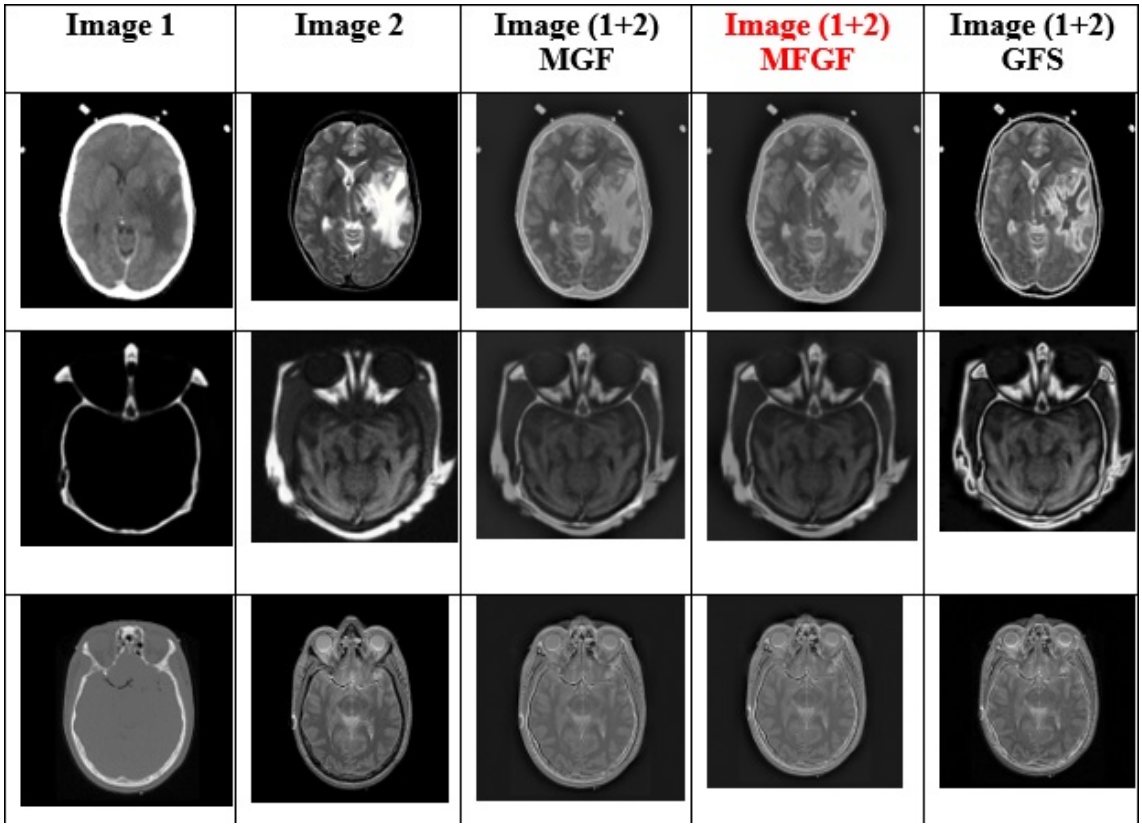


Figure 6.1: Utilisation d'un dataset d'images médical : medical brain dataset (CT et MRI) pour la validation de la technique MFGF avec r=8 et s=4

Des expériences ont été menées sur des ensembles de données et les résultats sont analysés en termes de qualité visuelle, de métriques de fusion et de temps d'exécution.
La figure 6.1, montre les résultats de la fusion d'images médical, c'est un exemple que nous avons pris avec r=8 et s=4. Les résultats visuels semblent très encourageants d'un point de vue quantitatif (temps de calcul) que du point de vue qualitatif (fusion visuellement correct comparée aux méthodes MGF et GFS).

	MGF	MFGF	GFS		MGF	MFGF	GFS
QAB/F	0.8163	0.8159	0.782	QAB/F	0.8163	0.8134	0.794
LAB/F	0.1467	0.1476	0.1734	LAB/F	0.1467	0.1498	0.1789
NAB/F	0.0371	0.0365	0.0446	NAB/F	0.0371	0.0368	0.0271
nfmi	0.8851	0.8851	0.8754	nfmi	0.8851	0.8851	0.8843
t_calcul	1.7122	0.4511	1.2536	t_calcul	1.7132	0.3369	1.2507
	r=8, s=2	r=8, s=2	r=8, s=2		r=8, s=4	r=8, s=4	r=8, s=4

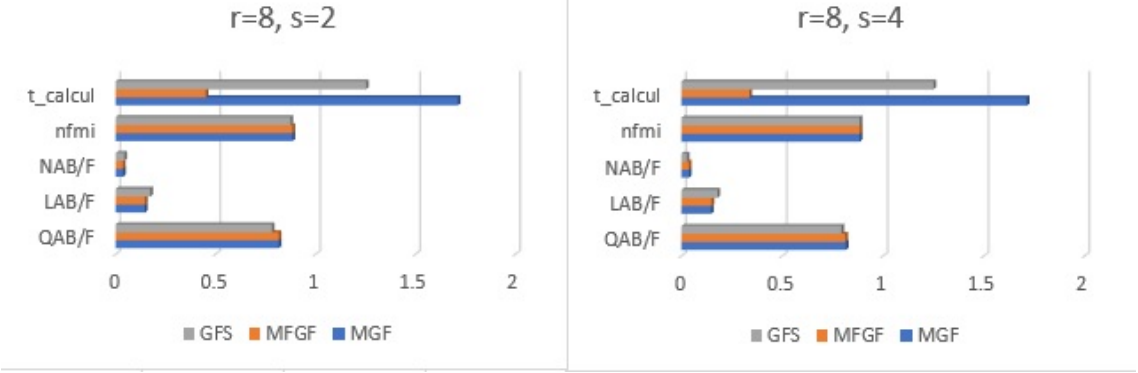
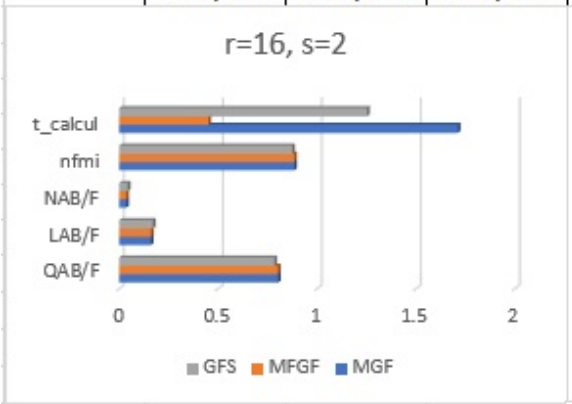


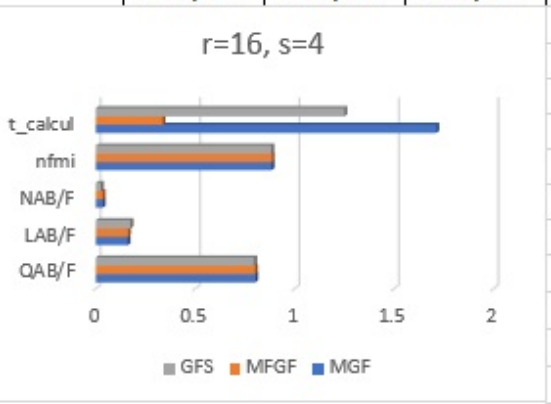
Figure 6.2: Utilisation d'un dataset d'images médical : medical brain dataset pour la validation de la technique MFGF avec r=8 et s=4

Comme le montre la figures 6.1, la qualité visuelle de la MFGF est similaire à celle du MGF (23) pour l'ensemble de données médicales en combinant les informations IRM et CT des images sources. Les deux méthodes sont gérées pour obtenir des images focalisées tout-en-un avec moins de perte d'informations visuelles et d'artefacts. La méthode proposée MFGF est donc capable d'obtenir l'image fusionnée avec une bonne qualité visuelle.

	MGF	MFGF	GFS
QAB/F	0.8035	0.803	0.7848
LAB/F	0.161	0.1617	0.1714
NAB/F	0.0355	0.0353	0.0438
nfmi	0.8851	0.8851	0.8755
t_calcul	1.7122	0.4511	1.2536
	r=16, s=2	r=16, s=2	r=16, s=2



	MGF	MFGF	GFS
QAB/F	0.8035	0.8014	0.7977
LAB/F	0.161	0.1632	0.1758
NAB/F	0.0355	0.0354	0.0265
nfmi	0.8851	0.8854	0.8838
t_calcul	1.7132	0.3369	1.2507
	r=16, s=4	r=16, s=4	r=16, s=4



	MGF	MFGF	GFS
QAB/F	0.8035	0.7971	0.7834
LAB/F	0.161	0.1685	0.2067
NAB/F	0.0355	0.0344	0.0099
nfmi	0.8851	0.8858	0.8881
t_calcul	1.7228	0.3001	1.2399
	r=16, s=8	r=16, s=8	r=16, s=8

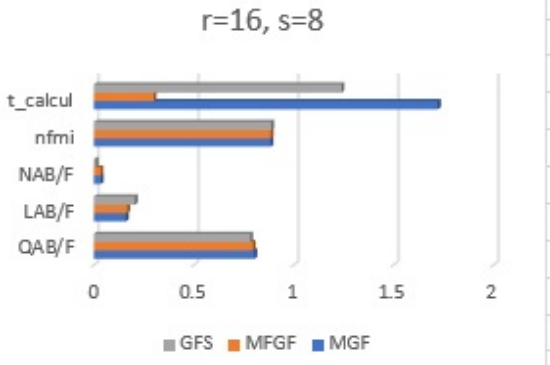
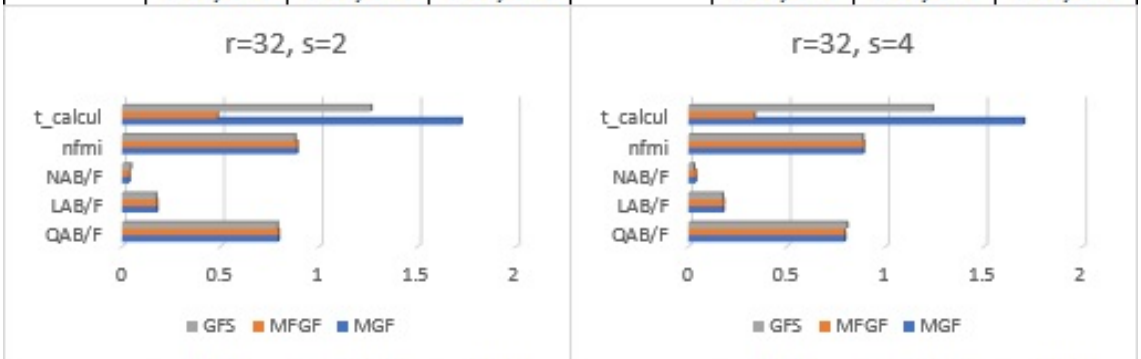


Figure 6.3: Utilisation d'un dataset d'images médical : medical brain dataset pour la validation de la technique MFGF avec r=8 et s=4

L'analyse quantitative du MFGF est proposé pour les ensembles de données médicales. Les performances de fusion de la méthode MFGF proposée sont comparées à la méthode GFS (5)et MFG (23) en utilisant les métriques sus citées. Comme on peut le voir dans les tableaux des figures, la méthode proposée montre des performances quantitatives presque identiques pour tous les ensembles de données d'images.

À partir des tableaux des figures 6.2, 6.3,6.4 nous pouvons également remarquer que la méthode proposée est capable d'exécuter des résultats en un temps d'exécution beaucoup plus réduit. Cela est dû à la règle adopter par le filtre guidé rapide sur l'image de départ, ainsi le fait d'échantillonner pour reconstruire à la fin prend moins de temps d'exécution.

	MGF	MFGF	GFS		MGF	MFGF	GFS
QAB/F	0.7915	0.7915	0.787	QAB/F	0.7915	0.7912	0.8032
LAB/F	0.1749	0.1751	0.1703	LAB/F	0.1749	0.1754	0.1713
NAB/F	0.0336	0.0335	0.0427	NAB/F	0.0336	0.0335	0.0256
nfmi	0.8865	0.8866	0.8756	nfmi	0.8865	0.8867	0.8835
t_calcul	1.7196	0.4825	1.2617	t_calcul	1.7005	0.3332	1.2373
	r=32, s=2	r=32, s=2	r=32, s=2		r=32, s=4	r=32, s=4	r=32, s=4



	MGF	MFGF	GFS		MGF	MFGF	GFS
QAB/F	0.7915	0.7882	0.7904	QAB/F	0.7915	0.7859	0.7364
LAB/F	0.1749	0.1787	0.1998	LAB/F	0.1749	0.1834	0.2626
NAB/F	0.0336	0.0331	0.0098	NAB/F	0.0336	0.0307	0.001
nfmi	0.8865	0.887	0.889	nfmi	0.8865	0.8872	0.8892
t_calcul	1.6889	0.2903	1.2249	t_calcul	1.7115	0.2886	1.217
	r=32, s=8	r=32, s=8	r=32, s=8		r=32, s=16	r=32, s=16	r=32, s=16

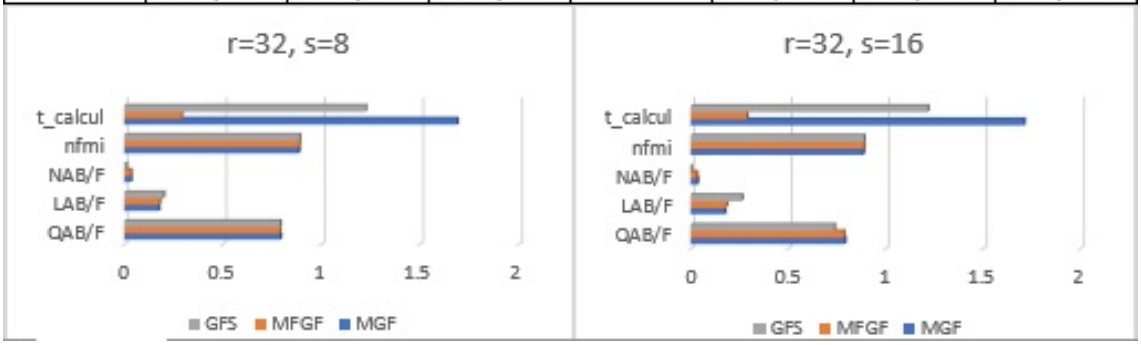


Figure 6.4: Utilisation d'un dataset d'images médical : medical brain dataset pour la validation de la technique MFGF avec r=8 et s=4

6.4 Conclusion

Le but de notre proposition est d'améliorer le temps de calcul de la méthode Multi-scale Guided Image en utilisant l'algorithme Fast Guided Filter (2015) donc d'après les résultats il y a une grande amélioration en terme de temps de calcul en garde la qualité des images "sans toucher à leur qualité".

Pour cette évaluation, nous avons utilisé les images médicales :

- Niveau de gris ou couleur
- Comparée notre méthode avec des méthodes similaires récentes MGF 2019 'Multi Guided filter', fusion method (GFS) based on GF and image statistics (GFS 2017), guided filter fusion GFF 2013

Nous pouvons constater que le temps d'exécution du GFS est énorme par rapport MFGF, GFF et MGF. Cela est dû au fait que la règle de fusion utilisée dans GFS est basée sur les propriétés statistiques du voisinage. Cependant, on peut observer que l'algorithme MFGF proposé est plus rapide que le GFF et MGF. De même, pour les autres classes, l'algorithme MFGF donne rapidement des résultats par rapport à MGF, GFF et GFS.

Chapitre 7

Conclusion générale

La principale préoccupation de la fusion d'images est de combiner les informations pertinentes de plusieurs images d'une scène en une seule image plus informative. Les méthodes de fusion d'images ont fait de grands progrès ces dernières années. L'objectif de cette thèse est d'étudier une approche de fusion d'images multi-focales basée sur l'intégration des techniques de perception visuelle via les cartes de saillance.

Nous avons donc élaboré lors de notre réalisation de cette thèse trois principales contributions (2), (3), la première concerne une étude comparative de la robustesse au bruit dans les modèles d'attention visuelle, l'objectif étant d'injecter un bruit blanc aux images initiales et d'étudier ensuite son influence sur la perception visuelle via les cartes de saillance.

Dans la deuxième contribution, nous avons développé une méthode de fusion d'images qui se base sur trois méthodes générales de de calcul de la saillance visuelle basées respectivement sur la détection de la saillance à l'aide de la technique de la carte de saillance humaine, la détection de la saillance à l'aide de la technique de transformation de contour (CT) et la détection de la saillance à l'aide de la technique de décomposition matricielle de faible rang et structurée (LSMD).

Enfin, la troisième contribution, qui est en cours traite du problème du temps de calcul et propose un algorithme de fusion d'images multi-focales, en utilisant un filtre d'image guidé rapide, ce qui a permis d'améliorer considérablement le temps et de proposer une technique de fusion en temps réel, qui pourrait répondre aux besoin de plusieurs applications notamment dans le domaine médical.

La méthodes que nous avons proposé dans le cadre de cette thèse, ont été testée sur différentes paires d'images. Alors qu'une évaluation qualitative a été menée pour évaluer les améliorations par inspection visuelle, des mesures de fusion objectives ont été utilisées pour évaluer les améliorations quantitatives. Les résultats de ces évaluations ont été comparés

avec SDMF et MGF, qui forment une combinaison de techniques de fusion multi-échelle de pointe. Les résultats de l'évaluation quantitative ont indiqué que la méthode proposée avait une prévalence plus élevée par rapport à différentes stratégies et à d'autres résultats d'analyse visuelle avec des niveaux de fiabilité élevés.

Les recherches futures devraient donc se concentrer sur la parallélisation de la méthode proposée en plus d'intégrer l'extension multi-échelle via des algorithmes pull-push (52). Nous prévoyons également d'intégrer la détection de saillance basée sur des méthodes d'apprentissage profond dans notre étude, afin de voir leur influence dans la fusion d'images multi-focales.

Pour le cadre général de la fusion d'images, plusieurs pistes d'améliorations des techniques proposées peuvent être envisagées :

Premièrement, des stratégies plus élaborées pour la fusion des régions frontières peuvent être étudiées plus. C'est un problème commun à toutes les catégories de méthodes. Les méthodes qui combinent domaine de transformation et domaine spatial ont fait quelques tentatives préliminaires sur ce sujet, mais les stratégies adoptées sont relativement simples. Le développement d'approches de détection de régions focales plus précises et de schémas de fusion plus efficaces pour les régions limites sera toujours une direction importante dans la fusion d'images multi-focales.

Deuxièmement, étude du problème de mauvais repérage causé par les objets en mouvement ou le bougé de l'appareil photo deviendra une direction active. Les méthodes conventionnelles de domaine de transformation et de domaine spatial ont rencontré des goulots d'étranglement techniques pour résoudre ce problème. Les méthodes d'apprentissage profond ont encore un potentiel suffisant pour résoudre ce problème en raison de leur forte capacité d'apprentissage.

Troisièmement, d'autres travaux se concentrant sur les applications spécifiques de la fusion d'images multi-focales sont attendus, ce qui permettra aux approches de fusion sur certaines applications spécifiques d'être étudiées de manière contrapuntique selon leurs propres caractéristiques.

Références Bibliographiques

[1] Ilyass Abouelaziz and Mohammed El Hassouni. New models of visual saliency: Contourlet transform based model and hybrid model. In *2015 Intelligent Systems and Computer Vision (ISCV)*, pages 1–5. IEEE, 2015. 55

[2] Sarra Babahenini, Foudil Cherif, and Fella Charif. Comparative study of noise robustness in visual attention models. In *International Conference on Electrical Engineering and Control Applications*, pages 1259–1269. Springer, 2019. 89

[3] Sarra Babahenini, Foudil Cherif, Fella Charif, Abdelmalik Taleb-Ahmed, and Yassine Ruichek. Using saliency detection to improve multi-focus image fusion. *International Journal of Signal and Imaging Systems Engineering*, in press, 2021. 89

[4] Durga Prasad Bavirisetti and Ravindra Dhuli. Multi-focus image fusion using multi-scale image decomposition and saliency detection. *Ain Shams Engineering Journal*, 9 (4):1103–1117, 2018. 68

[5] Durga Prasad Bavirisetti, Vijayakumar Kollu, Xiao Gang, and Ravindra Dhuli. Fusion of mri and ct images using guided image filter and image statistics. *International journal of Imaging systems and Technology*, 27(3):227–237, 2017. 86

[6] Durga Prasad Bavirisetti, Gang Xiao, Junhao Zhao, Ravindra Dhuli, and Gang Liu. Multi-scale guided image and video fusion: A fast and efficient approach. *Circuits, Systems, and Signal Processing*, 38(12):5576–5605, 2019. 45, 68, 79

[7] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2012. v, 9, 10, 15, 25, 26

[8] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational visual media*, 5(2):117–150, 2019. 57

[9] Dashan Gao and Nuno Vasconcelos. Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics. *Neural computation*, 21(1):239–271, 2009. 23

[10] Mohammad Bagher Akbari Haghighat, Ali Aghagolzadeh, and Hadi Seyedarabi. Real-time fusion of multi-focus images for visual sensor networks. In *2010 6th Iranian Conference on Machine Vision and Image Processing*, pages 1–6. IEEE, 2010. 49

[11] Mohammad Bagher Akbari Haghighat, Ali Aghagolzadeh, and Hadi Seyedarabi. A non-reference image fusion metric based on mutual information of image features. *Computers & Electrical Engineering*, 37(5):744–756, 2011. 67

[12] Kaiming He and Jian Sun. Fast guided filter. *arXiv preprint arXiv:1505.00996*, 2015. 80

[13] Sen He, Hamed R Tavakoli, Ali Borji, Yang Mi, and Nicolas Pugeault. Understanding and visualizing deep visual saliency models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10206–10215, 2019. 31

[14] Wei Huang and Zhongliang Jing. Evaluation of focus measures in multi-focus image fusion. *Pattern recognition letters*, 28(4):493–500, 2007. v, 40, 41

[15] Laurent Itti. *Models of bottom-up and top-down visual attention*. California Institute of Technology, 2000. 11, 21

[16] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, 2009. 22

[17] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12):1489–1506, 2000. 55

[18] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009. 30, 55

[19] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4:219–227, 1985. 15, 16

[20] Weiwei Kong, Yang Lei, and Minmin Ren. Fusion method for infrared and visible images based on improved quantum theory model. *Neurocomputing*, 212:12–21, 2016. 56

[21] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, 129:261–270, 2020. 31

[22] Shutao Li, James T Kwok, and Yaonan Wang. Combination of images with diverse focuses using the spatial frequency. *Information fusion*, 2(3):169–176, 2001. v, 39, 40

[23] Shutao Li, Xudong Kang, and Jianwen Hu. Image fusion with guided filtering. *IEEE Transactions on Image processing*, 22(7):2864–2875, 2013. 78, 79, 85, 86

[24] Yixiong Liang, Yuan Mao, Jiazhi Xia, Yao Xiang, and Jianfeng Liu. Scale-invariant structure saliency selection for fast image fusion. *Neurocomputing*, 356:119–130, 2019. 39

[25] Yu Liu, Shuping Liu, and Zengfu Wang. Multi-focus image fusion with dense sift. *Information Fusion*, 23:139–155, 2015. 39

[26] Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, 36:191–207, 2017. vi, 50

[27] Yu Liu, Lei Wang, Juan Cheng, Chang Li, and Xun Chen. Multi-focus image fusion: A survey of the state of the art. *Information Fusion*, 64:71–91, 2020. 34, 36, 40, 52

[28] Hermann J Müller and Joseph Krummenacher. Visual search and selective attention. *Visual Cognition*, 14(4-8):389–410, 2006. 12

[29] VPS Naidu and Jitendra R Raol. Pixel-level image fusion using wavelets and principal component analysis. *Defence Science Journal*, 58(3):338, 2008. 44

[30] Mansour Nejati, Shadrokh Samavi, and Shahram Shirani. Multi-focus image fusion using dictionary-based sparse representation. *Information Fusion*, 25:72–84, 2015. 65, 68

[31] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3): 145–175, 2001. 55

[32] Aude Oliva, Antonio Torralba, Monica S Castelhano, and John M Henderson. Top-down control of visual attention in object detection. In *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, volume 1, pages I–253. IEEE, 2003. 14

[33] Gonzalo Pajares and Jesus Manuel De La Cruz. A wavelet-based image fusion tutorial. *Pattern recognition*, 37(9):1855–1872, 2004. 43

[34] Sujoy Paul, Ioana S Sevcenco, and Panajotis Agathoklis. Multi-exposure and multi-focus image fusion in gradient domain. *Journal of Circuits, Systems and Computers*, 25(10):1650123, 2016. v, 47, 48

[35] Houwen Peng, Bing Li, Rongrong Ji, Weiming Hu, Weihua Xiong, and Congyan Lang. Salient object detection via low-rank and structured sparse matrix decomposition. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013. 56

[36] Robert J Peters and Laurent Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. v, 29

[37] Vladimir Petrovic and Costas Xydeas. Objective image fusion performance characterisation. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1866–1871. IEEE, 2005. 65, 67

[38] Gemma Piella. Image fusion for enhanced visualization: A variational approach. *International journal of computer vision*, 83(1):1–11, 2009. 46

[39] Feng Qi, Debin Zhao, Shaohui Liu, and Xiaopeng Fan. 3d visual saliency detection model with generated disparity map. *Multimedia Tools and Applications*, 76(2):3087–3103, 2017. 19

[40] Karsten Rauss and Gilles Pourtois. What is bottom-up and what is top-down in predictive coding? *Frontiers in psychology*, 4:276, 2013. 11

[41] Ruth Rosenholtz. A simple saliency model predicts a number of motion popout phenomena. *Vision research*, 39(19):3157–3163, 1999. 55

[42] Ron Sun. Introduction to computational cognitive modeling. *Cambridge handbook of computational psychology*, pages 3–19, 2008. 10

[43] John K Tsotsos and Albert Rothenstein. Computational models of visual attention. *Scholarpedia*, 6(1):6201, 2011. v, 9, 10

[44] Junle Wang, Matthieu Perreira Da Silva, Patrick Le Callet, and Vincent Ricordel. Computational model of stereoscopic 3d visual saliency. *IEEE Transactions on Image Processing*, 22(6):2151–2165, 2013. 19

[45] Wencheng Wang and Faliang Chang. A multi-focus image fusion method based on laplacian pyramid. *J. Comput.*, 6(12):2559–2566, 2011. v, 43, 44

[46] Jinhua Xu. Bayesian modeling of visual attention. In *International Conference on Neural Information Processing*, pages 92–99. Springer, 2012. 22

[47] Xiang Yan, Syed Zulqarnain Gilani, Hanlin Qin, and Ajmal Mian. Structural similarity loss for learning to fuse multi-focus images. *Sensors*, 20(22):6647, 2020. vi, 51

[48] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013. 56

[49] Cui Yang, Jian-Qi Zhang, Xiao-Rui Wang, and Xin Liu. A novel similarity based quality metric for image fusion. *Information Fusion*, 9(2):156–160, 2008. 67

[50] Hongpeng Yin, Zhaodong Liu, Bin Fang, and Yanxia Li. A novel image fusion approach based on compressive sensing. *Optics Communications*, 354:299–313, 2015. 45

[51] Weiling Yin, Wenda Zhao, Di You, and Dong Wang. Local binary pattern metric-based multi-focus image fusion. *Optics & Laser Technology*, 110:62–68, 2019. 39

[52] Ming Zhang and Bahadir K Gunturk. Multiresolution bilateral filtering for image denoising. *IEEE Transactions on image processing*, 17(12):2324–2333, 2008. 77, 90

[53] Xingchen Zhang. Deep learning-based multi-focus image fusion: A survey and a comparative study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 49