REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITE BADJI MOKHTAR –ANNABA-

FACULTE DES SCIENCES DE L'INGENIEUR - ANNEE 2019/2020-

Département d'Informatique

THESE

Présentée en vue de l'obtention du diplôme de

Doctorat en Sciences

Option : Intelligence Artificielle

Présentée par

M^ME SAMIRA CHEBBOUT

PROPOSITION D'UN SYSTEME D'ANALYSE ET D'INTERPRETATION D'IMAGES

DEVANT LE JURY

Directrice de Thèse

M^ME Hayet Farida MEROUANI Professeur à l'Université Badji Mokhtar, Annaba.

Présidente	
M^ME Nassira GHOUALMI	Professeur à l'Université Badji Mokhtar, Annaba
Examinateurs	
M^R Smaine MAZOUZI	MCA à l' Université 20 Août 1955, Skikda.
M^R Brahim FAROU	MCA à l'Université 8 Mai 1945, Guelma.
M^ME Ahlem MELHOUAH	MCA à l'Université Badji Mokhtar, Annaba

أَعُوذُ بِاللَّهِ مِنَ ٱلشَّيْطَانِ ٱلرَّجِيمِ

رَبَّنَا ظَلَمْنَآ أَنفُسَنَا

وَإِن لَّمْ تَغْفِرْ لَنَا وَتَرْحَمْنَا

لَنَكُونَنَّ مِنَ ٱلْخَٰسِرِينَ

Dédicace

A tous ceux que j'aime.

Remerciements

J'exprime mes profonds remerciements, à ma directrice de thèse : Madame Hayet Farida Merouani, Professeur à l'université Badji Mokhtar -Annaba- pour son soutien infaillible, sa disponibilité et ses encouragements continus. Je la remercie pour la confiance qu'elle a su m'accorder jusqu'à la dernière minute.

Je tiens également à remercier les membres de jury qui m'ont fait l'honneur de bien vouloir évaluer mon travail, et plus précisément

— Madame Ghoualmi Nassira à l'Université Badji Mokhtar -Annaba- pour l'honneur qu'il m'a fait, en acceptant de présider ce jury.

— Monsieur Mazouzi Smaine à l'Université 20 Août 1955 -Skikda- pour avoir accepté d'évaluer et d'examiner ce manuscrit.

— Monsieur Farou Brahim à l'Université 8 Mai 1945 -Guelma- d'avoir bien voulu faire partie de ce jury.

— Madame Melouah Ahlem à l'Université Badji Mokhtar -Annaba- pour l'honneur qu'elle m'a fait, en acceptant de faire partie de ce jury.

Résumé

Le travail présenté dans ce manuscrit se situe dans le domaine de l'analyse et l'interprétation d'images. Plus précisément, nous nous intéressons à la détection de la saillance et la segmentation d'objet saillant ainsi qu'à la reconnaissance de catégorie d'objets dans les images. Trois contributions se dégagent de cette étude. Dans la première contribution, nous proposons un modèle de saillance qui se base sur une approche neuronale. Nous procédons à l'extraction de caractéristiques visuelles telle que la couleur et la texture. La caractéristique de la couleur est exprimée dans l'espace de couleur CIE Lab, un espace de couleur perceptuellement uniforme et similaire à la perception humaine. La caractéristique de texture est extraite par l'intermédiaire des filtres (Log)-Gabor. Le réseau Self Organizing Tree procède au partitionnement des vecteurs caractéristiques des pixels de l'image en différents clusters. Pour chaque cluster, nous calculons la mesure de saillance basée sur l'indice spatial, spatial cue. En supposant que la vraisemblance de la saillance d'un pixel appartenant à un cluster satisfait une distribution gaussienne. La carte de saillance finale est obtenue en calculant la probabilité marginale de la saillance. L'évaluation quantitative et qualitative du modèle proposé sur la base d'images MSRA-1000 démontre des résultats satisfaisants comparé à d'autres modèles de saillance. Notre seconde contribution porte sur la mise en œuvre d'une méthode de segmentation d'objet qui se base sur le regroupement spectral des pixels. Cette méthode intègre la valeur de saillance des pixels, calculée à partir du modèle de saillance proposé, dans le calcul du graphe de similarité. L'évaluation qualitative de la méthode proposée sur la base d'images MSRA-1000 démontre des résultats satisfaisants comparés à d'autres méthodes de segmentation d'objet reconnues dans la littérature. Notre troisième contribution porte sur l'élaboration d'un modèle de dictionnaire hybride pour une tâche de catégorisation d'objets. Une approche de classification simultanée est appliquée aux descripteurs d'images pour générer deux variantes de modèles de dictionnaire. Ceux-ci sont utilisés séparément pour coder et représenter une image au travers un modèle de dictionnaire basé sur les patchs et un modèle de dictionnaire basé sur les caractéristiques. Nous testons et validons le modèle de dictionnaire proposé sur la base d'images Caltech-101. Les résultats expérimentaux démontrent sa performance vis-à-vis à d'autres modèles de dictionnaire adoptant une simple classification.
Mots-clés : perception visuelle, attention visuelle , détection de la saillance, segmentation d'objets, reconnaissance d'objet, dictionnaire visuel, modèle de codebook.

Abstract

The work presented in this manuscript is in the field of image analysis and interpretation. More precisely, we are interested in the detection of salience and segmentation of salient objects as well as the recognition of categories of objects in images. Three contributions emerge from this study. In the first contribution, we propose a saliency model based on a neural approach. We proceed to the extraction of visual characteristics such as color and texture. The color characteristic is expressed in the CIE Lab color space, a perceptually uniform color space similar to human perception. The texture characteristic is extracted via (Log)-Gabor filters. The Self Organizing Tree network partitions the characteristic vectors of the image pixels into different clusters. For each cluster, we compute the salience measure based on the spatial cue . Assuming that the likelihood of the saliency pixel belonging to a cluster satisfies a Gaussian distribution. The final saliency map is obtained by calculating the marginal probability of saliency. The quantitative and qualitative evaluation of the proposed model based on MSRA-1000 images shows satisfactory results compared to other saliency models. Our second contribution deals with the implementation of an object segmentation method based on the spectral clustering of pixels. This method integrates the pixel saliency value, calculated from the proposed saliency model, in the calculation of the similarity graph. The qualitative evaluation of the proposed method based on MSRA-1000 images shows satisfactory results compared to other object segmentation methods in the literature. Our third contribution deals with the development of a hybrid dictionary model for an object categorization task. A simultaneous classification approach is applied to image descriptors to generate two variants of dictionary models. These are used separately to code and represent an image through a patch-based dictionary model and a feature-based dictionary model. We test and validate the proposed dictionary model based on Caltech-101 images. Experimental results demonstrate its performance against other dictionary models that adopt a simple classification.
Keywords : visual perception, visual attention , saliency detection, object segmentation, object recognition, visual codebook, codebook model

ملخص

في السنوات الأخيرة ، ازداد حجم الصور المتاحة للجمهور بشكل مطرد بشكل مطرد بسبب تطوير الإنترنت وإضفاء الطابع الديمقراطي على أجهزة الاستحواذ الرقمية .ونتيجة لذلك ، اهتم كثير من الباحثين ، ولا سيما من مجتمع المنك الاصطناعي اهتمامًا كبيرًا بتطوير أنظمة تجعل من الممكن الحصول على الصور ومعالجتها وتحليلها وتفسيرها تلقائيًا أو شبه تلقائي تتعلق الدراسات التي أجريت في هذه الرسالة بتطوير الخوارزميات التي تجعل من الممكن تحليل الصور الرقمية وتفسيرها تلقائيًا .ركزنا بحثنا على الكشف عن الأشياء البارزة وتقسيمها إلى أجزاء، بالإضافة إلى التعرف البصري على الأشياء، على وجه الخصوص تصنيف الأشياء العامة في الصور.تعلق مساهمتنا الأولى بتنفيذ طريقة للكشف عن الأشياء البارزة في الصور والتي تستند إلى مزيج من نهجين مختلفين :أحدهما يعتمد على نهج عصبي و أخرى على نهج التردد نحن نستخرج خصائص بصرية منخفضة المستوى مثل اللون والتوجه .يتم التعبير عن خاصية اللون في فراغ اللون Lab ، وهي مساحة لونية موحدة بشكل ملحوظ تشبه الإدراك البشري .يتم استخراج خاصية التوجه عبر مرشحات جابور .أنها توفر أفضل توطين متزامن للمعلومات المكانية والتردد .ومع ذلك ، فإن الحد الأقصى لعرض النطاق الترددي لمرشح غابور يقتصر على حوالي أوكتاف واحد ومرشحات غابور ليست مثالية إذا كنا نبحث عن معلومات طيفية واسعة مع أقصى موقع مكاني .على النقيض من ذلك ، يتم إنشاء مرشحات Log-Gabor باستخدام نطاق ترددي تعسفي ويمكن تحسين عرض النطاق الترددي لإنتاج مرشح بأقل حد مكاني يتم إعطاء جميع هذه الخصائص كمدخلات لشبكة الخلايا العصبية ذاتية التنظيم .(SOTA) في نهاية مرحلة التعلم الخاضعة للإشراف ، يتم الحصول على خريطة طوبوغرافية للخلايا العصبية ، والتي تمثل نواتج إخراج النموذج الأولي مراكز التجمعات .k يتم حساب مؤشر الانتباه البصري الفردي ، جديلة مكانية ، لقياس قيمة بروز كل عنقود .نفترض أن احتمالية بروز بكسل ينتمي إلى عنقود يرضي التوزيع الغوسي .وبالتالي ، يتم الحصول على خريطة الملوحة النهائية عن طريق حساب الاحتمالية الهامشية للملوحة .يوضح التقييم التجريبي الكمي والنوعي على أساس صور MSRA-1000 نتائج واعدة من النموذج المقترح ، والذي يتجاوز بعض نماذج الكشف عن البروز الموجودة في الأدبيات يعامل التصنيف الطيفي كل بكسل صورة كعقدة رسم بياني وبالتالي يحول مشكلة التجميع إلى مشكلة تقسيم الرسم البياني .من المعروف أن التكتل الطيفي يتجاهل العلاقات المكانية بين بكسلات الصورة ، مما يعيق إنتاج مناطق غير متناسقة .بالإضافة إلى ذلك ، غالبًا ما تتطلب الحاجة إلى تحديد عدد المجموعات وتهيئة مراكزها إشرافًا بشريًا ، مما يحد من تطبيقها في تجزئة الصورة.

المساهمة الثانية لهذا العمل هي اقتراح طريقة تجزئة الكائن التي تجمع بين خريطة الملوحة المحسوبة من النموذج البارز المقترح سابقًا مع خوارزمية التصنيف الطيفي .في حالتنا ، لم نكن بحاجة إلى تحديد عدد المجموعات لأنه يمكن الحصول عليها من خريطة الكتلة التي تم إنشاؤها في نهاية مرحلة التعلم لشبكة .SOTA تحافظ هذه الخريطة الطوبوغرافية على علاقات الجوار لوحدات البكسل في الصورة نظرًا لأن SOTA تنفذ تجميعًا هرميًا يعتمد على الخرائط ذاتية التنظيم .(SOM) يوضح التقييم التجريبي الكمي والنوعي المستند إلى صور MSRA-1000 النتائج الواعدة للطريقة المقترحة مقارنةً بطرق تجزئة الصورة الثنائية الأخرى

تتعلق مساهمتنا الثالثة بتطوير نموذج كتاب هجين لمهمة تصنيف الكائنات .يتم تطبيق نهج التجميع المتزامن على واصفات الصور لإنشاء متغيرين من نماذج دفتر الكود .يتم استخدامها بشكل منفصل لتشفير وتمثيل صورة من خلال نموذج دفتر الرموز القائم على التصحيح ونموذج دفتر الرموز القائم على الميزات .تم اختبار نموذج الكودك المقترح والتحقق منه على أساس صور .Caltech-101 تظهر النتائج التجريبية أدائها مقارنة بنماذج دفتر الكودك الأخرى القائمة على التجميع البسيط

الكلمات المفتاحية :الإدراك البصري ، الانتباه البصري ، الكشف عن البروز ، تجزئة الأشياء ، التعرف على الأشياء ، القاموس المرئي نموذج دفتر الكود

Table des matières

Liste des tableaux

Liste des figures

Introduction

« *Celui qui veut réussir trouve un moyen.Celui qui ne veut rien faire trouve une excuse.*
»

Proverbe Français

L'INTRODUCTION de cette thèse présente d'abord son contexte global dans le domaine de l'analyse et l'interprétation des images. Puis, les deux problématiques de recherche abordées dans ce travail, la détection de la saillance ainsi que la reconnaissance d'objets dans les images, sont exposées et motivées. Elles sont suivies des objectifs visés et des contributions majeures apportées dans la thèse. Enfin, un aperçu de la structure du manuscrit est présenté.

Contexte de la recherche

La vision est le sens le plus important de l'espèce humaine. Près de la moitié de notre cerveau y est consacré. De nombreuses tâches de la perception visuelle humaine comme la détection, l'identification, la reconnaissance des objets sont accomplit de manière involontaire et automatique. Bien qu'une personne n'éprouve aucune difficulté à réaliser ces taches, des mécanismes cognitifs complexes se cachent derrière. Deux exemples concrets sont les aptitudes humaines à sélectionner et interpréter les informations visuelles.

— Le système visuel humain reçoit une quantité énorme d'informations visuelles, au niveau du nerf optique, estimée environ de 10^7 à 10^8 bits d'information chaque seconde. Environ 60 images constituées de millions de pixels chacune y sont traitées chaque seconde [KOCH et collab., 2006]. Afin de pouvoir gérer toute cette masse d'information en temps réel, un mécanisme capable de sélectionner les parties de l'image les plus pertinentes par rapport à la tâche à réaliser s'avère nécessaire pour réduire le temps de traitement. De nombreuses études ont démontré l'existence d'un tel mécanisme dans le système d'attention visuel humain.

— Une personne arrive à reconnaître plusieurs milliers de catégories d'objets dans des images ou des séquences vidéo. En effet, nous ne trouvons aucune difficulté à différencier des catégories d'objets similaires comme une moto et une bicyclette. Nous pouvons aussi identifier des objets de la même classe malgré des changements de position, de taille, de point de vue, d'éclairage, et même en présence

d'autres objets distrayants. En plus, de cette capacité impressionnante à interpréter les objets, il a été estimé qu'une personne arrive à reconnaître un objet en moins de 0,5 seconde.

Par ailleurs, la démocratisation des appareils numériques d'acquisition ainsi que l'essor d'Internet au cours de ces dernières années, a contribué potentiellement à l'explosion du volume d'images disponibles pour le grand public. Malgré que les ordinateurs et les téléphones intelligents actuels puissent gérer des tâches à forte intensité de calcul, ils ne sont pas encore en mesure d'effectuer certaines tâches cognitives de base qu'une personne normale effectue involontairement.

La vision artificielle est une discipline qui a vu le jour afin de reproduire la perception visuelle humaine sur un ordinateur, en automatisant des tâches que le système visuel humain est capable de réaliser. Il s'agit d'un processus qui consiste à acquérir, traiter, analyser et interpréter des images numériques et à extraire des données de grande dimension du monde réel pour produire des informations numériques/symboliques. Le processus d'acquisition d'images est effectué à l'aide de caméras, puis à travers de l'analyse d'images les images acquises sont traitées afin de produire une description des éléments qui composent la scène 3D. L'analyse d'images regroupe plusieurs disciplines que l'on peut classer selon deux niveaux.

— Les processus de bas niveau. Ils sont principalement axés sur le calcul des composants d'image et de leurs propriétés. Par exemple, la description de la forme, la détection de contour, l'extraction des caractéristiques de texture, la localisation des objets, la segmentation des objets.

— Les processus de haut niveau. Ils sont principalement axés sur le calcul de la sémantique des images et de la prise de décision basée sur l'extraction du contenu des images. Par exemple, la reconnaissance des objets, l'interprétation des scènes. Ils s'agit en général des processus cognitifs.

Les études menées dans cette thèse portent sur le développement d'algorithmes qui permettent d'analyser et d'interpréter automatiquement des images. Un système d'analyse d'images consiste à retourner en sortie de l'information extraite à partir d'images en entrée, d'un niveau sémantique intermédiaire, tandis qu'un système d'interprétation d'images consiste à renvoyer de l'information d'un niveau sémantique plus élevé. La frontière entre les processus d'analyse et d'interprétation d'images est toutefois parfois floue, et la terminologie utilisée varie d'un chercheur à l'autre. Dans notre travail, le champ de l'analyse d'image se verra réduit à la tâche de la détection de la saillance et la segmentation d'objet, tandis que le champ de l'interprétation d'image se verra réduit à celui de la reconnaissance d'objets, et plus particulièrement à la catégorisation des objets génériques.

La détection de la saillance

La détection de la saillance visuelle a suscité l'intérêt de nombreux chercheurs en vision artificielle. La façon dont le cerveau humain sélectionne les informations visuelles importantes est liée aux mécanismes d'attention du cerveau qui guident éventuellement les mouvements des yeux pour placer la fovéa, qui est une zone au centre de la rétine, sur les parties les plus saillantes et significatives de la scène perçue.

Selon des études réalisées en sciences cognitives, l'allocation de l'attention visuelle est réalisée selon deux mécanismes. Un mécanisme ascendant (bottom-up) qui autorise le traitement précoce de certaines informations permettant ainsi la réalisation de tâches complexes sans solliciter les ressources attentionnelles. Un mécanisme descendant (top-down) qui nécessite un effort cognitif important et l'allocation de la quasi-totalité des ressources attentionnelles. Ce mécanisme est déployé lorsqu'une tâche particulière doit être effectuée, par exemple reconnaître une personne sur une photo, chercher un objet bien précis. Toutefois, dans la réalité, les mécanismes ascendant et descendant ne sont pas considérés comme des processus indépendants, mais sont plutôt combinés pour diriger l'attention visuelle. Inspiré par la théorie de l'intégration des caractéristiques (FIT) et le modèle de recherche guidée, les premiers modèles d'attention visuelle ont vu le jour. En correspondance avec les mécanismes attentionnels ascendants et descendants, les modèles d'attention visuelle reposent soit sur des facteurs ascendants ou descendants. Toutefois, la complexité des interactions existantes entre ces deux mécanismes rend la modélisation de l'attention visuelle dans son ensemble une tâche difficile. Une voie réaliste est de modéliser l'attention visuelle ascendante, à travers des modèles de saillance appelés aussi des modèles pré-attentive, qui sont liés à des processus automatiques.

La détection de la saillance a déjà connu de nombreux domaines d'application comme la vidéo surveillance, la compression, la manipulation d'image, le rognage automatique d'image, la détection d'objet, la reconnaissance d'objet, l'extraction basée sur le contenu d'images/vidéos. Néanmoins, la détection de la saillance reste un sujet de recherche important et complexe à la fois. Cette difficulté est liée au fait qu'il tente d'accomplir la même tâche que le système visuel humain, un système complexe et difficile à reproduire. Le système visuel humain utilise un certain nombre d'indices visuels, tels que le contraste d'intensité, le contraste de couleur et l'emplacement spatial, combinés éventuellement à des connaissances préalables, afin de détecter et identifier des objets d'intérêts. La détection de la saillance peut modéliser de nombreuses caractéristiques semblables mais à des coûts de temps et de la complexité. Un autre aspect du défi de la détection de la saillance réside dans le fait qu'un objet peut être considéré comme saillant dans une image et par conséquent appartenir à son premier plan, comme il peut être considéré non saillant dans une autre image, et par conséquent appartenir à son arrière plan. Dans cette thèse, nous nous intéressons à la détection de la saillance qui considère principalement les facteurs ascendants c'est à dire l'influence des caractéristiques visuelles de bas niveau. Ce type de saillance dépend des données et des stimuli. Ainsi aucune connaissance n'est spécifiée au préalable. Au cours de ces dernières années, ce type de détection de saillance est devenu un domaine de recherche très actif dans la communauté de la vision par ordinateur.

Plusieurs taxonomies visant à catégoriser les modèles de saillance visuelle ont été proposé dans la littérature. Elles considèrent généralement que la définition de la saillance visuelle réside dans l'unicité visuelle, l'anormalité, la rareté, la maximisation de l'auto-information, la surprise, ou quelque chose de rare. De ce fait, notre premier objectif est d'étudier ces différentes taxonomies, leurs points communs et leur différences. Définissent t-elles de la même manière les catégories des modèles de saillance ? Qu'apporte une nouvelle catégorisation par rapport aux taxonomies précédentes ?

Indépendamment de la catégorie d'appartenance d'un modèle de saillance, ce dernier est censé générer une carte de saillance qui indique les zones les plus saillantes se

trouvant dans une image en entrée. Pour cela, différentes caractéristiques visuelles sont utilisées pour calculer des indices de saillance. Ainsi, notre deuxième objectif est d'étudier les différents aspects qui peuvent influencer sur la qualité d'une carte de saillance. En d'autres termes, cela revient à répondre aux questions suivantes :
Quels indices de saillance doit-on calculer pour obtenir une carte de saillance de bonne qualité, vue qu'elle est censé refléter au mieux les endroits/objets saillants dans une image ?
Sur quelles caractéristiques visuelles se basent les modèles de saillance pour calculer les mesures de saillance ?

Certaines des taxonomies étudiées catégorisent les modèles de saillance selon qu'ils visent à résoudre un problème de segmentation d'objet saillant. Ces modèles tentent, dans un premier temps, à détecter un objet saillant dans l'image puis, dans un deuxième temps, à le segmenter. La détection et la segmentation sont deux processus qui se déroulent successivement. Le résultat de la détection de la saillance est un masque binaire qui sépare l'objet saillant du premier plan de l'image de celui de l'arrière plan. Il est toutefois moins clair comment la détection et la segmentation objet saillant se rapporte à la tâche de segmentation d'objet. Ce nouveau champ de recherche a soulevé une ambiguïté concernant sa définition par rapport à la segmentation d'objet qui est vu comme un problème de séparation figure/fond. De ce fait, notre troisième objectif est de mettre en évidence les différentes approches de segmentation d'objets existantes dans la littérature. Il y a t-il un lien entre ces modèles de saillance et les méthodes de segmentation d'objet ?

Les modèles de saillance ayant été catégorisé selon qu'ils visent à résoudre un problème de détection et segmentation d'objet se basent souvent sur une étape de segmentation au préalable. Le choix de la méthode de segmentation s'avère ainsi critique vu que le calcul des indices de saillance en dépend largement. De ce fait, nous pensons que le développement d'un modèle de saillance basé sur une approche algorithmique, qui se base essentiellement sur une segmentation au préalable, et une approche connexioniste qui contribuerait directement dans la construction de la carte de saillance, serait avantageux et pourrait améliorer la qualité de la carte de saillance. De plus, nous pensons que l'utilisation de l'algorithme Self Organizing Tree(SOTA), qui a été utilisé avec succès dans l'analyse des données d'expression génétique peut s'avérer avantageux. Cette hypothèse s'appuie sur le fait qu'il combine les avantages de la classification hiérarchique et des cartes auto-organisatrice de Kohonen (SOM).

La reconnaissance d'objet

La reconnaissance d'objets est une discipline de la vision artificielle qui s'intéresse à programmer des ordinateurs pour qu'ils soient eux aussi capables de reconnaître des catégories d'objets ou des instances d'objets de la même manière que les être humains. La reconnaissance visuelle des objets a un large éventail d'applications possibles, comme l'annotation automatique des images, la surveillance vidéo, l'indexation et la récupération des images/vidéos basées sur le contenu. Malgré les nombreux efforts et progrès réalisés au cours des dernières années, la reconnaissance visuelle des objets demeure un problème ouvert et est toujours considérée comme l'un des problèmes les plus difficiles en vision par ordinateur. La raison principale réside dans les difficultés des ordinateurs à faire face aux diverses variations intra-classes, y compris la déformation de

l'apparence, l'occlusion, le fouillis de fond, les changements de point de vue, de pose, d'échelle et d'éclairage qui sont des problèmes beaucoup plus faciles pour les humains.

Parmi les approches de reconnaissance d'objets, une approche très populaire est celle qui se base sur les caractéristiques et les classificateurs, et ceci en raison du grand développement des caractéristiques/descripteurs d'image avancés et des algorithmes de reconnaissance des formes. En particulier, l'utilisation de descripteurs locaux, comme le descripteur SIFT, ainsi que la représentation sac à mots visuels suivie de classificateurs discriminants tels que les machines à vaste supports(SVM) est devenue le paradigme dominant depuis 2004. Une amélioration de la représentation sac à mots visuels a été de réintroduire l'information spatiale en superposant sur l'image une grille pyramidale assez grossière, par exemple 16 carrés, puis 4, puis 1, pour voir comment les éléments se répartissent dans l'image. La pyramide spatiale permet ainsi de prendre en considération des détails de l'image sur plusieurs niveaux de résolution. Cette méthode de reconnaissance des catégories d'objet est restée la meilleure en reconnaissance visuelle pendant plusieurs années. Elle a été exploité par plusieurs chercheurs et référencé comme le framework générique pour la catégorisation d'objets. Récemment, de nombreux travaux de recherche ont été proposé pour améliorer ce framework à travers de nouvelles méthodes avancées de génération de dictionnaire visuel, de codage et d'agrégation(pooling).

Une approche classique de création d'un dictionnaire visuel est réalisée par une approche de classification non supervisée(clustering). Les méthodes comme les k-moyennes, les k-moyennes hiérarchiques et le décalage moyen regroupent les vecteurs caractéristiques d'apprentissage d'une image en classe(clusters) et représentent le centre de chaque classe par un unique mot visuel. Cependant, une image peut être représentée de différentes manières et souvent un seul et unique dictionnaire visuel, qu'il soit basé sur des données ou annotations, ne suffit pas pour décrire complètement le contenu de l'image. Pour cette raison, plusieurs auteurs se sont concentrés sur la construction de plusieurs dictionnaires, soit à partir d'un ensemble d'apprentissage homogène caractérisé par des descripteurs d'images locaux de même type(descripteur SIFT uniquement), soit à partir d'un ensemble d'apprentissage hétérogène caractérisé par des descripteurs d'images locaux de différents types comme les descripteurs SIFT et LBP.

Toutefois, même si plusieurs dictionnaires visuels sont construits, ils se basent essentiellement sur une approche de clustering qui typiquement attribue une donnée à un seul cluster sur la base de similarités globales, c'est-à-dire des mesures de similarité calculées sur tous les attributs. Ces mesures de similarité sont généralement basées sur la notion de distance,comme la distance Euclidienne, la distance de Manhattan.

Contrairement à la classification qui vise à regrouper les variables, dans une matrice de données, appartenant à un modèle global dans les données, la co-classification (biclustering) est une méthode d'analyse de données conçue pour détecter les modèles locaux dans les données. Elle opère simultanément sur l'ensemble des objets et des attributs d'une matrice de données, à la recherche de sous matrices constituées de sous ensembles d'objets qui ont un modèle très cohérent sur un sous ensemble d'attributs. Ce paradigme a été utilisé dans l'analyse des données d'expression génétique. Nous pensons qu'employer la co-classification dans la construction de dictionnaire visuel serait avantageux.

Contributions

Notre première contribution dans le cadre de cette thèse porte sur la proposition et la mise en œuvre d'un modèle de saillance basé une approche connexioniste et plus particulièrement sur l'algorithme Self-Organizing Tree(SOTA). Ce dernier combine les avantages de la classification hiérarchique descendante et des cartes auto-organisation de Kohonen(SOM). Ayant une topologie d'un arbre binaire initialisé à un nœud parent et deux cellules descendantes, la croissance de ce réseau de neurone se fait de manière dynamique. Après une étape d'extraction des caractéristiques de couleur dans l'espace CIE LAB et de texture en utilisant les filtres de Gabor et Log-Gabor, le réseau SOTA procède au partitionnement des vecteurs caractéristiques des pixels de l'image en différents clusters. Pour chaque cluster, nous calculons la mesure de saillance basée sur l'indice spatial, spatial cue. En supposant que la vraisemblance de la saillance d'un pixel appartenant à un cluster satisfait une distribution gaussienne. La carte de saillance finale est obtenue en calculant la probabilité marginale de la saillance. L'évaluation quantitative et qualitative du modèle proposé sur la base d'images MSRA-1000 démontre des résultats satisfaisants comparé à d'autres modèles de saillance.

L'étude réalisée sur les différentes approches de segmentation d'objet existantes dans la littérature, nous a permis de mettre l'accent sur la théorie spectrale. Les méthodes de classification spectrale issues de cette théorie modélisent une image sous forme de graphe, transformant ainsi le problème de segmentation d'image en un problème de partitionnement de graphe. Notre seconde contribution porte sur la mise en œuvre d'une méthode de segmentation d'objet qui se base sur le regroupement spectral des pixels. Cette méthode intègre la valeur de saillance des pixels, calculée à partir du modèle de saillance proposé, dans le calcul du graphe de similarité. L'évaluation qualitative de la méthode proposée sur la base d'images MSRA-1000 démontre des résultats satisfaisants comparés à d'autres méthodes de segmentation d'objet reconnues dans la littérature.

Notre troisième contribution porte sur l'élaboration d'un modèle de dictionnaire hybride pour une tâche de catégorisation d'objets. Une approche de classification simultanée est appliquée aux descripteurs d'images pour générer deux variantes de modèles de dictionnaire. Ces derniers sont utilisés séparément pour coder et représenter une image au travers un modèle de dictionnaire basé-patchs et un autre modèle de dictionnaire basé-caractéristiques. Le modèle de dictionnaire proposé est testé et validé sur la base d'images Caltech-101. Les résultats expérimentaux démontrent sa performance comparé à d'autres modèles de dictionnaire se basant sur une simple approche de classification non supervisée dans la création du dictionnaire visuel.

Organisation du manuscrit

Le présent document est structuré en sept chapitres répartis en deux parties. Le présent chapitre préambule est une introduction générale qui précise le contexte de recherche, les problématiques de recherche, les contributions de la thèse ainsi que l'organisation du manuscrit. La première partie, contenant les chapitres 1, 2, 3 et 4 est consacrée à la présentation des principaux concepts, outils et travaux relatifs ainsi que nos deux premières contributions à l'étude entreprise. Dans la deuxième partie du mémoire, représentée par les chapitres 5, 6 et 7, nous présentons les principaux concepts,

outils et travaux relatifs à la reconnaissance visuelle des objets ainsi que notre troisième contribution.

Chapitre 1. La perception visuelle

Ce chapitre est consacré à la présentation des principales structures neuronales mises en jeu ainsi que leur rôle respectif dans la perception visuelle humaine. Nous présentons les différents concepts de base relatifs à l'attention visuelle ainsi que ses deux stratégies d'exploration selon un mécanisme ascendant et descendant. Nous expliquons d'une façon non exhaustive quelques phénomènes pertinents pour la compréhension des modélisations de l'attention visuelle présentées dans le chapitre 2.

Chapitre 2. Modèles de saillance : Un état de l'art

Dans ce chapitre, nous évoquons d'abord les différentes taxonomies/catégorisations des modèles de saillance. Par la suite, nous présentons les différents modèles de saillance visuelle selon la tache à réaliser, la prédiction des fixations oculaires et la détection d'objet saillant. Nous passons en revue les principales bases de données utilisées ainsi que les principales métriques utilisées pour l'évaluation de la qualité des cartes de saillance générées par les modèles de saillance. Nous terminons ce chapitre par une comparaison des modèles de saillance en définissant comme critères de comparaison, les caractéristiques de bas niveau utilisées, les cartes de caractéristiques calculés, la méthode de segmentation d'image éventuellement utilisée, le mécanisme adopté et la mesure de saillance visuelle calculée pour générer la carte de saillance.

Chapitre 3. Approches de segmentation d'objet : Un état de l'art

Ce chapitre est dédié au domaine de la segmentation d'images. Tout d'abord, nous distinguons trois niveaux de segmentation d'image, la segmentation d'image de bas niveau, la segmentation d'objet et la segmentation sémantique. Nous évoquons les principales bases de données utilisées. Nous mettons un accent particulier sur le problème de la segmentation d'objet connu sous le nom de la segmentation figure/fond, dont le but est de séparer un objet d'intérêt de l'arrière plan d'une image. Nous catégorisons les méthodes de segmentation d'objets selon le mode d'interaction(automatique, semi-automatique, totalement automatique) ainsi que le fait qu'elles se basent ou non sur l'information de la saillance visuelle.

Chapitre 4. Un Modèle de Détection de Saillance : application à la segmentation d'objet

Ce chapitre est consacré à nos deux premières contributions dans les domaines de la détection de la saillance et la segmentation d'objet. Notre première contribution porte sur la mise en œuvre d'une modèle de saillance d'images. L'évaluation expérimentale sur la base d'images MSRA-1000 démontre des résultats satisfaisants du modèle proposé comparé à certains modèles de détection de saillance de l'état de l'art. Notre deuxième contribution porte sur la mise en œuvre d'une méthode de segmentation d'objet qui combine la carte de saillance calculée à partir du modèle de saillance proposé à l'algorithme de classification spectrale.

Chapitre 5. La reconnaissance d'objets

Ce chapitre est consacré au domaine de la reconnaissance des objets dans les images. Nous distinguons la reconnaissance de catégories d'objets de la reconnaissance d'instances d'objets. Nous présentons les diverses approches de la reconnaissance de catégories d'objets ainsi que les principales bases d'images utilisées.

Chapitre 6. Dictionnaire Visuel et Modèles de Dictionnaires

Ce chapitre est dédié à l'état de l'art dans le domaine de la catégorisation des objets dans les images. Nous nous concentrons dans sa première partie sur la présentation des principaux travaux de recherche de création de dictionnaire visuelle. Nous comparons les différents travaux du domaine en définissant comme critères de comparaison, le type de dictionnaire visuel, l'approche et la méthode adoptée lors de sa construction. Ensuite, nous nous concentrons dans sa deuxième partie sur la présentation des principaux travaux de recherche qui portent sur la proposition de nouveaux modèles de dictionnaire visuel. Nous comparons les différents travaux réalisés en définissant comme critères de comparaison, le modèle de dictionnaire(représentation spatiale, codage, pooling) utilisé et l'approche de construction du dictionnaire utilisé. Cette étude nous a permis de proposer une nouvelle approche de catégorisation d'objet.

Chapitre 7. Une nouvelle approche de catégorisation d'objets

Ce chapitre est consacré à notre troisième contribution qui porte sur l'élaboration d'un modèle de dictionnaire hybride pour une tâche de catégorisation d'objets. Une approche de clustering simultanée est appliquée aux descripteurs d'images pour générer deux variantes de modèles de dictionnaire. Ceux-ci sont utilisés séparément pour coder et représenter une image au travers un modèle de dictionnaire basé patch et un modèle de dictionnaire basé caractéristiques. Le modèle de dictionnaire proposé est testé et validé sur la base d'images Caltech-101. Les résultats expérimentaux démontrent sa performance comparé à d'autres modèles de dictionnaire qui se basent sur une simple approche classification non supervisée pour la création du dictionnaire visuel.

A la fin de ce mémoire, nous émettons nos conclusions sur les recherches que nous avons menées dans le domaine de l'analyse et l'interprétation d'images. Nous présentons quelques perspectives envisageables pour faire évoluer les propositions que nous avons présentées dans ce document.

Première partie

DÉTECTION DE LA SAILLANCE ET SEGMENTATION D'OBJET

Chapitre 1

La perception visuelle

« Si tu essais tu as une chance de perdre.Si tu n'essaies pas, tu as déjà perdu. »

Sommaire

1.1 Introduction

La perception visuelle regroupe les mécanismes mis en œuvre pour la réception et la cognition de stimuli visuels. La partie réceptive se charge de capter et d'organiser les informations visuelles en provenance de l'environnement, tandis que la partie cognitive se préoccupe de l'interprétation de ces informations. La réception des stimuli visuels est effectuée par l'œil et la rétine en particulier. La quantité d'informations provenant de ces stimuli visuels étant trop importante pour être traitée dans sa totalité, un mécanisme de sélection des informations est nécessaire : c'est le rôle de l'attention visuelle. Avant de définir l'attention visuelle et de présenter les principaux concepts et théories qui lui sont liée, nous présentons tout d'abord quelques notions sur la biologie du système visuel humain. Nous détaillons les étapes de traitement de l'information visuelle au niveau de la rétine et du cortex visuel. L'explication de ces notions est fondamentale pour la compréhension de la modélisation de l'attention visuelle présentée dans le chapitre 2.

1.2 De l'œil aux voies visuelles

Le système visuel humain se compose de l'ensemble des deux yeux ainsi que des régions du cerveau qui participent au traitement des informations visuelles.

1.2.1 L'œil

L' œil ou le globe oculaire est l'organe de la vision , il nous permet de capter la lumière de notre environnement et de la convertir en message nerveux, lequel est transmis au cerveau qui l'analyse. Il est de faible volume $6.5\,cm^3$, pèse 7 grammes, et possède la forme d'une sphère d'environ 24 mm de diamètre, complétée vers l'avant par une autre demi sphère de 8 mm de rayon, la cornée.

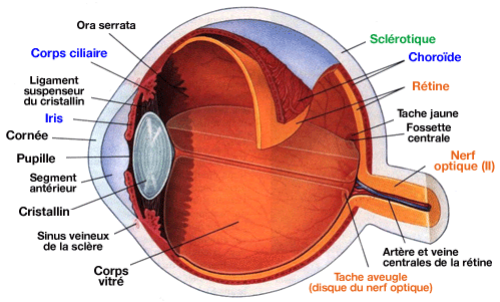


FIGURE 1.1 – Représentation de l'œil humain [1]

Le cristallin est la lentille de l'œil dont les fines modifications de forme se produisent, de manière automatique, en regardant un objet, afin de régler la distance de focalisation pour conserver l'image focalisée sur la rétine . Cela fait référence au phénomène d'accommodation qui est due à la contraction ou au relâchement des muscles ciliaires qui entourent la lentille, et elle joue un rôle important dans la vision des objets à toutes les distances[FERWERDA, 1998].

1. Source :http://vetopsy.fr/sens/vision/oeil.php

1.2.2 La rétine

La rétine est l'organe le plus important de l'œil. Elle mesure environ 0,5 mm d'épaisseur, et recouvre les trois quarts de l'intérieur du globe oculaire. La rétine appartient au système nerveux central, les autres parties de l'œil assurent des fonctions sensorielles dont le rôle est de focaliser les images sur la rétine. Cette dernière est composée d'un ensemble organisé de cellules nerveuses superposées, réparties verticalement et horizontalement, au travers lesquelles la lumière venant de l'extérieur est projetée [2]. Elle est constituée d'environ 150 millions de cellules nerveuses pouvant être décomposées en trois couches : La couche plexiforme externe(Outer Plexiform Layer, OPL), La couche plexiforme interne(Inner Plexiform Layer, IPL) et la couche ganglionnaire(Ganglionic Layer, GL) comme l'illustre la figure 1.2.

La couche plexiforme externe

La couche plexiforme externe(Outer Plexiform Layer ou OPL) se compose de cellules photo-réceptrices, horizontales et bipolaires. Elle capture la lumière incidente(les photons) et la transforme en un signal électrique [FERWERDA, 1998].

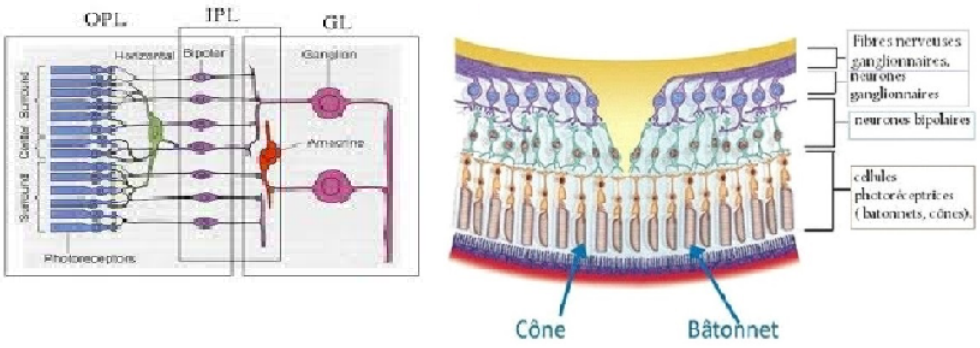


FIGURE 1.2 – Représentation de la coupe d'une rétine.

Les photo-récepteurs. La couche des cellules photo-récepteurs est la couche la plus éloignée du cristallin. Les photo-récepteurs sont sensibles à la lumière. Seules ces cellules sont en mesure de transformer l'information lumineuse en signaux nerveux. En d'autre terme, c'est dans ces neurones sensoriels que nait le message nerveux. Il existe deux types de photo-récepteur, les cônes et les bâtonnets.

— Les cônes. Ils représentent seulement 5 % des photo-récepteurs(5 millions) et sont concentrés au niveau de la fovéa, une zone au centre de la rétine qui couvre un disque d'environ 1.5 mm de diamètre de la surface de la rétine. Ils permettent la vision des couleurs ainsi que la perception des images détaillées.

— Les bâtonnets. Ils représentent 95 % des photo-récepteurs (120 millions)et peuvent réagir à des éclairements très faibles et sont donc utilisés pour distinguer différents niveaux de clarté. Ils ne sont pas dans la fovéa, où se situent les cônes, mais

2. http://www.savoirs.essonne.fr/

ils sont répartis dans la rétine. Ils perçoivent mal les couleurs car ils ont peu de liaisons directes avec le nerf optique, contrairement aux cônes.

Les photo-récepteurs sont répartis de manière hétérogène, et deux zones peuvent être distinguées.

— La rétine centrale. La fovéa est la région centrale de la rétine (5 degrés d'angle) où l'acuité visuelle et la résolution spatiale sont les meilleures. Les cônes y sont majoritairement localisés (150 000 cônes/m^2). La fovéola est une zone très petite(1 degré d'angle) située à l'intérieur de la fovéa qui correspond à la zone de fixation. Dans la rétine centrale, on constate un phénomène d'amplification de l'information, le phénomène de divergence. Une cellule photo-réceptrice va activer plusieurs cellules bipolaires, ce qui explique les bonnes performances de la zone centrale de la rétine.

— La rétine périphérique. La périphérie de la fovéa est la zone qui représente la quasi-totalité de la surface rétinienne, et contient majoritairement des bâtonnets. Il n'y a qu'une très faible densité de cônes. C'est la zone spécifique à la perception de faibles luminosités et il n'y a pas de perception détaillée. Dans la rétine périphérique, 30 à 50 bâtonnets rentrent en contact avec une cellule bipolaire c'est ce qui s'appelle le phénomène de convergence. Il conduit à une compression de l'information qui devient moins précise. Ceci implique une moindre acuité visuelle et une moins bonne performance spatiale de la zone périphérique de la rétine. En d'autres termes, la localisation précise de l'information ne peut pas être définit vue qu'elle a été regroupée avec plusieurs cellules. En effet, on ne peut pas connaitre lequel des 50 bâtonnets a reçu le signal.

Le champ récepteur d'un neurone est la portion de la rétine qui influence par excitation ou par inhibition l'activité du neurone lorsqu'elle est soumise à un stimulus visuel. L'intensité de la réponse du neurone dépend de la position du stimulus à l'intérieur du champ récepteur. La forme du champ récepteur correspond à l'intensité avec laquelle le neurone réagit en fonction de la position du stimilus dans le champ récepteur. Elle est importante pour la compréhension du système visuel car elle correspond à un filtrage rétinien []. La forme du champ récepteur peut aussi dépendre de la longueur d'onde du stimulus visuel. Le neurone peut être excité par certaines longueurs d'onde, et inhibées par d'autres [].

Le champ récepteur d'une cellule photo-réceptrice se limite au petit spot lumineux qui, dans le champ visuel, correspond à la localisation précise du photo-récepteur sur la rétine. Mais au fur et à mesure que l'on passe d'une couche de la rétine à l'autre, et aux neurones du cortex visuel, les champs récepteurs deviennent plus complexes.

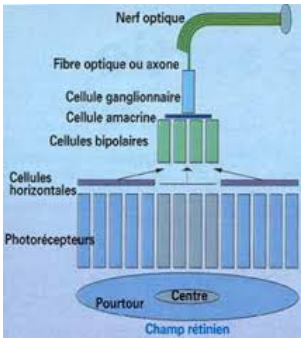


FIGURE 1.3 – Schéma d'un champ récepteur[3]

Les cellules horizontales Les cellules horizontales sont en contact avec les photo-récepteurs et les cellules bipolaires. Elles sont connectées latéralement à plusieurs cônes, bâtonnets et neurones bipolaires. Leur rôle est d'inhiber l'activité des cellules avoisinantes. Cette suppression sélective de certains signaux nerveux s'appelle l'inhibition latérale et son rôle principal est d'augmenter l'acuité d'un signal sensoriel. Dans le cas de la vision, quand une source lumineuse atteint la rétine, elle peut illuminer fortement certains photo-récepteurs et d'autres beaucoup moins. En supprimant le signal de ces photo-récepteurs moins illuminés, les cellules horizontales assurent que seul le signal des photo-récepteurs bien illuminés est transmis aux cellules ganglionnaires, améliorant ainsi le contraste et la définition du stimulus visuel.

Les cellules bipolaires Les cellules bipolaires relient un ou plusieurs photo-récepteurs à une cellule ganglionnaire. Les champs récepteurs des cellules bipolaires, qui correspondent aux régions du champ visuel où la présence d'un stimulus visuel modifie l'activité nerveuse de ces neurones de manière excitatrice ou inhibitrice sont circulaires et divisés en deux régions concentriques antagonistes le centre et le pourtour(ou la périphérie). On dit que les cellules bipolaires ont un champ récepteur de type centre-pourtour(center-surround en anglais). Les cellules bipolaires dites ON réagissent à une excitation des photo-récepteurs et une inhibition des cellules horizontales, ce qui se produit lorsque le signal incident est un spot de lumière entouré d'un pourtour sombre. Les cellules bipolaires dites OFF réagissent à une excitation des cellules horizontales et une inhibition des photo-récepteurs, ce qui correspond à un signal incident sombre au centre et lumineux au pourtour. Cette interaction antagoniste du centre sur le pourtour des cellules bipolaires est appelée mécanisme d'opposition centre-pourtour. Ce mécanisme permet aux cellules bipolaires d'être sensibles au contraste de luminance spatial(concept expliqué dans la section 1.4

Approximation de OPL. Une première approximation mathématique de la couche plexiforme externe a été proposé par [KUFFLER, 1952]. Une cellule bipolaire reçoit un signal d'entrée directement d'un ensemble de photo-récepteurs et/ou cellules horizontales comme l'illustre la figure 1.4 [DOUTSI et collab., 2018]. La figure de gauche illustre

3. Source :http://www.bioinformatics.org/oeil-couleur/dossier/images/champs-recepteurs.png

la structure rétinienne qui correspond à la couche OPL. La figure de droite illustre le champ récepteur d'une cellule bipolaire.

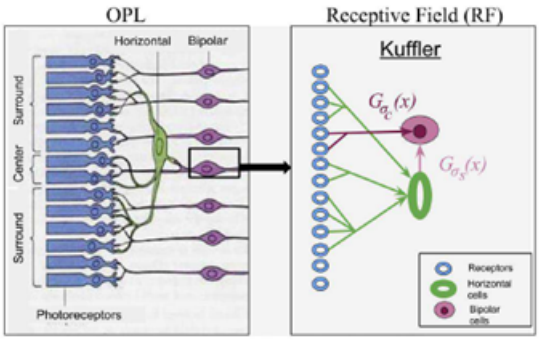


FIGURE 1.4 – La propagation du signal électrique vers le champ récepteur d'une cellule bipolaire selon [KUFFLER, 1952].

D'une part, la sortie de deux ou plusieurs photo-récepteurs est moyennée et transmise au centre du champ récepteur de la cellule bipolaire afin de l'exciter. Ceci est approximé par un filtre gaussien $G_{\sigma_c}(x)$ par l'équation 1.1

$$G_{\sigma c}(x) = \frac{1}{2\pi\sigma c^2} \exp(\frac{\|x\|^2}{2\sigma c^2}) \tag{1.1}$$

où $x \in R^2$, $\|x\|$ est la norme euclidienne de x. σ_c est l'écart-type qui règle la propagation du filtre gaussien.

D'autre part, un nombre identique ou supérieur de photo-récepteurs est lié aux cellules horizontales. Une cellule horizontale est fortement connectée aux cellules horizontales voisines, avec une moyenne de deux fois le signal entrant initial.

La sortie d'une ou plusieurs cellules horizontales est ensuite propagée vers le pourtour du champ récepteur de la cellule bipolaire afin de l'inhiber. Ce signal est modélisé par le filtre gaussien $G_{\sigma_c}(x)$ avec $\sigma_c \inf \sigma_s$. En conséquence, la cellule bipolaire reçoit deux signaux de signes opposés. Enfin, l'activité centre-pourtour à l'instant t $K(x, t)$ du champ récepteur de la cellule bipolaire est modélisée comme un filtre de Différence de Gaussienne(DoG) par l'équation 1.2.

$$DoG(x) = G_{\sigma_c}(x) - G_{\sigma_s}(x) \tag{1.2}$$

La couche plexiforme interne

La couche plexiforme interne(Inner Plexiform Layer ou IPL) se compose de cellules bipolaires et amacrines. La couche IPL est responsable de la rectification non linéaire du signal électrique.

Les cellules amacrines. Les cellules amacrines interviennent également dans la propagation latéralement du signal nerveux. Contrairement aux cellules bipolaires, qui sont sensibles au contraste spatial, les cellules amacrines sont notamment sensibles au contraste temporel et sont impliquées dans la détection du mouvement.

La couche ganglionnaire

La couche ganglionnaire(GL) constituée de cellules ganglionnaires est la dernière couche neuronale de la rétine. Leur axones se rejoignent pour former le nerf optique. Contrairement aux cellules précédemment décrites qui émettent des potentiels électriques gradués, les cellules ganglionnaires transmettent le signal nerveux sous forme de potentiels d'action. Les cellules ganglionnaires se catégorisent selon trois types : les cellules Parvocellulaire(P), les cellules Magnocellulaire(M) et les cellules Koniocellulaire(K).

Les cellules P réceptionnent et relayent une information visuelle issue des cônes du centre de la rétine, de haute résolution spatiale, de haute fréquence spatiale HFS), vu qu'elles sont faiblement soumise au phénomène de convergence. La diffusion de leur potentiel d'action est lent. Les cellules P sont plus sensibles à la forme et aux détails de la stimulation.

Les cellules M réceptionnent et relayent une information visuelle majoritairement issue des bâtonnets en périphérie de la rétine, de basse résolution spatial, de basse fréquence spatiale (BFS) car elles sont fortement soumise au phénomène de convergence. La propagation de leur potentiel d'action est rapide dans le nerf optique. Les cellules M sont particulièrement impliquées dans la détection du mouvement du stimulus.

Le champ récepteur des cellules ganglionnaires correspond à l'ensemble des cellules photo-récepteurs, horizontales et bipolaires qui sont en rapport avec elles. Chaque cellule ganglionnaire répond donc à la stimulation d'une petite zone de la rétine qui a la particularité d'être circulaire. Cette zone circulaire correspond a un petit disque qui comprend une partie centrale, excitatrice : cônes en contact synaptique direct et une partie périphérique, inhibitrice : cônes connectés à la cellule bipolaire via les cellules horizontales. Cette conformation permet d'éviter des interférences entre les informations. Ceci va créer du contraste ce qui permet de transmettre un signal le plus précis possible.

1.2.3 Voies nerveuses et CGL

De chaque œil partent deux demi nerfs optiques qui regroupent les stimuli de la partie gauche et de la partie droite du champ visuel et les deux demi nerfs optiques des champs situés vers le nez, s'entrecroisent dans le chiasma optique pour ensuite atteindre les corps genouillés latéraux(CGL). La rétine nasale gauche(qui est la partie de la rétine à gauche du nerf optique) et la rétine temporale droite(qui est la partie de la rétine à droite du nerf optique) sont aiguillées vers le CGL droit alors que la rétine temporale gauche et la rétine nasale droite sont envoyées vers le CGL gauche. Le rôle principal des CGL est une fonction de relais de l'information. Chaque CGL est constitué de 6 couches (la couche 1 étant la plus profonde et la 6 la plus superficielle). Les couches 1 et 2 sont magnocellulaires, elles reçoivent des axones de cellules ganglionnaires de type M. Les autres couches 3,4,5 et 6 sont parvocellulaires, elles reçoivent des axones de cellules ganglionnaires de type P. Les cellules des couches magnocellulaires et parvocellulaires ont des rôles différents dans la perception visuelle. Une lésion au niveau des couches magno provoquera une atteinte de la perception du mouvement, alors qu'une lésion au niveau des couches parvo provoquera une atteinte de la perception des couleurs, de la texture, de la profondeur et des contours. Les champs récepteurs des neurones du CGL ont la même configuration concentrique centre-pourtour que les cellules ganglionnaires.

1.2.4 Le cortex visuel

Arrivée au CGL, l'information visuelle traitée par la rétine se dirige vers le cerveau qui se compose de plusieurs aires(ou lobes) cérébrales comme le montre la figure 1.5. Le lobe occipital est le siège principal des processus visuels. Il existe en effet plusieurs aires visuelles, chacune ayant une fonction spécifique. C'est au niveau de l'aire V1 (ou cortex visuel primaire) que débute l'analyse de haut niveau de la scène visualisée. Néanmoins, l'aire V1 n'est que la première étape du traitement de l'information visuelle par le cerveau. Plus d'une trentaine d'aires corticales différentes contribuent à la perception visuelle. Les aires primaire(V1) et secondaire (V2) sont entourées de nombreuses autres aires visuelles tertiaires ou associatives V3, V4, V5 (ou MT), LO, etc. La figure 1.5 montre la répartition de ces aires.

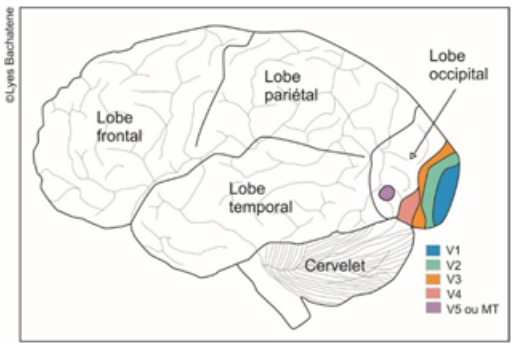
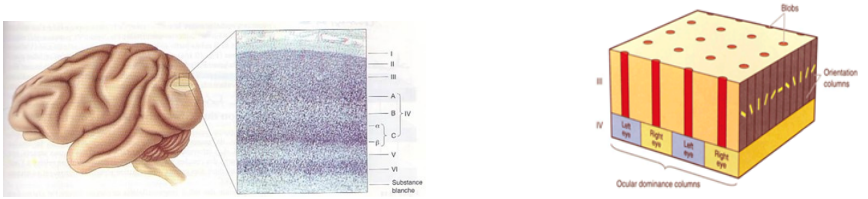


FIGURE 1.5 – Les aires cérébrales et les cinq aires visuelles du cortex visuel [4]

L'aire visuelle primaire V1(cortex strié, Aire 17 de Brodmann) est composée de six couches numérotées de I à VI. Les signaux provenant des CGN arrivent sur la couche du milieu IV et à partir de ces neurones, ils sont transmis vers les couches I à III qui vont analyser les signaux de chacun des deux yeux, et vers les couches V et VI qui regroupent les signaux des deux yeux pour permettre la vision en relief par l'appréciation des distances. Les figures 1.6a et 1.6b illustre cette organisation. En effet, chaque neurone de la couche IV du cortex strié réagit complètement à la stimulation lumineuse de l'œil droit ou gauche mais pas des deux yeux. La dominance pour l'œil gauche ou l'œil droit, se rencontre en bandes alternées de neurones de 800 µm de large, appelé colonne de dominance oculaire. Les neurones situés au-dessus et au-dessous de la couche IV peuvent répondre aux deux yeux mais tentent à être plus sensibles à l'œil qui active les neurones de la couche IV. [5]

Des études ont montré qu'au niveau du cortex strié, les colonnes d'orientation et de dominance oculaire se superposent l'une sur l'autre. Chaque colonne de dominance oculaire (800 um) contient une représentation ordonnée de toutes les orientations préférentielles possibles. Ainsi, à l'intérieur de deux colonnes de dominance adjacentes, toutes les orientations sont représentées deux fois, une pour chaque œil. Cette organisation montre plus précisément que *l'aire V1 procède à une analyse par bandes de fréquences et bandes d'orientations de la scène visualisée.* Cette sensibilité aux orientations a été mise en évidence par les travaux de [HUBEL et collab., 1962],[VALOIS et VALOIS, 1975]. Ces études ont également montré que les orientations verticales et horizontales

4. Source :Lyes Bachatene
5. http://lecerveau.mcgill.ca/



(a) Organisation en couches du cortex V1

(b) Organisation en colonnes corticales du cortex V1

FIGURE 1.6 – Organisation en couches et colonnes corticales du cortex V1

sont traitées différemment des orientations obliques, le cortex privilégie les directions horizontales et verticales. D'autres caractéristiques qualitatives du stimulus comme la sensibilité à la couleur, la sensibilité au mouvement, la fréquence spatiale, sont aussi organisés en colonne(module, bande), se répétant à intervalles réguliers dans tout le cortex strié. Les neurones sensibles à la couleur sont regroupés en région appelées taches de couleur [PRITCHARD et ALLOWAY, 2002].

Par ailleurs, l'analyse du cortex visuel montre que les neurones qui le constituent sont répertoriés comme simples, complexes et hypercomplexes avec des propriétés définies selon leur champ récepteur.

1.2.5 Voies visuelles : dorsale et ventrale

Après voir traversé les aires visuelles primaires, les signaux visuels se séparent et suivent deux voies parallèles de traitement de l'information relativement spécialisées et indépendantes. La voie ventrale majoritairement composée de cellules parvocellulaires est impliquée dans les formes et les couleurs, analyse les scènes visibles sans prendre en compte les mouvements. L'aire V4 est l'aire principale de cette voie. Elle est surnommé la « voie du quoi » ou la « voie du voir pour reconnaître ». La voie ventrale permet un traitement visuel pour l'identification sur la base d'une information en HFS.

Une voie dorsale, quasi-exclusivement composée de cellules magnocellulaires qui intervient dans les déplacements visuels des objets, et des mouvements. Elle est nécessaire pour éviter les obstacles, se rendre compte des reliefs avoisinants, saisir les objets. Elle oriente les yeux en fonction de l'environnement. Elle est surnommée la « voie du où » ou la « voie du voir pour agir ». Elle permet un traitement visuel pour l'action basé sur une intégration rapide des informations en BFS.

1.2.6 Mouvement des yeux

Pour détecter la présence d'objets dans l'environnement, mais surtout pour les reconnaître et les identifier, il est nécessaire de les placer au centre du champ visuel et de les suivre quand ils sont en mouvements. Ces mécanismes sont assurés par le système oculomoteur. On distingue plusieurs types de mouvements oculaires.

La fixation La fixation ne s'agit pas à proprement dit de mouvement oculaire. Elle représente l'activité des yeux lorsqu'ils restent plus ou moins un certain temps positionnés sur la même localisation spatiale(le même point). Plus une localisation comporte d'information, plus longue est la fixation qui y est faite. Pendant une fixation, une cible

est projetée sur la fovéa pour être traitée dans la plus haute résolution spatiale [YARBUS, 1967].

Les saccades L'observation d'un objet quelconque revient à fixer le regard en plusieurs endroits bien précis, dans un ordre déterminant. Le regard se déplace entre les endroits qui portent le plus d'information. [YARBUS, 1967] a relevé que l'observation d'objets stationnaires, comme des images par exemple, se traduit par une suite de saccades et de fixations sur des points-clés de l'objet observé. L'œil transite d'un point de fixation à un autre en effectuant des saccades. Les saccades peuvent être exécutées volontairement ou inconsciemment et aussi par réflexe chaque fois que les yeux sont ouverts. Elles sont utilisées pour l'exploration spatiale. La figure 1.7 montre bien que des localisation comme les yeux, la bouche, les oreilles ou les contours sont les plus observés par les sujets.

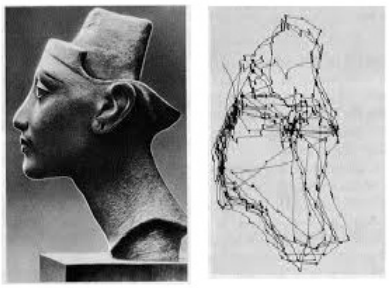


FIGURE 1.7 – Mouvements oculaires d'un sujet regardant une photographie d'un buste de la reine Néfertiti.Figure extraite de [YARBUS, 1967].

Les micro-saccades sont des mouvements qui se déroulent sans interruption. Sans ces mouvements, l'image observée ne seraient pas rafraîchie ou rechargée. Ces derniers permettent aux cellules photo-réceptrices de s'activer et de ne pas s'épuiser.

La poursuite oculaire La poursuite oculaire est un mouvement des yeux qui se produit lorsqu'ils continuent à fixer conjointement le même point alors que celui-ci se déplace sur la rétine. Le mouvement d'une cible sur la rétine peut être dû soit au déplacement réel de la cible, soit au déplacement de l'observateur. Ce type de mouvement est lent et ne peut être initié volontairement en l'absence de déplacement sur la rétine. Le but est de maintenir un point sur la fovéa alors que celui-ci se déplace. Puisque le mouvement des yeux est relativement lent, une grande résolution spatiale est conservée et il est toujours possible de tirer de l'information de la cible poursuivie.

1.3 Perception de la couleur

La perception humaine de la couleur dépend du signal couleur qui arrive au niveau du cortex cérébral, ce qui sous entend les aspects physiques et physiologiques du signal, ainsi que des aspects psychologiques, c'est à dire la connaissance à priori de l'environnement. Ainsi, le mécanisme de la perception des couleurs est un phénomène à la fois physique, physiologique et psychologique [].

1.3.1 La sensibilité à la couleur

Les cônes permettent de discriminer différentes longueurs d'onde. Avec les cellules P de la couche ganglionnaire, elles fournissent une vision précise des couleurs. Le système visuel humain est sensible aux couleurs rouges, vert et bleu. Ceux sont les trois couleurs auxquelles sont adaptés les différents types de cônes [VALOIS et VALOIS, 1975], mis en jeu dans la rétine qui génèrent trois pigments différents (voir la figure1.8).

— les cônes L sont sensibles aux ondes longues avec un pic de sensibilité vers 560 nm. Ils , sont dédiés à la vision du rouge.

— les cônes M sont sensibles aux ondes moyennes avec un pic de sensibilité vers 530 nm. Ils sont dédiés à la vision du vert.

— les cônes S sont sensibles aux ondes courtes avec un pic de sensibilité vers 420 nm. Ils sont dédiés à la vision du bleu.

Les bâtonnets ne permettent pas de discriminer différentes longueurs d'onde. Ils sont adaptés à toutes les longueurs d'onde mais, contrairement aux cônes, ils ne génèrent qu'un seul pigment, la rhodopsine. La vision des bâtonnets donne une perception en noir et blanc. Par conséquent, la limite de distinction des teintes décline la nuit ainsi qu'en périphérie [6].

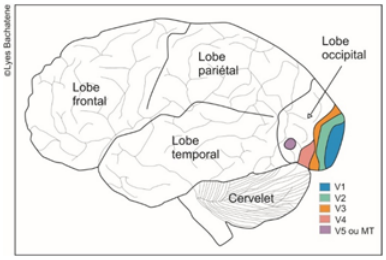


FIGURE 1.8 – Sensibilité des cellules photo-récepteurs de la rétine en fonction de la longueur d'onde [7]

1.3.2 Mécanismes de vison couleur

Mécanisme de trichromatie

Afin d'assurer la perception discriminative des couleurs, le système visuel humain s'appuie sur des propriétés structurelles et fonctionnelles bien définies. Au niveau des photo-récepteurs, il opère la première discrimination physique des différentes longueurs d'onde. En effet, les cônes se décomposent en trois groupes de photo-récepteurs dont les pigments propres présentent une probabilité définie d'absorber les différents photons du spectre de lumière, les cônes S, M et L étant sensibles aux longueurs d'onde courtes(420 nm bleu), moyennes (530 nm vert) et longues (560 nm rouge) [PIGNAT, 2005]. Ce codage trichromatique est maintenu au niveau des inter-neurones car ces derniers se dépolarisent préférentiellement pour une modalité de couleur et s'hyperpolarisent pour les autres [BOISSON, 2014].

Cependant, la stimulation des cônes par les photons de longueurs d'ondes différentes, mais relativement proches, n'assure pas une discrimination fine des couleurs.

6. http://serge.bertorello.free.fr/
7. Source :Serge Bertorello

Par exemple, une longueur d'onde intermédiaire autour de 540-550 nm aurait la même probabilité de stimuler un cône M qu'un cône L d'où la même probabilité de transmission d'une information codant pour la couleur verte ou jaune rouge.

Mécanisme d'opposition

Hering a relevé le problème de la couleur jaune parmi les couleurs fondamentales. Il a développé et mis en œuvre un mécanisme d'interaction par couples opposés de 4 couleurs : vert et rouge, bleu et jaune, et une opposition noir et blanc correspondant à la luminosité. Dans ce mécanisme, Hering fait l'hypothèse qu'il existe trois types de récepteurs : vert-rouge, bleu-jaune et noir-blanc. (voir la figure 1.9). Ils représentent des couleurs antagonistes dont le mélange crée du gris [8] et que la trichromatie n'explique pas.

— Le mécanisme antagoniste vert-rouge fournit une réponse antagoniste opposant les cônes L et les cônes M.

— Le mécanisme antagoniste bleu-jaune fournit une réponse antagoniste opposant les cônes S et la réponse additive issue des cônes M et L.

— Le mécanisme achromatique noir-blanc (ou canal de la luminosité) fournit une réponse additive des signaux issus des trois cônes S, M et L.

De ce fait, différent traitements intra-rétiniens permettent d'augmenter la discrimination des couleurs à travers trois mécanismes antagonistes de transformation du signal dénommés respectivement mécanisme vert-rouge, mécanisme bleu-jaune et mécanisme achromatique(noir-blanc). [9]

1.3.3 Adaptation visuelle en lumière

L'adaptation visuelle est le processus par lequel le système visuel adapte la perception aux propriétés de l'environnement lumineux. Elle permet la vision dans des intensités lumineuses très variées, et la reconnaissance de la couleur des objets vus dans des lumières de répartitions spectrales différentes. On peut distinguer deux processus semblables. L'adaptation au niveau lumineux, pour ce qui concerne les intensités lumineuses, et l'adaptation chromatique pour ce qui concerne la répartition spectrale de l'énergie lumineuse. [FLORU, 1996]

Adaptation en luminance

L'adaptation à la lumière et à l'obscurité décrivent les capacités du système visuel à s'adapter à des changements de luminance qui reposent sur le mécanisme d'ouverture et fermeture de la pupille. Dans un environnement très clair, la pupille se ferme pour réguler le flux de lumière reçu. Dans l'obscurité, la pupille s'ouvre plus. Cette adaptation n'est pas un phénomène instantané. Elle nécessite plusieurs secondes à plusieurs minutes de transitions entre deux états radicalement différents. La transition lumineux-sombre est plus lente que la transition sombre-lumineux.

8. http://www.toutes-les-couleurs.com/couleurs-primaires.php.
9. http://www.profil-couleur.com/lc/010-couleur-opposees.php

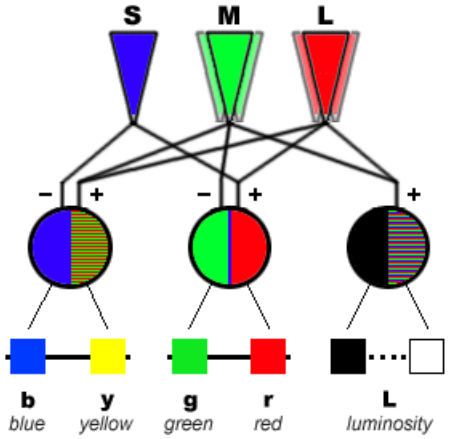


FIGURE 1.9 – Codage antagoniste des signaux dans la rétine [10]

Adaptation chromatique

L'adaptation chromatique(constante couleur) décrit la faculté du système visuel à s'adapter à des changements de couleur d'illumination. L'adaptation chromatique est le phénomène qui permet de réajuster la couleur de l'illumination globale, afin de garantir la constance des couleurs, la constante chromatique.

1.4 Perception du contraste

La vision des contrastes sous-entend les capacités du système visuel à percevoir et reconnaître des motifs lumineux. Ces capacités différent selon le domaine de vision(diurne, nocturne).[CAMPBELL et collab., 1966],[CAMPBELL et ROBSON, 1968] ont proposé que l'atome de la perception visuelle sont des grilles sinusoïdales(sin-wave gratings, en anglais). L'idée sous-jacente à cette théorie est que qu'elle que soit une image, elle peut être décomposée sans perte d'information en une somme unique de grilles sinusoïdales. Chacune est caractérisée uniquement par quatre paramètres : une fréquence spatiale, une orientation, une phase et un contraste. La figure 1.10 illustre une grille sinusoïdale ayant une fréquence spatiale, une phase, un contraste et une orientation spécifique. La même grille est représenté avec différentes valeurs de fréquence, phase, contraste et orientation.

10. Source :https://sites.google.com/site/enactionvarela/2-enaction/l-exemple-de-la-couleur/la-manifestation-de-la-couleur-1/modelisation-theorie-des-processus-opposants

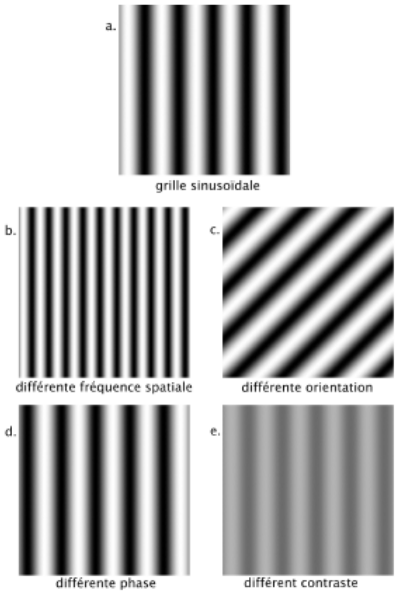


FIGURE 1.10 – (a)Une grille sinusoïdale ayant une fréquence spatiale, une phase, un contraste et une orientation spécifiques.(b)la même grille avec une fréquence différente.(c) la même grille avec une phase différente.(d)la même grille avec un contraste différent. (e) la même grille avec une orientation différente [FISET et GOSSELIN, 2009].

1.4.1 Réseaux sinusoïdal et fréquence spatiale

Selon [CAMPBELL et collab., 1966],[CAMPBELL et ROBSON, 1968], le système visuel humain décompose une image en une somme de réseaux sinusoïdaux et la résume en un spectre de fréquences spatiales. Selon cette vision, chacun de ces réseaux sinusoïdaux peut alors être caractérisé par une alternance périodique, dans une orientation donnée, de bandes sombres et claires nommées cycles. La fréquence spatiale représente alors le nombre de cycles par degré d'angle visuel et reflète ainsi un indice de la taille de l'image sur la rétine. La mesure de la fréquence spatiale dépendant alors de la distance entre l'œil de l'observateur et le stimulus visuel. Plus l'observateur s'éloigne de l'objet, ou inversement plus l'objet s'éloigne de l'observateur, plus le nombre de cycles par degré d'angle visuel augmente et plus la fréquence spatiale augmente également. Pour percevoir une fréquence spatiale de 30 cycles par degré(cpd), il faut un pouvoir de résolution (angle de résolution minimal) d'une minute d'arc, soit une acuité visuelle équivalente de 10/10e.

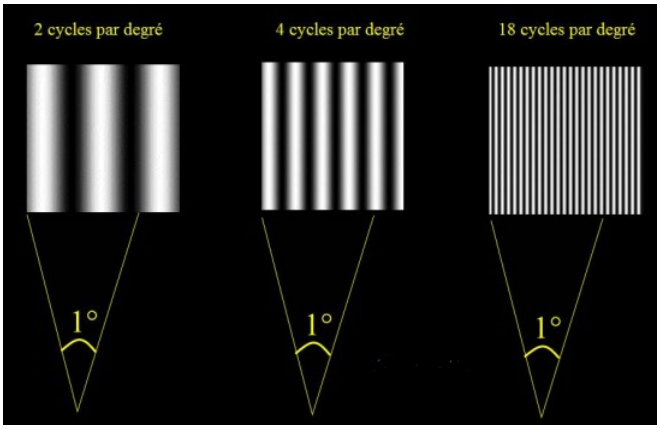


FIGURE 1.11 – Exemples de réseaux sinusoïdaux à différentes fréquences spatiales. [11]

Les réseaux sinusoïdaux ont été utilisé dans plusieurs travaux en psychophysique chez les humains. Dans leur expérimentations, les chercheurs ont mesuré les temps de réaction pour détecter des réseaux sinusoïdaux à différentes fréquences spatiales. Les observations ont démontré que les réseaux en basse fréquence spatiale(BFS = 0,5 cpd) étaient détectés plus rapidement que les réseaux en haute fréquence spatiale(HFS= 11 cpd).

De plus, la sensibilité du système visuel humain à l'orientation de grilles sinusoïdales a elle aussi été étudié, et il a été observé que notre système visuel est plus sensible aux grilles verticales et horizontales qu'aux grilles obliques.

1.4.2 Réseaux sinusoïdal et contraste

En plus de la fréquence spatiale, un réseau sinusoïdal se caractérise également par son amplitude(ou son contraste). Le contraste globale d'un réseau sinusoïdal correspond à la plus petite différence d'intensité lumineuse perçue entre les bandes sombres et les bandes claires au sein du réseau [PROST, 2016]. Cette capacité de discrimination de notre système visuel détermine l'acuité visuelle. Classiquement, le contraste global C_{Global} d'un réseau sinusoïdal se calcule à partir des luminances. Il se définit par la loi de Mickelson selon l'équation 1.3 qui représente la somme de la luminance de l'objet le plus lumineux moins celle du moins lumineux sur leur somme.

$$C_{Global} = \frac{L_{max} - L_{min}}{L_{max} + L_{min}} \tag{1.3}$$

Ce rapport donne un chiffre compris entre 0 et 1, souvent exprimé en pourcentage et utilisé pour mesurer la fonction de sensibilité aux contrastes. C = 90% étant un contraste très élevé et C= 10% un contraste très faible devenant à peine visible.

Par ailleurs, le contraste local, utilisé en photographie, correspond à la perception des contours d'une image. Il se définit par la loi de Weber-Fechner selon l'équation 1.4

$$C_{Local} = \frac{L_{zone} - L_{fond}}{L_{fond}} \tag{1.4}$$

11. https ://www.gatinel.com/recherche-formation/acuite-visuelle-definition/frequence-spatiale/

où L_{zone} est la luminance de l'objet dans une l'image et L_{fond} la luminance du fond.

1.4.3 Sensibilité au contraste

Il existe deux types de sensibilité au contraste : la sensibilité au contraste spatial et la sensibilité au contraste temporel. La sensibilité au contraste spatial représente la capacité du système visuel à détecter des différences de luminance sur des éléments de dimensions variées, statiques. Elle dessine l'enveloppe du domaine visible et les possibilités de discrimination du contraste. La sensibilité au contraste temporel permet de distinguer des différences de luminance sur des images mobiles, l'œil reste fixe et l'image en mouvement se déplace sur la rétine.

La sensibilité au contraste spatiale est maximale pour les fréquences spatiales intermédiaires. L'œil humain arrive donc particulièrement bien à discriminer les bandes sombres des bandes claires d'un réseau sinusoïdal avec une faible valeur de contraste pour cette gamme de fréquence. A l'inverse pour une même valeur de contraste, dans des gammes de fréquences plus extrêmes, le système visuel devient moins sensible à cette alternance de bandes sombres et claires et il tend vers la perception d'un tout unifié.

Par ailleurs, en raison de l'hétérogénéité histologique de la rétine, la sensibilité aux contrastes tend également à diminuer avec l'excentricité rétinienne. Ainsi, la plus haute fréquence spatiale perceptible, à un contraste maximum, varie d'environ 60 cycles par degré (cpd) en présentation fovéale à seulement 2 cpd en vision périphérique de 30° d'excentricité rétinienne.

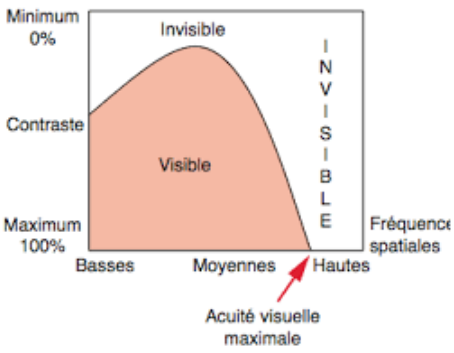


FIGURE 1.12 – Sensibilité au contraste, issu du cours du Dr BONNIN.

Notre système visuel décompose un stimulus visuel en différentes fréquences spatiales, qui correspondent à une variation plus ou moins rapide du contraste. Les basses fréquences spatiales qui correspondent aux variations lentes du contraste, permettent de percevoir l'aspect global du stimulus. Les hautes fréquences spatiales qui correspondent aux variations rapides du contraste, permettent de percevoir les détails du stimulus visuel ainsi que la bordure des objets ou des parties d'objets avec une grande précision [12].

12. https://theses.univ-lyon2.fr/documents/getpart.php?id=lyon2.2001.baudouin_jy&part=37504

1.4.4 Adaptation visuelle en contraste

D'une manière similaire à l'adaptation en luminance, l'œil humain peut aussi s'adapter à un contraste. Deux motifs dont la fréquence spatiale est identique, peuvent ne pas paraitre avoir la même fréquence spatiale, l'explication de ce phénomene revient au fait que l'environnement semble avoir une forte influence sur la vision des objets, d'où la définition du contraste de luminosité et de couleur [FARRUGIA, 2012].

Contraste de luminosité Le contraste de luminosité se définit par le fait que la même couleur est perçue plus foncée sur un fond clair que sur un fond sombre. Dans la figure 1.13, le carré central semble d'un gris plus foncé à droite qu'à gauche. The inner squares have the same intensity, but they appear proressively darker as the background becomes lighter.

FIGURE 1.13 – Illustration du contraste de luminosité.

Contraste couleur Le contraste couleur(ou simultané) est le phénomène d'apparence couleur qui correspond au changement de couleur d'un stimulus lorsque la couleur ou la structure spatiale de l'arrière-plan change. La couleur du stimulus tend à suivre la théorie des couleurs opposées : un arrière-plan rouge induit un changement dans le vert, un arrière-plan vert induit un changement dans le rouge, et de même pour le jaune et le bleu [TABART, 2010]. Dans la figure 1.14, les cercles du haut apparaissent différents des cercles du bas alors que dans la réalité ils sont identiques.

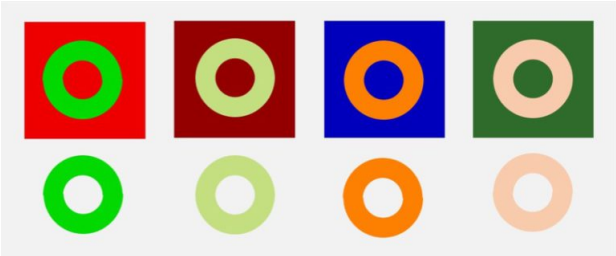


FIGURE 1.14 – Illustration du contraste simultané des couleurs [13]

1.5 Attention et saillance visuelle

A partir de l'aire V1, les informations issues des cellules P sont envoyées vers la voie ventrale et les informations issues des cellules M sont envoyées vers la voie dorsale. La

13. https://scilogs.fr/questions-de-couleurs/pourquoi-la-couleur-nous-trompe-t-elle-continuellement-

voie ventrale, appelée aussi la voie du quoi, reçoit principalement des stimuli chromatiques qui sont impliqués dans la forme, la texture, la couleur, les détails et la taille afin d'identifier et reconnaitre les objets. La voie dorsale, appelée aussi la voie du où, reçoit principalement des stimuli de luminance qui sont impliqués dans le déplacement visuel des objets, le mouvement, la transformation spatiale, les relations spatiales et l'attention pour effectuer la vision spatiale [PRAHARA et collab., 2020].

Par ailleurs, on estime que les données visuelles qui circulent dans nos yeux sont d'environ 10^8 à 10^9 bits par seconde [KOCH et collab., 2006]. La gestion de ce flux de données en temps réel est une tache incroyablement lourde pour le système visuel humain(SVH). Seule une partie des données est sélectionnée et traitée plus en détail par le système visuel humain par un mécanisme sélectif appelé attention visuelle. La première définition de l'attention visuelle fut donnée par le père de la psychologie, William James. Sa définition, dans sa version anglaise originale, est la suivante : « Everyone knows what attention is. It is the taking possession of the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration of consciousness are of its essence … ». La traduction en langue française donne : « L'attention est la prise de possession par l'esprit, sous une forme claire et vive, d'un objet ou d'une suite de pensées parmi plusieurs qui semblent possibles …. Elle implique le retrait de certains objets afin de traiter plus efficacement les autres ».

L'attention visuelle se définit comme étant la capacité du cerveau à sélectionner l'information visuelle pertinente en rejetant ce qui ne l'est pas dans un contexte particulier. Un des rôles de ce mécanisme est d'accélérer le processus de vision, en réduisant sensiblement la quantité d'informations visuelles qui sera traitée par les tâches de plus haut niveau. La réduction se fait par la suppression d'informations redondantes et par la sélection rapide des informations les plus pertinentes en se focalisant sur les éléments les plus saillants. Un élément visuellement saillant, est un élément qui ressort prioritairement lors de la perception visuelle d'une scène, au point de prendre une importance cognitive particulière.

En plus, des études réalisées en neuro-physiologie sur l'attention visuelle ont démontré l'existence d'une carte de saillance, qui consisterait en une représentation topographique des stimuli pertinents. Cette carte serait située au niveau de plusieurs zones cérébrales comme le cortex pariétal postérieur(PP),le colliculus supérieur(SC), la zone intra-pariétale latérale(LIP), le champ oculaire frontal(FEF), le cortex visuel primaire(V1), l'aire V4 [LI, 2002]. Cependant, les corrélations entre ces différentes zones reste encore mal déterminées jusqu'à présent [FRINTROP, 2006],[PRAHARA et collab., 2020].

1.5.1 Recherche visuelle

Un outil important dans le domaine de recherche sur l'attention visuelle est la recherche visuelle. La recherche visuelle est une tâche qui consiste à détecter une cible visuelle spécifique à partir de divers autres stimuli, les distracteurs, ce qui est largement accepté comme clarifiant la perception visuelle humaine. La question générale de la recherche visuelle est la suivante : étant donné une cible et une image de test, y a-t-il une instance de la cible dans l'image de test? De nombreuses études, y compris les travaux pionniers entrepris par [NEISSER et collab., 1996], ont examiné les mécanismes de perception visuelle via la recherche visuelle avec diverses relations de cibles et de dis-

tracteurs. Les travaux de [TREISMAN et GELADE, 1980] reposent sur des expériences de recherche visuelle, mesurant le temps de réaction nécessaire pour distinguer un objet cible enfoui parmi d'autres objets, distracteurs. Ces objets peuvent être caractérisés par une seule dimension visuelle comme la couleur, l'orientation, la forme ou plusieurs à la fois. La recherche visuelle peut être classée dans les deux types suivants en fonction de la relation entre la cible et les distracteurs.

La recherche de caractéristiques (Effet pop-out)

La recherche de caractéristiques ou l'effet du pop-out utilise une cible qui peut être distinguée des distracteurs par une caractéristique unique telle que la couleur, l'orientation, la taille. Le phénomène du pop-out visuel se traduit par le fait qu'un objet saute aux yeux parce qu'il a une apparence qui diffère de tous les autres objets de son environnement.

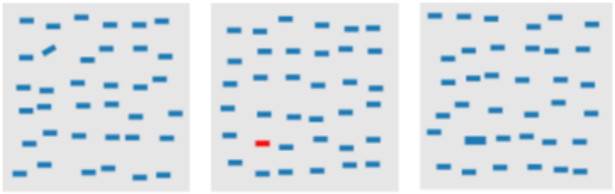


FIGURE 1.15 – Exemples de quelques caractéristiques visuelles, pré-attentives, qui sautent aux yeux.

Il a été démontré que le temps nécessaire pour trouver la cible ne dépend pas du nombre de distracteurs. Ainsi pour trouver un trait rouge parmi des traits bleus, un sujet est toujours aussi rapide qu'il y ait 5, 10 ou 100 distracteurs. L'effet de pop-out indique un traitement en parallèle sur l'ensemble de la scène visuelle. [TREISMAN et GELADE, 1980] a rapporté un certain nombre d'expériences qui identifient les caractéristiques visuelles qui saute aux yeux, appelées aussi caractéristiques pré-attentives, dans le système visuel humain. Les caractéristiques visuelles les plus efficaces sont : la couleur, l'orientation, la taille et le mouvement.

La recherche de conjonction

La recherche de conjonction utilise une cible avec une cible impliquant plusieurs caractéristiques différentes des distracteurs. Dans cette tâche de recherche visuelle, un sujet humain ne peut pas utiliser un seul trait pour trouver la cible. Par exemple, s'il cherche les objets rouges, les distracteurs carrés rouges vont rendre plus difficile sa recherche, de même avec les cercles bleus lorsqu'il cherche les objets circulaires. Il est donc nécessaire d'engager l'attention car les traitements pré-attentionnels ne sont pas suffisants pour détecter directement la cible, ce qui impose de faire une recherche sérielle en portant attention sur les différents objets. Le temps nécessaire pour trouver la cible est donc nettement plus long et dépend directement du nombre de distracteurs.

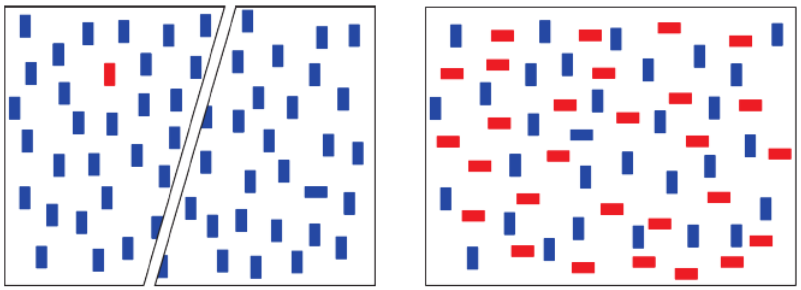


FIGURE 1.16 – Exemples de recherche de caractéristiques et de conjonction [BAJLEKOV, 2012].

1.5.2 Théories de l'attention visuelle

Théorie d'intégration des traits

La théorie d'intégration des traits(Feature Integration Theory(FIT) en anglais) de [TREISMAN et GELADE, 1980] est l'une des théories de l'attention visuelle les plus connues et les plus citées dans la littérature. Cette théorie divise le processus d'attention en deux étapes : un processus pré-attentif et un processus de focalisation. Selon [TREISMAN et GELADE, 1980], dans la théorie d'intégration des traits, différentes caractéristiques sont enregistrées tôt, automatiquement et en parallèle à travers le champ visuel, tandis que les objets sont identifiés séparément et seulement à un stade ultérieur qui nécessite l'attention". Cette théorie est basée sur le fait que le processus d'attention dans le cerveau fournit plusieurs cartes de caractéristiques en fonction des attributs physiques du stimulus. Ces cartes de caractéristiques sont ensuite combinées en une carte principale qui permet de distinguer les régions importantes de la scène visuelle. La figure 1.17 montre un schéma de FIT.

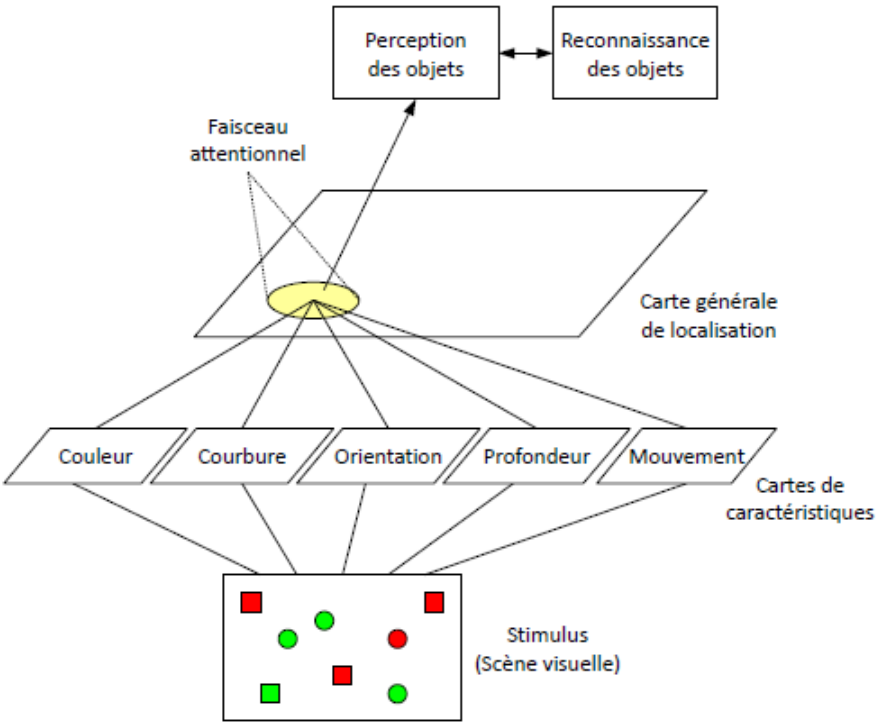


FIGURE 1.17 – La théorie d'intégration des attributs [TREISMAN et GELADE, 1980]

Modèle de la recherche guidée

Outre la théorie FIT, le modèle de recherche guidée proposé par Wolf [WOLFE et collab., 1989], est l'un des modèles psychologiques d'attention les plus connus. Ce modèle a également évolué au cours des années et plusieurs versions de sa simulation informatique sont disponibles [WOLFE, 1994], [WOLFE et GANCARZ, 1997]. Le modèle, dans plusieurs aspects, est similaire au FIT, mais plus détaillé pour être simulé par ordinateur. La figure 1.18 représente un schéma de ce modèle. Le principal aspect qui différencie ce modèle de FIT est qu'au lieu des types de caractéristique (rouge, vert, etc.), la dimension de caractéristique (couleur, orientation, etc.) forme chaque carte de caractéristique. De plus, le modèle associe une carte descendante(top-down) à chaque carte de caractéristiques.

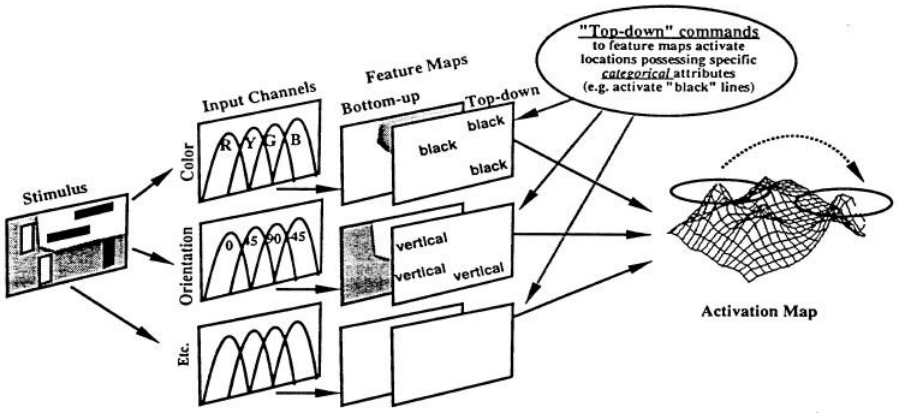


FIGURE 1.18 – Schéma du modèle de la recherche guidée.

1.5.3 Mécanismes d'attention visuelle

Dans le mécanisme de l'attention visuelle, la sélection est une notion très importante qui repose sur plusieurs facteurs(attention ascendante/descendante, attention explicite/implicite, et attention basée sur l'espace/l'objet)[SOTO et BLANCO, 2004].

Attention sélective ascendante/descendante

L'attention ascendante. L'attention ascendante (appelée aussi exogène, bottom-up) représente l'ensemble des processus automatiques, déclenchés par les stimuli externes captés par notre système visuel. Elle est considéré comme un mécanisme relativement éphémère piloté par les données de notre champ visuel et faisant référence à l'attention involontaire. Cette attention est déployée par exemple quand une personne tourne la tête lorsqu'elle perçoit un mouvement soudain à la périphérie de son champ visuel.

L'attention descendante. L'attention descendante (aussi appelée endogène,top-down), est déployée volontairement, activement et consciemment à un endroit de l'espace pour y attendre un événement spécifique. Elle dépend de nos attentes et objectifs.

Attention sélective explicite/implicite

L'attention explicite. Lorsque qu'une personne porte son attention sur un objet, ses yeux se déplacent afin de fixer cet objet, c'est ce qu'on appelle l'attention explicite(appelée aussi l'attention ouverte ou déclarée). En d'autres termes, l'attention visuelle explicite correspond à un déplacement de la fovéa sur le stimulus par le biais d'un mouvement oculaire. L'un des pionniers à avoir étudier l'attention visuelle explicite et sa relation avec le mouvement des yeux est [YARBUS, 1967]. Il a montré que la trajectoire du regard lors de l'exploration d'une scène dépendait de la tâche demandée, montrant ainsi que l'allocation de l'attention explicite n'est pas un processus uniquement bottom-up, mais aussi top-down.

L'attention implicite. L'attention implicite(appelée l'attention cachée ou couverte) correspond à la faculté à focaliser l'attention visuelle sur une cible(objet ou position)

sans déplacer les yeux. Cela signifie que l'attention implicite n'est pas obligatoirement liée aux mouvements oculaires et qu'elle peut se déplacer indépendamment. Ce type d'attention peut être défini comme la mise en exergue d'une région distincte du champ visuel sans le déplacement des yeux vers celle-ci. Puisqu'elle n'implique aucun mouvement oculaire, son observation est particulièrement difficile. L'attention implicite précède l'attention explicite. Elle permet de guider les mouvements oculaires à des endroits où l'information semble pertinente. Cependant, contrairement à l'attention sélective explicite, les seuls moyens de savoir que l'attention est bien dirigée sur une cible spécifique sont les changements comportementaux comme la réaction de l'individu, ou bien l'observation du fonctionnement cérébral [SHAHRBABAKI, 2015].

Par ailleurs, des études semblent montrer une relation entre l'attention explicite et implicite. D'après les travaux de [KLEIN, 1980] [SILVA, 2010], l'attention explicite et implicite sont indépendantes, elles se déploient simultanément car elles sont déclenchées par les informations visuelles. D'après les travaux de [RIZZOLATTI et collab., 1987] [SILVA, 2010], l'attention implicite est considérée comme un mécanisme préparatoire aux saccades oculaires. Elle devient ainsi un produit dérivé du système moteur.

Attention basée sur espace/objet

Attention basée sur l'espace. La majorité des études réalisées en psychophysique et neurobiologie portent sur l'attention basée sur l'espace, également appelée attention basée sur la localisation. L'attention visuelle spatiale est déployée lorsque l'observateur dirige son attention vers un endroit précis de son champ visuel. Si un stimulus apparait à l'endroit où son attention visuelle est dirigée, alors l'information de ce stimulus sera transmis plus rapidement que si le stimulus était apparu dans une autre localisation.

Attention basée objet. Contrairement à l'attention basée sur l'espace, un autre point de vue suggère que l'attention est dirigée vers des objets entiers ou des groupes perceptifs de la scène visuelle, qui seraient le résultat d'un regroupement(grouping) effectué pré-attentivement. La théorie de l'attention orienté objet permettrait de faire le lien entre attention et théorie de la forme. Cette dernière est élaborée par les psychologues Gestaltistes qui présuppose l'existence de différentes lois, la continuité, la proximité, la similitude, utilisées par notre cerveau pour structurer la scène visuelle [SILVA, 2010].

1.6 Tâches de la perception visuelle

Dans leur travaux, [BUFFAT, 2007] ont proposé une hiérarchie des tâches de la perception visuelle. Ces tâches se différencient en termes de processus impliqués et de mécanismes mis en jeu par la perception visuelle. [CHARBONNEAU, 2010],

La détection. Dans la tache de détection, une personne est censé répondre à la question suivante : l'objet en question est-il présent ou absent dans l'image ?. Cette tâche nécessite le moins d'information spécifique à l'objet d'où sa considération comme étant la tache de plus bas niveau dans la hiérarchie des tâches de la perception visuelle. La tâche de détection correspond au premier niveau de traitement dans le modèle de [TREISMAN et GELADE, 1980].

La catégorisation. Dans cette tâche, une personne est censé classer un objet cible dans une catégorie, par une procédure de choix forcé. On peut ainsi demander de classer les images naturelles selon le critère suivant : artificiel ou naturel. La tâche de catégorisation permet d'évaluer la capacité d'organisation des connaissances car, dans l'environnement dans lequel nous vivons, les objets et évènements prennent du sens par leur organisation. Elle constitue une des tâches essentielle de l'intelligence humaine. Cependant, [GRILL-SPECTOR et KANWISHER, 2005] ont réalisé plusieurs expériences mettant en évidence que le temps de traitement entre la détection et la catégorisation sont identiques, ce qui remet en question cette hiérarchie.

La reconnaissance. Dans cette tâche, une personne est censé identifier si l'objet cible, a déjà été rencontrée ou non. Ce sont des tâches classiquement utilisées pour les études sur la mémoire où l'on demande à une personne de retrouver parmi une liste d'objets ceux qui sont nouveaux et/ou connus, et donc qui font parties des éléments mémorisés. Elle appartient au deuxième niveau de traitement dans le modèle de [TREISMAN et GELADE, 1980]. Selon[BIEDERMAN, 1987], la reconnaissance d'un objet ne repose pas sur la reconnaissance de tous les éléments qui le composent. Certains éléments sont plus ou moins stratégiques dans la reconnaisse des objets. Ainsi, la perception visuelle d'un objet se décompose en un certain nombre d'éléments primaires : cylindres, blocs, cônes, ... , etc. Ces éléments primaires sont appelés des géons. Une quarantaine de géons sont nécessaire pour reproduire la totalité des formes perçues.

L'identification. Dans cette tâche, une personne doit préciser l'identité de l'objet cible, le plus souvent par dénomination. Comme la tâche de reconnaissance, l'identification appartient au deuxième niveau du modèle de Treisman [TREISMAN et GELADE, 1980] et nécessité la mise en jeu de l'attention.

1.7 Conclusion

Dans ce chapitre, nous avons présenté dans un premier temps, l'anatomie et la physiologie du système visuel humain. Nous avons décrit les traitements se produisant sur l'information visuelle au niveau de la rétine et de l'aire V1 du cortex visuel. L'information visuelle est captée par les photo-récepteurs de la rétine puis les différentes couches neuronales rétiniennes pré-traitent l'information visuelle et la séparent en canaux qui véhiculent parallèlement des informations de nature différente. Deux voies importantes se distinguent : une voie dédiée à l'analyse spatiale décrite par la voie parvocellulaire et une autre voie dédiée à l'analyse du mouvement et des transitions temporelles décrite par la voie magnocellulaire. Ces deux voies sont transmises à l'aire V1 du cortex visuel dont les cellules corticales décomposent les informations issues des deux types de cellules rétiniennes, magnocellulaire et parvocellulaire, en différentes orientations et différentes fréquences spatiales. L'information continue d'être véhiculée vers des aires corticales de plus haut niveau pour aboutir à la reconnaissance d'objets et à la détection du mouvement. Dans un deuxième temps, nous avons évoqué l'attention visuelle sélective ainsi que ses fondements théoriques. Le comportement attentionnel est guidé par deux types de mécanismes, un mécanisme ascendant qui est axé sur les stimuli et un mécanisme descendent qui est axé sur les attentes. Dans le domaine de la vision par ordinateur, l'attention visuelle porte principalement sur le mécanisme attentionnel ascendant en raison de sa simplicité. Dans le chapitre suivant, nous présenterons un état

de l'art sur les différents modèles de d'attention visuelle proposés dans la littérature.

Références

BAJLEKOV, G. 2012, «Theories of visual search and their applicability to haptic search. thèse de doctorat.», . ix, 29

BIEDERMAN, I. 1987, «Recognition by components : a theory of human image understanding.», *Psychological review*, vol. 94, n° 2, p. 115. 33

BOISSON, L. A. 2014, *Etude et optimisation d'un système d'éclairage efficace énergétiquement et adapté aux besoins de ses utilisateurs (santé, sécurité et qualité de vie). Thèse de doctorat en Physique de la lumière et perception visuelle*, thèse de doctorat, Université de Toulouse, Université Toulouse III-Paul Sabatier. 20

BUFFAT, S. 2007, *Reconnaissance d'objets dans des environnements bruités et fusion d'informations visuelles naturelles. Thèse de doctorat en Sciences Cognitives.*, thèse de doctorat, Université de Paris VI. 32

CAMPBELL, F., J. J. KULIKOWSKI et J. LEVINSON. 1966, «The effect of orientation on the visual resolution of gratings», *The Journal of Physiology*, vol. 187. 22, 23

CAMPBELL, F. et J. ROBSON. 1968, «Application of fourier analysis to the visibility of gratings», *The Journal of Physiology*, vol. 197. 22, 23

CHARBONNEAU, M. 2010, *Approche méthodologique et comparative des critères de qualité d'image, de perception et d'exploitabilité opérationnelle Application aux systèmes d'aide à la vision nocturne en aéronautique. Thèse de doctorat en Sciences Cognitives*, thèse de doctorat, Université de Bordeaux 2. 32

DOUTSI, E., L. FILLATRE, M. ANTONINI et J. GAULMIN. 2018, «Retina-inspired filter», *IEEE Transactions on Image Processing*, vol. 27, p. 3484–3499. 14

FARRUGIA, J. P. 2012, *Modèles de vision et synthèse d'images. Thèse de doctorat en Informatique.*, thèse de doctorat, Ecole Nationale Supérieure des Mines de Saint-Etienne et Université Jean Monnet de Saint-Etienne. 26

FERWERDA, J. 1998, «Fundamentals of spatial vision», *Applications of visual perception in computer graphics*, vol. 140. 11, 12

FISET, D. et F. GOSSELIN. 2009, «L'information visuelle efficace pour la reconnaissance des visages», *Traitement et reconnaissance des visages : du percept à la personne*, vol. 145, n° 164, p. 125. ix, 23

FLORU, R. 1996, «Eclairage et vision. rapport de recherche.», cahier de recherche, Institut National de Recherche et de Sécurité (INRS). 21

FRINTROP, S. 2006, *2 Background on Visual Attention*, chap. VOCUS : A Visual Attention System for Object Detection and Goal-Directed Search, Springer Berlin Heidelberg, p. 7–31. 27

GRILL-SPECTOR, K. et N. KANWISHER. 2005, «Visual recognition», *Psychological Science*, vol. 16, p. 152–160. 33

KLEIN, R. 1980, «Does oculomotor readiness mediate cognitive control of visual attention ?», *Academic Press*, p. 259–276. 32

KOCH, K., J. MCLEAN, R. SEGEV, M. FREED, M. BERRY, V. BALASUBRAMANIAN et P. STERLING. 2006, «How much the eye tells the brain», *Current Biology*, vol. 16, p. 1428–1434. 27

KUFFLER, S. W. 1952, «Neurons in the retina; organization, inhibition and excitation problems.», *Cold Spring Harbor symposia on quantitative biology*, vol. 17, p. 281–292. ix, 14, 15

LI, Z. 2002, «A saliency map in primary visual cortex», *Trends in Cognitive Sciences*, vol. 6, p. 9–16. 27

PIGNAT, J. M. 2005, *Etude de la perception visuelle du mouvement et de la couleur par IRMf. Thèse de doctorat*, thèse de doctorat, University of Geneva. 20

PRAHARA, A., M. MURINTO et D. P. ISMI. 2020, «Bottom-up visual attention model for still image : a preliminary study», *International Journal of Advances in Intelligent Informatics*, vol. 6, p. 82–96. 27

PRITCHARD, T. C. et K. D. ALLOWAY. 2002, *Neurosciences médicales : Les bases neuroanatomiques et neurophysiologiques*, chap. Chapitre 10, Blackwell Science Limited, Oxford, p. 337–373. 18

PROST, M. 2016, *Evolution de la vision des contrastes et chirurgie réfractive. Quels sont les effets de la chirurgie réfractive sur notre vision et plus précisément sur la vision des contrastes ?Mémoire de fin d'étude en capacité d'orthoptiste*, thèse de doctorat, Université de Clermont-Ferrand I. 24

RIZZOLATTI, G., L. RIGGIO, I. DASCOLA et C. U. BIEDERMAN. 1987, «Reorienting attention across the horizontal and vertical meridians : evidence in favor of a premotor theory of attention», *Neuropsychologia*, vol. 25, n° 1A, p. 31–40. 32

SHAHRBABAKI, S. T. 2015, *Contribution of colour in guiding visual attention and in a computational model of visual saliency. These de doctorat en Signal and Image processing*, thèse de doctorat, Université Grenoble Alpes. 32

SILVA, M. P. D. 2010, *Modèle computationnel d'attention pour la vision adaptative*, thèse de doctorat, Université de La Rochelle. 32

SOTO, D. et M. BLANCO. 2004, «Spatial attention and object-based attention : a comparison within a single task», *Vision Research*, vol. 44, p. 69–81. 31

TABART, G. 2010, *Méthodes et outils pour l'aide à la conception et la vérification du rendu graphique des systèmes interactifs. Thèse de doctorat en Informatique.*, thèse de doctorat, Université Toulouse 3 Paul Sabatier. 26

TREISMAN, A. et G. GELADE. 1980, «A feature-integration theory of attention», *Cognitive Psychology*, vol. 12, n° 1, p. 97–136. ix, 28, 29, 30, 32, 33

VALOIS, R. D. et K. D. VALOIS. 1975, *Handbook of Perception, Volume V : Seeing*, chap. Neural coding of color, In Carterette E.C. and Friedman M.P (Eds), p. 117–168. 17, 20

WOLFE, J. 1994, «Guided search 2.0 a revised model of visual search», *Psychonomic Bulletin & Review*, vol. 1, p. 202–238. 30

WOLFE, J., K. CAVE et S. L. FRANZEL. 1989, «Guided search : an alternative to the feature integration model for visual search.», *Journal of experimental psychology. Human perception and performance*, vol. 15, n° 3, p. 419–433. 30

WOLFE, J. M. et G. GANCARZ. 1997, «Guided search 3.0», dans *Basic and clinical applications of vision science*, Springer, p. 189–192. 30

YARBUS, A. L. 1967, *Eye Movements and Vision*, Plenum. New York. ix, 19, 31

Chapitre 2

Modèles de saillance : Un état de l'art

« *Je ne perds jamais. Soit je gagne, soit j'apprends.* »

Nelson Mandela

Sommaire

2.1 Introduction

Les théories de l'attention visuelle présentées dans le chapitre précédent ont eu une grande influence sur la modélisation informatique(computationnelle)de l'attention visuelle. En effet, au cours des dernières années, un intérêt croissant a été porté sur les systèmes informatiques qui modélisent l'attention visuelle humaine. Par exemple, en vision par ordinateur, ces systèmes sont couramment utilisés comme des mécanismes de sélection des objets les plus relevant dans une scène visuelle, dont dépendent d'autres tâches de plus haut niveau comme la reconnaissance des objets et la compréhension des scènes visuelles. En correspondance avec les mécanismes d'attention ascendants et descendants évoqués au chapitre 1, les modèles d'attention visuelle se basent soit sur des facteurs ascendants de bas niveau. Ces modèles sont connus sous le nom de modèles de saillance. Soit, ils se basent sur des facteurs descendants de haut niveau. Ces facteurs de nature cognitive peuvent être des buts, des connaissances ou des attentes. Toutefois, en raison de la complexité à inclure les facteurs descendants dans les modèles d'attention, la majorité des chercheurs se sont intéressés à proposer des modèles de saillance qui se basent sur des facteurs ascendants seulement. Par ailleurs, plusieurs taxonomies ont été proposé dans la littérature catégorisant ces modèles de saillance selon différentes catégories. Nous présentons d'abord ces différentes taxonomies. Ensuite, nous présentons les différents modèles de saillance visuelle selon la catégorisation de [Borji et collab., 2015]. Nous passons en revue les principales bases de données utilisées ainsi que les principales métriques utilisées pour l'évaluation des cartes de saillance.

2.2 Taxonomies des modèles de saillance

Taxonomie de Achanta(2008)

[Achanta, 2011] ont proposé une taxonomie des modèles de saillance en 3 familles, composée des modèles de saillance biologiques, purement computationnels et hybrides. Les modèles de saillance biologique [Itti et collab., 1998], [Frintrop et collab., 2007] tentent d'imiter les mécanismes attentionnel du système visuel humain pour détecter la saillance. Les modèles purement computationnels reposent principalement sur les principes de la théorie de l'information, du traitement du domaine spectral ou du traitement du signal pour atteindre cet objectif. Les méthodes hybrides [Harel et collab., 2006] intègrent partiellement les principes sur lesquels reposent les modèles biologiques et les modèles purement computationel.

En plus, [Achanta, 2011] catégorisent les modèles de saillance selon qu'ils détectent la saillance sur plusieurs échelles [Itti et collab., 1998], [Achanta et collab., 2008], ou sur une seule échelle [Ma et Zhang, 2003a], [Hu et collab., 2004]. Dans certains modèles, des cartes de caractéristiques individuelles sont créées séparément, puis combinées pour obtenir la carte de saillance finale [Itti et Koch, 1999], [Hu et collab., 2004], [Frintrop et collab., 2007] tandis que dans d'autres modèles de saillance, une carte de saillance est obtenue directement [Ma et Zhang, 2003a], [Achanta et collab., 2008].

Taxonomie de Silva (2010)

Dans leur travaux, [da Silva, 2010] ont proposé une taxonomie des modèles de saillance divisée en 5 familles.

— Les modèles de saillance hiérarchiques [ITTI et collab., 1998],[ITTI, 2000] qui construisent à partir d'une image initiale, une hiérarchie de différentes cartes de caractéristiques, qui sont progressivement combinées jusqu'à obtenir la carte de saillance.

— Les modèles basés sur les statistiques et les probabilités [ITTI et BALDI, 2005] à partir desquels la saillance d'un objet se définit comme étant sa singularité par rapport aux autres. Des théories probabiliste ou statistique sont utilisée afin d'associer la saillance aux éléments les moins probables d'une scène. Par exemple, la théorie bayésienne est exploitée pour estimer la probabilité que différentes régions d'une image soient ou pas des éléments saillants d'une scène visuelle segmentée auparavant.

— Les modèles basés sur la théorie de l'information [BRUCE et TSOTSOS, 2009] qui postulent que le cerveau humain utilise les mécanismes d'attention visuelle afin de maximiser la quantité d'information acquise. Estimée localement, celle-ci peut servir à définir la saillance d'une image.

— Les modèles connexionnistes [VITAY et collab., 2005] exploitent essentiellement des réseaux de neurones artificiels qui ne rentrent pas dans la construction proprement dite de la carte de saillance mais plutôt dans la génération d'un ensemble de focalisations attentionnelles [OLIVA et collab., 2003]. Ils travaillent généralement à partir de cartes de saillances fournies.

— Les modèles algorithmiques [ORABONA et collab., 2007], [AZIZ et MERTSCHING, 2008] qui proposent diverses méthodes difficilement classables dans les précédentes.

Taxonomie de Toet(2011)

[TOET, 2011] catégorisent les modèles de saillance selon deux catégories fondamentales, les modèles computationnels et psychophysiques. Les modèles computationnels se définissent comme tout modèle transformant une image en une carte de saillance. Les modèles psychophysiques se référencent à une estimation directe dérivées des statistiques de œil humain et qui sont utilisables dans l'évaluation d'un modèle de saillance computationnel .

Taxonomie de Judd et al(2012)

[JUDD et collab., 2012] ont catégorisé les modèles de saillance selon qu'ils soient : des modèles inspirés des modèles computationnel de l'attention visuelle comme celui de [KOCH et ULLMAN, 1985]. Des modèles basés sur des facteurs descendants qui intègrent une information de haut niveau pouvant être contextuelle, basée sur un détecteur d'objet. Les modèles basés sur la théorie de Fourier qui traitent une image dans le domaine de Fourier. Des modèles de saillance basés sur les régions considérant un certain regroupement ou segmentation des pixels de l'image. Des modèles qui nécessite d'apprendre certains paramètres à travers des algorithmes d'apprentissage automatique. Des modèles basé sur un biais de centre(center-bias) qui tirent profit des phénomènes de tendance de fixation oculaire vers le centre d'une image.

Taxonomie de Borji et Itti (2013)

[BORJI et collab., 2013] ont proposé une taxonomie des modèles de saillance en 8 familles en fonction du mécanisme adopté pour calculer la carte de saillance. Les mo-

dèles cognitifs se basent sur les théories psychologiques de l'attention visuelle, les modèles spectrales se basent sur l'analyse spectrale estiment la saillance dans le domaine fréquentiel, les modèles probabilistes se basent sur les théories des probabilités, les modèles basés sur la théorie de l'information, les modèles graphiques, les modèles basés sur la classification de patterns qui apprennent les modèles de saillance en utilisant les mouvements oculaires d'apprentissage et d'autres modèles de saillance qui incorporent différentes sources d'informations autres que celles cités.

Taxonomie de Tavakoli(2014)

Similaire à [BORJI et collab., 2013], [TAVAKOLI, 2014] ont catégorisé les modèles de saillance comme suit. Les modèles centre-périphérie reposent sur les principes de la théorie des traits caractéristiques(FIT) de [TREISMAN et GELADE, 1980] et le modèle d'attention de [KOCH et ULLMAN, 1985]. Ces modèles passent souvent par trois étapes principales : extraction des caractéristiques, comparaison des caractéristiques de type centre-périphérie et fusion des cartes de saillance pour obtenir la carte de saillance finale. Les modèles basés sur la théorie de l'information sont fondés sur la théorie de l'information. Dans cette taxonomie, ces modèles gardent la même définition apportée dans les taxonomies précédentes qui considère la saillance comme la quantité d'information par exemple une entropie, une information mutuelle la plus informative dans une image. Les modèles basés sur le domaine de fréquentiel et sur et gardent la même définition apportée dans les taxonomies précédentes. Les modèles basés sur la connexion considèrent une relation structurelle entre les pixels d'une image. Ils modélisent la saillance comme l'émergence d'interactions entre les pixels inter-connectés dans une image. Ces connexions peuvent être représentées par des réseaux de neurones ou des modèles de graphe. Les modèles basés sur l'apprentissage tentent d'établir une relation entre les caractéristiques de bas niveau et les statistiques des mouvements oculaires humain. Cela peut se produire en apprenant un classificateur, certains paramètres et/ou certains connaissances à priori. [TAVAKOLI, 2014] ajoute aussi une catégorie qui regroupe tout modèle de saillance pouvant incorporer différentes sources d'informations autres que celles cités.

Catégorisation de Li et al(2014)

Selon [LI et collab., 2014], les modèles de saillance peuvent être classés selon deux tâches spécifiques à saillance visuelle : la prédiction de fixation des yeux et la segmentation d'objet saillant. Dans une tache de fixation, la saillance s'exprime sous forme de regard. Les sujets sont invités à visualiser chaque image pendant quelques secondes pendant que leurs fixations oculaires sont enregistrées. Dans ce cas, le but d'un modèle de saillance est de calculer une carte de saillance, de probabilité, d'une image pour prédire le regard actuel des yeux. Alternativement, dans une tache de segmentation d'objet saillant, des sujets annotent une image en dessinant des silhouettes d'objets qu'ils voient saillantes. Dans ce cas, le but d'un modèle de saillance est de générer une carte de saillance qui correspond au masque de l'objet saillant annoté.

Catégorisation de Bordji et al(2015)

Une autre catégorisation des modèles de saillance repose sur la nature du problème à résoudre [BORJI et collab., 2015]. Les modèles de saillance qui visent à résoudre un

problème de prédiction de la fixation du regard, également appelé problème de fixation des yeux, où la carte de la saillance est généralement interprétée comme une carte de la distribution de la probabilité de fixation du regard humain pour chaque point indépendant de la carte. La solution recherchée est fournit sous la forme de la carte de saillance. Les modèles de saillance qui visent à résoudre un problème de détection et segmentation d'objet saillant dans une scène, où l'estimation d'une carte de saillance est primordiale mais la solution recherchée doit être fournit sous la forme d'une partie de la scène visuelle initiale, qui représente l'objet le plus saillant.

Dans les sections suivantes, nous présentons des modèles de saillance selon qu'ils traitent un problème de fixation du regard ou de détection et segmentation d'objet saillant. Nous évoquons les modèles de saillance qui nous ont le plus inspirés dans nos travaux.

2.3 La prédiction de fixation

2.3.1 Modèles de prédiction de fixation

Les premières versions des modèles de saillance sont inspirés de la théorie de l'intégration des caractéristiques de [TREISMAN et GELADE, 1980]. L'objectif des modèles de prédiction de fixation est de calculer une carte de saillance qui simule les comportements des mouvement oculaires.

Modèle de Koch et Ullman (1985)

Le modèle d'attention sélective proposé par [KOCH et ULLMAN, 1985] est l'un des premiers modèles d'attention biologiquement plausible. Ce modèle a été conçu sur la base de la théorie de [TREISMAN et GELADE, 1980]. La figure 2.1 décrit le schéma de ce modèle. Tout d'abord, un ensemble de caractéristiques élémentaires, telles que la couleur, l'orientation, la direction du mouvement sont extraites de l'image d'entrée. Ensuite, ces caractéristiques sont traitées en parallèle et forment des cartes de caractéristiques différentes, en respectant la topologie de l'image d'entrée. L'inhibition latérale dans les cartes de caractéristiques, qui simulent les cellules photo-réceptrices de la rétine, améliore les régions de l'image qui sont différentes de leur voisinage. Les cartes de caractéristiques sont fusionnées à une carte de saillance. La notion de carte de saillance a d'abord été introduite par [KOCH et ULLMAN, 1985], selon qui, une carte de saillance donne une vue biaisée de l'environnement visuel, en mettant l'accent sur les zones intéressantes ou bien visibles dans le champ visuel". La carte de saillance représente les zones saillantes de la scène visuelle, mais l'ordre dans lequel elles sont axées est déterminé par une approche WTA (Winner-takes-All).

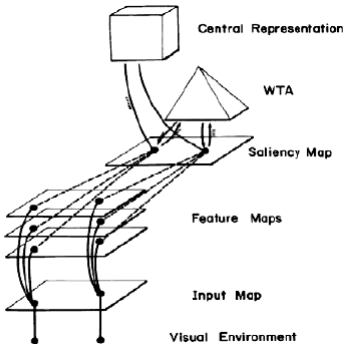


FIGURE 2.1 – Modèle original de la carte de saillance de [KOCH et ULLMAN, 1985].

Modèle de Itti et al (1998)

Le modèle de [KOCH et ULLMAN, 1985] a fourni une architecture de base à de nombreux modèles de saillance qui ont été proposés plus tard, comme le modèle de [ITTI et collab., 1998]. La figure 2.2 illustre l'architecture générale de ce modèle.

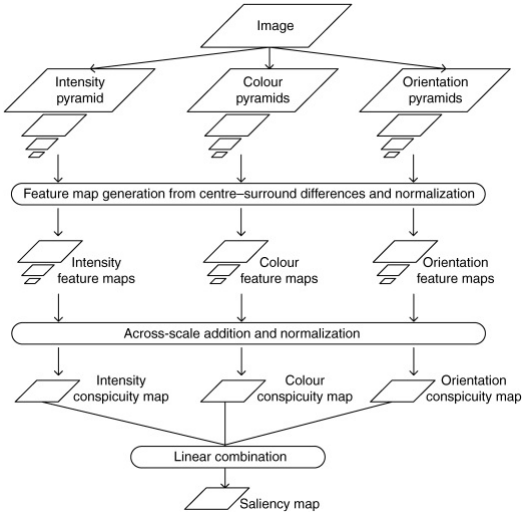


FIGURE 2.2 – Modèle de saillance visuelle proposé par [ITTI et collab., 1998].

A partir d'une image couleur d'entrée rgb, trois canaux de traitement sont extraits

Un canal d'intensité I tel que $I = \frac{r+g+b}{3}$, I est utilisé pour créer une pyramide gaussienne d'images d'intensité I(σ), tel que $\sigma \in \{0 \cdots 8\}$.

Un canal de couleur qui inclut quatre images de couleur : $R = r - \frac{g+b}{2}$ pour le rouge, $G = g - \frac{r+b}{2}$ pour le vert, $B = b - \frac{r+g}{2}$ pour le bleu et $Y = \frac{r+g}{2} - \frac{|r-g|}{2} - b$ pour le jaune.

Un canal d'orientation qui est extrait à partir des images d'intensité. Une pyramide de filtres de Gabor orientés O(σ,θ),tel que θ ∈ {0,45,90,135} est utilisé pour extraire quatre images d'orientation pour chaque niveau de la pyramide, σ ∈ {0···8}.

Ensuite, les cartes de caractéristiques associées à chaque canal de caractéristiques sont calculées en utilisant les différences centre-périphérie entre une échelle fine centrale c et une échelle plus grossière périphérique. Au total, 24 cartes de caractéristiques sont calculées : 6 pour l'intensité, 12 pour la couleur et 24 pour l'orientation. Pour chaque canal, les cartes de caractéristiques sont normalisées et linéairement combinées pour créer une carte de saillance unique pour chaque canal, N(I), N(C) et N(O). La fusion linéaire de ces trois cartes fournit une carte de saillance qui met en valeur les régions les plus attirantes de l'image d'entrée.

Enfin, l'ordre dans lequel la focalisation de l'attention est déplacée sur les régions saillantes de l'image d'entrée est calculée par une approche Winner Takes All(WTA). Celle-ci est combinée à un mécanisme d'inhibition du retour (IOR) pour empêcher de revenir immédiatement à la position saillante visitée. Le modèle de [ITTI et collab., 1998] est probablement le modèle de vision par ordinateur le plus influant de l'attention visuelle. Il est aussi connu sous le nom de Neuromorphic Vision Toolkit(NVT). La première version de NVT n'a effectué qu'une analyse ascendante de l'attention, mais une amélioration ultérieure rapportée dans [NAVALPAKKAM et ITTI, 2006] invoque l'effet de la sélection descendante pendant la recherche visuelle.

Travaux de Ma et Zhang (2003)

[MA et ZHANG, 2003b] ont proposé un modèle computationnel de la saillance qui n'est basée sur aucun modèle biologique. Dans leur travaux, ils ont proposé une mesure de saillance basée sur le contraste. De nombreuse méthodes permettent de calculer le contraste de couleur et le contraste de luminance. Cependant, un contraste générique est préférable. Par conséquence, les auteurs ont défini une zone percevant le stimulus comme un champ de perception, par analogie à la notion de champ récepteur qu'on retrouve dans le système visuel humain. Une image d'entrée est ainsi redimensionnée et transformée de l'espace RVB à l'espace LUV. La perception visuelle humaine est généralement plus sensible aux changements dans les zones lisses que dans les zones de texture. Afin de rendre les couleurs lisses dans les zones de texture, les auteurs emploient une quantification des couleurs en utilisant la méthode de filtrage de groupe de pairs(peer group filtering) de [DENG et collab., 1999]. Afin de lisser davantage les zones de texture et réduire le coût de calcul, l'image est divisée en blocs de $n \times n$ pixels tel que chaque unité de perception du champ de perception possède $n \times n$ pixels. Dans chaque unité de perception, les moyennes des pixels LUV sont calculées séparément. De cette façon, une image de bloc quantifiée est obtenue. Étant donnée une image de bloc quantifiée de taille M × N pixels , celle-ci est considérée comme un champ de perception de M × N unités de perception, si chaque unité de perception contient un pixel. La valeur de contraste $C_{i,j}$ sur une unité de perception(i,j) se calcule comme suit

$$C_{i,j} = \sum_{q \epsilon \Theta} d(p_{i,j}, q) \tag{2.1}$$

où $p_{i,j}, (i \in [0,M], j \in [0,N])$ et q dénote le stimulus perçu par les unités de perception, comme la couleur. Θ est le voisinage de l'unité de perception(i,j). La taille de Θ contrôle la sensibilité du champ de perception. Plus la taille de Θ est petite, plus le champ de

perception est sensible. d est la différence entre $p_{i,j}$ et q, qui est calculée par la distance gaussienne. Tous les contrastes $C_{i,j}$ sur les unités de perception sont normalisés dans l'intervalle [0,255] et forment la carte de saillance. Afin d'extraire les zones d'attention de la carte de saillance, les points observés sont d'abord détectés directement de la carte de saillance par l'approche WTA (Winner-takes-All)adopté auparavant par [ITTI et collab., 1998]. Ils correspondent aux points ayant un contraste local maximum. Ensuite, les zones d'attention sont extraites en utilisant une segmentation par croissance floue. En considérant les points observés dont la valeur de contraste dépasse la valeur du paramètre déterminant la forme de la fonction d'appartenance à l'ensemble floue comme des germes, une segmentation par croissance floue des zones d'attention est réalisée jusqu'à ce qu'aucun candidat des unités de perception ne puisse être regroupé.

Travaux Harel et al (2006)

Dans les travaux de [HAREL et collab., 2006], les auteurs introduisent un modèle de saillance qui se base sur la représentation de graphe. Tout d'abord, trois cartes de caractéristiques de bas niveau(couleur, d'intensité et d'orientation) correspondant à un espace de caractéristique F sont calculés à différentes échelles. Ensuite, un graphe G entièrement connecté est construit en connectant tous les deux pixels de chaque carte de caractéristique, et un poids w est attribué à l'arc connectant le pixel(i,j) au pixel(p,q) comme suit

$$w((i,j),(p,q)) = |log\frac{\mathrm{F}(i,j)}{\mathrm{F}(p,q)}|.\exp(\frac{(i-p)^2+(j-q)^2}{2\sigma^2}) \tag{2.2}$$

le terme $|log\frac{\mathrm{F}(i,j)}{\mathrm{F}(p,q)}|$ représente la dissimilarité entre F(i,j)et F(p,q)
le terme $\exp(\frac{(i-p)^2+(j-q)^2}{2\sigma^2})$ est une fonction gaussienne pour augmenter le poids de deux pixels proches et diminuer le poids des pixels qui sont éloignés les uns des autres.

Enfin, chaque graphe est traité comme une chaîne de Markov pour construire une carte d'activation où des nœuds très différents de leur nœuds avoisinants se verront attribuer des valeurs élevées. Toutes les cartes d'activation sont fusionnées dans la carte de saillance finale.

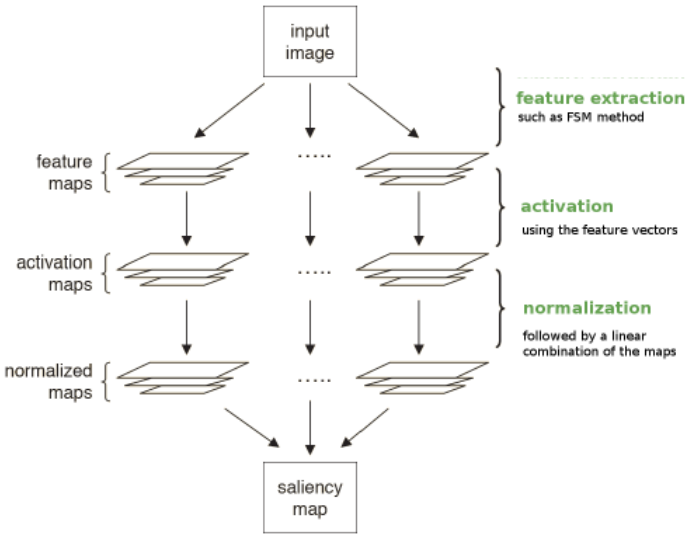


FIGURE 2.3 – Modèle de saillance visuelle proposé par [HAREL et collab., 2006].

Modèle de Hou et Zhang (2007)

Dans les travaux de [HOU et ZHANG, 2007], les auteurs ont proposé le modèle de spectre de résidu(SR) qui ne se base sur aucune caractéristique de bas niveau. A partir d'une image d'entrée, le spectre de Fourier est calculé, ce qui sous entend le calcul des cartes d'amplitude et de phase. Ensuite, le spectre logarithmique de la carte d'amplitude est aussi calculé. Le filtrage de la carte d'amplitude est également calculé en multipliant le spectre logarithmique de la carte d'amplitude par un filtre local moyen. La carte résiduelle spectrale est obtenue en soustrayant le spectre logarithmique de la carte d'amplitude et la carte d'amplitude filtrée. La carte de saillance est obtenue en appliquant la transformée inverse de Fourier. Ce modèle est principalement basé sur la propriété générale des images naturelles, décrite par la loi $\frac{1}{f}$. Cette loi stipule que l'amplitude du spectre moyen de Fourier, A(f), de l'ensemble des images naturelles est proportionnelle à 1/f, dans lequel f est la fréquence. Par conséquent, le spectre de phase est préservé. L'idée sous-jacente est que si le spectre logarithmique de l'image est loin de 1/f des images naturelles, le spectre filtré de l'image, il y a quelque chose de anormal qui mérite attention.

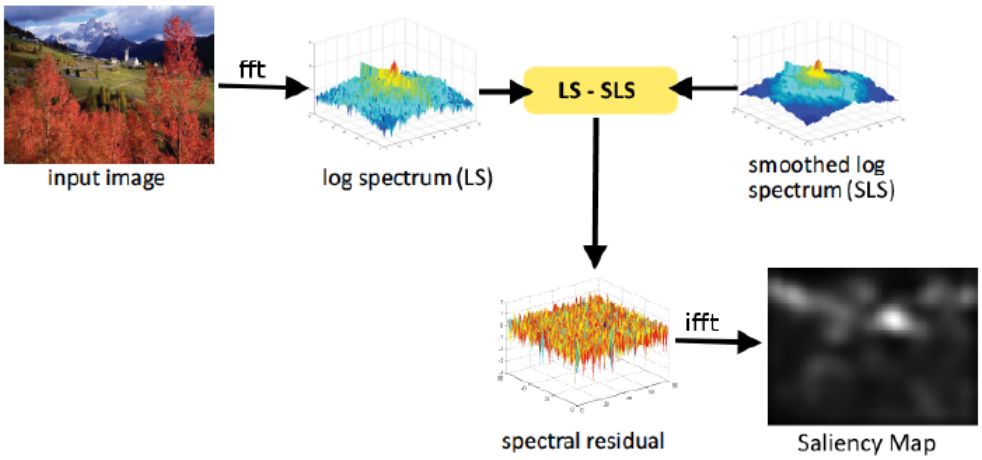


FIGURE 2.4 – Modèle de saillance visuelle proposé par [HOU et ZHANG, 2007].

Travaux de Guo et al (2008,2010)

Dans les travaux de [GUO et collab., 2008], les auteurs proposent une méthode qui se base sur le modèle de [HOU et ZHANG, 2007] qui utilise le résidu spectral du spectre d'amplitude pour obtenir la carte de saillance. [GUO et collab., 2008] ont proposé d'utiliser le spectre de phase au lieu de celui de l'amplitude. L'idée clé est que le spectre d'amplitude spécifie la quantité de chaque composante sinusoïdale présente dans une image tandis que les informations de phase spécifient où chacune des composantes sinusoïdales réside dans l'image. Ils ont montré que les endroits présentant moins de périodicité, ou moins d'homogénéité, dans une image saute au yeux dans le spectre de phase de l'image reconstruite.En conséquence, ils ont proposé un modèle de détection de saillance pour des images d'intensité, qui se base sur le spectre de phase de la transformée de Fourier(PFT). En comparaison avec le modèles de [HOU et ZHANG, 2007], le modèle PFT s'avère aussi moins coûteux en temps d'exécution. En ce qui concernent les cartes de saillance générées, elles leur sont très similaires.

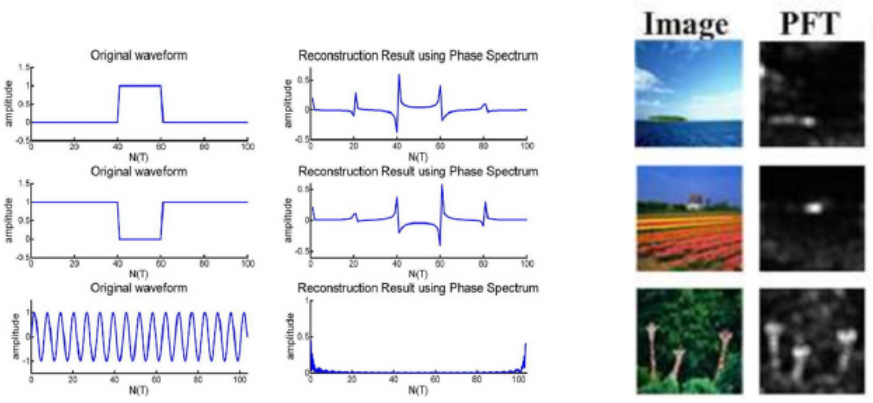


FIGURE 2.5 – Modèle de saillance visuelle proposé par [GUO et collab., 2008].

Comme extension à leur travaux, [GUO et ZHANG, 2010] ont introduit le modèle du spectre quaternion de la transformée de Fourier (PQFT) qui analyse la couleur, l'orientation et le mouvement, en plus de l'intensité (en PFT), pour calculer la carte de la saillance. Le modèle PQFT est indépendant de toutes connaissances à priori et a démontré son efficacité pour répondre aux exigences en temps réel.

Travaux de Hou et al (2012)

Dans les travaux de [HOU et collab., 2012], les auteurs ont proposé de modéliser la saillance à travers le calcul de la signature d'image qui permet d'approximer le premier plan d'une image dans le cadre théorique du mixage des signaux parcimonieux. Étant donné une image d'entrée, tout d'abord, trois canaux de couleur sont extraits. Les deux espaces de couleur RVB ou CIE LAB peuvent être utilisés. Ceci dit, les auteurs ont choisi d'utiliser l'espace de couleur CIE LAB vu qu'il imite de près la façon dont la vision humaine perçoit la couleur. Ensuite, la signature d'image est calculée sur chaque canal pour supprimer l'arrière-plan et détecter le premier plan d'une image. Pour ce faire, une transformation discrète en cosinus (DCT) est appliquée à chaque canal. Le signe de chaque composante DCT, équivalent à la phase pour une décomposition de Fourier, est stocké et inversement retransformé dans le domaine spatial. Les informations d'amplitude sur toute la fréquence sont rejetées. Un filtrage gaussienne est ensuite appliqué pour brouiller les résultats et la carte de saillance finale est obtenue en additionnant les résultats des trois canaux.

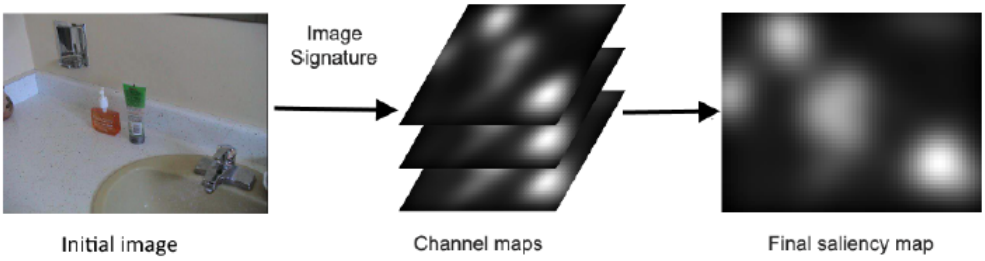


FIGURE 2.6 – Modèle de saillance visuelle proposé par [HOU et collab., 2012].

2.3.2 Principales bases d'images

Bruce Database

Une des premières bases d'images dédiées à la prédiction de la fixation a été introduite par [BRUCE et TSOTSOS, 2005]. Cette base a été dérivée à partir des expériences de eye-tracking par des sujets qui ont observé 120 images de couleurs différentes. Les images sont présentées dans un ordre aléatoire et chaque image est présentée pendant 4 secondes avec un masque entre chaque paire d'images. Les données ont été collectées à partir de 20 sujets pour un groupe de 120 images. Bien que cette base d'images ait été initialement conçu pour la prédiction de la fixation, elle a été récemment complété par [BORJI et collab., 2013] avec 70 sujets sous l'instruction d'étiqueter l'objet le plus saillant de l'image.

MIT Database

Dans leur travail, [JUDD et collab., 2009] ont introduit un ensemble de données beaucoup plus grand avec 1003 images collectées de Flickr Creative Commons et LabelMe. Les données du eye-tracking ont été enregistré par 15 utilisateurs qui ont consulté ces images. La plus grande dimension de chaque image est de 1024 pixels et l'autre dimension varie de 405 à 1024 avec une nombre majoritaire de 768 pixels. Cet ensemble contient 779 images en mode paysage et 228 images en mode portrait . Un eye tracker enregistre le trajet du regard de l'observateur sur un ordinateur pendant 3 secondes séparées par 1 seconde de visualisation d'un écran gris.

Li Database

La base de données public Microsoft de [LI et collab., 2013] est constitué de 5000 images couleurs. Elle contient aussi des vérités de terrain qui correspondent à des régions d'attention c'est à dire des objets d'intérêt dans les scènes perçues par les sujets. Celles-ci sont mise en évidence par des boites de sélection créés par 9 sujets. L'évaluation de la qualité d'un modèle de détection de saillance est réalisé quantitativement en vérifiant la cohérence entre la vérité de terrain étiquetée par des humains et la carte de saillance calculée à partir de n'importe quel modèle de saillance.

PASCAL-S Database

La base de données PASCAL-S de [LI et collab., 2014] a été construite sur l'ensemble de validation de la base de données PASCAL VOC 2010 [EVERINGHAM et collab., 2010]. Ce sous ensemble contient 850 images naturelles. Dans les expérimentations de prédiction de fixation, 8 sujets ont été invités à effectuer la tâche de visualisation pour explorer les images. Chaque image a été présentée pendant 2 secondes et un re-calibrage du suivi oculaire a été effectué toutes les 25 images. Les données sur le regard ont été échantillonnées à l'aide d'un eye-tracker Eyelink 1000, à 125 Hz.

DUT-OMRON Database

Un appareil de suivi oculaire, Tobbii X1 Light Eye tracker, est utilisé pour enregistrer les fixations oculaires pendant qu'un participant visualise une image affichée sur un moniteur. Chaque image est affichée pendant 2 secondes sans aucun intervalle entre des images successives. Cinq données de suivi oculaire sont enregistrées pour chaque image de la base d'images.

2.3.3 Métriques d'évaluation

Dans la littérature, de nombreuses mesures sont utilisées pour comparer les cartes de saillance générées automatiquement avec les cartes de fixation oculaire enregistrées par des observateurs humains. Dans ce qui suit, nous présentons les mesures d'évaluation les plus utilisées.

Coefficient de corrélation(Pearson)

Le coefficient de corrélation de Pearson (PCC) fournit des informations sur l'existence d'une relation linéaire entre la carte de saillance calculée S_a et la carte de fixation humaine S_h qui est considérée comme une vérité terrain. Il se définit par :

$$\mathrm{PCC}(\mathrm{S}_a,\mathrm{S}_h) = \frac{cov(\mathrm{S}_a,\mathrm{S}_h)}{\sigma_{\mathrm{S}_a}\sigma_{\mathrm{S}_h}} \tag{2.3}$$

tel que σ_{S_a} et σ_{S_h} correspondent respectivement à l'écart type des valeurs des cartes S_a et S_h. $cov(\mathrm{S}_a,\mathrm{S}_h)$ correspond à la covariance entre S_a et S_h. Ce coefficient est normalisé dans l'intervalle [-1 1]. Ainsi, une valeur égale à 0 signifie l'absence de relation linéaire entre ces deux cartes. Plus la valeur est proche de 1, plus la valeur de la saillance élevée dans S_a correspond aux zones les plus observées(valeurs élevées dans S_h). Inversement, plus la valeur est proche de -1, plus la valeur de la saillance élevée de S_a correspond aux zones les moins observées (valeurs faibles dans S_h).

L'aire sous la courbe ROC (score AUC)

Une autre métrique pour évaluer une carte de saillance est l'indicateur AUC-Judd[JUDD et collab., 2009]. Cette métrique interprète les fixations comme une tâche de classification où un pixel de la carte peut être soit saillant ou non saillant en appliquant un seuil sur la valeur d'intensité de la carte de saillance. Chaque pixel saillant correspondant au fixation humaine sur la carte est considéré comme une valeur positive réelle, tandis que les pixels saillants sur les zones de non fixation sont classés comme des valeurs faussement positives. Le score final de l'AUC est ensuite calculé et tracé comme un compromis entre les valeurs vraies et fausses positives. Le score le plus élevé possible peut être 1, alors qu'un score de 0,5 est considéré comme aléatoire.

Trajet de scannage de saillance normalisé

Une autre mesure largement utilisée pour comparer deux cartes de saillance est le trajet de scannage de saillance normalisé(NSS). Cette mesure a été introduite comme une simple mesure de correspondance entre les cartes de saillance et les vérités de terrain, calculée comme la moyenne des valeurs de saillance normalisée sur les endroits fixés [PETERS et collab., 2005], [MARAT et collab., 2012]. La valeur NSS(i,j) à la position (i,j) de la carte de saillance est calculée comme suit :

$$\mathrm{NSS}_{(i,j)} = \frac{\mathrm{M}_m.\mathrm{M}_h - \overline{x}_m}{s_m} \tag{2.4}$$

M_m est la valeur de la saillance à la position (i,j).
M_h est la carte de densité de la position humaine à la position (i,j).
$\overline{x}_m$ est la moyenne empirique de la carte de saillance M_m.
s_m est l'écart-type empirique de la carte de saillance M_m.

La valeur de NSS est nulle, quand il n'y aucun lien entre les positions expérimentales des yeux et les régions saillantes. Elle est négative quand les positions sont sur des régions non saillantes, et positive lorsqu'elles sont projetés sur les régions saillantes. Plus le valeurs positives de NSS sont élevées, plus les régions saillantes sont observées.

2.4 La détection d'objet salient

La détection d'objets saillants vise à mettre en évidence les objets saillants dans les images. En comparaison avec la tache de prédiction de fixation oculaire, plusieurs modèles de sailance ont été proposé pour détecter et segmenter les objets saillants dans les images.

2.4.1 Modèles de détection d'objet saillant

Travaux de Achanta et al (2008)

Dans leur travaux, [ACHANTA et collab., 2008] ont définit la saillance par le calcul du contraste local d'une région d'image par rapport à son voisinage, effectué à travers différentes échelles. Autrement, il s'agit de calculer la distance entre le vecteur moyen des pixels d'une sous région d'image et le vecteur moyen des pixels de son voisinage. De cette façon, une carte de saillance est obtenue à une échelle donnée en utilisant des vecteurs de caractéristiques pour chaque pixel, au lieu de combiner différentes cartes de saillance individuelles, de valeurs scalaires pour chaque caractéristiques.

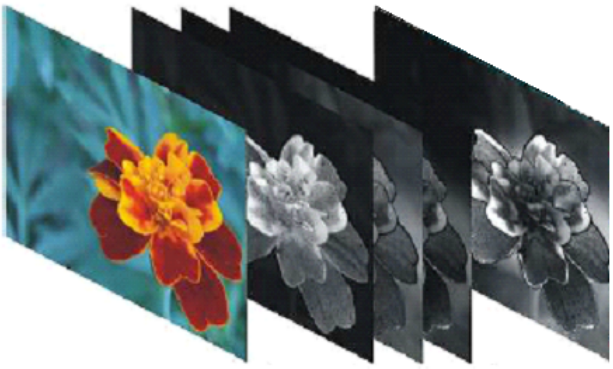


FIGURE 2.7 – Modèle de saillance visuelle proposé par [ACHANTA et collab., 2008].

À une échelle donnée, la valeur de la saillance basée sur le contraste $c_{i,j}$ pour un pixel se trouvant à la position (i, j) dans une image se calcule comme la distance D entre les vecteurs moyens des caractéristiques des pixels de la région intérieure R_1 et celle de la région extérieure R_2 comme suit

$$c_{i,j} = \mathrm{D}[(\frac{1}{\mathrm{N}_1}\sum_{p=1}^{\mathrm{N}_1} v_p), (\frac{1}{\mathrm{N}_2}\sum_{q=1}^{\mathrm{N}_2} v_q)] \tag{2.5}$$

où N_1 et N_2 sont le nombre de pixels dans R_1 et R_2 respectivement, et v est le vecteur de caractéristiques correspondant au pixel. La distance D est une distance Euclidienne si v est un vecteur de caractéristiques non corrélés, et toute autre mesure de distance appropriée si les caractéristiques du vecteur sont corrélés.

[ACHANTA et collab., 2008] ont utilisé l'espace de couleur CIE Lab, en supposant des images couleurs RVB, pour générer des vecteurs de caractéristiques pour la couleur et la luminance. Puisque les différences perceptuelles dans l'espace de couleur du CIE Lab sont approximativement euclidiennes, la distance D dans l'équation 2.6 se calcule comme suit :

$$c_{i,j} = \| v_1 - v_2 \| \tag{2.6}$$

tel que $v_1 = [\mathrm{L}_1, a_1, b_1]^{\mathrm{T}}$ et $v_2 = [\mathrm{L}_2, a_2, b_2]^{\mathrm{T}}$ sont les vecteurs moyens des régions R_1 et R_2 respectivement.

Vue que seulement les valeurs moyennes des vecteurs de caractéristiques de R_1 et R_2 sont à calculer, les auteurs ont utilisé l'approche des images intégrales pour son efficacité de calcul.

Un changement d'échelle est affecté par la mise à l'échelle de la région R_2 au lieu de la mise à l'échelle de l'image. La mise à l'échelle du filtre au lieu de l'image permet de générer des cartes de saillance de la même taille et de la même résolution que l'image d'entrée. La région R_1 est généralement choisi pour être un pixel.

Pour une image de largeur w pixels et de hauteur h pixels, la largeur de la région R_2 notée w_{R_2} varie comme dans l'équation 2.7

$$\frac{w}{2} \geq w_{R_2} \geq \frac{w}{8} \tag{2.7}$$

Pour chaque image, un filtrage est effectué à trois échelles différentes, selon l'équation 2.7, et la carte de saillance finale est calculée comme la somme des valeurs de saillance à travers les échelles S

$$m_{i,j} = \sum_{S} c_{i,j} \tag{2.8}$$

$\forall i \in [1, w], j \in [1, h]$ tel que $m_{i,j}$ est un élément de la carte combinée de la saillance combinée M obtenue par sommation des valeurs de saillance à travers les différentes échelles.

Travaux de Achanta et al (2009)

Dans leur travaux, [ACHANTA et collab., 2009] ont proposé un modèle de saillance à fréquence réglable qui exploitent presque tout le contenu de basse fréquence et la plupart du contenu de haute fréquence pour obtenir des cartes de saillance de haute qualité en utilisant les caractéristiques de couleur et d'intensité.

La carte de saillance est obtenue en calculant la distance euclidienne du vecteur CIE LAB moyen de tous les pixels d'une image d'entrée avec chaque pixel (également un vecteur CIE LAB) d'une version gaussienne floue de la même image d'entrée

$$S(x, y) = |I_\mu - I_f(x, y)| \tag{2.9}$$

où S(x,y) est la valeur de la saillance des pixels à la position (x,y), I_μ est la moyenne de tous les vecteurs de pixels CIE LAB de l'image, $I_f(x, y)$ est le vecteur de pixels d'image CIE LAB correspondant dans la version filtrée gaussienne de l'image originale, et || est la norme L_2, la distance euclidienne dans l'espace de couleur CIE LAB. L'espace de couleur CIELAB est utilisé car les distances euclidiennes dans cet espace de couleur sont perceptuellement uniformes.

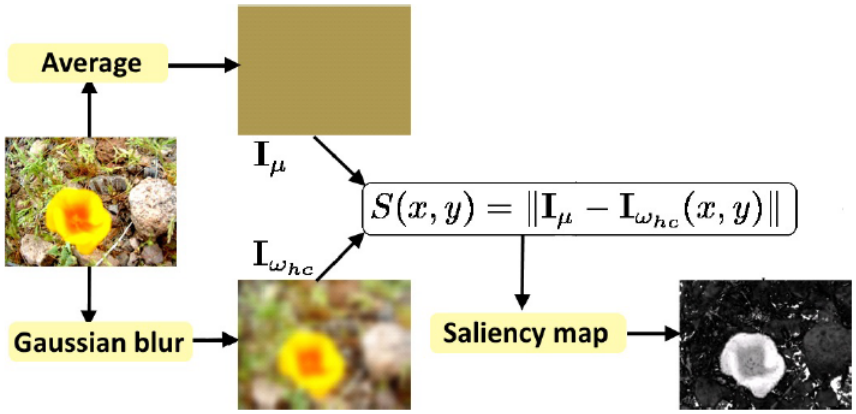


FIGURE 2.8 – Modèle de saillance visuelle proposé par [ACHANTA et collab., 2009].

Les cartes de saillance résultantes possèdent des régions saillantes uniformément mises en évidence avec des contours bien définies. Cependant, dans les images où la région saillante est très large, ou le fond est complexe, les cartes de saillance en tendance à mettre en évidence le fond à la place de la région saillante. D'après les auteurs, cela revient au fait que dans le calcul du vecteur CIE LAB moyen de l'image dans l'équation 2.9, la région saillante contribue plus à la moyenne de l'image que le reste de l'image, générant ainsi des valeurs $S(x,y)$ inférieures aux pixels de l'arrière-plan.

Travaux de Achanta et Susstrunk (2010)

Dans leur travaux [ACHANTA et SÜSSTRUNK, 2010], les auteurs ont amélioré leur modélisation de la saillance en choisissant un pourtour symétrique pour chaque pixel. De cette façon, ils traitent implicitement chaque pixel comme étant au centre de sa propre sous région. Ceci diffère du principe adopté par [ACHANTA et collab., 2009] où l'image entière est utilisée comme un pourtour global commun, pour un pixel donné. Ce pourtour global représente le vecteur de couleur CIE LAB de l'image moyenne. Par conséquent, tous les pixels qui ne sont pas au centre de l'image ont un pourtour asymétrique comme le montre la figure 2.9.

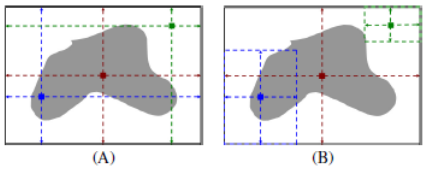


FIGURE 2.9 – Principe du pourtour symétrique proposé par [ACHANTA et SÜSSTRUNK, 2010].

Ainsi, pour une image d'entrée de largeur w et de hauteur h, la valeur de saillance de pourtour symétrique d'un pixel donné $S_{ss}(x,y)$ est obtenue comme suit :

$$S_{ss}(x,y) = I_\mu(x,y) - I_f(x,y) \tag{2.10}$$

où $\mathrm{I}_\mu(x,y)$ est le vecteur de couleur moyen CIE LAB de la sous image dont le pixel central se trouve à la position (x,y) indiquée comme suit

$$\mathrm{I}_\mu(x,y) = \frac{1}{\mathrm{A}} \sum_{i=i-x_0}^{x+x_0} \sum_{j=j-y_0}^{j+j_0} \mathrm{I}(i,j) \tag{2.11}$$

Avec les décalages x_o, y_o, et la zone A de la sous image calculée comme suit :

$$\begin{aligned} x_o &= \min(x, w-x) \\ y_o &= \min(y, h-y) \\ \mathrm{A} &= (2x_o+1)(2y_o+1) \end{aligned} \tag{2.12}$$

$\mathrm{I}_f(x,y)$ est le vecteur du pixel de l'image CIE LAB correspondant au filtrage de l'image originale avec un filtre gaussien, et $\|\|$ est la norme L_2 (distance euclidienne dans l'espace couleur CIE LAB).

Les sous images obtenues dans l'équation 2.11 en utilisant les paramètres de l'équation 2.12 sont les régions pourtour symétriques maximales possibles pour un pixel donné au centre. Par conséquent, plus un pixel est proche des bords, plus poutour est plus étroit. Pour calculer les moyennes CIE LAB de ces sous images, l'approche des images intégrales est également utilisée.

Travaux de Goferman et al (2010)

Dans les travaux de [GOFERMAN et collab., 2012], les auteurs proposent un nouveau type de saillance, la saillance contextuelle, qui vise à détecter les régions de l'image qui représentent la scène. Cette définition diffère des définitions précédentes dont leur but est soit d'identifier des points de fixation soit de détecter l'objet dominant. La saillance contextuelle suit quatre principes de base de l'attention visuelle humaine, qui sont soutenus par des preuves psychologiques. Les considérations locales de bas niveau, qui incluent les facteurs de contraste et de couleur. Les considérations globales qui suppriment les caractéristiques fréquentes tout en maintenant les caractéristiques qui s'écartent de la norme. Les règles d'organisation visuelle qui stipulent que les formes visuelles peuvent posséder un ou plusieurs centres de gravité autour desquels la forme est organisée. Les facteurs de haut niveau tels que les indices sur l'emplacement de l'objet saillant et la détection d'objet. Les travaux connexes ne suivent généralement que certains de ces principes et, par conséquent, peuvent ne pas fournir les résultats souhaités. Cette méthode définit une nouvelle mesure de caractère distinctif qui combine 3 principes de base de l'attention visuelle humaine.Le dernier principe est considéré comme un post-traitement.

Travaux de Cheng et al (2011),(HC)

Dans leur travaux [CHENG et collab., 2011], les auteurs ont proposé un modèle de saillance (HC) qui se base sur le calcul de contraste à partir de l'histogramme d'une image d'entrée. Plus précisément, la saillance d'un pixel I_k dans l'image I se définit en utilisant son contraste de couleur avec tous les autres pixels comme suit

$$\mathrm{S}(\mathrm{I}_k) = \sum_{\forall \mathrm{I}_i \in \mathrm{I}} \mathrm{D}(\mathrm{I}_k, \mathrm{I}_i) \tag{2.13}$$

tel que $\mathrm{D}(\mathrm{I}_k, \mathrm{I}_i)$ est la distance de couleur entre les pixels I_k et I_i dans l'espace CIE LAB. Cette équation peut être ré-écrite comme suit,

$$S(I_k) = D(I_k, I_1) + D(I_k, I_2) + \cdots + D(I_k, I_N) \tag{2.14}$$

où N est le nombre de pixels de l'image I.

Selon cette définition, les pixels qui ont la même valeur de couleur ont la même valeur de saillance, vue que la mesure ne considère pas les relations spatiales entre les pixels. En ré-écrivant l'équation 2.14 de telle sorte que les termes avec la même valeur de couleur c_j soient regroupés dans le même ensemble, la valeur de saillance pour chaque couleur se définit comme suit :

$$S(I_k) = S(c_l) = \sum_{j=1}^{n} f_j D(c_l, c_j) \tag{2.15}$$

avec c_l la valeur de la couleur du pixel I_k, n le nombre de couleurs de pixel distinctes, et f_j la probabilité de la couleur du pixel c_l dans l'image I.

Une quantification des couleurs est réalisée dans l'espace de couleur RVB selon 123 couleurs en divisant de manière uniforme chaque canal de couleur en 12 niveaux différents. Tandis que la quantification des couleurs est réalisée dans l'espace de couleur RVB, les différences de couleur sont calculées dans l'espace de couleur CIE LAB, vu qu'il s'agit d'un espace perceptuellement uniforme. La valeur de saillance de chaque couleur c est remplacée par la moyenne pondérée des valeurs de saillance de couleurs similaires (mesurée par la distance CIE LAB). Il s'agit en fait d'un processus de lissage dans l'espace des caractéristiques de couleur.

Travaux de Cheng et al (2011),(RC)

Dans leur travaux [CHENG et collab., 2011], les auteurs ont proposé un autre modèle de saillance qui repose sur l'analyse du contraste basé région(Region based contrast ou RC), où une image d'entrée est segmentée en régions à travers une méthode de segmentation basée sur les graphes. Pour chaque région r_k, un histogramme de couleur est construit et un contraste de couleur est calculé. La saillance de chaque région $S(r_k)$ se définit comme la somme pondérée des contrastes de la région par rapport à toutes les autres régions de l'image. Les poids sont affectées en fonction des distances spatiales, les régions les plus éloignées se voient affecter des poids plus faibles.

$$S(r_k) = \sum_{r_k \neq r_i} w(r_i) D_r(r_k, r_i), \tag{2.16}$$

où $w(r_i)$ est le poids de la région r_i et $D_r(,)$ est la distance de couleur entre les deux régions. Ils utilisent le nombre de pixels en r_i comme $w(r_i)$ pour accentuer le contraste des couleurs vers les régions les plus larges.

La distance de couleur entre deux régions r_1 et r_2 est définit comme suit

$$D_r(r_1, r_2) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} f(c_{1,i}) f(c_{2,j}) D(c_{1,i}, c_{2,j}) \tag{2.17}$$

où $f(c_{k,i})$ est la probabilité de la $i^{ième}$ couleur $c_{k,i}$ parmi tous les n_k couleurs dans la région k r_k, $k = \{1,2\}$. Le auteurs utilisent la probabilité d'une couleur dans la fonction de densité de probabilité (histogramme de couleur normalisé) de la région comme le poids de cette couleur pour accentuer plus les différences de couleur entre les couleurs dominantes.

Par ailleurs, les auteurs incorporent l'information spatiale en introduisant un terme de pondération spatiale dans l'équation 2.18 pour augmenter les effets des régions les plus proches et diminuer les effets des régions les plus éloignées. Plus précisément, pour toute région r_k, la saillance basée sur le contraste de la région pondérée spatialement est définit comme suit :

$$S(r_k) = \sum_{r_k \neq r_i} \exp(\frac{D_s(r_k, r_i)}{\sigma_s^2}) w(r_i) D_r(r_k, r_i) \tag{2.18}$$

où $D_s(r_k, r_i)$ est la distance spatiale entre les régions r_k et r_i, et σ_s contrôle la force de la pondération spatiale. La distance spatiale entre deux régions est définit comme la distance euclidienne entre leurs centres. Dans leur travail, les auteurs ont utilisé $\sigma_s^2 = 0,4$ avec des coordonnées de pixels normalisées à [0,1].

Travaux de Perazzi et al (2012)

Dans les travaux de [PERAZZI et collab., 2012], les auteurs utilisent une adaptation de l'approche superpixels SLIC pour segmenter l'image en régions perceptuellement uniformes. Les superpixels SLIC segmentent une image en utilisant l'algorithme des K-moyennes dans l'espace RVB. Une modification de cette approche a été apporté en utilisant l'algorithme des k-moyennes avec la distance géodésique dans l'espace CIE LAB. Ceci permet de garantir une meilleur connectivité, tout en conservant la localisation, la compacité et les contours des superpixels SLIC. Ensuite, deux mesures de contraste ont été définit pour calculer la valeur de saillance de chaque pixel au travers une série d'opération de filtrage gaussien. La mesure d'unicité U_i d'un élément se définit comme la rareté d'un segment i étant donné sa position p_i et sa couleur c_i dans l'espace CIE LAB par rapport à tous les autres segments j. La mesure de distribution spatiale D_i d'éléments pour un segment i est calculée en utilisant la variance spatiale de sa couleur c_i, autrement dit cela revient à mesurer son occurrence ailleurs dans l'image. Les mesures d'unicité U_i et de distribution D_i sont ensuite normalisées dans l'intervalle[0 1], puis combinées pour calculer la valeur de saillance S_i de chaque élément/ Celle-ci est intégrée dans un espace à 5 dimensions en utilisant sa position p_i et sa valeur de couleur c_i dans l'espace RVB.

Travaux de Fu et al (2013)

Dans leur travaux, [CHENG et collab., 2011] ont procédé à la quantification des canaux caractéristiques des pixels sous forme d'histogramme couleur afin de mesurer la dissimilarité du contraste spatial et d'évaluer la saillance d'un pixel par rapport aux autres pixels de l'image. Toutefois, les distributions de caractéristiques estimées à l'aide de l'histogramme sont des discontinuités sur les bords des bins. Inspiré par le principe de mesurer la saillance en se basant sur le calcul du contraste global à partir de l'histogramme d'une seule image, et afin de résoudre le problème de discontinuité des distributions de caractéristique sur les bords des bins, [FU et collab., 2013] ont proposé de mesurer la saillance en se basant sur le calcul du contraste local à partir de partitions(clusters) de l'image. Dans leur travaux, les auteurs ont utilisé la méthode des k-moyennes avec k=6. Une carte de saillance basé sur le contraste est obtenue. En plus, les auteurs ont également mesuré la saillance en se basant sur le calcul de l'indice spatial, spatial cue, à partir des clusters de l'image. En supposant que la vraisemblance de la saillance d'un pixel appartenant à un cluster satisfait une distribution gaussienne.

La carte de saillance basée sur l'indice spatial est obtenue en calculant la probabilité marginale de la saillance. La carte finale de saillance est calculé en combinant les deux cartes de saillance précédemment obtenues.

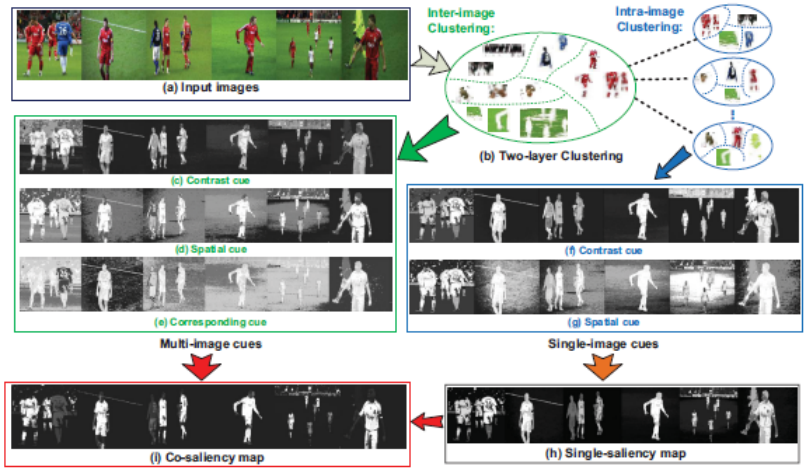


FIGURE 2.10 – Modèle de saillance visuelle proposé par [FU et collab., 2013].

Travaux de Zhang et al (2013)

Dans leur travaux [ZHANG et collab., 2013], ont proposé un modele de saillance, SDSP, qui se base sur la définition de trois mesures de saillance : la fréquence spatiale, la couleur et la localisation spatiale comme l'illustre la figure 2.11.

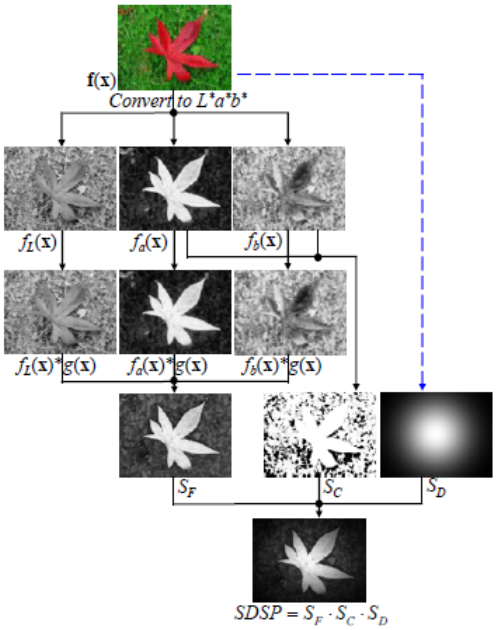


FIGURE 2.11 – Modèle de saillance visuelle proposé par [ZHANG et collab., 2013].

Inspiré par le travail de [ACHANTA et collab., 2009], les auteurs utilisent également le filtrage par bande pour détecter la saillance. Cependant, ils adoptent le filtre de Log-Gabor au lieu d'une différence de gaussienne(DoG). Partant du constat que les couleurs chaudes, comme le rouge et le jaune, attirent plus le système visuel humain que les couleurs froides, comme le vert et le bleu, les auteurs proposent une nouvelle mesure de couleur pour modéliser cette observation.

Pour une image d'entrée f(x) dans l'espace de couleur RVB, sa conversion dans l'espace de couleur CIELab est d'abord réalisée. $f_{\mathrm{L}}(x), f_a(x)$ et $f_b(x)$ représentent respectivement les trois canaux L, a et b. L'espace CIELab est un système de couleur antagoniste dans lequel le canal-a représente l'information vert-rouge tandis que le canal-b représente l'information bleu-jaune. Si un pixel a une valeur plus petite(plus grande), il semblerait verdâtre(rougeâtre). De la même manière, si un pixel a une valeur b plus petite(plus grande), il semblerait bleuâtre(jaunâtre). Par conséquent, si un pixel a une valeur a ou b plus élevée, il semblerait plus chaud, sinon, il semblerait plus froid.

Partant aussi du constat qu'une personne est plus susceptible de faire attention au centre d'une image. La localisation spatiale est modélisée par une carte gaussienne. En supposant que c soit le centre de l'image d'entrée f(x), la saillance spatiale de l'image est exprimée comme une carte gaussienne tel que σ_{D} est un paramètre de la gaussienne.

$$\mathrm{S_D}(x) = \exp(-\frac{\|x-c\|_2{}^2}{\sigma_{\mathrm{D}}^2}) \tag{2.19}$$

Par conséquent, trois cartes de saillance $\mathrm{S_F}(x), \mathrm{S_C}(x)$ et $\mathrm{S_D}(x)$ sont respectivement calculées. La carte de saillance finale de l'image est obtenue comme suit

$$\mathrm{S}_{Finale} = \mathrm{S_F}(x).\mathrm{S_D}(x).\mathrm{S_C}(x) \tag{2.20}$$

Travaux de İmamoğlu et Lin 2013

Dans les travaux de [IMAMOGLU et collab., 2013], les auteurs proposent un modèle de détection de saillance en utilisant des caractéristiques de bas niveau obtenues à partir du domaine de la transformée en ondelettes. Le modèle proposé convertit d'abord une image de couleur RVB en une image de couleur CIE LAB suivi d'un filtrage gaussien. Par la suite, une transformée en ondelettes avec des largeurs de bande de fréquences croissantes est utilisée pour créer des cartes de caractéristiques à différentes échelles, ce qui permet de représenter différentes caractéristiques de bord et de texture. Une fois ces cartes de caractéristiques obtenues, les auteurs calculent la distribution globale des caractéristiques locales pour obtenir à la fois une carte de saillance globale et une carte de saillance locale en fusionnant les cartes de caractéristiques à chaque niveau sans procéder à une opération de normalisation. La carte de saillance finale est une combinaison linéaire de ces deux cartes.

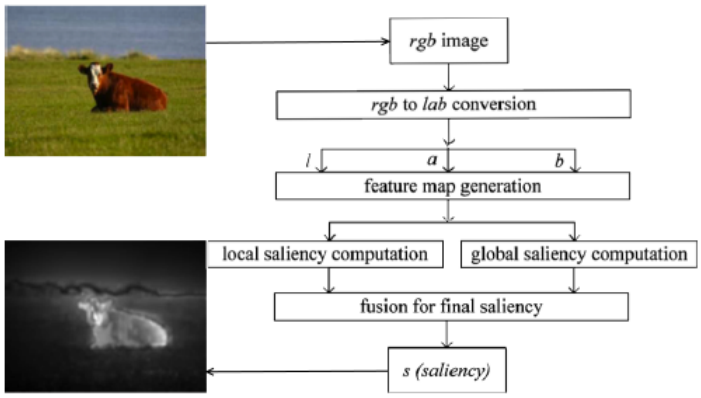


FIGURE 2.12 – Modèle de saillance visuelle proposé par [IMAMOGLU et collab., 2013].

Travaux de Yang et al (2013)

Dans les travaux de [YANG et collab., 2013], les auteurs proposent une méthode de détection de la saillance basée sur un modèle probabiliste universel, qui mesure la saillance en combinant des caractéristiques de bas niveau (saillance visuelle basée sur les caractéristiques) et l'indice de localisation spatiale (saillance visuelle basée sur la localisation). Ensuite, ils simulent un processus de décalage du centre du champ visuel (appelé décalage central), puis procèdent à une analyse à différentes échelles.

Dans leur travaux, les auteurs [YANG et collab., 2013] considèrent une image comme un ensemble de régions R, $R_1, R_2, \ldots, R_N$ où N est le nombre total de régions segmentées obtenues par une méthode de segmentation basée sur les graphes. Pour chaque région r_i, la valeur de saillance S_{r_i} est calculée, et toutes les valeurs correspondantes à différentes régions composent la carte de saillance de l'image entière. En supposant qu'une variable aléatoire V_{r_i} indique que la région r_i est saillante, et B_{r_i} est un ensemble de facteurs qui se traduit par la saillance S_{r_i}.

Par ailleurs, S_{r_i} peut être une probabilité conditionnelle de V_{r_i} sachant B_{r_i}, c'est-à-dire, $S_{r_i} = P(V_{r_i} vert B_{r_i})$. Il s'agit d'un modèle universel pour estimer la saillance visuelle.

Étant donné la variable aléatoire F_{r_i} dénotant les caractéristiques visuelles observées dans la région r_i, et la variable aléatoire L_{r_i} indiquant la localisation spatiale de la région S_{r_i}. La valeur de saillance de la région S_{r_i} est définit comme suit

$$S_{r_i} = P(V_{r_i}|F_{r_i}, L_{r_i}). \tag{2.21}$$

En assumant que ces deux variables aléatoires sont indépendantes, l'équation 2.21 peut être ré-écrite comme suit

$$S_{r_i} = P(V_{r_i}|F_{r_i})P(V_{r_i}|L_{r_i}). \tag{2.22}$$

Le premier terme de l'équation 2.22 représente la probabilité conditionnelle de la valeur de saillance possible sachant les caractéristiques observées F_{r_i}. Dans leur travaux, les auteurs ont utilisé la dissimilarité, mesurée par le contraste de couleur pondéré et la distance spatiale entre les régions, pour former la description des caractéristiques de bas niveau.

Étant donné une région r_i, les auteurs estiment le contraste de dissimilarité avec toutes les autres régions en utilisant la distance de couleur $D_c(r_i, r_j)$ et la distance spatiale $D_s(r_i, r_j)$.

La second terme décrit la pertinence entre l'emplacement et la saillance, connu sous le nom de emplacement avant(location prior). Il est indépendant des caractéristiques visuelles et reflète les emplacements qu'un observateur remarque probablement sans aucune influence sur le contenu de l'image. Les auteurs adoptent la théorie du biais central[43] pour montrer que les observateurs accordent plus d'attention aux régions plus proches du centre du champ visuel. La distance spatiale pondérée $D_s(r_i, r_j)$ est définit comme suit

$$D_s(r_i, r_j) = \exp(-\frac{\|C_{ri} - C_{rj}\|}{\sigma}) \tag{2.23}$$

tel que C_{r_i} et C_{r_j} est le centre de la région r_i et r_j respectivement, $\|C_{r_i} - C_{r_j}\|$ est la distance euclidienne entre les deux centres, et σ contrôle la force de la distance spatiale qui est égale à 0.4.

Les auteurs ont considéré deux échelles différentes avec des facteurs d'échelle $\sigma_l = 1$ et $\sigma_l = 0,2$, ce qui signifie que les mêmes opérations sont mises en œuvre sur l'image originale et $1/5$ de sa taille (à la fois la hauteur et la largeur). Tout d'abord, la carte de la saillance initiale S_i basée sur le centre initial du champ visuel est estimée à partir de la petite échelle, puis le centre décalé est calculé. Enfin, deux cartes de saillance à deux échelles différentes sont mesurées. La carte de la saillance finale est obtenue comme suit :

$$S = \alpha S_l + (1 - \alpha)S_s \tag{2.24}$$

2.4.2 Métriques d'évaluation

L'évaluation de la qualité du modèle de saillance est effectuée selon l'approche adoptée par [POWERS, 2011]. Elle peut être interprétée comme un problème de classification, où chaque pixel de l'image d'entrée peut être classé comme saillant ou non saillant. Selon [POWERS, 2011], quatre métriques peuvent être calculés

FP(faux positifs) est le nombre de pixel non saillant, classés comme saillants.
FN(faux négatifs) est le nombre de pixel saillant, classés comme non saillants.

TP(true positives) est le nombre de pixels saillants, correctement classés.
TN(vrais négatifs) est le nombre de pixels non saillants, correctement classés.

A partir de ces mesures classiques, d'autres métriques peuvent être déduites :

La précision(PR). Elle indique la précision de la classification. Cette métrique évalue la qualité de la carte de saillance dans son ensemble.

$$\mathrm{PR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2.25}$$

Le rappel(RE). Le rappel indique la qualité de la détection des objets saillants.

$$\mathrm{RE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.26}$$

La mesure-F. Aussi connue sous le nom de F1. Cette mesure est un compromis de la précision et du rappel.

$$\mathrm{F} = 2 \times \frac{\mathrm{PR} \times \mathrm{RE}}{\mathrm{PR} + \mathrm{RE}} \tag{2.27}$$

Courbe Précision-Rappel(P-R Curve)

De nombreux travaux segmentent la carte de saillance avec un seuil allant de 0 à 1 avec un intervalle de 0,05. Ils calculent la précision et le rappel à chaque valeur du seuil pour tracer la courbe Précision-Rappel. En utilisant la carte de saillance A, la mesure convertit A en un masque binaire B. Ensuite, la précision et le rappel sont calculés en comparant B avec la vérité terrain GT. La principale étape de ce processus est la conversion de A en B. Les méthodes de seuil fixe et de seuil adaptatif sont utilisées pour effectuer la binarisation.

Erreur Absolue Moyenne (EAM)

Les métriques de précision et de rappel ne prennent pas en compte les véritables affectations de la saillance négative, c'est à dire le nombre de pixels correctement marqués comme non saillants. Cela favorise les méthodes qui attribuent avec succès la saillance aux pixels saillants mais ne parviennent pas à détecter les régions non saillantes par rapport aux méthodes qui détectent avec succès les pixels non saillants mais font des erreurs dans la détermination de celles saillantes. Pour une comparaison plus équilibrée qui tient compte de ces effets, l'erreur absolue moyenne(MAE) peut être calculée entre la carte de saillance et la vérité terrain, ces deux cartes doivent être normalisées dans l'intervalle [0,1].

.

2.4.3 Principales bases de données

SED Database

La base SED [ALPERT et collab., 2007] comprend un sous-ensemble à objet unique SED1 et un sous-ensemble à deux objets SED2. Chacun de ces deux sous ensembles

contient 100 images avec des annotations pixel par pixel. Les objets présents dans les images de cette base diffèrent de leur environnement par divers indices de bas niveau comme l'intensité et la texture. Chaque image a été segmenté par trois sujets. Un pixel est considéré appartenant au premier plan si au moins deux sujets sont d'accord sur ce point.

ASD Database

La base ASD [ACHANTA et collab., 2009] est la base de d'images la plus utilisée. Elle est relativement simple. La base ASD contient 1000 images avec des masques binaires. Les images sont sélectionnées depuis la base d'image MSRA-A [LIU et collab., 2007], où seules les zones de délimitation autour des régions saillantes sont fournies. Les masques des objets saillants de la base ASD sont créés en fonction des contours des objets.

SOD Database

La base SOD [MOVAHEDI et ELDER, 2010] contient 300 images de la base d'image de Berkeley BSDS300. Chaque image est étiquetée par sept sujets. Plusieurs images possèdent plus d'un objet saillant de faible contraste de couleur par rapport à l'arrière-plan ou qui touche les bords des images. Les annotations des pixels sont disponibles.

MSRA-5000 Database

La base de données MSRA (ou MSRA-5000) [LIU et collab., 2007] contient 5 000 images de l'ensemble d'images de la base d'objets saillants de Microsoft Research Asia. Ces images possèdent une large variation entre elles. Les masques binaires pixel par pixel des objets saillants [JIANG et collab., 2013]sont fournit comme vérité de terrain.

THUS10K Database

La base THUS10K [CHENG et collab., 2011], connue aussi sous le nom de MSRA10K, contient 10 000 images sélectionnées de la base MSRA. Elle contient toutes les 1 000 images de la base ASD. Les images sont dotées d'annotations pixel par pixel délimitant les objets saillants de manière cohérente. Cette base est largement utilisée pour entraîner les modèles de saillance basés sur les réseaux de neurones profonds.

PASCAL-S Database

La base de données PASCAL-S [LI et collab., 2014] est construit sur l'ensemble de validation de la base PASCAL VOC 2010 [EVERINGHAM et collab., 2010]. Ce sous-ensemble contient 850 images naturelles. Afin d'évaluer des modèles dédiés à la segmentation d'objets saillants, les auteurs effectuent d'abord manuellement une segmentation complète afin de délimiter tous les objets de l'image. Ensuite, ils construisent la vérité terrain de la segmentation complète. Étant donné une image, un sujet est invité à sélectionner les objets saillants par un simple clic. Aucune contrainte de temps n'est imposée sur le nombre d'objets qu'une personne peut choisir. La valeur de saillance finale de chaque segment est le nombre total de clics qu'il reçoit, divisé par le nombre de sujets.

DUT-OMRON Database

La base de données DUT-OMRON est constitué de 5 168 images de haute qualité qui sont sélectionnées manuellement parmi plus de 140 000 images. Les images contiennent un ou plusieurs objets saillants et un arrière-plan relativement complexe. 25 participants ont participé à la collecte des vérité terrain. Cette base est muni à la fois des fixations oculaires, des boîtes englobantes et des vérités terrain au niveau des pixels. Comparée à d'autres bases d'images comme ASD et MSRA , les images de cette base sont considérées plus difficiles et offrent plus d'espace d'amélioration pour des recherches connexes sur la détection de la saillance.

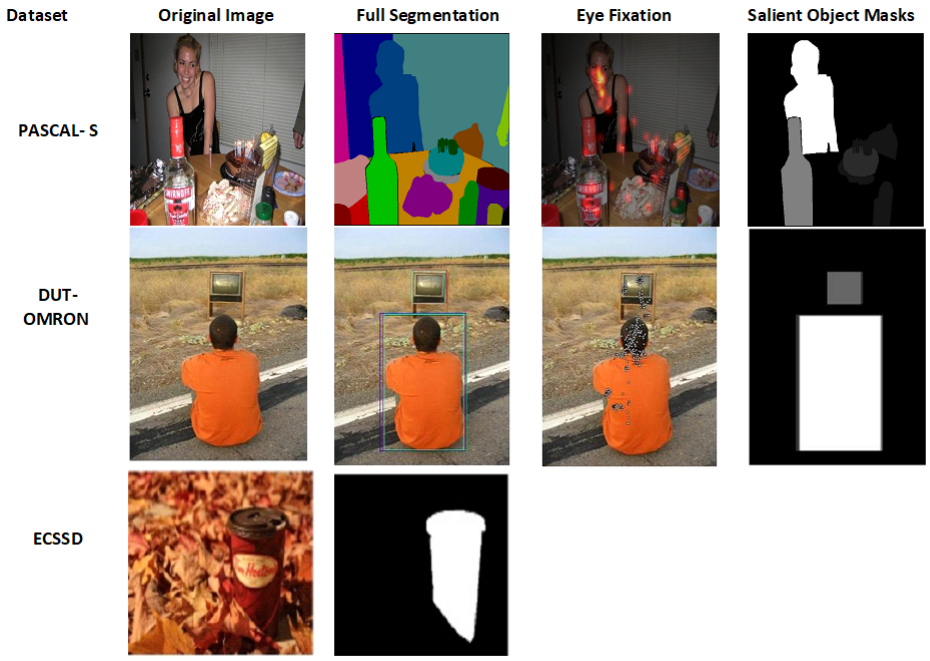


FIGURE 2.13 – Exemples de diverses bases d'images et de leur annotations associées.

Le tableau 2.1 récapitule les différents bases d'images dédiées à la détection de la saillance. Nous mettons en évidence les divers aspects qui peuvent caractériser une base d'image selon quelle soit dédiée à la tache de prédiction de fixation/segmentation d'objet saillant,

TABLEAU 2.1 – Synthèse des principales bases d'images dédiée à la détection de la saillance.

Base d'images	Tâche	Année de Année de Publication	Nombre d'image	Résolution Maximale	Objet Saillant Single/Multiple	Vérité terrain
MSRA	Détection d'objet(OD)	[LIU et collab., 2007]	25000		Single	Boites Englobantes
MSRA-5000	Détection d'objet saillant SOD	[LIU et collab., 2007]	5000		Single	Masque pixelique
MSRA-1000(ASD)	SOD	[ACHANTA et collab., 2009]	1K	400 × 400	Single	Masque pixelique
SOD	SOD	[MOVAHEDI et ELDER, 2010]			Single	Masque pixelique
SED	SOD	[ALPERT et collab., 2007]	SED1(100) SED2(100)		Single Two	Masque pixelique
ECSSD	SOD	[YAN et collab., 2013]	1K	400 × 400	Single Multiple	Masque pixelique
DUT-OMRON	SOD FP	[YANG et collab., 2013]	5168	400 × 400	Single	Bounding boxes Carte de fixation des yeux Masque pixelique
MIT	FP	[JUDD et collab., 2009]	900	1024 × 768	Free-viewing Single	Carte de fixation des yeux
OSIE	FP	[XU et collab., 2014]	700	800 × 600	Free-viewing Multiple	Carte de fixation
MSRA10K (THUS10K)	SOD	[CHENG et collab., 2011]	10 K	400 × 400	Single	Boites Englobantes Masque pixelique
THUR15K	SOD	[CHENG et collab., 2014]	15000		Single	Masque pixelique
PASCAL-S	SOD FP	[LI et collab., 2014]	1 K	500 × 500	Multiple	Full segmentation Carte de fixation des yeux Masque pixelique

2.5 Bilan et critiques

Notre étude bibliographique nous a permis de constater que plusieurs catégorisations des modèles de saillance ont été proposé dans la littérature. Toutefois, nous croyons qu'elles reprennent implicitement la même catégorisation de [ACHANTA et collab., 2008] qui classe les modèles de saillance selon trois approches majeures : les modèles biologiques, les modèles computationnels et les modèles hybrides.

La catégorie de modèles biologiques regroupe les modèles centre-périphérie qui reposent sur les principes de la théorie des traits caractéristiques(FIT) de [TREISMAN et GELADE, 1980] et le modèle d'attention de [KOCH et ULLMAN, 1985]. D'après [TAVAKOLI, 2014], ces modèles passent souvent par les trois étapes d'extraction des caractéristiques, de comparaison des caractéristiques de type centre-périphérie et de fusion des cartes de saillance pour obtenir la carte de saillance finale. Ces modèles ont été désigné sous d'autres noms comme les modèles cognitifs d'après [BORJI et collab., 2013], les modèles hiérarchiques d'après [DA SILVA, 2010], les modèles psychophysiques d'après [TOET, 2011].

Les modèles de saillance biologiques tentent de reprendre les principes de base des théories psychologiques ainsi que des modèles d'attention visuelle. Les mesures de saillance les plus utilisées sont le contraste de couleur/intensité, la localisation spatiale. Le calcul du contraste porte généralement sur les trois caractéristiques visuelles de bas niveau : l'intensité, la couleur(Lab) et l'orientation(filtre de gabor). Le contraste local est modélisé par une différence de gaussienne(DoG), à différentes échelles, et pour différentes caractéristiques visuelles, afin d'imiter le champ récepteur des cellules de type centre-périphérie du système visuel humain (cartes de caractéristiques). Les différentes cartes de caractéristiques sont normalisées et linéairement combinées pour créer la carte de saillance finale.

Les modèles de saillance purement computationnel introduisent différentes théories issues de domaines différents de traitement de signal comme la théorie de Fourier, du domaine des probabilités, Théorie de l'information, l'apprentissage automatique. Le calcul des mesures de saillance porte généralement sur les trois caractéristiques visuelles de bas niveau : l'intensité, la couleur(Lab). Les mesures de saillance les plus utilisées pour calculer la carte de saillance sont le contraste d'intensité/de couleur, la localisation spatiale, la fréquence. La mesure de contraste local est calculée par la différence des vecteurs des caractéristiques d'intensité/de couleur d'une région r_i par rapport à sa région pourtour r_j. La mesure de contraste global est calculée par la différence des vecteurs des caractéristiques d'intensité/de couleur d'une région r_i par rapport à son voisinage, c'est à dire tout les pixels de l'image qui n'appartiennent pas à la région r_i. La mesure de localisation spatiale est calculé par la distance spatiale, généralement Euclidienne, entre les deux centres de régions C_{r_i} et C_{r_j} dans une image. Cette mesure de saillance est utilisé pour montrer que les observateurs accordent plus d'attention aux régions les plus proches du centre du champ visuel. Les différentes cartes de saillance préliminaires sont normalisées et linéairement combinées pour créer la carte de saillance finale.

Les modèles de saillance reposent sur des principes biologiques issus des théories psychologiques et des neurosciences combinés à des concepts purement computationnel que l'on retrouve dans la théorie du signal, la théorie de l'information, la théorie des probabilité, la théorie de graphe, l'apprentissage automatique.

Les catégorisations de [LI et collab., 2014],[BORJI et collab., 2015] classent les mo-

dèles de saillance selon qu'ils soient des modèles de prédiction de fixation ou des modèles de détection d'objet saillant. Cependant, nous avons constaté que ces deux catégories de modèles de saillance passent par une étape d'extraction des caractéristiques, une étape de calcul d'une carte de saillance, une étape de normalisation de cette carte ainsi que la fusion des éventuelles cartes de saillance pour obtenir la carte de saillance finale, à l'exception d'une étape de segmentation antérieure et postérieure qu'on retrouve dans les modèles de détection d'objet saillant. Toutefois, le choix de la méthode de segmentation est critique car il peut influencer sur le calcul des indices de saillance et par conséquent sur la qualité de la carte de saillance.

Le tableau 2.2 récapitule les différents modèles de saillance étudiés dans ce chapitre. Nous mettons en évidence les divers aspects qui peuvent influencer la qualité des modèles de saillance comme les caractéristiques visuelles utilisées, les mesures de saillance calculées(contraste d'intensité/de couleur, centre prior),le mécanisme adopté dans leur calcul, la méthode de segmentation adoptée, le nombre de cartes de saillance calculées.

Toutes ces constatations ont soulevé notre attention dans la mesure de proposer un modèle de saillance qui d'une part se baserait sur une segmentation en considérant un certain regroupement des pixels de l'image, et d'autre part, se baserait sur une approche connexioniste en utilisant un réseau de neurone dans la construction de la carte de saillance.

TABLEAU 2.2 – Aperçu des méthodes proposées dans le domaine de la détection de la saillance.

Publication	Segmentation	Caractéristiques Bas Niveau	Carte de caractéristiques	Approche	Mécanisme	Mono/Multi Échelle	Local Global	Carte de saillance
ITTI et collab. [1998] (IT)	Aucune	Color Intensité Orientation	Couleur Intensité Orientation	Biologique	Centre-Périphérie (DoG)	Multiple	Local	1.Carte de saillance de contraste couleur 2.Carte de saillance de contraste d'intensité 3.Carte de saillance de contraste d'orientation Combinaison linéaire
[HAREL et collab., 2006] (GB)	Aucune	Couleur Intensité Orientation	Couleur Intensité Orientation	Hybride	1.Centre-Périphérie(DoG) 2.Théorie des Graphes	Multiple	Local	1.Carte de saillance de contraste couleur 2.Carte de saillance de contraste d'intensité 3.Carte de saillance de contraste d'orientation Combinaison linéaire
[HOU et ZHANG, 2007] (SR)	Aucune	Intensité	Log-spectrum	Purement Computationnel	Analyse Spectrale (Transformée de Fourier)	Mono	Global	1.Carte de résidu spectral
[MA et ZHANG, 2003a]	Aucune	Couleur Intensité Orientation	Aucune	Purement Computationnel	Centre-Périphérie (Distance de caractéristiques)	Mono	Local	1.Carte de saillance de contraste couleur 2.Carte de saillance d'intensité couleur 3.Carte de saillance d'orientation Combinaison linéaire
[ACHANTA et collab., 2008]	Aucune	Couleur(Lab) Intensité	Filtrage	Purement Computationnel	Center-Périphérie (Distance de caractéristiques)	Multiple	Local	Trois cartes de saillance basée filtrage Combinaison linéaire
[ACHANTA et collab., 2009] (FT)	Aucune	Couleur(Lab) Intensité	Image Floue Gaussienne Image de couleur moyenne	Purement Computationnel	Centre-Périphérie (Distance de caractéristiques) Analyse de Fréquence	Mono	Global	1.Carte de saillance de contraste couleur

TABLEAU 2.2 – Aperçu des méthodes proposées dans le domaine de la détection de la saillance.

Publication	Segmentation	Caractéristiques Bas Niveau	Carte de caractéristiques	Approche	Mécanisme	Mono/Multi Échelle	Local Global	Carte de saillance
[ACHANTA et SÜSSTRUNK, 2010] (SS)	Aucune	Couleur(Lab) Intensité	Image Floue Gaussienne Image de couleur moyenne	Purement Computationnel	Centre-Périphérie Symétrique (Distance de caractéristique) Analyse de Fréquence	Mono	Local	1.Carte de saillance de contraste couleur
[CHENG et collab., 2011](HC)	Aucun	Couleur(Lab)	Histogramme de couleur	Purement Computationnel	Distance de caractéristiques	Mono	Global	1.Carte de saillance de contraste couleur
[CHENG et collab., 2011](RC)	Basée sur les graphes	Couleur(Lab)	Aucune	Purement Computationnel	Centre-Périphérie (Distance de caractéristiques)	Mono	Global	1.Carte de saillance de contraste couleur
[YANG et collab., 2013]	Basée sur les graphes	Couleur(Lab)		Purement Computationnel	Distance spatiale pondérée (Euclidienne)	Multiple	Local	1.Carte de saillance de contraste couleur 2.Carte de saillance de location Combinaison linéaire
[HOU et collab., 2012](IS)	Aucune	Couleur(CIELab) Intensité	Aucune	Purement Computationnel	Analyse Spectrale (DCT)	Mono	Global	1.Carte spectrale de phase(Canal L) 2. Carte spectrale de phase(Canal a) 3.Carte spectrale de phase(Canal b) Combinaison linéaire
[PERAZZI et collab., 2012]	SLIC Superpixel	Intensité	Aucune	Purement Computationnel	Basée sur filtrage Gaussien	Mono	Global	1.Carte de saillance de contraste d'intensité
[FU et collab., 2013]	K-Moyenne	1.Couleur(Lab) 2.Texture (Filtre de Gabor)	None	Purement Computationnal	Center-Périphérie (Distance de caractéristiques) 2. Fonction de Gaussienne 2D	Mono	Local	1.Carte de saillance de contraste couleur 2.Carte de saillance de location spatiale Combinaison linéaire

TABLEAU 2.2 – Aperçu des méthodes proposées dans le domaine de la détection de la saillance.

Publication	Segmentation	Caractéristiques Bas Niveau	Carte de caractéristiques	Approche	Mécanisme	Mono/Multi Échelle	Local Global	Carte de saillance
[ZHANG et collab., 2013]	Aucune	Couleur(Lab) Intensité Texture (Filtre de Log-Gabor)	Aucune	Purement Computationnel	1. Couleurs Chaudes/Froides 2.Analyse de fréquence 3.Fonction de Gaussienne 2D	Mono	Global	1.Carte de saillance couleur 2.Carte de saillance de fréquence 3. Carte de saillance de location spatiale Combinaison linéaire
[IMAMOGLU et collab., 2013](TMM)	Aucune	Couleur(Lab) Intensité	Ondelette	Purement Computationnel	Transformée d'ondelette	Multiple	Local Global	1.Carte de saillance globale 2.Carte de saillance local Combinaison linéaire

2.6 Conclusion

Dans ce chapitre, nous avons présenté une revue des travaux les plus répandus qui se sont intéressé à la modélisation de la saillance visuelle dans le domaine de la vision artificielle. Nous avons essayée de diversifier l'analyse afin de montrer les aspects importants caractérisant ces modèles. Au terme de cette revue, nous avons analysé certains travaux selon la catégorisation proposé par [BORJI et collab., 2013] et qui les divisent en deux catégories. Ceci nous a permis de ressortir les caractéristiques communes de chacune de ces catégories, mais également de bien identifier les points qui les différencient. Cette étude nous a permis de mieux distinguer les caractéristiques qui décrivent les modèles biologiques de ceux purement computationnel, dans un but de pouvoir proposer un modèle de saillance qui tient compte de ces caractéristiques.

Références

ACHANTA, R. 2011, *Finding Objects of Interest in Images using Saliency and Superpixels*, thèse de doctorat, École polytechniue fédérale de Lausanne, Suisse. Thèse de doctorat en Informatique, Communications et Information. 39

ACHANTA, R., F. J. ESTRADA, P. WILS et S. SÜSSTRUNK. 2008, «Salient Region Detection and Segmentation», dans *Proceedings of the 6th International Conference on Computer Vision Systems (ICVS'08)*, vol. 5008, Springer Lecture Notes in Computer Science, Springer,Berlin,Heidelberg, p. 66–75. ix, 39, 51, 65, 67

ACHANTA, R., S. HEMAMI, F. ESTRADA et S. SUSSTRUNK. 2009, «Frequency-tuned Salient Region Detection», dans *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition(CVPR'09)*, IEEE, p. 1597–1604. ix, 52, 53, 58, 62, 64, 67

ACHANTA, R. et S. SÜSSTRUNK. 2010, «Saliency detection using maximum symmetric surround», dans *Proceedings of the 17th IEEE International Conference on Image Processing(ICIP'10)*, IEEE, p. 2653–2656. ix, 53, 68

ALPERT, S., M. GALUN, R. BASRI et A. BRANDT. 2007, «Image segmentation by probabilistic bottom-up aggregation and cue integration», *IEEE Conference on Computer Vision and Pattern Recognition*, p. 1–8. 61, 64

AZIZ, M. Z. et B. MERTSCHING. 2008, «Fast and robust generation of feature maps for region-based visual attention», *IEEE transactions on image processing*, vol. 17, nº 5, p. 633–644. 40

BORJI, A., M. CHENG, H. JIANG et J. LI. 2015, «Salient object detection : A benchmark», *IEEE Transactions on Image Processing*, vol. 24, p. 5706–5722. 39, 41, 65

BORJI, A., D. N. SIHITE et L. ITTI. 2013, «What stands out in a scene? a study of human explicit saliency judgment», *Vision Research*, vol. 91, p. 62–77. 40, 41, 48, 65, 70

BRUCE, N. D. B. et J. K. TSOTSOS. 2005, «Saliency based on information maximization», dans *Advances in neural information processing systems*, p. 155–162. 48

BRUCE, N. D. B. et J. K. TSOTSOS. 2009, «Saliency, attention, and visual search : An information theoretic approach», *Journal of Vision*, vol. 9, nº 3, p. 1–24. 40

CHENG, M., G. ZHANG, N. MITRA, X. HUANG et S. HU. 2011, «Global contrast based salient region detection», dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*, IEEEComputer Society, p. 409–416. 54, 56, 62, 64, 68

CHENG, M.-M., N. MITRA, X. HUANG et S.-M. HU. 2014, «Salientshape : group saliency in image collections», *The Visual Computer*, vol. 30, nº 4, p. 443–453. 64

DENG, Y., C. KENNEY, M. S. MOORE et B. S. MANJUNATH. 1999, «Peer group filtering and perceptual color image quantization», *ISCAS'99. Proceedings of the 1999 IEEE International Symposium on Circuits and Systems VLSI (Cat. No.99CH36349)*, vol. 4, p. 21–24 vol.4. 44

EVERINGHAM, M., L. GOOL, C. K. WILLIAMS, J. WINN et A. ZISSERMAN. 2010, «The pascal visual object classes (voc) challenge», *International Journal Computer Vision*, vol. 88, nº 2, p. 303—-338. 49, 62

FRINTROP, S., M. KLODT et E. ROME. 2007, «A real-time visual attention system using integral images», dans *International Conference on Computer Vision Systems : Proceedings (2007)*. 39

FU, H., X. CAO et Z. TU. 2013, «Cluster-based co-saliency detection», *IEEE Transactions on Image Processing (TIP)*, vol. 22, nº 10, p. 3766–3778. ix, 56, 57, 68

GOFERMAN, S., L. ZELNIK-MANOR et A. TAL. 2012, «Context-aware saliency detection», *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI'10*, vol. 34, nº 10, p. 1915–1926. 54

GUO, C., Q. MA et L. ZHANG. 2008, «Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform», dans *Proccedings of IEEEConference on Computer Vision and Pattern Recognition (CVPR'08)*, IEEEComputer Society, p. 1–8. ix, 47

GUO, C. et L. ZHANG. 2010, «A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression», *IEEE Transactions on Image Processing*, vol. 19, p. 185–198. 48

HAREL, J., C. KOCH et P. PERONA. 2006, «Graph-based visual saliency», dans *Advances in Neural Information Processing Systems 19(NIPS)*, p. 545—-552. ix, 39, 45, 46, 67

HOU, X., J. HAREL et C. KOCH. 2012, «Image signature : Highlighting sparse salient regions», *IEEETransactions on Pattern Analysis and Machine Intelligence(PAMI'12)*, vol. 34, nº 1, p. 194–201. ix, 48, 68

HOU, X. et L. ZHANG. 2007, «Saliency detection : A spectral residual approach», dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR'07*, IEEE, p. 1–8. ix, 46, 47, 67

HU, Y., X. XIE, W.-Y. MA, L.-T. CHIA et D. RAJAN. 2004, «Salient region detection using weighted feature maps based on the human visual attention model», dans *Pacific-Rim Conference on Multimedia*, Springer, p. 993–1000. 39

IMAMOGLU, N., W. LIN et Y. FANG. 2013, «A saliency detection model using low-level features based on wavelet transform», *IEEE transactions on multimedia*, vol. 15, nº 1, p. 96–105. ix, 59, 69

ITTI, L. 2000, *Models of Bottom-Up and Top-Down Visual Attention*, thèse de doctorat, California Institute of Technology. 40

ITTI, L. et P. BALDI. 2005, «A principled approach to detecting surprising events in video», dans *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, p. 631–637. 40

ITTI, L. et C. KOCH. 1999, «Comparison of feature combination strategies for saliency-based visual attention systems.», dans *Human Vision and Electronic Imaging(HVEI)*, vol. 3644, p. 473–482. 39

ITTI, L., C. KOCH et E. NIEBUR. 1998, «A model of saliency-based visual attention for rapid scene analysis», *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI'98)*, vol. 20, p. 1254–1259. ix, 39, 40, 43, 44, 45, 67

JIANG, Z., Z. LIN et L. DAVIS. 2013, «Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, nº 11, p. 2651–2664. 62

JUDD, T., F. DURAND et A. TORRALBA. 2012, «A benchmark of computational models of saliency to predict human fixations.», cahier de recherche, Massachusetts institute of technology. 40

JUDD, T., K. A. EHINGER, F. DURAND et A. TORRALBA. 2009, «Learning to predict where humans look», *2009 IEEE 12th International Conference on Computer Vision*, p. 2106–2113. 49, 50, 64

KOCH, C. et S. ULLMAN. 1985, «Shifts in selective visual attention : Towards the underlying neural circuitry», *Human Neurobiology*, vol. 4, p. 219–227. ix, 40, 41, 42, 43, 65

LI, J., M. D. LEVINE, X. AN, X. XU et H. HE. 2013, «Visual saliency based on scale-space analysis in the frequency domain», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, p. 996–1010. 49

LI, Y., X. HOU, C. KOCH, J. M. REHG et A. L. YUILLE. 2014, «The secrets of salient object segmentation», *2014 IEEE Conference on Computer Vision and Pattern Recognition*, p. 280–287. 41, 49, 62, 64, 65

LIU, T., Z. YUAN, J. SUN, J. WANG, N. ZHENG, X. TANG et H. SHUM. 2007, «Learning to detect a salient object», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, p. 353–367. 62, 64

MA, Y. F. et H. J. ZHANG. 2003a, «Contrast-based image attention analysis by using fuzzy growing», dans *Proceedings of the Eleventh ACMInternational Conference on Multimedia*, ACM, p. 374–381. 39, 67

MA, Y. F. et H. J. ZHANG. 2003b, «Contrast-based image attention analysis by using fuzzy growing», dans *Proceedings of the 11th ACMInternational Conference on Multimedia*, ACM, p. 374–381. 44

MARAT, S., A. RAHMAN, D. PELLERIN, N. GUYADER et D. HOUZET. 2012, «Improving visual saliency by adding 'face feature map' and 'center bias'», *Cognitive Computation*, vol. 5, p. 63–75. 50

MOVAHEDI, V. et J. H. ELDER. 2010, «Design and perceptual validation of performance measures for salient object segmentation», *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, p. 49–56. 62, 64

NAVALPAKKAM, V. et L. ITTI. 2006, «An integrated model of top-down and bottom-up attention for optimizing detection speed», *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, p. 2049–2056. 44

OLIVA, A., A. TORRALBA, M. S. CASTELHANO et J. M. HENDERSO. 2003, «Top-down control of visual attention in object detection», dans *Proceedings of IEEE International Conference on Image processing ICIP'03*, vol. 1, IEEE, p. I–253. 40

ORABONA, F., G. METTA et G. SANDINI. 2007, «A proto-object based visual attention model», dans *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, p. 198–215. 40

PERAZZI, F., P. KRÄHENBÜHL, Y. PRITCH et A. HORNUNG. 2012, «Saliency filters : Contrast based filtering for salient region detection», dans *2012 IEEE Conference on Computer Vision and Pattern Recognition*, p. 733–740. 56, 68

PETERS, R. J., A. IYER, L. ITTI et C. KOCH. 2005, «Components of bottom-up gaze allocation in natural images», *Vision Research*, vol. 45, p. 2397–2416. 50

POWERS, D. 2011, «Evaluation : from precision, recall and f-measure to roc, informedness, markedness and correlation», *ArXiv*, vol. abs/2010.16061. 60

DA SILVA, M. P. 2010, *Modèle computationnel d'attention pour la vision adaptative.*, thèse de doctorat, Université de La Rochelle. 39, 65

TAVAKOLI, H. 2014, *Visual saliency and eye movement : modeling and applications.Thése en informatique.*, thèse de doctorat, University of Oulu, Finland. 41, 65

TOET, A. 2011, «Computational versus psychophysical bottom-up image saliency : A comparative evaluation study», *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, nº 11, p. 2131–2146. 40, 65

TREISMAN, A. et G. GELADE. 1980, «A feature-integration theory of attention», *Cognitive Psychology*, vol. 12, nº 1, p. 97–136. 41, 42, 65

VITAY, J., N. P. ROUGIER et F. ALEXANDRE. 2005, «A distributed model of spatial visual attention», dans *Biomimetic Neural Learning for Intelligent Robots*, Springer, p. 54–72. 40

XU, J., M. JIANG, S. WANG, M. KANKANHALLI et Q. ZHAO. 2014, «Predicting human gaze beyond pixels.», *Journal of vision*, vol. 14 1, nº 28, p. 1–20. 64

YAN, Q., L. XU, J. SHI et J. JIA. 2013, «Hierarchical saliency detection», *IEEE Conference on Computer Vision and Pattern Recognition*, p. 1155–1162. 64

YANG, C., L. ZHANG, H. LU, X. RUAN et M. H. YANG. 2013, «Saliency detection via graph-based manifold ranking», dans *IEEE Computer Vision and Pattern Recognition (CVPR)*, p. 3166—-3173. 59, 64, 68

ZHANG, L., Z. GU et H. LI. 2013, «Sdsp : A novel saliency detection method by combining simple priors», dans *Proceedings of the 20th IEEE International Conference on Image Processing(ICIP'13)*, IEEE, p. 171–175. ix, 57, 58, 69

Chapitre 3

Approches de segmentation d'objet : Un état de l'art

« *La connaissance s'acquiert par l'expérience tout le reste n'est que de l'information.* »

Albert Einstein

Sommaire

3.1 Introduction

La segmentation d'image est un domaine de recherche important et difficile à la fois. Plusieurs systèmes de vision artificielle dépendent d'une étape de segmentation. D'après notre étude sur la détection de la saillance dans le chapitre 2, nous avons constaté que certains auteurs définissent cette tâche comme un problème de segmentation binaire dont l'objectif est de générer un masque binaire qui sépare l'objet de premier plan de l'arrière-plan d'une image. A cette issue, plusieurs modèles de détection et de segmentation d'objet(s) saillant(s) ont vu le jour. Cependant, il est moins clair comment cette nouvelle définition se rapporte au domaine de la segmentation d'image et plus particulièrement à celui de la segmentation d'objet. Dans ce chapitre, nous abordons le domaine de la segmentation d'image que nous décrivons selon trois niveaux différents : bas niveau, niveau intermédiaire et haut niveau. Nous nous intéressons plus particulièrement au niveau intermédiaire à savoir la segmentation d'objet. Nous présentons les différentes approches de segmentation d'objet associées à des travaux réalisés dans ce domaine.

3.2 Niveaux de segmentation

Dans cette section, nous décrivons la segmentation d'image selon trois niveaux : la segmentation d'image bas-Niveau, la segmentation d'objet et la segmentation haut-Niveau(ou sémantique).

3.2.1 Segmentation bas-Niveau

La segmentation d'image est abordée par la communauté de la vision artificielle comme une opération purement bas-niveau. Elle se définit par le partitionnement de l'ensemble des pixels constituant une image en différentes partitions. Quand il s'agit de partitionner les pixels en deux partitions distinctes seulement, on parle alors de segmentation binaire. La tâche de segmentation revient à découper une image, en régions deux à deux disjointes et dont l'union permet de reconstituer toute l'image. En se basant sur cette définition, on constate que le concept région est l'élément primordial de la description délivrée par une segmentation. Cette démarche a donné naissance aux approches de segmentation dites orientées-région. Formellement, soit I une image et E une partition de I constituée de sous régions connexes tel que

$$\begin{aligned} &\mathrm{E} = \{\mathrm{R}_1, \mathrm{R}_2, \ldots, \mathrm{R}_n\}, \mathrm{R}_i \neq \emptyset, \mathrm{R}_i \; connexes \; \forall i = 1, \cdots, n \\ &\mathrm{R}_i \cap \mathrm{R}_j = \emptyset, \forall (i, j), i \neq j, \\ &\mathrm{I} = \cup \mathrm{R}_i, \forall i = 1, \ldots, n. \end{aligned} \tag{3.1}$$

Selon [Fu et Mui, 1981],[Freixenet et collab., 2002], il est possible de distinguer quatre familles principales de méthodes de segmentation d'image :(1)les méthodes orientées régions, (2)les méthodes orientées contours, (3)les méthodes de seuillage et (4)les méthodes de classification. Les méthodes orientées régions utilisent des critères d'homogénéités. Soit P un prédicat d'homogénéité appliqué aux pixels. E est une segmentation de I, selon le prédicat P, si seulement si

$$\begin{aligned} \mathrm{P}(\mathrm{R}_i) &= vrai, \forall i = 1, \ldots, n \\ \mathrm{P}(\mathrm{R}_i \cup \mathrm{R}_j) &= faux, \forall i \neq j, \mathrm{R}_i \; et \; \mathrm{R}_j \; sont \; adjacents. \end{aligned} \tag{3.2}$$

Les méthodes basées sur la croissance des régions [ADAMS et BISCHOF, 1994] font elles aussi parties des méthodes orientées région vue qu'elles s'appuient sur des prédicats d'homogénéité pour construire les régions d'intérêt. A partir d'un ensemble de pixels initialement sélectionnés(ces pixels sont connus sous le nom de germes ou graines), des régions commencent à se former et croître en accumulant des pixels autour de ces graines. L'ajout d'un pixel dans une région se fait selon un critère d'homogénéité. La croissance des régions s'arrête une fois qu'aucun pixel ne peut être ajouté à aucune des régions. Il existe d'autres méthodes orientées régions comme les méthodes basées sur la division et/ou fusion des régions. Le principe de base des méthodes basées sur la division des régions [OHLANDER et collab., 1978]est qu'une image est divisée récursivement en sous images de plus petite taille jusqu'à vérifier un certain critère d'homogénéité. La division peut s'effectuer avec différentes méthodes comme celle des quadtree [KELKAR et GUPTA, 2008]. Les méthodes basées sur la fusion des régions [HONG et ROSENFELD, 1984] considèrent initialement chaque pixel de l'image comme une région. À chaque itération, des paires de régions sont fusionnées pour former un région plus large. La décision de fusion est basée sur un critère local d'homogénéité qui permet de calculer une mesure de similarité entre les régions voisines d'une image. Si la valeurs de la mesure de similarité est inférieure à un seuil alors les régions sont fusionnées.

La segmentation d'image revient également à délimiter les régions d'intérêt qui la composent. Cette définition a donné naissance aux approches orientées contour/frontière qui admettent que le passage d'une région à une autre région adjacente se traduit par une variation (ou transition) rapide dans l'intensité lumineuse de ces deux régions. Les premiers détecteurs de contours se sont intéressés à l'approximation du gradient d'une image discrète à travers la convolution de l'image avec des filtres de petites dimensions. Ces détecteurs dépendent de la taille des objets traitées et sont fortement sensibles au bruit. Le détecteur de Canny a été proposé afin de définir des critères d'optimalité pour la détection de contours à travers une opération de filtrage.

Malgré le grand développement qu'ont connu les détecteurs de contours, ils produisent des contours incomplets et non fermés [YOUSFI, 2008]. Par conséquence, des approches globales ont émergé introduisant la notion de modèle de contour déformable, comme le contour actif(snake en anglais). Il est formé d'une série de points mobiles, répartis sur une courbe en deux dimensions. La courbe (qui peut être fermée) est placée dans la zone d'intérêt de l'image ou autour d'un objet. Plusieurs équations décrivent son évolution : la courbe se déplace et délimite progressivement les contours des objets en fonction de plusieurs paramètres comme l'élasticité, la tolérance au bruit, ..., etc. Le déplacement du contour initial vers l'objet d'intérêt se fait grâce à la minimisation d'une énergie (somme d'une énergie interne et d'une énergie externe) sous certaines contraintes qui garantissent la régularité de la courbe tout en autorisant des déformations. Cependant, ces approches dépendent de plusieurs paramètres rendant la phase d'initialisation assez contraignante et par conséquent non-automatique. De plus, ces méthodes sont dédiées à la segmentation d'un seul objet ou d'un ensemble d'objets de nombre réduit.

Les méthodes de seuillage quand à elles procèdent à un traitement global de l'image qui consiste à faire un seuillage sur les intensité de l'image afin de déterminer les objets qui possèdent une intensité d'image proche. Les méthodes de classification(plus souvent appelées méthodes de clustering en anglais) consistent à regrouper en sous en-

sembles les pixels ayant des caractéristiques proches. Citons par exemple les méthodes de k-moyennes(k-means en anglais) qui consistent à séparer les pixels en k groupes en minimisant une distance entre un pixel donné et le représentant d'un groupe. Cependant, ces méthodes fournissent une partition de l'image mais ne permettent pas de distinguer quelles sont les classes faisant parties de l'objet d'intérêt et quelles sont les classes qui font partie du fond.

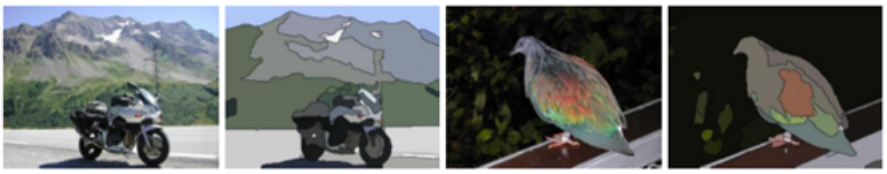


FIGURE 3.1 – Exemples de segmentation d'images bas-Niveau en utilisation une approche basée région [ZOU et collab., 2014].

Généralement, les méthodes de segmentation de bas-Niveau ne prennent pas en considération la notion d'objet explicitement sans extension adaptée. Elles ont du mal à traiter de l'aspect sémantique et sont confronté à la subjectivité des résultats de segmentation.
Parmi les bases d'images dédiées à l'évaluation des méthodes de segmentation de bas-niveau, nous pouvons citer : La base d'images Berkeley Segmentation Data Set BSDS300 [MARTIN et collab., 2001]contient 300 images associées à leur vérité terrain avec au moins 4 segmentations humaines par image. Les segments correspondent généralement à des zones de l'image avec une couleur ou une texture homogène mais ne définissent pas nécessairement des régions correspondant à des objets. De plus, il n'y a pas de distinction entre les segments d'objet et d'arrière-plan. La base d'images Berkeley Segmentation Data Set BSD500 comprend 500 images avec 5 à 10 vérités de terrain étiquetées manuellement pour chaque image. Les bases d'images BSDS300 et BSDS500 sont toutes les deux destinés à la détection des contours et à la segmentation générale de l'image.

3.2.2 Segmentation d'objet

La segmentation d'objet, connue aussi sous le nom de segmentation figure-fond vise à extraire le premier plan de l'arrière-plan d'une image, où le premier plan désigne la région contenant l'objet le plus significatif de l'image. Cependant, bien qu'il soit simple pour une personne d'extraire le premier plan, cette tâche est vraiment complexe à gérer par un ordinateur. Un premier plan peut être une région de couleur uniforme ou une région texturée.

FIGURE 3.2 – Exemples de segmentation d'objet dans des images(séparation figure/fond)[ZOU et collab., 2014]

D'après la littérature, trois approches différentes de segmentation d'objet (ou fi-

gure/fond) se distinguent : approche automatique, approche semi-automatique(interactive) et approche totalement automatique que nous évoquons dans les sections qui suivent.
Parmi les bases d'images dédiés à la segmentation d'objet, nous pouvons citer les bases d'images suivantes :
La base d'images de [McGuinness, 2009] contient 100 images sélectionnées de l'ensemble des images de la base de segmentation d'image Berkeley BSD-300. Les masques d'objets ont été créés manuellement à l'aide d'une tablette graphique et du logiciel de traitement d'images GNU(GIMP).
La base d'images WSED [Alpert et collab., 2007] contient 100 images avec des objets de premier plan qui diffèrent de leur environnement en termes d'intensité et de texture. Les images représentent clairement l'objet de premier plan et chaque image possède trois vérités de terrain fondamentales qui sont segmentées par trois personnes différentes.
La base d'images Object Segmentation Database(OSD) [Richtsfeld, 2012] contient 111 images avec des annotations par pixel pour chacune d'elle afin d'évaluer les approches de segmentation d'objets. Cette base d'image a été crée pour segmenter des objets à partir de scènes génériques même sous des occlusions partielles. Cependant, vue que la base OSD ne différencie pas entre les catégories d'objets différents,ses classes d'objets ont été réduites à deux classes seulement, objet et non objet.

3.2.3 Segmentation sémantique

Contrairement à la segmentation d'objet, la segmentation sémantique vise à extraire plusieurs objets d'intérêt de l'arrière plan d'une image. Elle associe à chaque pixel d'une image une étiquette(label) qui indique la classe des objets présents dans l'image. De nombreuses applications de la vision artificielle telle que la reconnaissance d'objets, l'interprétation des images dépendent d'une bonne segmentation sémantique d'objets.

FIGURE 3.3 – Exemples de segmentation sémantique d'image [Zou et collab., 2014].

Parmi les bases d'images dédiés à la segmentation sémantiques, nous pouvons citer par exemple, la base d'images PASCAL dont la plupart de ses annotations possède un large arrière plan rendant ainsi la segmentation sémantique presque équivalente à la segmentation d'objet. Comparée à la base d'image PASCAL, la base d'image Cityscapes fournit des annotations plus détaillées. Cependant, elle ne contient que des scènes de rue urbaines et n'est pas appropriée pour une segmentation sémantique générique. La base d'images ADE20K qui contient 150 classes sémantiques et 20 000 images de différents types de scène, est dont ses annotations sont denses est devenu dernièrement une base de référence dans la segmentation sémantique des images.

Le tableau 3.2 récapitule les différents bases d'images dédiées à la segmentation d'image selon les trois niveaux de segmentation évoqués ci-dessus .

TABLEAU 3.1 – Aperçu des bases d'images proposées selon les trois niveaux de segmentation : bas-Niveau, Intermédiaire(Objet) et haut-Niveau(Sémantique).

Base d'images	Année	Tache	Nombre d'images	Vérité Terrain
BSDS300	2001	Segmentation d'image and détection de contour	300	Annotation humaine
BSDS500	2011	Segmentation d'image and détection de contour	500	Annotation humaine
BSDS300	2001	Segmentation d'objet	10	Annotation Pixelique
iCoseg		Segmentation d'objet	10	Annotation Pixelique
Flower		Segmentation d'objet		Annotation Pixelique
Corel-1000		Segmentation d'objet	1000 (10 classes)	Annotation Pixelique
GrabCut	2004	Segmentation d'objet	50	Annotation Pixelique
Microsoft Research Asia	2007	Segmentation d'objet	5000	Annotation Pixelique
[MCGUINNESS, 2009]	2009	Segmentation d'objet	100	Annotation Pixelique
OSD	2012	Segmentation d'objet	111	Annotation Pixelique
MSRC		Segmentation sémantique	660 (21 classes)	Annotation Pixelique
PASCAL VOC	2010	Segmentation sémantique Segmentation d'objet	20 classes 9963 images	Annotation Pixelique
Microsoft COCO		Segmentation sémantique	91 classes 250.000	Annotation Pixelique
Cityscapes		Segmentation sémantique	Urban street scenes	Annotation Pixelique
ADE20K	2017	Segmentation sémantique	150 classes(stuff) 20.000	Annotation Pixelique (masques de partie et d'objet)

3.3 Segmentation d'objet interactive

La segmentation interactive, connu aussi sous le nom de segmentation semi-automatique a été un vaste domaine de recherche en analyse d'image et plus particulièrement dans le domaine de la segmentation d'objet. Elle a suscité l'intérêt de beaucoup de chercheurs en vision par ordinateur. Plusieurs travaux ont été réalisés et plusieurs algorithmes ont été conçus pour la segmentation interactive d'image jusqu'à présent. Dans ce qui suit, nous présentons certains travaux réalisés dans ce contexte.

Travaux de Boykov et al (2001)

Dans leur travaux, [BOYKOV et JOLLY, 2001] ont proposé une méthode de coupe de graphe interactive pour la segmentation d'objet. À l'aide de la souris, un utilisateur marque certain pixels, appelés graines, comme appartenant à la classe objet(avec une étiquette 1) et d'autres appartenant à la classe arrière-plan(avec une étiquette 0). Cet étiquetage impose des contraintes dures pour la segmentation de sorte que les pixels sélectionnés ne peuvent changer de classe. En considérant le processu de labelisation binaire des pixels d'une image $L = \{L_i, x_i \in I\}$, $L_i = 0$ si x_i est un pixel qui appartient à la classe arrière plan et $L_i = 1$ si x_i est un pixel qui appartient à la classe objet.

Un graphe orienté G =(V,E) composé d'un ensemble de nœuds V et d'un ensemble d'arêtes E reliant les nœuds est construit. Les pixels germes sélectionnés par l'utilisateur pour désigner l'objet et l'arrière-plan sont respectivement représentés par les nœuds source s et puits t. Chaque pixel non marqué est associé à un nœud dans le plan image 2D. Par conséquent, V se compose de deux nœuds terminaux, s et t, et d'un ensemble de nœuds non terminaux dans le graphe G dénoté par I. Les paires de nœud sélectionnées sont connectées par des arêtes et chaque arête lui est attribuée un coût positif. Le coût de l'arête qui relie les nœuds x_i et x_j est désigné par $c(x_i, x_j)$. Une arête est appelé une liaison-t, si elle relie un nœud non terminal en I au nœud terminal t ou s. Une arête est appelé une liaison-n si elle relie deux nœuds non terminaux en I.

Les intensités des pixels germes marqués sont utilisées pour estimer les distributions d'intensité de l'objet de premier plan et de l'arrière-plan, désignées respectivement par $P_r(I|F)$ et $P_r(I|B)$.

Ensuite, les contraintes souples imposées sur les propriétés des frontières et régions de L sont décrites par la fonction de coût E(L)

$$E(L) = \lambda \sum_{x_i \in I} D_i(L_i) + \sum_{(x_i,x_j)\in N} V_{i,j}(L_i, L_j) \tag{3.3}$$

où $L = \{L_i, x_i \in I\}$ est un procédé d'étiquetage binaire pour les pixels de l'image. $L_i = 0$ if x_i est un pixel de l'arrière plan et $L_i = 1$ si x_i est un objet de premier plan $\lambda \geq 0$ est défini pour spécifier une importance relative du terme régional D_i par rapport au terme frontière $V_{i,j}$

L'algorithme de coupe de graphe interactif définit le terme régional avec des log-vraisemblances négatives sous la forme suivante :

$$D_I(L_i = 1) = -\ln P_r(I|F), \tag{3.4}$$

$$D_I(L_i = 0) = -\ln P_r(I|B), \tag{3.5}$$

où I(i) es l'intensité du pixel x_i et $P_r(I(i)|L)$ est calculé en se basant sur l'histogramme d'intensité.

Le terme de frontière est défini comme suit

$$V_{i,j}(L_i, L_j) \propto |L_i - L_j| \exp(-\frac{(I(i) - I(j))^2}{2\sigma^2}).\frac{1}{d(i,j)} \tag{3.6}$$

où $d(i,j)$ est la distance spatiale entre les pixels x_i et x_j et la déviation, σ, est un paramètre lié au niveau de bruit de la caméra. La similarité des pixels x_i et x_j est calculée sur la base de la distribution gaussienne.

Enfin, le résultat de l'étiquetage L, segmentation d objet, est obtenu en minimisant la fonction d'énergie de l'équation 3.3.

Travaux de Rother et al (2004)

Dans leurs travaux [ROTHER et collab., 2004], les auteurs définissent la segmentation d'objet comme un problème d'optimisation d'une fonction de cout énergétique, qui peut être résolu par l'application de l'algorithme de coupe de graphe sur un graphe orienté, pondéré qui modélise l'image d'entrée. Un étiquetage de l'image définit l'appartenance de chaque pixel au premier ou arrière plan de l'image. L'objectif de cet algorithme est de réduire le cout de la fonction énergétique pour le meilleur étiquetage. Pour ce faire, ils encouragent, en abaissant le coût, les pixels voisins similaires de même couleur à avoir la même étiquette, et vice versa. De plus, ils encouragent les pixels à correspondre à un certain modèle de distribution de couleur, en fonction de leurs valeurs.

Soit $k = (k_1, \ldots, k_n, \ldots, k_N), k_n \in 1, \ldots, K$, tel que N est le nombre total de pixels dans la région marquée par l'utilisateur. Le vecteur k assigne à chaque pixel un composant unique d'un modèle de mélange de gaussienne. Le modèle objet et le modèle d'arrière-plan d'un pixel selon l'index n sont dénoté par $\alpha_n = 0$ et $\alpha_n = 1$, respectivement. La fonction énergétique constituée des deux termes "données" et "lissage" s'écrit comme suit :

$$E(\alpha, k, \theta, z) = U(\alpha, k, \theta, z) + V(\alpha, z) \tag{3.7}$$

où z est la couleur du pixel, α = 0 ou 1 est l'étiquette du pixel, θ représente les paramètres du modèle de mélange de gaussiennes(GMM)

$$\theta = \{\pi(\alpha, k), \mu(\alpha, k), \Sigma(\alpha, k)\} \tag{3.8}$$

où π est le poids de mélange.μ et Σ sont la moyenne et la matrice de covariance d'une composante gaussienne.

Le terme données. Le terme données $U(\alpha, k, \theta, z)$ mesure l'adéquation d'un certain modèle tel que la distribution des couleurs. L'étiquette α et la couleur z de chaque pixel de l'image sont utilisées pour vérifier si le pixel correspond à un modèle de distribution de couleur, dans ce cas un modèle de mélange de K-gaussiennes, en utilisant la fonction D() de l'équation 3.9.

$$U(\alpha, k, \theta, z) = \sum_n D(\alpha_n, k_n, \theta_n, z_n) \tag{3.9}$$

où $D(\alpha, k, \theta, z) = -\log\pi(\alpha_n, k_n) + \frac{1}{2} log|\Sigma(\alpha_n, k_n)| + \frac{1}{2}(z_n - \mu(\alpha_n, k_n))^T(\Sigma(\alpha_n, k_n))^{-1}(z_n - \mu(\alpha_n, k_n))$

Le terme de lissage. Le terme de lissage mesure le degré de lissage de l'étiquetage sur des pixels voisins similaires/dis-similaires. Pour chaque paire de pixels voisins qui n'ont pas la même étiquette, la fonction énergétique croit selon un paramètre β qui se trouve dans l'exposant de l'équation 3.10 et qui détermine effectivement le degré de lissage de l'étiquetage.

$$V(\alpha, z) = \gamma \sum_{(m,n)\in E} \Psi(\alpha_n \neq \alpha_m) \exp(-\beta \|z_m - z_n\|^2) \tag{3.10}$$

où la fonction $\Psi(.)$ retourne 1 si l'énoncé est vrai et 0 sinon.

La première itération commence par les contraintes imposées sur l'étiquetage manuel de l'arrière plan. La création de graphe commence par l'attribution des poids des "inter-pixels voisins". Ceux-ci sont calculés en utilisant le gradient horizontal et vertical de l'image, en le manipulant pour obtenir des poids plus élevés sur les bords à faible gradient, et vice versa, en utilisant une fonction d'exposant avec le paramètre β.

Afin d'obtenir les poids d'étiquetage d'arrière/avant plan, un modèle de mélange de deux gaussiennes est utilisé, une gaussienne selon la distribution de couleur de certains pixels d'arrière-plan (en dehors du rectangle dessiné par l'utilisateur) et une gaussienne selon les pixels de premier plan (à l'intérieur du rectangle dessiné par l'utilisateur).

Une segmentation initiale est réalisée en choisissant les vecteurs d'appartenance k et α. Ensuite, le paramètre θ est déterminé en minimisant la fonction énergie dans l'équation 3.7. Une fois le paramètre θ fixé, le résultat de segmentation α et l'appartenance à la composante gaussienne k sont affiné en minimisant également la fonction énergie dans 3.7. Les deux étapes ci-dessus sont répétées itérativement jusqu'à ce que l'algorithme espérance-maximisation(EM) converge.

Travaux de Freedman and Zhang (2005)

[BOYKOV et JOLLY, 2001] ont proposé une approche de segmentation interactive basée sur les coupes de graphes. Avec une interaction relativement faible de la part de l'utilisateur, l'algorithme de la coupe de graphe interactif arrive à segmenter avec succès une variété d'objets dans des images médicales et naturelles. Toutefois, cette approche ne garantit pas toujours un optimum global. L'absence de fortes frontières et la présence d'un certain nombre d'objets avec des profils d'intensité similaires, a tendance à créer une confusion au niveau des termes de frontière et de région de la fonction énergétique. Pour palier à ce problème, [**?**] ont proposé un algorithme qui incorpore les connaissances au préalable de la forme dans une segmentation interactive basée sur les coupes de graphe. Ces connaissances sont intégrées dans les poids sur les arcs du graphe, en utilisant une formulation des méthodes d'ensemble des niveaux(level-set). Cela permet aux arcs du graphe de transmettre des informations à la fois sur l'image (en terme de frontières et de régions) ainsi que sur la connaissance au préalable de la forme. Les transformations du gabarit(template) de la forme sont également prises en considération, où une transformation particulière est choisit en fonction de l'intervention de l'utilisateur.

3.4 Segmentation d'objet automatique

La segmentation d'une image selon deux segments différents uniquement sous entend une segmentation binaire de celle-ci. En d'autres terme, cela revient à séparer les pixels appartenant à un objet de premier plan(étiquetés comme "figure") de ceux appartenant à l'arrière plan(étiquetés comme "fond") d'une image. Nous supposons que toute méthode de segmentation binaire de bas-Niveau peut être considérée comme une méthode de segmentation d'objet(figure/fond). Dans ce qui suit, nous présentons une liste non exhaustives de méthodes de segmentation d'objet dites automatiques, à travers lesquelles aucune intervention d'un utilisateur n'est envisagée.

Travaux de Wu et Leahy (1993)

Étant donné un graphe pondéré G =(V,E) avec V l'ensemble des nœuds et E l'ensemble des arcs reliant ces nœuds. Les pixels d'une image sont considérés comme les nœuds du graphe G et les arcs modélisant les relations d'adjacence entre les pixels sont exprimées par une valeur(poids) de dissemblance. Une matrice de poids W est alors construite tel que $\Omega(\mu,\nu)$ est le poids entre le pixel μ et le pixel ν. La segmentation de l'image est ainsi vue comme un problème de partitionnement du graphe G dont la solution est de trouver une coupe qui divise le graphe G en deux sous graphes complémentaires A et B tel que $A \cup B = G$ and $A \cap B = \emptyset$ par la minimisation du critère de la coupe minimale comme l'indique l'équation 3.11. Un degré de dissimilarité entre ces deux sous graphes est calculé par le poids total des arcs qui ont été retirés.

$$cut(A,B) = \sum_{\mu \in A, \nu \in B} \Omega(\mu,\nu) \tag{3.11}$$

La coupe optimale peut être calculée de manière efficace en utilisant l'algorithme de Ford Fulkerson [FORD et FULKERSON, 1956]. [WU et LEAHY, 1993] ont aussi proposé une méthode de regroupement basée sur le critère de coupe minimale. Toutefois, [FORD et FULKERSON, 1956], [WU et LEAHY, 1993] ont constaté que le critère de la coupe minimale favorise le regroupement de petits ensembles de nœuds isolés dans le graphe vue qu'il ne contient aucune information intra-groupe. En d'autres termes, la coupe minimale aboutit en générale à une sur-segmentation de l'image lorsque la coupe de graphe est appliquée de façon récursive. Pour éviter ce biais non naturel de partitionnement de petits ensembles de pixels, [SHI et MALIK, 2000] propose une nouvelle mesure de dissociation entre deux groupes appelée la coupe normalisée(Ncut).

Travaux de Shi et al (2000)

Dans la segmentation par coupe normalisée, une image est considérée comme un graphe pondéré non-orienté G =(V,E) avec V l'ensemble des nœuds et E l'ensemble des arcs reliant ces nœuds. Contrairement au critère de la coupe minimale qui a un biais en faveur de la coupe de petits ensembles de nœuds, le critère de la coupe normalisée Ncut est non biaisé. L'algorithme de coupe normalisée [SHI et MALIK, 2000] tente de partitionner le graphe en segments équilibrés, normalisés par la somme des poids des arcs dans chaque segment. Ce nouveau critère global mesure à la fois la dissimilarité totale entre les segments ainsi que la similarité à l'intérieure de ces segments. La coupure normalisée (NCut), pour le problème de définition d'une partition de V en deux segments A et B, s'écrit comme suit :

$$\mathrm{NC}ut(\mathrm{A},\mathrm{B}) = \frac{\mathrm{C}ut(\mathrm{A},\mathrm{B})}{\mathrm{A}ssoc(\mathrm{A},\mathrm{V})} + \frac{\mathrm{C}ut(\mathrm{A},\mathrm{B})}{\mathrm{A}ssoc(\mathrm{A},\mathrm{V})} \tag{3.12}$$

Assoc(A,V) se définit comme étant la somme des mesures des arcs entre les nœuds du segment A et tous les autres nœuds de V (les nœuds de A inclus par conséquent).

$$\mathrm{A}ssoc(\mathrm{A},\mathrm{V}) = \sum_{i\in\mathrm{A}}\sum_{j\in\mathrm{V}} \omega_{i,j} \tag{3.13}$$

Cut(A,B) se définit comme étant la somme des mesures des arcs reliant les nœuds de A et B, ou encore la somme des mesures des arcs que l'on enlèverait si on devait séparer les deux segments A et B.

$$\mathrm{C}ut(\mathrm{A},\mathrm{B}) = \sum_{i\in\mathrm{A}}\sum_{j\in\{\mathrm{V}-\mathrm{A}\}} \omega_{i,j} \tag{3.14}$$

En autre, l'algorithme de coupe normalisée [SHI et MALIK, 2000] se base sur la théorie spectrale de graphe, à travers laquelle le problème de la segmentation revient à trouver des vecteurs propres d'une matrice Laplacienne. Il a été démontré que la minimisation de la coupe normalisée $\min_x = \mathrm{NC}ut(x)$ revient à minimiser l'expression suivante

$$\min_y = \frac{y^{\mathrm{T}}(\mathrm{D}-\mathrm{W})_y}{y^{\mathrm{T}}\mathrm{D}_y} \tag{3.15}$$

où D est la matrice diagonale de W.

La minimisation de l'équation 3.15 peut être résolue par le système généralisé des vecteurs propres [5]

$$\mathrm{D}^{\frac{-1}{2}}(\mathrm{D}-\mathrm{W})\mathrm{D}^{\frac{1}{2}} = \lambda y \tag{3.16}$$

La matrice associée à L=D-W correspond au Laplacien du graphe.

L'algorithme de coupe normalisé utilise généralement la deuxième plus petite valeur propre λ_2 qui contient toutes les informations sur les petites coupes de graphe pour partitionner une image. Toutefois, une limitation majeure de cet algorithme est ses exigences de calcul élevées lorsque l'image est de grande taille. En plus, le nombre de segments(clusters) souhaités doit être fixé à l'avance.

Travaux de Cour et al (2005)

[COUR et collab., 2005] construisent un graphe G = (V,E,W) qui modélise une image I, tel que les pixels de V se trouvant à l'intérieur d'une distance $\leq \mathrm{G}_r$ sont connectés par un arc de E. Une valeur de poids W(i,j) mesure la probabilité que les pixels i et j appartiennent à la même région d'image. La valeur idéal du rayon G_r de connexion du graphe G est un compromis entre le coût de calcul et le résultat de segmentation. Dans leur travaux, les auteurs ont montré que ce compromis peut être atténuer au travers un algorithme de segmentation spectrale d'image à échelles multiples qui peut effectivement utiliser une très grande valeur de G_r en décomposant un graphe de connexion à long terme en des sous graphes indépendants.

Au niveau de la première échelle du graphe W_1, chaque pixel de l'image est considéré comme un nœud du graphe, et les pixels à l'intérieur d'une distance r sont connectés à part, par un arc.

Pour la deuxième échelle du graphe W_2, des pixels à l'intérieur d'une distance (2r+1) peuvent être échantillonnés à part dans la grille de l'image originale en tant que nœuds

représentatifs. En appliquant cette procédure récursivement, à l'échelle s, les pixels représentatifs à la distance $(2r+1)s1$ sont échantillonnés sur la grille de l'image originale.

Les pixels représentatifs dans chaque échelle sont désignés par I_s, et la matrice d'affinité compressée avec des connexions entre les pixels représentatifs en I_s par W_c^s . Les différentes échelles du graphe sont définit sur les différentes couches de la pyramide, chacun comme un sous échantillon de l'image originale. $W_c s$ est crée en sous-échantillonnant le graphe original W_s qui encode les indices de contour/luminosité.

De cette façon, [COUR et collab., 2005] ont décomposé un graphe de segmentation à différentes échelles $(W_s)_{s=1,...,S}$, où chaque échelle W_s peut être compressé à l'aide d'un sous-échantillonnage récursif des pixels de l'image. Comparé à d'autres méthodes de segmentation à échelles multiples où les différentes échelles sont traitées séquentiellement, ce graphe est traité en parallèle de sorte que l'information se propage d'une échelle à l'autre. Ceci a pu être concrétisé en spécifiant une contrainte de coupe normalisée à échelles multiples pour le partitionnement de ce graphe. Par conséquent, les auteurs ont pu démontrer que de grands graphes d'images peuvent être compressés en plusieurs échelles capturant la structure de l'image à un voisinage de plus en plus grand. Cet algorithme s'avère aussi efficace en termes de calcul, permettant ainsi de segmenter des images de grandes tailles.

Travaux de Nagahashi et al (2007)

Dans leur travaux, [NAGAHASHI et collab., 2007] ont présenté un processus de segmentation d'image à travers lequel ils ont utilisé des coupes de graphes itératives basées sur un lissage à échelles multiples. L'idée est de segmenter successivement les régions d'un objet en commençant par une segmentation globale jusqu'à obtenir une segmentation locale en utilisant des itérations de coupe de graphe et de lissage gaussien. Les expérimentations de cette méthode de segmentation ont porté sur 50 images de 3 catégories différentes(humains, animaux et paysages) de la base d'images GrabCut [1]. Une comparaison avec les méthodes de segmentation de coupe de graphe interactive [BOYKOV et JOLLY, 2001] et GrabCut [ROTHER et collab., 2004] en calculant les taux d'erreur de la sur/sous segmentation a démontré l'efficacité de cette méthode.

Travaux de Yang and Rosenhahn (2016)

Dans les travaux de [YANG et ROSENHAHN, 2016], les auteurs proposent une nouvelle méthode de segmentation d'objet basée sur les graphes selon un nouveau critère de coupure, la coupe de superpixel(superpixel cut). L'idée clé est de formuler la segmentation d'un objet de premier plan comme la recherche d'un sous ensemble de superpixels qui partitionne un graphe sur des superpixels. Le problème est formulé comme celui de la coupe minimal(Min-Cut). Par conséquent, [YANG et ROSENHAHN, 2016] proposent une nouvelle fonction de coût qui minimise simultanément la similarité interclasse tout en maximisant la similarité intra-classe. Cette fonction de coût est optimisée à l'aide d'une programmation paramétrique. Après une simple phase d'apprentissage, l'approche est entièrement automatique et ascendante, ce qui ne nécessite aucune connaissance au préalable sur la forme et le contenu de l'image. Cette méthode permet de récupérer des composants cohérents de l'image, qui fournissant un ensemble d'hypothèses à échelles multiples pour un raisonnement de haut niveau.

1. http://research.microsoft.com/vision/cambridge/i3l/segmentation/GrabCut.htm

Les auteurs ont comparé cette méthode à d'autres méthodes de segmentation comme celle proposée par [COUR et collab., 2005]. Pour l'évaluation , ils ont utilisé les 50 dernières images de la base d'images WHD(Weizmann Horse Database)pour l'apprentissage de la matrice poids (for learning the weight matrix). Les images restantes de la base WHD ainsi que certaines images de la base d'images BSDS500 ont été réservé pour la phase de test.

Travaux de Comaniciu et Meer(2002)

La méthode de décalage moyen(mean shift en anglais) commence par la définition de fenêtres avec une bande passante prédéfinie à chaque point (ou emplacements aléatoires) dans l'espace des caractéristiques. Les centres des fenêtres forment les premières estimations des régions les plus denses. Pour chaque fenêtre, le centre de masse est calculé ,il est égal à la moyenne pondérée (moyenne) pour tous les points de données à l'intérieur de la fenêtre. Après cela, chaque fenêtre est déplacée vers son centre de masse. Le calcul des moyennes est calculé de nouveau pour déplacer les fenêtres jusqu'à la convergence des vecteurs de décalage moyens vers 0. Cependant, déterminer une bande passante appropriée n'est pas une tâche facile. Des bandes passantes trop grandes ou trop petites peuvent entraîner une sur/sous segmentation. Pour traiter efficacement ce problème, des auteurs ont combiné différents algorithmes au décalage moyen afin de trouver une solution optimale.

Travaux de Felzenszwalb(2004),GB)

Dans leur travaux, [FELZENSZWALB et HUTTENLOCHER, 2004] se servent d'une approche basée sur les graphes pour segmenter une image. Ils considèrent un graphe non orienté G = (V, E) avec $v_i \in$ V, l'ensemble des nœuds à segmenter, et $(v_i, v_j) \in$ E les paires des nœuds voisins(arcs). Chaque arc $(v_i, v_j) \in$ E a un poids non négatif qui lui correspond $w((v_i, v_j))$ mesurant la dissimilarité entre les nœuds voisins v_i et v_j. Dans le cas d'une segmentation d'images les nœuds dans V sont les pixels de l'image et le poids d'un arc est une mesure de dissimilarité entres les deux pixels connectés par cet arc, la différence en intensité, couleur, texture ou autre attribut local [HEDJAM, 2008]. Dans cette approche, une segmentation S est la partition de V en plusieurs composantes (ou régions) telle que chaque région C $\in$ S correspond à une composante connectée dans le graphe G$'$ = (V, E$'$), où E$'$ $\subseteq$ E. Le but de cette approche est d'avoir des éléments beaucoup plus similaires dans une même région et plus différents s'ils appartiennent à des régions différentes. Cela signifie que les arcs entre deux nœuds dans la même région devraient être relativement de faibles poids, et les arcs entre les nœuds de différentes composantes devraient avoir des poids forts. Les auteurs de cette approche ont défini un prédicat D, pour évaluer s'il y a, ou non, une frontière évidente entre deux régions dans la segmentation. Ce prédicat est basé sur une mesure de la similarité entre les éléments le long de la frontière séparant deux régions relativement à une mesure de dissimilarité entre des pixels voisins dans chacune de ces régions. Ce prédicat compare la différence inter-composantes (inter-régions) à la différence intra-composantes (à l'intérieur de chaque région) en respectant certaines caractéristiques locales des données.

La différence interne d'une composante C $\subseteq$ V est définie comme le poids le plus fort dans l'arbre de recouvrement minimal MST(C,E) :

$$Int(\mathrm{C}) = \max_{e \in \mathrm{MST(C,E)}} \omega(e) \tag{3.17}$$

Une autre différence entre deux composantes C1etC2 $\subseteq$ V , est aussi considérée pour être le poids le plus faible de l'arc connectant ces deux composantes, c'est :

$$Dif(C_1, C_2) = \min_{v_i \in C_1, v_j \in C_2, (v_i, v_j) \in E} \omega(v_i, v_j) \tag{3.18}$$

Le prédicat D, se défini comme suit :
Où la différence interne minimal M Int, est défini comme

$$MInt(C_1, C_2) = (\min Int(C_1) + \tau(C_1), Int(C_2) + \tau(C_2)) \tag{3.19}$$

Le seuil τ contrôle le degré pour lequel la similarité intra-régions doit être grande que la similarité inter-régions, dans l'ordre que la frontière entre ces deux régions soit évidente (D est vrai).

Travaux de Han et al (2006)

Dans les travaux de [HAN et collab., 2006], les auteurs développent une méthode d' extraction automatiquement d'objets d'attention sur la base des mécanismes d'attention visuelle humaine, sans avoir besoin de la compréhension sémantique complète de l'image. Cette méthode est réalisée selon deux étapes. La première étape consiste à générer la carte de saillance par le modèle [ITTI et collab., 1998] qui code la valeur de saillance à chaque emplacement de l'image. Dans la deuxième étape, seules quelques germes de saillance sont d'abord sélectionnées selon la carte de saillance. Ensuite, un modèle de champ aléatoire de Markov (MRF) intégrant la valeur d'attention et les caractéristiques de bas niveau est utilisé pour faire croître séquentiellement les objets d'attention à partir de ces graines d'attention sélectionnées. Un avantage important de ce travail est qu'il extrait des objets d'une manière analogue à l'homme. Des études sur les mouvements oculaires ont montré qu'un sujet humain peut ne pas accorder une attention égale à tous les objets de l'image et ne s'occupe généralement que d'un petit nombre d'objets. Au début, la personne jette généralement un coup d'œil rapide sur l'image pour localiser plusieurs foyers d'attention (FOA), puis se concentre sur ces FOA pour les traiter pour plus de détails. Dans ce modèle, la génération de la carte de saillance peut simuler le comportement de l'homme en regardant l'image. La sélection des semences d'attention peut imiter le processus de localisation des FOA. L'objet d'attention croissant peut être traité comme la dernière opération de concentration.

Travaux de Achanta et al (2008)

À partir d'une image d'entrée, des cartes de saillance à différentes échelles sont calculées, additionnées pixel par pixel, et normalisées pour obtenir la carte de saillance finale. L'image est sur-segmentée à l'aide de l'algorithme des k-moyennes. Les centres k pour la segmentation k-moyennes sont automatiquement déterminées en utilisant l'algorithme de montée(hill-climbing algorithm) à partir de l'histogramme tridimensionnel CIE Lab de l'image. L'algorithme de montée peut être vu comme une fenêtre de recherche qui est exécutée à travers l'espace de l'histogramme de dimension d pour trouver la plus grande case(bin) dans cette fenêtre.

Vue que l'espace de caractéristiques CIE LAB est tridimensionnel, chaque case de l'histogramme de couleur possède $3^d - 1 = 26$ voisins où d est le nombre de dimensions

de l'espace de caractéristiques. Le nombre de pics obtenus indique la valeur de K, et les valeurs de ces cases forment les centres initiaux. Vue que l'algorithme des k-moyennes regroupe les pixels dans l'espace de caractéristiques CIE LAB, un algorithme de 8 voisins à composantes connectées est exécuté pour connecter spatialement les pixels de chaque groupe. Une fois que les régions segmentées r_k pour $k = 1, 2, \cdots, \mathrm{K}$ sont trouvés, la valeur moyenne de saillance V par région segmentée est calculée en additionnant les valeurs dans la carte finale de saillance M correspondant aux pixels de l'image segmentée

$$\mathrm{V}_k = \frac{1}{|r_k|} \sum_{i,j \in r_k} m_{i,j} \tag{3.20}$$

où $|r_k|$ est la taille de la région segmentée en pixels. Une méthode simple de seuillage peut être utilisée, dans laquelle les segments ayant une valeur moyenne de saillance supérieure à un certain seuil T sont retenues tandis que les autres sont rejetés. Il en résulte une sortie contenant uniquement les segments qui constituent l'objet saillant (voir la figure 3.4.

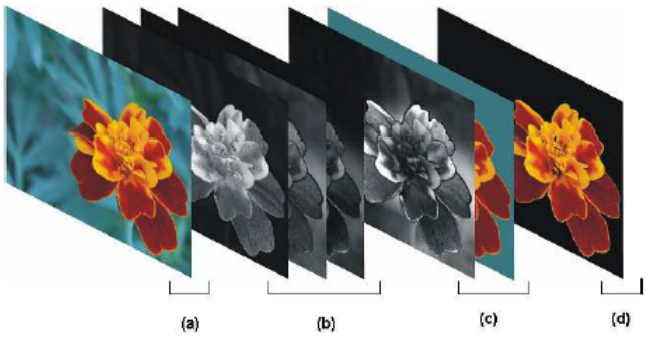


FIGURE 3.4 – Méthode de segmentation de régions saillantes (Variante 1)

La méthode de segmentation des régions saillantes illustrée dans 3.5 est une version modifiée de la variante 1. Il y a deux différences par rapport à la variante précédente : au lieu de la segmentation par les k-moyennes, les auteurs utilisent une segmentation par décalage de la moyenne(mean shift algorithm), et le seuil de sélection des régions saillantes est adaptatif (à la saillance moyenne de l'image d'entrée).

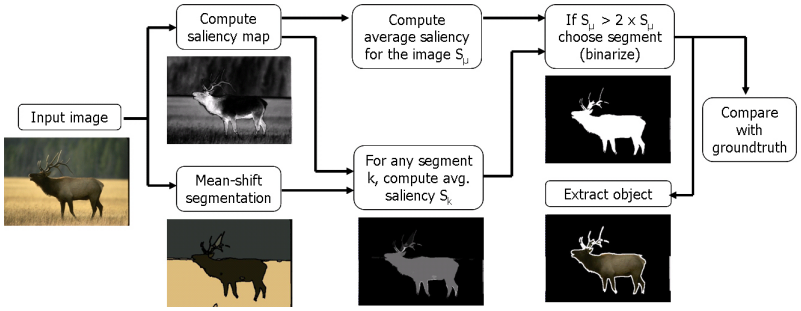


FIGURE 3.5 – Méthode de segmentation de régions saillantes (Variante 2).

Travaux de Achanta et al (2010)

[ACHANTA et SÜSSTRUNK, 2010] ont présenté un algorithme de détection de la saillance basé sur l'idée de pourtour symétrique maximal. L'algorithme utilise les caractéristiques de bas niveau de couleur et luminance. Il est efficace sur le plan des calculs, facile à mettre en œuvre et fournit des cartes de saillance à pleine résolution qui suppriment avec succès l'arrière-plan. Les auteurs ont démontré l'utilisation de leurs cartes de saillance dans la segmentation des objets en utilisant l'approche de la coupe de graphe. Ils utilisent une approche similaire à celle de [BOYKOV et JOLLY, 2001], cependant, au lieu de marquer les pixels de l'arrière-plan et de premier plan à l'aide de la souris, l'utilisateur utilise la carte de saillance pour marquer ces pixels automatiquement. Comme dans la formulation de la méthode de coupe de graphe proposée par [BOYKOV et JOLLY, 2001], ils attribuent des valeurs binaires de saillant ou de non saillant à un vecteur $V = [V_1, V_2, \cdots, V_{|P|}]$ de taille $|P|$, le nombre de pixels dans une image. Ils recherchent une coupe optimale entre les pixels appartenant aux régions saillantes et non saillantes. Ils utilisent des coupes de graphe pour minimiser l'énergie $E(V)$:

$$E(V) = \lambda E_1(V) + E_2(V) \tag{3.21}$$

où $E_1(V)$ compte pour la valeur de saillance et $E_2(V)$ 3.22 favorise la cohérence entre les pixels voisins similaires. $\lambda \geq 0$ spécifie l'importance relative de la valeur de saillance par rapport à la similarité des pixels. $E_2(V)$ pénalise l'attribution d'étiquettes différentes aux pixels voisins ayant des vecteurs de couleur CIE LAB similaires.

$$E_2(V) = \sum_{\{p,q\} \in N} \exp \frac{-(\|I(p) - I(q)\|)}{2\sigma^2} \times \frac{1}{dist(p,q)} \tag{3.22}$$

où N est l'ensemble des 8 pixels voisins connectés q autour de chaque pixel p de l'image et dist est la distance spatiale entre les pixels. Le processus de segmentation dépend fortement de la qualité de la carte de saillance. Le résultat est meilleur si les contours sont bien définies, si la région saillante est bien mise en évidence et si l'arrière-plan est bien supprimé.

Travaux de Fukuda et al (2008)

Dans les travaux de [FUKUDA et collab., 2008], le modèle de saillance de [ITTI et collab., 1998] est intégré à une coupe de graphe vu que certaines régions d'objet semblent

attirer l'attention visuelle plus que les régions d'arrière-plan. Par conséquent, ils utilisent la carte de saillance comme une probabilité antérieure du modèle objet (informations spatiales). Tout d'abord, AdaBoost détermine un emplacement approximatif de l'objet d' intérêt à l'aide d'une fenêtre rectangulaire pour apprendre les informations sur la couleur de l'objet et l'arrière-plan à l'aide de deux GMM, et le théorème de Bayes donne ensuite une probabilité postérieure en utilisant la probabilité précédente. La probabilité postérieure est utilisée comme un coût de liaison t dans les coupes de graphe, où aucun étiquetage manuel des régions d'image n'est requis. Les auteurs ont utilisé 150 images de fleurs dans leur expérimentations et ont évalué leur approche en calculant le taux d'erreur tel que $Err = (E_O/P + E_B/P) \times 100$, où E_O et E_B sont le nombre de pixels de détection d'erreur dans l'objet et les régions d'arrière-plan, respectivement. Le nombre total de pixels dans une image entière est donné par P. La méthode de coupe de graphe en utilisant AdaBoost de [HAN et collab., 2006] entraîne un taux d'erreur plus élevé, comparé à cette méthode, ceci est du au fait que [HAN et collab., 2006] utilisent la région rectangulaire de l'objet fleur détecté pour former les informations de couleur de l'objet, où la région rectangulaire comprend une région large de l'arrière-plan par rapport à l'objet, ce qui représente une petite proportion de la zone fermée.

Travaux de Jung et al (2010)

Dans les travaux de [JUNG et collab., 2010], les auteurs présentent une méthode automatique pour extraire les objets saillants des images naturelles. La segmentation des objets saillants est formulée comme un problème global de minimisation de l'énergie dans un framework itératif d'auto-adaptation. Afin d'estimer la coupe optimale globale, une fonction de coût de segmentation est définit en terme de propriétés de contour et région de segmentation.

En utilisant une méthode de détection de la saillance, les graines d'objets et d'arrière-plan sont déduites automatiquement. Le problème de cette étape est que les graines générées automatiquement peuvent ne pas être positionnées de manière fiable. Une méthode itérative réversible de coupe de graphe est introduite pour surmonter le problème inhérent à la méthode d'extraction des graines basée sur la saillance. Dans le framework itératif auto-adaptatif, des transitions d'état bidirectionnelles sont impliquées de manière itérative pour réduire les pixels mal classés. Afin d'évaluer qualitativement la performance du système proposé, la base de données [LIU et collab., 2007] fournie par Microsoft Research Asia a été utilisé.

Travaux de Fu et al (2011)

Dans leur travaux, [FU et collab., 2011] ont proposé une approche de segmentation d'objets automatique surnommé coupe de saillance qui combine la détection de la saillance et les coupes de graphes. Ils utilisent le modèle de saillance proposé par [HOU et ZHANG, 2007] comme processus de détection de saillance en raison de son faible coût de calcul. [FU et collab., 2011] ont également exploré les effets des étiquettes sur la segmentation basée sur les graphes et leur évaluation en introduisant le concept des étiquettes professionnelles qui sont générées automatiquement. Tout d'abord, une segmentation grossière basée sur l'étiquetage de la détection de saillance est calculée à un faible niveau de résolution. Grâce à cette segmentation grossière, les étiquettes professionnelles sont construites à un haut niveau de résolution. Les auteurs ont démontré l'efficacité de leur méthode comparée à celle de GrabCuts [ROTHER et collab., 2004],

en menant leur expérimentation sur la base d'images de [LIU et collab., 2007]. Toutefois, cette méthode peut échouer lorsque la région de l'objet s'étend vers la gauche et la droite de l'image, ce qui mène à considérer ces régions comme des germes de l'arrière-plan de l'image.

Travaux de Tang et al (2010)

Dans leur travaux, [TANG et collab., 2010] ont proposé une méthode de segmentation d'objet automatique. Ils utilisent le modèle de saillance proposé par [LIU et collab., 2007] pour détecter automatiquement l'objet de premier plan tout en le localisant par une boîte englobante. Ils utilisent une boîte englobante plus large au lieu du résultat de segmentation binaire de [LIU et collab., 2007] pour localiser l'objet, car une boîte englobante plus large garantit que l'objet d'avant-plan soit entièrement entouré à l'intérieur de celle-ci. Après la détection de la saillance, les auteurs construisent une carte de probabilité initiale qui indique la probabilité que chaque pixel appartienne à l'avant-plan. Ensuite, l'algorithme Weighted kernel density estimation(WKDE)est utilisé pour l'affinement de cette carte. Toutefois, WKDE est un algorithme itératif dont le nombre d'itérations dépend largement de la précision de la carte de probabilité initiale. De ce fait, les auteurs ont proposé d'initialiser la carte de probabilité en utilisant conjointement l'estimation de Bayes et la différence de couleurs des pixels. Finalement, l'algorithme de coupe de graphe est employé pour segmenter l'image. Le terme de probabilité de la fonction d'énergie est définit comme étant la probabilité estimée de chaque pixel. Cet algorithme a été testé sur un ensemble de 80 images et comparé à l'algorithme de [BOYKOV et JOLLY, 2001] en calculant le taux d'erreur.

3.5 Segmentation d'objet totalement automatique

La segmentation d'objet totalement automatique nécessite en générale l'apprentissage de modèles de segmentation à partir d'un ensemble d'apprentissage annoté.

Travaux de Kuettel et al (2012)

Dans les travaux de [KUETTEL et FERRARI, 2012], les auteurs proposent une méthode de segmentation binaire qui sépare toutes les classes d'objets de toutes sortes d'arrière plan. Cependant, ils ne font pas la distinction entre les différentes classes d'objets. Cette méthode de segmentation d'objet n'apprend pas de modèles explicites pour chaque classe, mais contrairement à ce qui a été évoqué jusqu'à présent, elle transfère directement des masques de segmentation d'instances d'apprentissage individuelles basées purement sur la similarité visuelle. Bien que cette méthode de segmentation s'appuie sur une formulation énergétique, elle est entièrement automatique. L'entrée de utilisateur est remplacée par un nouveau mécanisme de transfert de segmentation.

Travaux de Fu et al (2016)

Dans leur travaux, [FU et collab., 2016] ont proposé un algorithme de segmentation de figure-fond qui se base sur l'algorithme de segmentation GrabCut. Toutefois, au lieu d'interagir avec un véritable utilisateur, un réseau de neurones à convolution profond pré-entrainé(DCNN) est utilisé pour interagir avec l'algorithme GrabCut. Les tâches que le DCNN doit effectuer incluent la spécification de l'objet du premier plan dans l'image

d'entrée, l'estimation de l'emplacement approximatif du premier plan et l'interaction avec l'algorithme GrabCut.

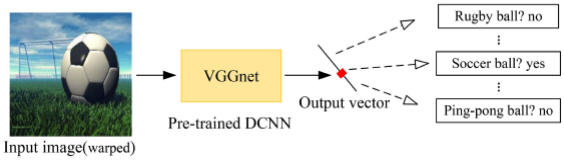


FIGURE 3.6 – Le processus de traitement du réseau DCNN.

Le DCNN pré-entrainé est dérivé du réseau profond pré-entrainé VGG sur la base d'images Imagenet. Dans ce réseau DCNN, l'image d'entrée doit avoir la même dimension que celle dans le réseau profond VGG(224×224 pixels. La sortie générée par le réseau DCNN est représenté sous forme d'un vecteur de 1000 dimension, où chaque élément du vecteur représente une catégorie particulière. En observant la distribution de la valeur du vecteur, la catégorie de l'image d'entrée peut être déterminée. La figure 3.6 illustre un objet du premier plan, un ballon de soccer, dans l'image d'entrée. Si le réseau arrive à reconnaître correctement cette image, l'élément particulier qui représente le ballon de soccer dans son vecteur de sortie devrait obtenir la valeur maximale. Une fois que le réseau reconnaît une image, il se comporte comme si un utilisateur avait spécifié l'objet de premier plan dans une image.

Pour initialiser la carte trimap de GrabCut [ROTHER et collab., 2004], le réseau DCNN doit estimer l'emplacement approximatif de l'objet de premier plan. Ceci peut être réalisé via la méthode de recherche sélective qui génère des boîtes rectangulaires candidates pouvant contenir les objets présents dans une image. Soit une image I ayant subit une déformation, alors le vecteur de sortie du réseau DCNN pour I atteint la valeur maximale dans son nième élément indiquant que l'objet de premier plan dans I appartient à la nième catégorie. Avec la méthode de recherche sélective, N boîtes englobantes seront calculées dans I. D'abord, N régions englobantes $C_k = (1,\ldots,N)$ sont extraites de I. Ensuite, elles subissent une déformation de leur taille comme dans l'étape précédente puis chaque région déformée est envoyée au réseau pré-entrainé DCNN. Par conséquence, les N vecteurs de sortie de dimension 1000 sont calculés. La valeur du nième élément dans chaque vecteur est déterminé par le score de sa région respective $s_k = (1,\ldots,N)$. Plus le score s_k est élevé, plus la distance entre la région C_k et l(objet de e premier plan est proche. Pour améliorer la robustesse cette algorithme, les auteurs définissent un hotspot(point sensible) comme suit

$$H(x,y) = \sum_{\{k|(x,y)\in C_k\}} s_k \tag{3.23}$$

où H est le hotspot. (x,y) est un pixel de l'image d'entrée et $C_{k|(x,y)\in C_k}$ représente une région couvrant le pixel (x,y). Typiquement, il y a des centaines de régions englobantes N. Afin de minimser le temps de calcul et améliorer l'efficacité de cet algorithme, les auteurs ont gardé seulement les M scores des régions les plus élevés de l'équation 3.23 tout en éliminant les scores des régions restantes. Les auteurs ont testé leur algorithme sur la base d'images WSED [ALPERT et collab., 2007] et ont évalué la cohérence des résultats de segmentation par rapport à des vérités de terrain en calculant la mesure F.

Travaux de Kim et al (2017)

Dans les travaux de KIM et collab. [2017], les auteurs proposent une nouvelle méthode d'initialisation du modèle de mélange de gaussiennes(GMM) pour une segmentation d'objet entièrement automatique. Alors que les méthodes classiques emploient un sous ensemble de pixels obtenu de manière inexacte, les auteurs utilisent tous les pixels de l'image pour initialiser chaque GMM. En outre, une mise à jour adaptative des paramètres des GMMs est réalisé pendant les itérations EM en utilisant les valeurs de salience originales sans faire appel à une binarisation. Tout d'abord, les paramètres représentant les objets de premier et arrière plan, GF et GB respectivement, sont initialisés. Les vecteurs moyens sont initialisés aléatoirement. Les matrices de covariance sont initialisées en prenant $[128,128,128]^{\mathrm{T}}$ comme vecteur moyen fixe. Les coefficients de mélange sont initialisés uniformément. Ensuite, la probabilité à postériori du $k^{ième}$ composant Gaussian à l'itération t EM est calculée. Soit s_j la valeur normalisée de la saillance du $j^{ième}$ pixel dans l'intervalle $[0,1]$, alors s_j et $(1-s_j)$ peuvent être considérés comme les probabilités que le $j^{ième}$ pixel appartient aux objets de premier et d'arrière plans respectivement. Par conséquent, nous introduisons un facteur de pondération w_j est introduit à chaque pixel pour mettre à jour les paramètres de deux GMMs de manière adaptative. À l'aide de la probabilité postérieure et du facteur de pondération, les vecteurs moyens, les matrices de covariance et les coefficients de mélange des GMMs sont mis à jour pour la prochaine itération.

3.6 Bilan

D'après l'étude réalisée dans ce chapitre, nous pouvons constater que différentes approches peuvent être adopté pour la segmentation d'objet.

Approche interactive. Dans une approche de segmentation d'objet interactive, un utilisateur intervient dans la sélection de certains pixels appartenant à l'objet de premier plan (avec une étiquette 1) et à d'autres pixels appartenant à l'arrière plan(avec une étiquette 0) de l'image. Cette sélection est faite en réalisant des cliques de souris [BOYKOV et JOLLY, 2001], une boite englobante [ROTHER et collab., 2004] où les pixels correspondants à l'objet de premier plan se trouvent à l'intérieure de la boite, tandis que ceux à l'extérieur correspondent aux pixels de l'arrière plan.

Nous avons constaté que les travaux qui portent sur la segmentation interactive d'objet se basent essentiellement sur l'approche de coupe de graphe. L'image à segmenter est modélisée sous la forme d'un graphe G(V,E) où les pixels de l'image sont représentés par un ensemble de nœuds, qui sont connectées par des arcs dotés de valeurs (poids) reflétant le degré de dissimilarité entre les pixels. La tache de segmentation d'objet revient à résoudre un problème de partitionnement de graphe en deux sous graphes distincts représentant chacun, soit l'objet de premier plan, soit l'arrière plan de l'image. La solution à ce problème fait généralement appel à des méthodes d'optimisation dont l'objectif est de minimiser le cout d'une fonction énergétique à travers les coupes de graphes.

Approche automatique. Dans une approche de segmentation d'objet automatique, la sélection des pixels qui appartiennent à l'objet de premier et d'arrière plan de l'image à segmenter se fait automatiquement, sans aucune intervention de l'utilisateur.

Dans notre étude, nous avons essayé de diversifier et étendre notre recherche sur les travaux qui s'intéressent aussi à la segmentation binaire d'une image. Ainsi, toute méthode de segmentation bas-Niveau qui segmente une image en deux classes seulement peut être considérée comme une méthode de segmentation d'objet, qui sépare un objet du fond d'une image. Un intérêt particulier a été porté sur les méthodes qui se basent sur la théorie des graphes, vue que la plupart des méthodes de segmentation d'objets interactives [BOYKOV et JOLLY, 2001],[ROTHER et collab., 2004]se basent sur cette théorie dans la modélisation et la segmentation des images, ainsi que d'autres méthodes de segmentation [COMANICIU et MEER, 2002] reconnues pour leur efficacité.

[WU et LEAHY, 1993] sont les premiers auteurs a avoir utilisé le critère de la coupe minimale de graphe pour la segmentation d'objet. Toutefois, ce critère favorise la création de petit segments de pixels. Pour résoudre ce problème [SHI et MALIK, 2000]ont proposé d'utiliser la coupe normalisée NCut. [COUR et collab., 2005] ont également utilisé le critère de la coupe normalisée afin de partitionner un graphe à différentes échelles.

Nous avons constaté que des travaux sur la segmentation d'objet ont porté sur la combinaison de l'approche des coupes de graphe avec d'autres approches comme les super-pixels [YANG et ROSENHAHN, 2016], la détection de contour [NAGAHASHI et collab., 2007], [HSIAO et CHANG, 2015], les probabilités [TANG et collab., 2010], l'apprentissage automatique avec le classificateur Adaboost [FUKUDA et collab., 2008], les mélanges de gaussiennes [KIM et collab., 2017].

Par ailleurs, d'autres travaux sur la segmentation d'objet automatique ont porté sur la combinaison de l'approche des coupes de graphe avec la détection de la saillance [FUKUDA et collab., 2008],[JUNG et collab., 2010], [TANG et collab., 2010], [FU et collab., 2011]. Ainsi, les pixels saillants et non saillants de la carte de saillance sont utilisés pour étiqueter les pixels qui appartiennent à l'objet de premier et d'arrière plan de l'image, avant que l'image soit segmentée via une approche de coupe de graphe. Toutefois, ces méthodes de segmentation différent dans le modèle de saillance exploité. [JUNG et collab., 2010], [FU et collab., 2011] ont utilisé le modèle de saillance de [HOU et ZHANG, 2007], [TANG et collab., 2010] ont utilisé le modèle de saillance de [LIU et collab., 2007], tandis que [FUKUDA et collab., 2008] ont utilisé celui de [ITTI et collab., 1998].

Totalement automatique. Dans une approche de segmentation d'objet totalement automatique, nous avons constaté qu'au lieu d'interagir avec un véritable utilisateur, les méthodes de segmentation d'objet remplacent l'entrée de l'utilisateur par un mécanisme de transfert de modèles d'apprentissage de segmentation [KUETTEL et FERRARI, 2012], [FU et collab., 2016] utilisent un réseau de neurones à convolution profond pré-entrainé(DCNN) pour interagir avec l'algorithme GrabCut . [KIM et collab., 2017] mettent à jour itérativement les GMMs des objets de premier et arrière en utilisant le modèle de saillance de [ZHU et collab., 2014].

Nous récapitulons les différents travaux de segmentation d'objet étudiés tout au long de ce chapitre. Comme l'illustre le tableau 3.2, nous proposons trois critères de catégorisation des méthodes de segmentation d'objet, selon le mode d'interaction avec l'utilisateur, selon les approches adoptées afin d'extraire l'objet de premier plan, et selon que la méthode de segmentation d'objet utilise ou non un modèle de saillance visuelle.

TABLEAU 3.2 – Aperçu des méthodes proposées dans le domaine de la segmentation d'objet.

Publication	Mode Interaction	Approche	Salience	Base d'images
[SHI et MALIK, 2000] (NCut)	Automatique	Coupure de graphe		
[BOYKOV et JOLLY, 2001]	Semi-Automatique	Coupure de graphe		GrabCut(50 images)
[COMANICIU et MEER, 2002] (MS)	Automatique	Clustering		
[ROTHER et collab., 2004] (GrabCut)	Semi-Automatique	Coupure de graphe itérative		GrabCut(50 images)
[FELZENSZWALB et HUTTENLOCHER, 2004] (GB)	Automatique	Graphe		
[COUR et collab., 2005] (NCut-MS)	Automatique	Coupure de graphe Échelle Multiples		
[HSIAO et CHANG, 2015]	Automatique	Coupure de Graphe Détection de contour		Berkeley(10 images) iCoseg (10 images)
[HAN et collab., 2006]	Automatique	Coupure de Graphe Apprentissage(AdaBoost)		Flower
[HAN et collab., 2006]	Automatique	Croissance de région	Modèle de ITTI et collab. [1998]	Corel(100 images)
[NAGAHASHI et collab., 2007]	Semi-Automatique	Coupure de Graphe Échelle multiple		GrabCut(50 images)
[FUKUDA et collab., 2008]	Automatique	Coupure de Graphe Apprentissage(AdaBoost)	Modèle de ITTI et collab. [1998]	Flower

TABLEAU 3.2 – Aperçu des méthodes proposées dans le domaine de la segmentation d'objet.

Publication	Mode Interaction	Approche	Salience	Base d'images
[JUNG et collab., 2010]	Automatique	Coupure de Graphe Itérative Réversible	Modèle de HOU et ZHANG [2007]	Microsoft Research Asia LIU et collab. [2007]
[TANG et collab., 2010]	Automatique	Coupure de Graphe Théorème de Bayes	Modèle deLIU et collab. [2007]	(80 images)
[FU et collab., 2011] (Saliency Cut)	Automatique	Coupure de Graphe Résolution Multiple	Modèle de [HOU et ZHANG, 2007]	Microsoft Research Asia LIU et collab. [2007]
[FU et collab., 2011]	Semi-Automatique	Coupure de Graphe Étiquettes Professionnelles Manuelles		[?] database
[KUETTEL et FERRARI, 2012]	Totalement Automatique	Transfert d'apprentissage		
[YANG et ROSENHAHN, 2016]	Automatique	Coupure de Graphe Super Pixel		WHD(train set, 50 images) BSDS500(ensemble de test) WHD(ensemble de test)
[FU et collab., 2016]	Totalement Automatique	Coupure de Graphe CNN		SED1(100 images)
[KIM et collab., 2017]	Totalement Automatique	GMM Coupure de Graphe	Modèle de[ZHU et collab., 2014]	MSRC(660images) iCoseg(642 images) PASCAL(2913 images)
[ACHANTA et collab., 2008]	Automatique	Seuillage	Modèle de [ACHANTA et collab., 2008]	
[ACHANTA et SÜSSTRUNK, 2010]	Automatique	Coupure de Graphe	Modèle de [ACHANTA et collab., 2008]	

Références

ACHANTA, R., F. J. ESTRADA, P. WILS et S. SÜSSTRUNK. 2008, «Salient Region Detection and Segmentation», dans *Proceedings of the 6th International Conference on Computer Vision Systems (ICVS'08)*, vol. 5008, Springer Lecture Notes in Computer Science, Springer,Berlin,Heidelberg, p. 66–75. 97

ACHANTA, R. et S. SÜSSTRUNK. 2010, «Saliency detection using maximum symmetric surround», dans *Proceedings of the 17th IEEE International Conference on Image Processing(ICIP'10)*, IEEE, p. 2653–2656. 90, 97

ADAMS, R. et L. BISCHOF. 1994, «Seeded region growing», *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, p. 641–647. 77

ALPERT, S., M. GALUN, R. BASRI et A. BRANDT. 2007, «Image segmentation by probabilistic bottom-up aggregation and cue integration.», dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 79, 93

BOYKOV, Y. Y. et M. P. JOLLY. 2001, «Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images», dans *Proceedings of "Internation Conference on Computer Vision", Vancouver, vol.I*, p. 105–112. 81, 83, 86, 90, 92, 94, 95, 96

COMANICIU, D. et P. MEER. 2002, «Mean shift : A robust approach toward feature space analysis», *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI'02)*, vol. 42, n° 5, p. 603–619. 95, 96

COUR, T., F. BENEZIT et J. SHI. 2005, «Spectral segmentation with multiscale graph decomposition», dans *IEEEConference on Computer Vision and Pattern Recognition(CVPR'05)*, vol. 2, IEEEComputer Society, p. 1124–1131. 85, 86, 87, 95, 96

FELZENSZWALB, P. et D. HUTTENLOCHER. 2004, «Efficient graph-based image segmentation», *International Journal of Computer Vision(IJCV)*, vol. 59, n° 2, p. 167—-181. 87, 96

FORD, L. et D. FULKERSON. 1956, «Maximal flow through a network.», . 84

FREIXENET, J., X. MUÑOZ, D. RABA, J. MARTÍ et X. CUFÍ. 2002, «Yet another survey on image segmentation : Region and boundary information integration», dans *Computer Vision — ECCV 2002 : 7th European Conference on Computer Vision*. 76

FU, K. S. et J. K. MUI. 1981, «A survey on image segmentation», *Pattern Recognition*, vol. 13, n° 1, p. 3–16. 76

FU, R., B. LI, Y. GAO et P. WANG. 2016, «Fully automatic figure-ground segmentation algorithm based on deep convolutional neural network and grabcut», *IET Image Processing*, vol. 10, p. 937–942. 92, 95, 97

FU, Y., J. CHENG, Z. LI et H. LU. 2011, «Saliency cuts : An automatic approach to object segmentation», dans *2008 19th International Conference on Pattern Recognition*, IEEE, p. 1–4. 91, 95, 97

FUKUDA, K., T. TAKIGUCHI et Y. ARIKI. 2008, «Automatic segmentation using graph cuts based on adaboost and saliency map», dans *Proc. Meeting on Image Recognition and Understanding*, p. 796–801. 90, 95, 96

HAN, J., K. N. NGAN, M. LI et H. J. ZHANG. 2006, «Unsupervised extraction of visual attention objects in color images», *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, nº 1, p. 141–145. 88, 91, 96

HEDJAM, R. 2008, *Segmentation non-supervisée d'images couleur par sur-segmentation Markovienne en régions et procédure de regroupement de régions par graphes pondérés.*, thèse de doctorat, Université de Montréal. 87

HONG, T. et A. ROSENFELD. 1984, «Compact region extraction using weighted pixel linking in a pyramid», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, p. 222–229. 77

HOU, X. et L. ZHANG. 2007, «Saliency detection : A spectral residual approach», dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR'07*, IEEE, p. 1–8. 91, 95, 97

HSIAO, Y.-M. et L.-W. CHANG. 2015, «Unsupervised figure-ground segmentation using edge detection and game-theoretical graph-cut approach», dans *14th IAPR International Conference on Machine Vision Applications (MVA), Tokyo, Japan*, p. 353–356. 95, 96

ITTI, L., C. KOCH et E. NIEBUR. 1998, «A model of saliency-based visual attention for rapid scene analysis», *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI'98)*, vol. 20, p. 1254–1259. 88, 90, 95, 96

JUNG, C., B. KIM et C. KIM. 2010, «Automatic segmentation of salient objects using iterative reversible graph cut», *2010 IEEE International Conference on Multimedia and Expo*, p. 590–595. 91, 95, 97

KELKAR, D. et S. GUPTA. 2008, «Improved quadtree method for split merge image segmentation», *2008 First International Conference on Emerging Trends in Engineering and Technology*, p. 44–47. 77

KIM, G., S. YANG et J. SIM. 2017, «Saliency-based initialisation of gaussian mixture models for fully-automatic object segmentation», *Electronics Letters*, vol. 53, nº 25, p. 1648–1649. 94, 95, 97

KUETTEL, D. et V. FERRARI. 2012, «Figure-ground segmentation by transferring window masks», dans *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, p. 558–565. 92, 95, 97

LIU, T., Z. YUAN, J. SUN, J. WANG, N. ZHENG, X. TANG et H. SHUM. 2007, «Learning to detect a salient object», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, p. 353–367. 91, 92, 95, 97

MARTIN, D., C. FOWLKES, D. TAL et J. MALIK. 2001, «A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics», dans *Proc. 8th Int'l Conf. Computer Vision*, p. 416–423. 78

MCGUINNESS, K. 2009, *Image Segmentation, Evaluation, and Applications*, thèse de doctorat, School of Electronic Engineering. 79, 80

NAGAHASHI, T., H. FUJIYOSHI et T. KANADE. 2007, «Image segmentation using iterated graph cuts based on multi-scale smoothing», dans *Computer Vision–ACCV*, p. 806–816. 86, 95, 96

OHLANDER, R., K. PRICE et R. REDDY. 1978, «Picture segmentation using a recursive region splitting method», *Computer Graphics and Image Processing*, vol. 8, p. 313–333. 77

RICHTSFELD, A. 2012, «The object segmentation database», (osd). 79

ROTHER, C., V. KOLMOGOROV et A. BLAKE. 2004, «"grabcut" : Interactive foreground extraction using iterated graph cuts», dans *ACM SIGGRAPH*, Association for Computing Machinery, p. 309—-314. 82, 86, 91, 93, 94, 95, 96

SHI, J. et J. MALIK. 2000, «Normalized cuts and image segmentation», *IEEETransactions on Pattern Analysis and Machine Intelligence(PAMI'00)*, vol. 22, nº 8, p. 888—-905. 84, 85, 95, 96

TANG, Z., Z. MIAO, Y. WAN et J. LI. 2010, «Automatic foreground extraction for images and videos», dans *2010 IEEE International Conference on Image Processing*, IEEE, p. 2993–2996. 92, 95, 97

WU, Z. et R. LEAHY. 1993, «An optimal graph theoretic approach to data clustering : Theory and its application to image segmentation», *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, p. 1101–1113. 84, 95

YANG, M. Y. et B. ROSENHAHN. 2016, *Superpixel cut for figure-ground image segmentation*, chap. Proceedings of the XXIII ISPRS Congress : From human history to the future with spatial information, 12-19 July 2016, Prague, Czech Republic. Peer reviewed Annals, Volume III-3, International Society for Photogrammetry and Remote Sensing(ISPRS), p. 387–394. 86, 95, 97

YOUSFI, K. 2008, *Segmentation Hiérarchique Optimale par Injection d'A Priori : radiométrique, géométrique ou spatial. Thèse de Doctorat. Spécialité Technologies de l'Information et des Systèmes.*, thèse de doctorat, Université de Technologie de Compiègne HeuDiaSyC. 77

ZHU, W., S. LIANG, Y. WEI et J. SUN. 2014, «Saliency optimization from robust background detection», *IEEE Conference on Computer Vision and Pattern Recognition*, p. 2814–2821. 95, 97

ZOU, W., C. BAI, K. KPALMA et J. RONSIN. 2014, «Online glocal transfer for automatic figure-ground segmentation», *IEEE transactions on image processing*, vol. 23, nº 5, p. 2109–2121. x, 78, 79

Chapitre 4

Un Modèle de Détection de Saillance : application à la segmentation d'objet

« *Tout ce qui en vaut la peine prend du temps.* »

Sommaire

4.1 Introduction

Dans les chapitres précédents, nous avons constaté que des progrès considérables dans le domaine de la détection des objets saillants ont ouvert de nouvelles directions de recherches durant la dernière décennie. En effet, plusieurs chercheurs proposent, dans des travaux assez récents, de combiner différents types d'indices de saillance, d'introduire de nouveaux indice de saillance, d'intégrer des indices de haut niveau(top-down), d'introduire des caractéristiques visuelles profondes(deep features). Afin de développer un modèle de saillance, il est indispensable de calculer une carte de saillance qui repose sur le calcul d'un certain nombre de mesure de saillance comme le contraste d'intensité/de couleur/ d'orientation qui permettent de caractériser un objet saillant et de le distinguer de son voisinage. Par ailleurs, d'après notre étude bibliographique réalisée dans le chapitre 2, nous avons constaté que plusieurs modèles de saillance adoptent une segmentation au préalable pour calculer des indices de saillance. Le choix de la méthode de segmentation s'avère crucial, vu qu'il peut avoir une influence sur le calcul des indices de saillance ainsi que sur la qualité de la carte de saillance. Dans [CHEBBOUT et MEROUANI, 2012], nous avons étudié et démontré l'efficacité des cartes auto-organisatrice de Kohonen dans la segmentation des images couleurs par rapport aux méthodes des k-moyennes et des k-médoids. Nous pensons que l'utilisation de l'algorithme Self Organizing Tree(SOTA), qui a été utilisé avec succès dans l'analyse des données d'expression génétique peut s'avérer avantageux. Cette hypothèse s'appuie sur le fait qu'il combine les avantages de la classification hiérarchique et des cartes auto-organisatrice de Kohonen (SOM).

4.2 Base de données utilisée

Dans notre travail, nous avons utilisé la base d'images MSRA-1000 [ACHANTA et collab., 2009] qui contient 1000 images de couleur avec des segmentations précises des contours de l'objet. Les images de cette base ont été sélectionné à partir de la base d'images de [LIU et collab., 2011] dont les annotations d'objets sont sous la forme de boites rectangulaires. Ces images contiennent au moins un objet saillant ou un objet distinct de premier plan dans des scènes simples ou complexes.

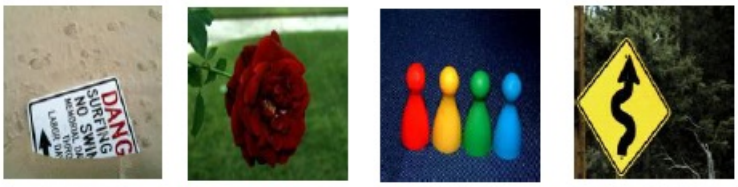


FIGURE 4.1 – Des exemples d'images de la base MSRA-1000.

4.3 Extraction de caractéristiques

Caractéristique de couleur

Divers espaces de couleur sont utilisés dans la littérature afin de caractériser les pixels d'une image. Les propriétés de l'espace de couleur CIE LAB le rendent approprié pour l'extraction des caractéristiques chromatiques globales d'une image numérique. Dans notre travail, nous convertissons l'image couleur originale de l'espace RVB à l'espace CIE LAB. Ce dernier est un espace perceptuellement uniforme et similaire à la perception humaine, dans lequel le canal de luminance et les deux canaux chromatiques R-G et B-Y sont bien décorrélés.

Ensuite, nous appliquons un filtre Gaussien dans l'espace de couleur Lab afin d'éliminer le bruit, chacun des trois canaux résultants de l'image transformée est normalisé dans la plage des valeurs de [0-255] afin d'éviter la suppression de tout canal dominant possible. Enfin, chaque pixel de l'image est définit par un vecteur $3-D$ dont les éléments sont les valeurs de couleur normalisées L, a, b .

Caractéristique de texture

Dans les travaux de [ACHANTA et collab., 2009], les auteurs reportent que le mécanisme de détection de la saillance repose sur le calcul des réponses d'un filtre passe-bande, une différence de gaussiennes, à des canaux de couleur opposés comme les canaux de couleur CIE Lab. Inspirés par les travaux de [ACHANTA et collab., 2009], nous utilisons un filtre passe-bande pour la détection de la saillance. Cependant, nous optons pour le filtre de Gabor, au lieu d'une différence de gaussienne.

Construction du filtre de Gabor. Le filtre de Gabor bidimensionnel décompose une image en composantes correspondantes à différentes fréquence et orientations. Les représentations de fréquence et d'orientation des filtres de Gabor sont similaires à celles du système visuel humain. Elles semblent bien assimiler les champs récepteurs des cellules simples du cortex visuel chez les mammifères. Le filtre de Gabor s'avère particulièrement approprié pour la représentation et la discrimination des textures.

Une fonction de Gabor bidimensionnelle consiste en une onde plane sinusoïdale d'une certaine fréquence et orientation, modulée par une gaussienne bidimensionnelle. Soit $g(x, y, f, \theta)$ la fonction définissant un filtre de Gabor centré à l'origine avec f comme fréquence spatiale et θ comme orientation. Dans le domaine spatial, la famille des filtres de Gabor 2D se définit comme suit :

$$g(x, y, f, \theta) = \exp^{-\frac{1}{2}(\frac{x_\theta^2}{\sigma_x^2}+\frac{y_\theta^2}{\sigma_y^2})} \exp^{j2\pi f x_\theta} \tag{4.1}$$

$x_\theta = x\cos\theta + y\sin\theta$ and $y_\theta = -x\sin\theta + y\cos\theta$.

λ est la longueur d'onde.

$f = \frac{1}{\lambda}$ est la fréquence spatiale du filtre.

θ représente l'angle entre la direction de l'onde sinusoïdale et l'axe x du domaine spatial.

L'enveloppe gaussienne Les paramètres σ_x et σ_y représentent les écarts types de l'enveloppe gaussienne respectivement dans la direction de l'onde et perpendiculaire à celle-ci.
σ_x est la largeur de la bande le long de l'axe majeur gaussien(direction de l'onde)
σ_y est la largeur de la bande le long de l'axe mineur(perpendiculaire à l'onde).
Ces deux paramètres, nommés parfois paramètres de lissage, représentent le facteur de forme de la surface gaussienne. Ils déterminent la plus ou moins grande sélectivité du filtre dans le domaine spatial.

Les valeurs de l'enveloppe gaussienne σ_x et σ_y sont déterminées par les équations 4.2 et 4.3 après avoir fixé les largeurs de bande de fréquence B_f à des valeurs constantes correspondant aux données psychovisuelles. Inspirée par des expériences qui ont montré que la largeur de bande de fréquence des cellules simples dans le cortex visuel est d'environ une octave. La largeur de bande de fréquence B_f est fixée à 1 et une seule échelle est extraite.

$$\sigma_x = \frac{\sqrt{\ln 2}(2^{B_f}+1)}{\sqrt{2} f_0 (2^{B_f}-1)} \tag{4.2}$$

$$\sigma_y = \frac{\sigma_x}{\gamma} \tag{4.3}$$

γ est le rapport d'aspect spatial et il spécifie l'elliptique du support de la fonction Gabor. Dans ce travail, une gaussienne elliptique est choisi en définissant σ_x et σ_y selon les équations 4.2 et 4.3 afin d'avoir une couverture spatiale identique dans toutes les directions.

Le paramètre Ψ. Le paramètre Ψ représente le décalage de phase en degrés. $\Psi = 0^\circ$ et $\Psi = 90^\circ$ retourne la partie réelle et la partie imaginaire du filtre de Gabor respectivement. Dans le cas bidimensionnel, cette onde sinusoïdale est la somme de deux fonctions sinusoïdales, la première paire et réelle, et la deuxième impaire et imaginaire.

La partie réelle du filtre de Gabor est un filtre symétrique pair avec $\psi = 0$. Il a la forme générale suivante

$$gb_e = \exp^{-\frac{1}{2}\left(\frac{x_\theta^2}{\sigma_x^2}+\frac{y_\theta^2}{\sigma_y^2}\right)} \cos(2\pi/\lambda x_\theta + \Psi \tag{4.4}$$

La partie imaginaire du filtre de Gabor est un filtre de Gabor symétrique impaire avec $\psi = \frac{\pi}{2}$. Il a la forme générale suivante

$$gb_o = \exp^{-\frac{1}{2}\left(\frac{x_\theta^2}{\sigma_x^2}+\frac{y_\theta^2}{\sigma_y^2}\right)} \cos(2\pi/\lambda x_\theta + \Psi \tag{4.5}$$

Dans notre travail, nous utilisons des paires de composantes impaires et paires de filtres de Gabor avec la relation de phase en quadrature. Chaque paire de filtres de Gabor est réglé sur une seule fréquence spatiale $f = \frac{1}{8}$ et 8 orientations à 22.5° d'intervalle tel que $\theta \in \{0, \frac{\pi}{8}, \frac{2\pi}{8}, \frac{3\pi}{8}, \frac{4\pi}{8}, \frac{5\pi}{8}, \frac{6\pi}{8}, \frac{7\pi}{8}\}$

Extraction de caractéristiques avec le filtre de Gabor. Soit $I_{gray}(x,y)$ une image en niveau de gris de taille M × N et soit $\psi_{\mu,\nu}(x,y)$ un filtre de Gabor défini par sa fréquence centrale f_μ et son orientation θ_ν. L'opération de filtrage ou la procédure d'extraction de caractéristiques peut alors être écrite comme la convolution de l'image $I_{gray}(x,y)$ avec l'ondelette de Gabor(filtre,noyau) $\psi_{\mu,\nu}(x,y)$

Dans le domaine spatial, le filtrage est effectué par la convolution complexe d'une image par le filtre de Gabor comme suit :

$$G_{\mu,\nu}(x,y) = I(x,y) * \psi_{\mu,\nu}(x,y) \tag{4.6}$$

Les filtres de Gabor représentent des filtres complexes qui combinent une partie paire (de type cosinus) et une partie impaire (de type sinus). Dans l'expression ci-dessus, $G_{\mu,\nu}(x,y)$ représente la sortie de convolution complexe qui peut être décomposée en ses parties réelle(ou paire) et imaginaire(ou impaire). En se basant sur ces résultats, $G_{mu,\nu}(x,y)$ peut être décomposé en ses parties réelle(paire) et imaginaire(impaire) comme suit :

$$E_{\mu,\nu}(x,y) = \Re[G_{\mu,\nu}(x,y)] \tag{4.7}$$

$$O_{\mu,\nu}(x,y) = \Im[G_{\mu,\nu}(x,y)] \tag{4.8}$$

A partir de ces résultats, les réponses de l'amplitude $A_{\mu,\nu}(x,y)$ de chaque pixel de l'image(et qui définissent les valeurs de texture des pixels) est la somme des carrés des parties réelle et imaginaire, et se calculent comme suit :

$$t = A_{\mu,\nu}(x,y) = \sqrt{E^2_{\mu,\nu}(x,y) + O^2_{\mu,\nu}(x,y)} \tag{4.9}$$

Extraction de caractéristiques avec le filtre Log-Gabor. Après la conversion de l'image couleur de l'espace RVB à l'espace CIE LAB, nous optons pour un autre filtrage de type passe-bande mais cette fois-ci en utilisant le filtre de Log-Gabor[FIELD, 1987]. Comparé au filtre de Gabor, il peut être construit avec une bande passante arbitraire. En plus, il assure une meilleure couverture dans le domaine de Fourier ce qui le rend mieux adapté au codage des images naturelles que les autres filtres passe-bande. La fonction de transfert d'un filtre de log-Gabor $g(x)(x = (x,y) \in \mathbb{R}^2$ dans le domaine fréquentiel s'exprime comme suit :

$$G(u) = e(-(\log \frac{\|u\|_2}{\omega_0})^2 / 2\sigma_F^2) \tag{4.10}$$

où $u = (\mu,\nu) \in \mathbb{R}^2$ est la coordonnée dans le domaine fréquentiel, ω_0 est la fréquence centrale du filtre et σF contrôle la bande passante du filtre. La fonction $g(x)$ ne peut pas être exprimée analytiquement en raison de la singularité de la fonction log à l'origine. Alternativement, la fonction $g(x)$ ne peut être approximativement obtenue qu'en effectuant une transformée de Fourier inverse en $G(u)$.

Soit une image couleur I de taille M × N dans l'espace de couleur CIE Lab et soit ses trois canaux de couleur de luminance et de chrominance noté respectivement $I_L, I_a et I_b$. Nous modélisons par un filtre passe-bande de Log-Gabor la valeur de fréquence de chaque pixel de l'image I , que nous notons $f(x,y)$, comme suit :

$$f(x,y) = \sqrt{(I_L(x,y) * g(x,y))^2 + (I_a(x,y) * g(x,y))^2 + (I_b(x,y) * g(x,y))^2} \tag{4.11}$$

où $*$ dénote l'opération de convolution.

4.4 Calcul de la carte de saillance

Dans cette section ,nous expliquons l'algorithme d'apprentissage du réseau de neurones Self Organizing Tree)[DOPAZO et CARAZO, 1997; HERRERO et collab., 2001] qui implémente un algorithme de classification hiérarchique descendante combiné aux cartes auto-organisatrices de Kohonen.

Apprentissage de SOTA

Initialement, un arbre binaire est composé de deux cellules externes reliées par l'intermédiaire d'un nœud interne parent. Chaque nœud/cellule i est initialisé avec un vecteur prototype aléatoire $w_i \in \mathrm{R}^n$.

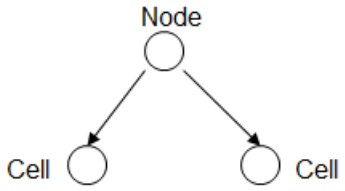


FIGURE 4.2 – Architecture initiale du réseau.

Chaque époque consiste à présenter tous les vecteurs caractéristique d'apprentissage. Une présentation implique deux étapes. La première étape consiste à trouver la cellule i la plus proche (la cellule gagnante) pour chaque vecteur d'apprentissage. La distance entre la cellule i, et le vecteur d'apprentissage j peut être calculer en utilisant la distance euclidienne.

Une fois que la cellule gagnante i a été trouvée pour un vecteur j donné, le vecteur de référence de la cellule gagnante i et son voisinage sont mis à jour comme suit

Le vecteur poids w_i de la cellule gagnante i est mis à jour avec la valeur α_{winner}

$$w_i(t+1) = w_i(t) + \alpha_{winner}(x_i(t) - w_i(t)) \tag{4.12}$$

Le vecteur poids du nœud parent(mère) de la cellule gagnante i est mis à jour avec la valeur α_{mother}

$$w_i(t+1) = w_i(t) + \alpha_{mother}(x_i(t) - w_i(t)) \tag{4.13}$$

Le vecteur poids de la cellule sœur de la cellule gagnante i est mis à jour avec la valeur α_{sister}

$$w_i(t+1) = w_i(t) + \alpha_{sister}(x_i(t) - w_i(t)) \tag{4.14}$$

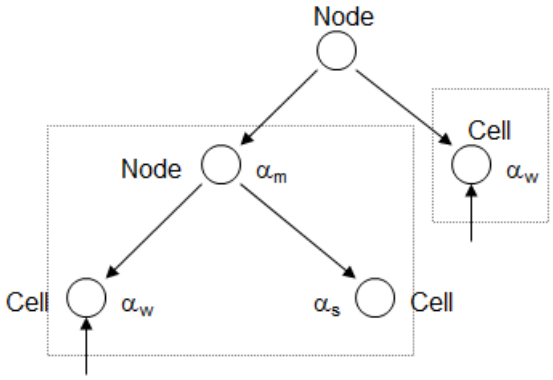


FIGURE 4.3 – Les deux mécanismes de mise à jour du vecteur prototype.

$$|\frac{\varepsilon_t - \varepsilon_{t-1}}{\varepsilon_{t-1}}| > \mathrm{E} \tag{4.15}$$

Dès qu'un cycle se termine et converge, la taille du réseau augmente en rattachant deux nouvelles cellules descendantes à la cellule ayant la valeur de ressource la plus élevée. La ressource R_i d'une cellule i se calcule par la moyenne des distances entre une cellule i et les vecteurs caractéristiques d'apprentissage qui lui sont associés selon l'équation 4.16.

$$\mathrm{R}_i = \frac{\sum_{j=1}^{\mathrm{K}} d(x_i, w_i)}{\mathrm{K}} \tag{4.16}$$

où K est le nombre total des vecteurs caractéristiques d'apprentissage associés à la cellule i .

La cellule développée est transformée en un nœud. Le processus de croissance du réseau se termine lorsque la valeur la plus élevée de ressource des cellules atteint un certain seuil.

Dans notre travail, le réseau a été entrainée pendant un maximum de 10 époques. Le temps d'itération initial est de 100. La valeur du seuil 0,001 . La valeur de la ressource finale est R = 0,001. Les valeurs de mise à jour des paramètres pour le gagnant et ses voisins directs (cellules mère et sœur) sont respectivement 0,1, 0,05 et 0,01 .

Ainsi, étant donné une image I = pi, j avec $i = 1, \cdots, \mathrm{M}$ et $j = 1, \cdots, \mathrm{N}$, on obtient après formation du réseau SOTA, K clusters $\mathrm{C}^k_{k=1,\cdots,\mathrm{K}}$. Les clusters sont désignés par un ensemble de vecteurs dimensionnels $\mu^k_{k=1,\cdots,\mathrm{K}}$ dans lesquels μ_k désigne le prototype (centre du cluster) associé au cluster C_k.

Soit $\mathrm{Z}_{i,j}$ la position normalisée du pixel $p_{i,j}$ dans l'image I et la fonction $b : \mathrm{R}^2 \rightarrow 1, \ldots, \mathrm{K}$ associe le pixel $p_{i,j}$ et l'index de cluster $b(p_{i,j})$.

Calcul de Spatial Cue (Prior Location)

Similaire à [FU et collab., 2013], nous calculons l'indice spatial $w(k)$ pour chaque cluster C^k comme suit :

$$w(k) = \frac{1}{n^k} \sum_{i=1}^{N_j} [\mathrm{N}(\|\mathrm{Z}_{i,j} - \mathrm{O}^j\|^2 | 0, \sigma^2).\delta[b(p_{i,j}) - \mathrm{C}^k]] \tag{4.17}$$

$\delta(.)$ est la fonction delta de Kronecker qui vaut 1 si $b(p_{i,j})$ et C^k sont égaux et 0 sinon. O^j désigne le centre de l'image I^j, Le noyau gaussien N(.) calcule la distance Euclidienne entre le pixel Z_i^j et le centre de l'image O^j, la variance σ^2 est le rayon normalisé de l'image. Le coefficient de normalisation n^k est le nombre de pixels du cluster C^k.

Après cela, la carte spatiale est normalisée à une gaussienne standard en utilisant la distribution des scores dans tous les clusters.

La valeur de saillance du cluster qui fournit l'affectation discrète est calculée en utilisant la probabilité de saillance de cluster $p(k)$ du cluster C^k as

$$p(\mathrm{C}^k) = \prod_i w_i(k) \tag{4.18}$$

où $w_i(k)$ indique un indice de saillance.

Ensuite, la valeur de saillance pour chaque pixel est lissée. La vraisemblance de saillance du pixel x appartenant au cluster C_k satisfait une distribution gaussienne $\aleph$ comme suit

$$p(x|\mathrm{C}^k) = \aleph(\|\nu_x, \mu^k\|_2 | \mathrm{O}, \sigma_k^2) \tag{4.19}$$

où ν_x désigne le vecteur caractéristique du pixel x, et la variance σ_k de gaussien utilise la variance du cluster C_k. Par conséquent, la probabilité de saillance marginale $p(x)$ est obtenue en additionnant la saillance conjointe $p(\mathrm{C}^k)p(x|\mathrm{C}^k)$ sur tous les clusters comme suit

$$p(x) = \sum_{k=1}^{K} p(x|\mathrm{C}^k) = \sum_{k=1}^{K} p(\mathrm{C}^k).p(x|\mathrm{C}^k) \tag{4.20}$$

4.5 Évaluation du modèle de saillance

Dans cette section, nous évaluons le modèle de saillance proposée sur la base d'images MSRA-1000. Nous le comparons à 12 autres modèles de saillance : le modèle (IT) [Itti et collab., 1998], le modèle (MZ) [Ma et Zhang, 2003], le modèle (LC) [Weibin et collab., 2013], le modèle (GB) [Harel et collab., 2006],le modèle (SR) [Hou et Zhang, 2007],le modèle (AC) [Achanta et collab., 2008], le modèle(IG) [Achanta et collab., 2009], le modèle(MSSS) [Achanta et Süsstrunk, 2010], le modèle(CA) [Goferman et collab., 2012], le modèle(HC) [Cheng et collab., 2011], le modèle (IS) [Hou et collab., 2012] et le modèle (TMM) [Imamoglu et collab., 2013].

Similaire à [Achanta et collab., 2008; Walther et collab., 2002; Weibin et collab., 2013], nous évaluons la qualité des cartes de saillance générées par le modèle de détection de saillance proposé dans le contexte de la segmentation d'objet. Pour une carte de saillance donnée dont les valeurs appartiennent à l'intervalle [0-255], la manière la plus simple de segmenter l'objet saillant est de procéder à un seuillage de la carte de saillance avec un seuil $\mathrm{T}_f \in [0-255]$. Lorsque T_f varie de 0 à 255, différentes paires de précision-rappel sont obtenues et une courbe précision-rappel peut être tracée. La

courbe moyenne de rappel de précision est générée en faisant la moyenne des résultats de toutes les images de test. Les courbes résultantes sont représentées par les figures 4.4 and 4.5.

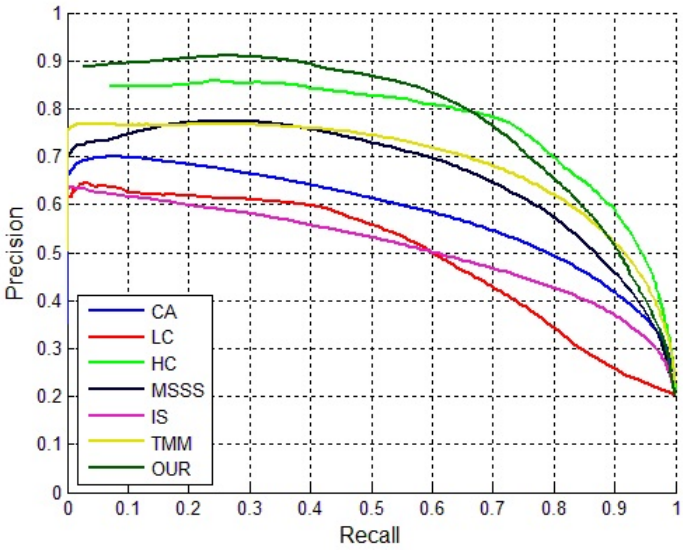


FIGURE 4.4 – Les courbes rappel-précision pour la binarisation des cartes de saillance sur la base d'images MSRA-1000

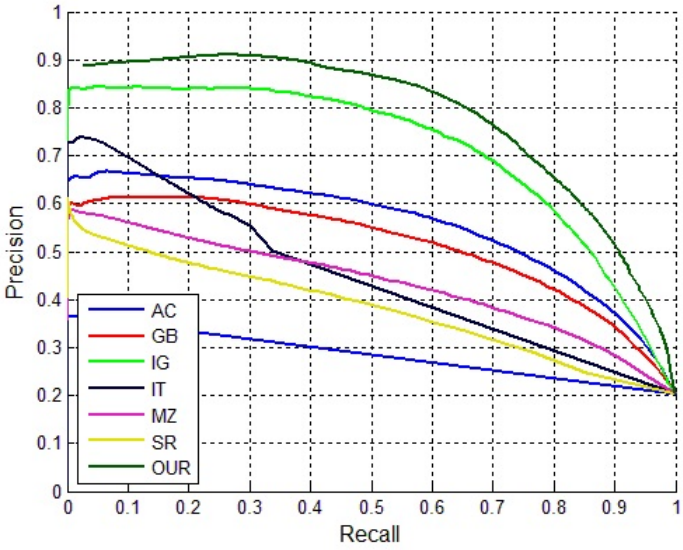


FIGURE 4.5 – Les courbes rappel-précision pour la binarisation des cartes de saillance sur la base d'images MSRA-1000.

Nous évaluons également la qualité des cartes de saillance obtenues par le modèle

proposé en procédant à un seuillage adaptatif afin de segmenter les objets dans l'image. Plus précisément, un seuil adaptatif se définit par le double de la saillance moyenne de l'image en utilisant l'équation 4.21

$$T_a = \frac{2}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} S(x, y) \tag{4.21}$$

où W et H sont respectivement la largeur et la hauteur de la carte de saillance, et $S(x, y)$ est la valeur de saillance du pixel à la position (x, y). Si la saillance dans ce segment est supérieur à deux fois la valeur de saillance moyenne globale, le segment est marqué comme premier plan. Les valeurs de précision (P) et de rappel (R) sont ensuite calculées, et la mesure F (F_β) est également obtenue pour l'évaluation. Pour mettre en évidence la précision plus que le rappel, $\beta^2 = 0,3$ [ACHANTA et collab., 2009; CHENG et collab., 2011].Les barres résultantes sont représentées par la figure 4.6 et la figure 4.7.

La mesure F reflète aussi la précision globale de la prédiction d'un algorithme donné. Nous calculons la mesure F moyenne sur 1000 images pour chaque algorithme de détection de saillance. Les résultats sont listée dans le tableau 4.1.

TABLEAU 4.1 – Calcul de la mesure-F pour chaque algorithme.

Publication	Acronyme	Précision	Rappel	Mesure-F
[ACHANTA et collab., 2008]	AC	0.6009	0.4761	0.5666
[HAREL et collab., 2006]	GB	0.5682	0.4551	0.5374
[ACHANTA et collab., 2009]	IG	0.7619	0.5842	0.7119
[ITTI et collab., 1998]	IT	0.5901	0.2378	0.4397
[MA et ZHANG, 2003]	MZ	0.4624	0.4300	0.4545
[HOU et ZHANG, 2007]	SR	0.4371	0.3233	0.4043
[GOFERMAN et collab., 2012]	CA	0.5977	0.5467	0.5851
[WEIBIN et collab., 2013]	LC	0.5496	0.4873	0.5339
[CHENG et collab., 2011]	HC	0.7444	0.7096	0.7361
[ACHANTA et SÜSSTRUNK, 2010]	MSSS	0.6997	0.5870	0.6700
[HOU et collab., 2012]	IS	0.5522	0.4155	0.5132
[IMAMOGLU et collab., 2013]	TMM	0.7375	0.4634	0.6489
[FU et collab., 2013] (Contrast Cue)				0.755
[FU et collab., 2013] (Contrast,Spatial Cues				0.854
[CHEBBOUT et MEROUANI, 2015]		**0.8363**	**0.5271**	**0.7366**

Nous pouvons constaté que les résultats obtenus en utilisant le modèle de saillance proposé surpassent considérablement les autres méthodes avec $F_\beta = 73,66\%$, précision = 83,63 % et rappel = 52,71 % , et un résultat sous-optimal est obtenu par HC avec $F_\beta = 73,61\%$, precision = 74,44 % et rappel = 70,96 % , par TMM avec $F_\beta = 64,89\%$, précision = 73,75 % et rappel = 46,34 % . De plus, la majorité des modèles obtiennent une précision plus élevée que le rappel, car nous avons définit $\beta^2 = 0,3$, ce qui met davantage l'accent sur la précision. La figure 4.8 illustre les résultats obtenus.

En utilisant ces cartes de saillance, nous arrivons à extraire clairement les objets tout en ignorant les détails inutiles. Cette efficacité du modèle est peut être du i) à l'utilisation d'une approche d'apprentissage non supervisée se basant sur le réseau SOTA, ii)à l'extraction des caractéristiques de texture en utilisant les filtres Log-Gabor qui procèdent à un filtrage passe bande de l'image dans le domaine fréquentiel.

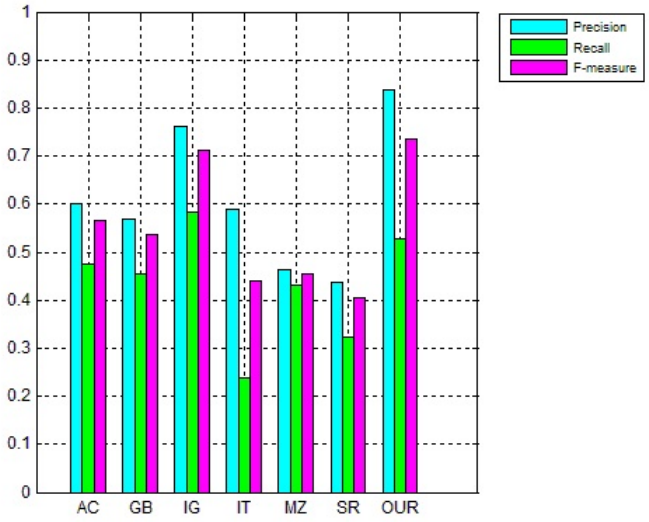


FIGURE 4.6 – Les barres rappel-précision pour la binarisation des cartes de saillance sur la base d'images MSRA-1000 .

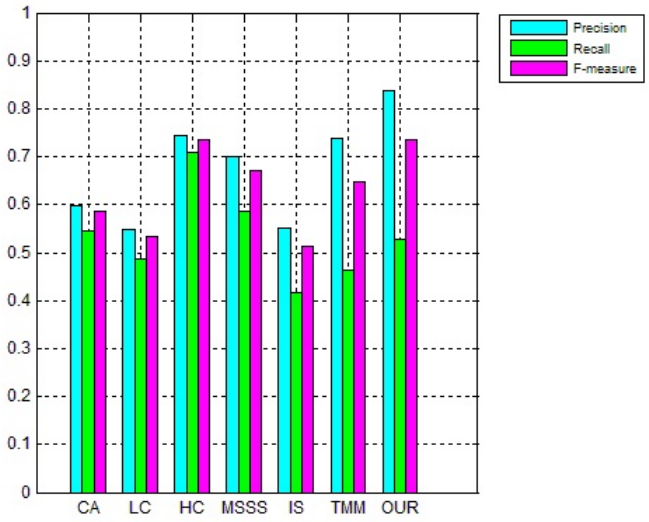


FIGURE 4.7 – Les barres rappel-précision pour la binarisation des cartes de saillance sur la base d'images MSRA-1000.

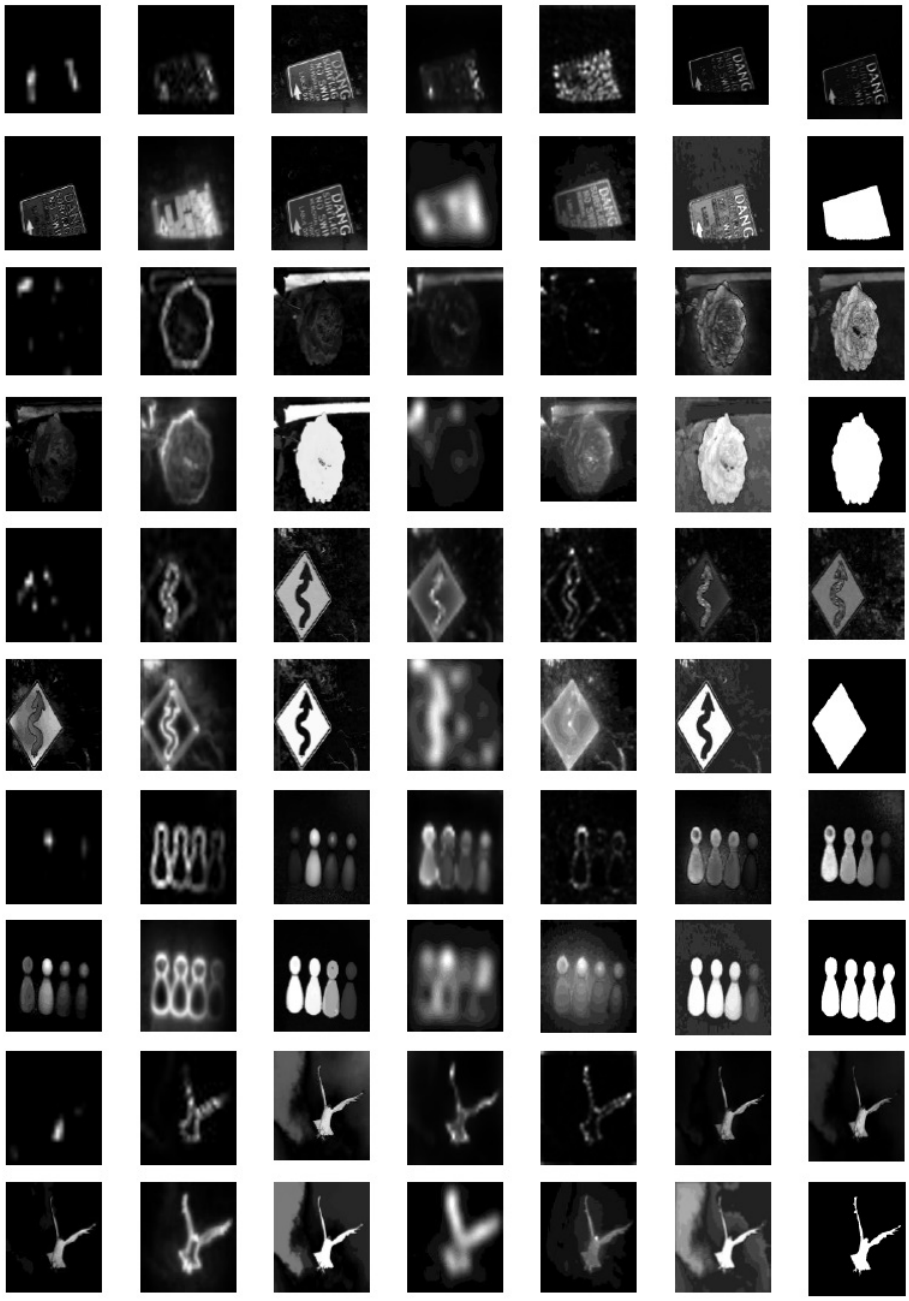


FIGURE 4.8 – Comparaison des résultats obtenus avec le modèle de saillance proposé et d'autres modèles sur la base d'images MSRA-1000. Les images sont présentées de gauche à droite selon les colonnes dans l'ordre suivant : image originale, IT [ITTI et collab., 1998], MZ [MA et ZHANG, 2003], LC [WEIBIN et collab., 2013], GB [HAREL et collab., 2006], SR [HOU et ZHANG, 2007],AC [ACHANTA et collab., 2008], FT ou IG [ACHANTA et collab., 2009], MSSS [ACHANTA et SÜSSTRUNK, 2010],CA [GOFERMAN et collab., 2012], HC [CHENG et collab., 2011],IS [HOU et collab., 2012], TMM [IMAMOGLU et collab., 2013] puis le modèle proposé, et la vérité de terrain à la fin.

4.6 Application à la segmentation d'objet

Dans cette section, nous incorporons les valeurs de saillance des pixels obtenues à partir du modèle de saillance proposé dans le calcul du graphe de similarité de la méthode de classification spectrale.

4.6.1 Classification Spectrale

Soit une image couleur dans l'espace de couleur R-V-B convertit en une image dans l'espace de couleur CIE LAB, nous procédons à l'extraction du vecteur de caractéristique de couleur, de texture en utilisant le filtre Log-Gabor, pour chaque pixel. Soit S la carte de saillance correspondante à l'image couleur. Nous définissons pour chaque pixel x_i de l'image un vecteur de caractéristique de dimension 5-D $x_i = \{\mathrm{L}, a, b, t, s\}$ qui combine les valeurs de couleur L, a, b, de texture t et de saillance s.

Soit G =(V,E) un graphe non orienté, pondéré, constitué d'un ensemble de sommets V. Le graphe G est totalement connecté tel que l'ensemble des N pixels de l'image V = $x_1, \ldots, x_\mathrm{N}$ avec $x_i \in \mathrm{R}^p$ sont reliés entre eux. Nous définissons la similarité $s_{ij} \geq 0$ entre l'ensemble des vecteurs caractéristiques des pixels $\{x_i\}$ à travers les voisinages dans R^p comme suit

$$s_{ij} = \exp(-\frac{\|x_i - x_j\|}{2\sigma^2}, \forall i \neq j \tag{4.22}$$

Où $\|\|$ représente une mesure de distance euclidienne. σ est un paramètre d'échelle qui contrôle la taille des voisinage dans R^p.

Similaire [Zelnik-manor et Perona, 2004], nous adoptons une approche locale qui consiste à définir un scalaire pour chaque couple de point x_i, x_j. Ils assignent un paramètre scalaire σ_i différent à chaque point x_i de l'ensemble des vecteurs caractéristiques de l'image. σ_i est égal à la distance entre le point x_i et son k ième voisin le plus proche. Dans notre travail, nous avons utilisé K =7 .

Les algorithmes de partitionnement spectral utilisent l'information contenue dans les vecteurs propres d'une matrice Laplacienne, construite à partir de la matrice de similarités S.

Soit D, la matrice diagonale des degrés

$$\mathrm{D}_{(i,i)} = \sum_{j=1}^{n} \mathrm{S}_{(i,j)} \tag{4.23}$$

La matrice Laplacienne [Ng et collab., 2001] est définit comme une matrice Laplacienne normalisée, L_{norme}, normalisation par division symétrique, basée sur S et D

$$\mathrm{L}_{norm} = \mathrm{D}\ -1/2\ \mathrm{S}\ .\ \mathrm{D}\ -1/2 \tag{4.24}$$

Ensuite, les premiers k vecteurs propres $e_1, e_2, \ldots, e_\mathrm{K}$ de L_{norme}, associés aux K plus grandes valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_\mathrm{K}$ sont calculés

La matrice $\mathrm{E} = [e_1, e_2, \ldots, e_\mathrm{K}]$ de taille N × K est construite et nous obtenons la matrice U de taille N × K en normalisant les lignes de E pour avoir une norme unitaire comme l'indique l'équation ci-dessous

$$\mathrm{U}_{(i,j)} = \frac{e(i,j)}{\sqrt{\sum_k e_{(i,k)^2}}} \tag{4.25}$$

Enfin, les lignes N de U sont regroupées en K groupes, en utilisant l'algorithme des k-moyennes. Puis, nous affectons le nœud i au cluster k si et seulement si la ligne i de la matrice U est affecté au cluster k.

4.6.2 Résultats expérimentaux

Nous évaluons la méthode proposée sur des images de taille 300×400 de la base MSRA-1000. Quelques résultats de la segmentation d'objet sont illustrés dans la figure 4.9. Quelques comparaisons visuelles de la segmentation d'objet en utilisant les méthodes de segmentation NCut-MS [COUR et collab., 2005], Ncut [SHI et MALIK, 2001], MS[COMANICIU et MEER, 2002], GB [FELZENSZWALB et HUTTENLOCHER, 2004] et la méthode proposée sont illustrées dans les figures 4.10, 4.11, 4.12, 4.13.

FIGURE 4.9 – Quelques résultats de la segmentation d'objet en utilisant la méthode proposée sur des images de la base MSRA-1000.

D'après nos expérimentations, nous remarquons que l'objet saillant(en blanc) dans la figure 4.10 a pu être correctement segmenté à partir de l'arrière-plan(en noir) par la méthode proposée, tandis qu'il a été segmenté en noir par les 4 autres méthodes de segmentation de comparaison. Autrement dit, il n'a pas été segmenté en tant qu'un objet saillant mais plutôt en tant qu'un objet d'arrière plan.

Les résultats de segmentation obtenus dans la figure 4.11 par les quatre méthodes de comparaison sont tous de mauvaise qualité. Les images segmentées sont bruitées et ne contiennent que certaines parties de l'objet saillant avec des bords discontinus.

Par ailleurs, la méthode proposée a pu générer une meilleur segmentation de l'objet saillant.

Dans la figure 4.12, nous observons qu'aucune des méthodes de segmentation n'a pu segmenter l'objet saillant. Ceci revient peut être au fait que l'objet saillant dans cette image est représenté par 3 joueurs différents de rugby. Par ailleurs, nous remarquons que l'objet saillant a été segmenté en noir par la méthode proposée.

Dans la 4.13, nous remarquons que l'objet saillant(en blanc) a pu être correctement segmenté à partir de l'arrière-plan(en noir) par la méthode proposée, tandis qu'il a été segmenté en noir par les 4 autres méthodes de segmentation de comparaison.

L'objectif principal du développement de cette méthode a été d'étudier l'influence de la saillance visuelle en tant que caractéristique visuelle quantifiable dans la segmentation d'objet, et ceci au travers l'utilisation de la carte de saillance générée par le modèle de saillance. Nous avons introduit la saillance dans le calcul du graphe de similarité adopté par l'algorithme de la classification spectrale. Comparé aux quatre méthodes de segmentation, la méthode de segmentation proposée donne de meilleurs résultats.

Le développement de cette méthode nous a amené à une remarque très importante : un des limites de la méthode de classification spectrale est le nombre de classes à fixer au préalable. Une solution envisageable serait de le déduire automatiquement à partir du modèle de saillance proposé, vu que l'algorithme Self Organizing Tree fournit en sortie le nombre et la taille des clusters.

Théoriquement, la méthode de segmentation d'objet proposée peut être appliquée à la segmentation sémantique vu que l'algorithme de classification spectrale procède à un partitionnement (clustering) sur k vecteurs propres en utilisant n'importe quel algorithme de partitionnement tel que les k-moyennes, la classification hiérarchique. Afin de valider cette hypothèse, l'évaluation expérimentale doit se faire sur une base d'image comportant plusieurs objets saillants, vue que la base MSRA-1000 ne contient que des images simples contenant un seul objet saillant. Par ailleurs, la re-génération de la carte de saillance est indispensable.

FIGURE 4.10 – Comparaison des résultats de segmentation d'objet obtenus avec différentes méthodes de segmentation sur la base d'images MSRA-1000. Les images sont présentées de gauche à droite selon les lignes dans l'ordre suivant : image originale, méthode NCut-MS [COUR et collab., 2005], méthode GB [FELZENSZWALB et HUTTENLOCHER, 2004], méthode Ncut [SHI et MALIK, 2001], méthode MSCOMANICIU et MEER [2002], méthode proposée et vérité terrain.

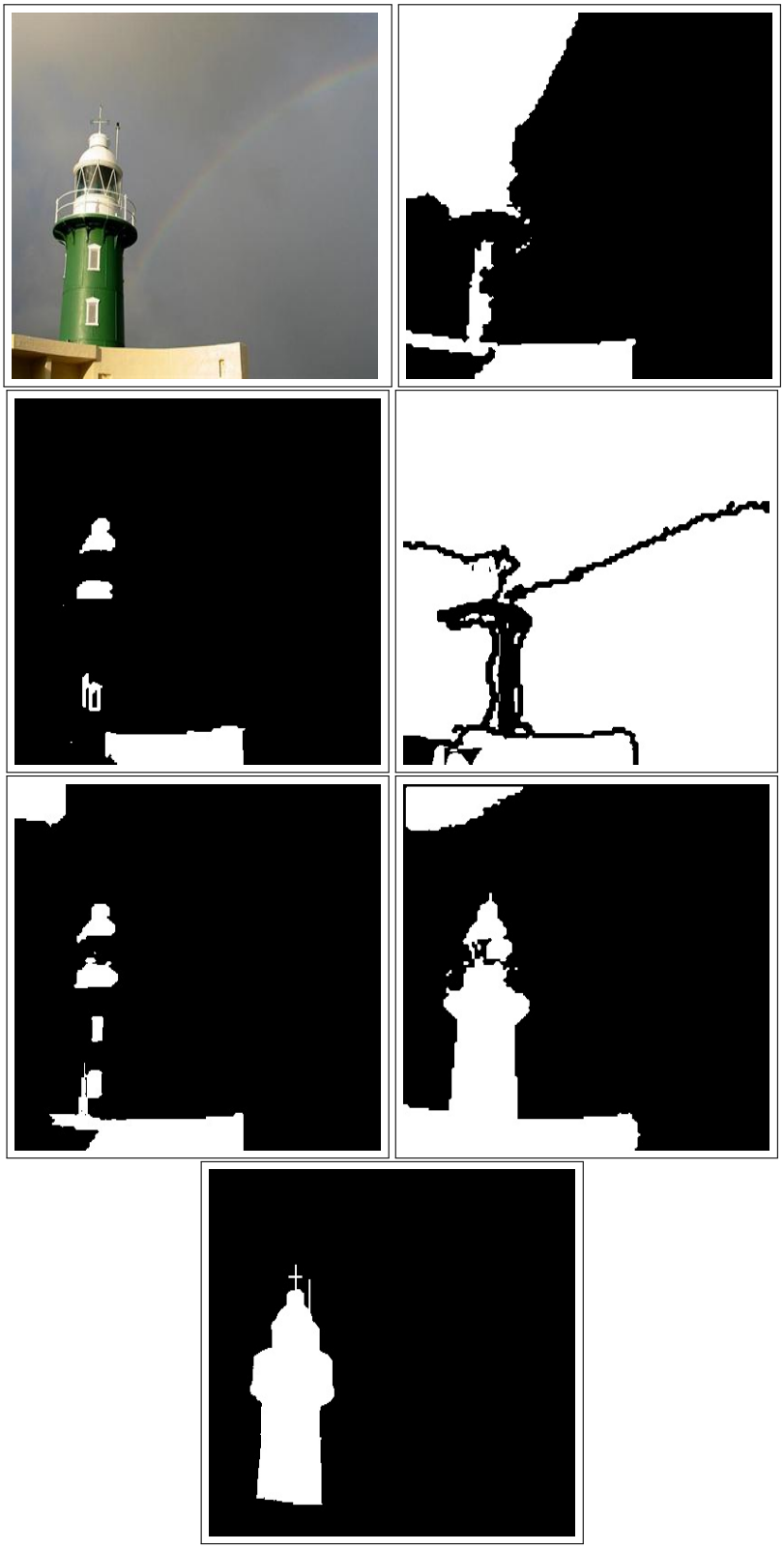


FIGURE 4.11 – Comparaison des résultats de segmentation d'objet obtenus avec différentes méthodes de segmentation sur la base d'images MSRA-1000. Les images sont présentées de gauche à droite selon les lignes dans l'ordre suivant : image originale, méthode NCut-MS [COUR et collab., 2005], méthode GB [FELZENSZWALB et HUTTENLOCHER, 2004], méthode Ncut [SHI et MALIK, 2001], méthode MSCOMANICIU et MEER [2002], méthode proposée et vérité terrain.

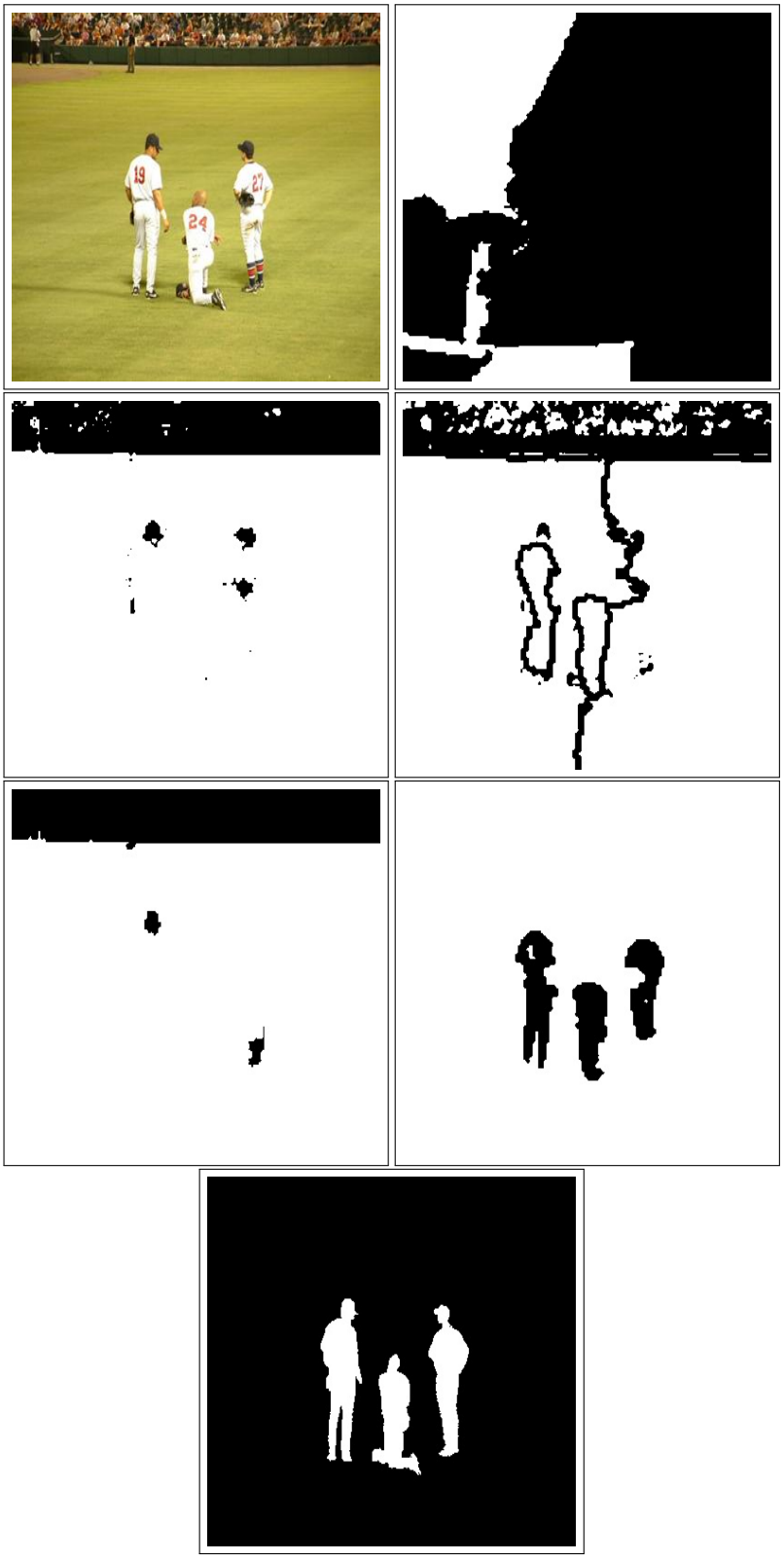


FIGURE 4.12 – Comparaison des résultats de segmentation d'objet obtenus avec différentes méthodes de segmentation sur la base d'images MSRA-1000. Les images sont présentées de gauche à droite selon les lignes dans l'ordre suivant : image originale, méthode NCut-MS [COUR et collab., 2005], méthode GB [FELZENSZWALB et HUTTENLOCHER, 2004], méthode Ncut [SHI et MALIK, 2001], méthode MSCOMANICIU et MEER [2002], méthode proposée et vérité terrain.

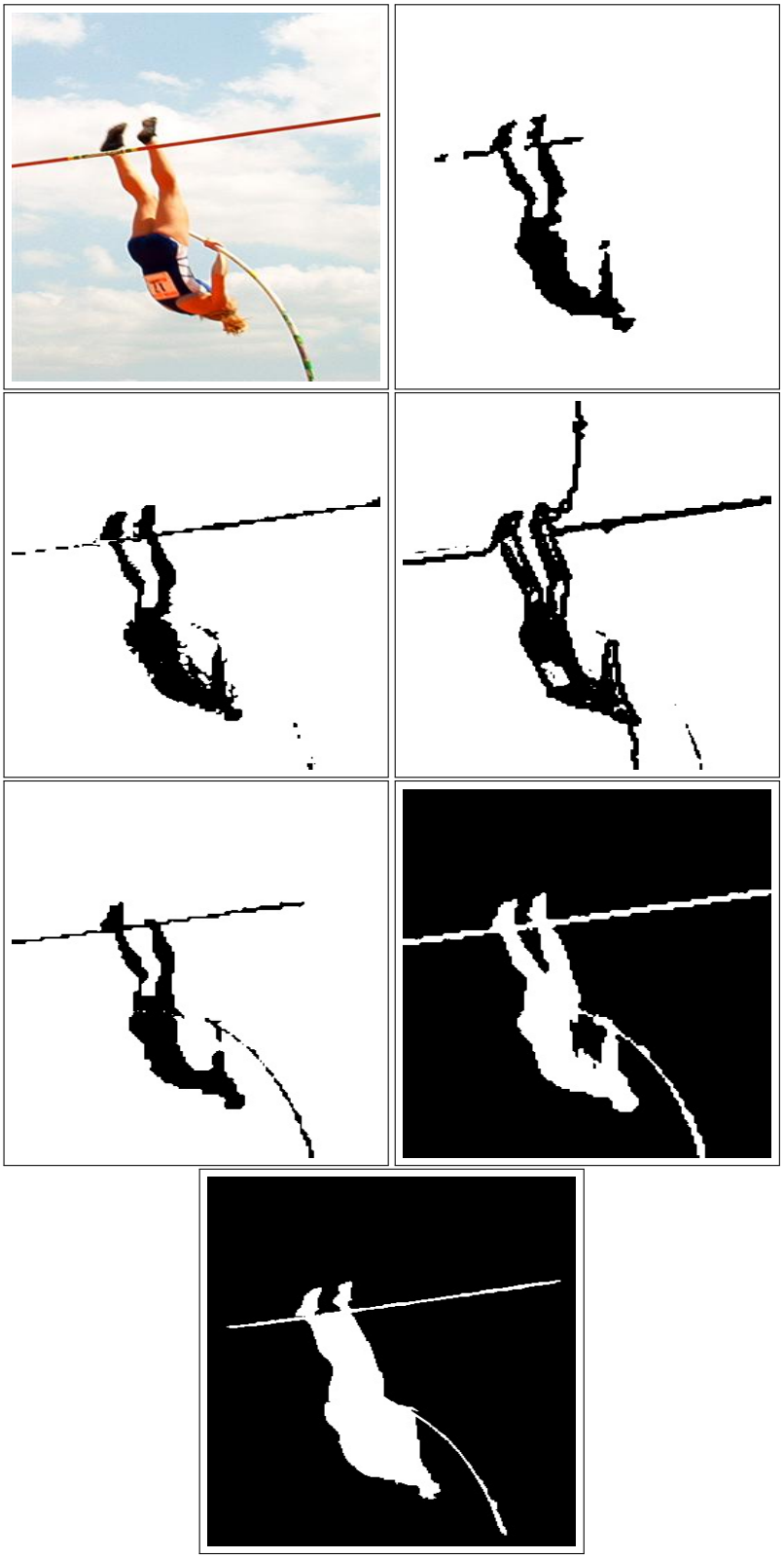


FIGURE 4.13 – Comparaison des résultats de segmentation d'objet obtenus avec différentes méthodes de segmentation sur la base d'images MSRA-1000. Les images sont présentées de gauche à droite selon les lignes dans l'ordre suivant : image originale, méthode NCut-MS [COUR et collab., 2005], méthode GB [FELZENSZWALB et HUTTENLOCHER, 2004], méthode Ncut [SHI et MALIK, 2001], méthode MSCOMANICIU et MEER [2002], méthode proposée et vérité terrain.

4.7 Conclusion

Dans ce chapitre, nous avons proposé un modèle de saillance visuelle qui se base sur une approche connexioniste et plus particulièrement sur l'algorithme Self-Organizing Tree(SOTA).Il combine les avantages de la classification hiérarchique descendante et des cartes auto-organisation de Kohonen(SOM). Ayant une topologie d'un arbre binaire initialisé à un nœud parent et deux cellules descendantes, la croissance de ce réseau de neurone se fait de manière dynamique. Le nœud possédant la plus grande diversité est divisé en deux cellules descendantes.

Après une étape d'extraction des caractéristiques de couleur dans l'espace CIE LAB et de texture en utilisant les filtres Gabor et Log-Gabor, le réseau SOTA procède au partitionnement des vecteurs caractéristiques des pixels de l'image en différents clusters. Similaire à [FU et collab., 2013], pour chaque cluster, nous calculons la mesure de saillance basée sur le prior center. En supposant que la vraisemblance de la saillance d'un pixel appartenant à un cluster satisfait une distribution gaussienne. La carte de saillance finale est obtenue en calculant la probabilité marginale de la saillance.

D'après les travaux étudiés dans le chapitre 2, tous les modèles de saillance proposés ont utilisé une mesure de saillance qui se base sur le contraste d'intensité/de couleur combiné ou pas à d'autres mesures de saillance. Comparé aux travaux étudiés, aucune mesure de saillance se basant sur le contraste n'a été utilisé. Seulement la mesure de saillance basé sur le prior center a été calculé, localement, au niveau de chaque cluster.

Pour évaluer la qualité de la carte de saillance obtenue , nous avons adopté le même protocole d'évaluation que [ACHANTA et collab., 2009]. Le modèle de saillance proposé est évalué sur la base d'images MSRA-1000, en calculant la courbe de précision-rappel et la mesure-F. Les résultats obtenus ont démontré l'efficacité du modèle proposé comparé à 12 autres modèles de saillance de l'état de l'art.

Références

ACHANTA, R., F. J. ESTRADA, P. WILS et S. SÜSSTRUNK. 2008, «Salient Region Detection and Segmentation», dans *Proceedings of the 6th International Conference on Computer Vision Systems (ICVS'08)*, vol. 5008, Springer Lecture Notes in Computer Science, Springer,Berlin,Heidelberg, p. 66–75. x, 108, 110, 113

ACHANTA, R., S. HEMAMI, F. ESTRADA et S. SUSSTRUNK. 2009, «Frequency-tuned Salient Region Detection», dans *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition(CVPR'09)*, IEEE, p. 1597–1604. x, 102, 103, 108, 110, 113, 122

ACHANTA, R. et S. SÜSSTRUNK. 2010, «Saliency detection using maximum symmetric surround», dans *Proceedings of the 17th IEEE International Conference on Image Processing(ICIP'10)*, IEEE, p. 2653–2656. x, 108, 110, 113

CHEBBOUT, S. et H. F. MEROUANI. 2012, «Comparative study of clustering based colour image segmentation techniques», *2012 Eighth International Conference on Signal Image Technology and Internet Based Systems*, p. 839–844. 102

CHEBBOUT, S. et H. F. MEROUANI. 2015, «A novel saliency detection model based on self organizing tree algorithm», dans *Proceedings of the International Conference on Intelligent Information Processing, Security and Advanced Communication*. 110

CHENG, M., G. ZHANG, N. MITRA, X. HUANG et S. HU. 2011, «Global contrast based salient region detection», dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*, IEEEComputer Society, p. 409–416. x, 108, 110, 113

COMANICIU, D. et P. MEER. 2002, «Mean shift : A robust approach toward feature space analysis», *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI'02)*, vol. 42, n° 5, p. 603–619. x, xi, 115, 118, 119, 120, 121

COUR, T., F. BENEZIT et J. SHI. 2005, «Spectral segmentation with multiscale graph decomposition», dans *IEEEConference on Computer Vision and Pattern Recognition(CVPR'05)*, vol. 2, IEEEComputer Society, p. 1124–1131. x, xi, 115, 118, 119, 120, 121

DOPAZO, J. et J. CARAZO. 1997, «Phylogenetic reconstruction using and unsupervised growing neural network that adopts the topology of a phylogenetic tree», *Journal of Molecular Evolution*, vol. 44, n° 2, p. 226–233. 106

FELZENSZWALB, P. et D. HUTTENLOCHER. 2004, «Efficient graph-based image segmentation», *International Journal of Computer Vision(IJCV)*, vol. 59, n° 2, p. 167—-181. x, xi, 115, 118, 119, 120, 121

FIELD, D. 1987, «Relations between the statistics of natural images and the response properties of cortical cells», *Journal of Optical Society of America,A*, vol. 4, n° 12, p. 2379–2394. 105

FU, H., X. CAO et Z. TU. 2013, «Cluster-based co-saliency detection», *IEEE Transactions on Image Processing (TIP)*, vol. 22, n° 10, p. 3766–3778. 107, 110, 122

GOFERMAN, S., L. ZELNIK-MANOR et A. TAL. 2012, «Context-aware saliency detection», *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, p. 1915–1926. x, 108, 110, 113

HAREL, J., C. KOCH et P. PERONA. 2006, «Graph-based visual saliency», dans *Advances in Neural Information Processing Systems 19(NIPS)*, p. 545—-552. x, 108, 110, 113

HERRERO, J., A. VALENCIA et J. DOPAZO. 2001, «A hierarchical unsupervised growing neural network for clustering gene expression patterns», *Bioinformatics*, vol. 17, nº 2, p. 126–136. 106

HOU, X., J. HAREL et C. KOCH. 2012, «Image signature : Highlighting sparse salient regions», *IEEETransactions on Pattern Analysis and Machine Intelligence(PAMI'12)*, vol. 34, nº 1, p. 194–201. x, 108, 110, 113

HOU, X. et L. ZHANG. 2007, «Saliency detection : A spectral residual approach», dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR'07*, IEEE, p. 1–8. x, 108, 110, 113

IMAMOGLU, N., W. LIN et Y. FANG. 2013, «A saliency detection model using low-level features based on wavelet transform», *IEEE transactions on multimedia*, vol. 15, nº 1, p. 96–105. x, 108, 110, 113

ITTI, L., C. KOCH et E. NIEBUR. 1998, «A model of saliency-based visual attention for rapid scene analysis», *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI'98)*, vol. 20, p. 1254–1259. x, 108, 110, 113

LIU, T., Z. YUAN, J. SUN, J. WANG, N. ZHENG, X. TANG et H. Y. SHUM. 2011, «Learning to detect a salient object», *IEEETransaction Pattern Analysis Machine Intelligence (PAMI'11)*, vol. 33, nº 2, p. 353–367. 102

MA, Y. F. et H. J. ZHANG. 2003, «Contrast-based image attention analysis by using fuzzy growing», dans *Proceedings of the 11th ACMInternational Conference on Multimedia*, ACM, p. 374–381. x, 108, 110, 113

NG, A. Y., M. I. JORDAN et Y. WEISS. 2001, «On spectral clustering : Analysis and an algorithm», dans *Proceedings of the 14th International Conference on Neural Information Processing Systems : Natural and Synthetic(NIPS'01)*, MIT Press, p. 849–856. 114

SHI, J. et J. MALIK. 2001, «Normalized cuts and image segmentation», *IEEETransactions on Pattern Analysis and Machine Intelligence(PAMI'00)*, vol. 22, nº 8, p. 888—-905. x, xi, 115, 118, 119, 120, 121

WALTHER, D., L. ITTI, M. RIESENHUBER, T. POGGIO et C. KOCH. 2002, «Attentional selection for object recognition—a gentle way», dans *Biologically motivated computer vision*, Springer, p. 251–267. 108

WEIBIN, Y., T. Y. YAN, F. BIN, S. ZHAOWEI et L. YUEWEI. 2013, «Visual saliency detection with center shift», *Journal of Neurocomputing*, vol. 103, nº 1, p. 63–74. x, 108, 110, 113

ZELNIK-MANOR, L. et P. PERONA. 2004, «Self-tuning spectral clustering», dans *Proceedings of the 17th International Conference on Neural Information Processing Systems(NIPS'04)*, MIT Press, p. 1601–1608. 114

Deuxième partie

RECONNAISSANCE VISUELLE

Chapitre 5

La reconnaissance d'objets

« Plus l'épreuve s'alourdit contre toi, plus tu es sur que la solution venant d'Allah approche »

Ibn Al Qayyim

Sommaire

5.1 Introduction

La reconnaissance d'objets dans les images est un problème difficile en vision par ordinateur. Les principales raisons sont dues à la fois aux fortes variations intra-classe et aux similitudes inter-classes. Les objets d'une même catégorie peuvent sembler très différents, tandis que les objets de différentes catégories peuvent sembler assez similaires. De plus, en fonction du point de vue, de l'échelle et de l'éclairage différents, le même objet peut même sembler différent sur les images. L'encombrement du fond et l'occlusion partielle augmentent également les difficultés de reconnaissance des objets. Le terme objet a une large et vaste définition, c'est pourquoi ce domaine de recherche est aussi connu sous le nom de reconnaissance de formes(RdF). Au cours de ces dernières décennies, les chercheurs de la communauté de la vision par ordinateur ont accordé beaucoup d'attention à ce champ de recherche et de nombreuses approches ont vu le jour. Dans ce chapitre, nous abordons le problème de la reconnaissance d'objet génériques/spécifiques. Nous présentons un aperçu des représentations appropriées pour la reconnaissance des catégories d'objets génériques. Ensuite, nous présentons les différentes approches de reconnaissance de catégories d'objets génériques. Enfin, nous passons en revue les principales bases d'images utilisés dans la reconnaissances de catégories d'objets.

5.2 Reconnaissance d'objets génériques/spécifiques

La reconnaissance visuelle d'objets est abordée dans la communauté des chercheurs de vision selon deux types différents : la reconnaissance d'instances d'objet et la reconnaissance de classes ou catégories d'objet.

La reconnaissance d'instances d'objet se définit en la tâche de reconnaître des instances d'objets spécifiques, tels que des endroits, monuments et personnes particulière(e)s. Par exemple, la figure 5.1 illustre différentes instances de la place Makam El Chahid ainsi qu' El Masjid El-Nabawi ElCharif.

Contrairement à la reconnaissance d'instances d'objet, la reconnaissance de classes/catégories d'objet traite des catégories d'objets génériques comme les voitures, les visages, les avions. Par exemple, la figure 5.2 illustre différentes instances de catégories génériques de mosquées et d'avions.

FIGURE 5.1 – Des instances différents d'objets particuliers.

FIGURE 5.2 – Des instances différentes de catégories d'objet génériques.

5.3 Représentation des catégories d'objet

La représentation des catégories d'objets dans les images peut se faire selon une représentation globale ou locale.

5.3.1 Représentation globale

Représenter de manière globale des catégories d'objets repose sur l'extraction de caractéristiques globales qui porte sur la totalité des pixels des catégories d'images. La représentation globale la plus simple est une concaténation directe des intensités de pixels en un seul vecteur de caractéristiques, qui peut ensuite être éventuellement réduit par des méthodes comme l'ACP, ou bien l'histogramme des intensités des pixels. Une autre représentation globale consiste à décrire la distribution de la couleur à travers les histogramme de couleur. Ainsi, chaque case(bin) de l'histogramme est associé à la fréquence d'une valeur de couleur dans l'image. Les histogrammes présentent l'avantage d'être invariants aux transformations géométriques et d'avoir une certaine tolérance aux occlusions partielles. Les moments de couleur représentent également la distribution globale de la couleur dans une catégorie d'image. Le moment d'ordre 1 calcule la valeur couleur moyenne d'une image, le moment d'ordre 2 fournit une information sur la variance des valeurs de couleur et le moment d'ordre 3 calcule le degré d'asymétrie de la distribution de la couleur,le skewness.

5.3.2 Représentation locale

Les images sont considérées comme un ensemble de régions locales plutôt qu'une entité entière. La détermination de la localisation de ces régions locales donne deux types de caractéristiques locales : dense et parcimonieuse.

Échantillonnage dense

L'échantillonnage dense, également appelé échantillonnage en grille, est une méthode populaire d'échantillonnage d'image et qui est utilisée pour sa simplicité. L'image est segmentée en sous-régions par quelques lignes horizontales et verticales selon une grille régulière. Les caractéristiques sont extraites dans chaque sous-région.

FIGURE 5.3 – Échantillonnage dense d'une image.

Échantillonnage parcimonieux

L'échantillonnage parcimonieux définit le contenu informatif des images comme la partie autour des points d'intérêt locaux, en se basant sur le fait que les détecteurs de points d'intérêt sont puissants pour localiser les points locaux informatifs dans les images. Cette méthode d'échantillonnage dépend beaucoup des détecteurs de points d'intérêt utilisés comme le détecteur de Moravec [MORAVEC, 1977], le détecteur de Harris [HARRIS et STEPHENS, 1988], la différence de Gaussian(DoG)[LOWE, 2004], le détecteur de Gabor-Wavelet [**?**], le détecteur SIFT [LOWE, 2004] .

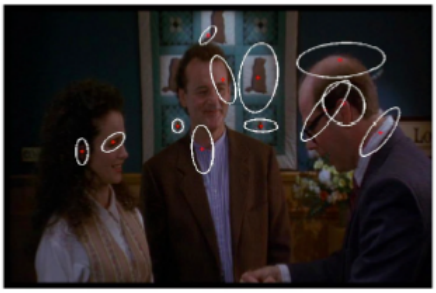


FIGURE 5.4 – Échantillonnage parcimonieux d'une image.

Après avoir localisé les régions locales parcimonieuses/denses, ces régions sont représentées en termes de caractéristiques descriptives, ce qui est souvent représenté par des vecteurs de caractéristiques. Ces descripteurs doivent être hautement discriminants et faciles à générer.On peut citer par exemple, les descripteurs SIFT, SURF, PCA-SIFT.

5.3.3 Représentation sac à mots visuels

L'approche de sac à mots(Bag of words, BoW) est à l'origine une hypothèse simplificatrice dans le traitement du langage naturel. Ce groupe de méthodes est populaire dans la classification de documents. Il représente le texte comme un sac, qui contient la collection de mots du dictionnaire et ignore l'ordre des mots et la grammaire comme l'illustre la figure 5.5.

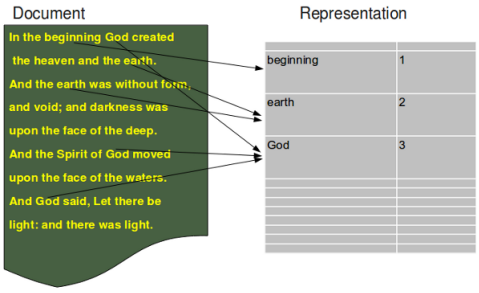


FIGURE 5.5 – La représentation de document basée sur le modèle de sac de mots, image issue du cours de L.Fei-Fei 2009.

En vision par ordinateur, il existe un traitement similaire. Une image est représentée sous forme d'un histogramme de vecteur caractéristiques qui sont extraites à partir

d'une grille régulière ou d'un ensemble de points clés. Chaque caractéristique est un mot visuel qui représente une entrée dans le dictionnaire visuel. Ce dernier est généralement normalement généré par la méthode des k-moyennes. Dans cette approche de sac à mots visuels illustrée par la figure 5.6, toutes les caractéristiques sont indépendantes et ce modèle de représentation considère l'image comme une collection de ces caractéristiques.

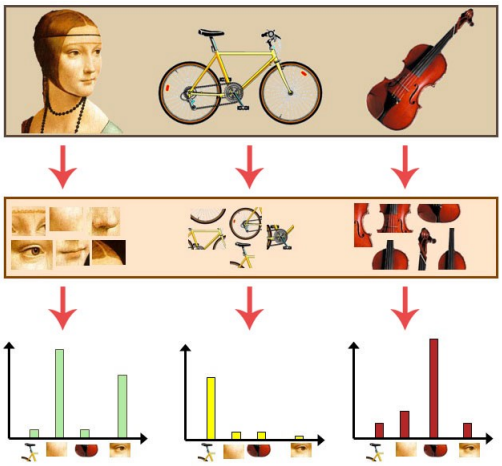


FIGURE 5.6 – La représentation d'image basée sur le modèle de sac de mots visuels.

5.4 Approches de reconnaissance d'objets génériques

Dans cette section, nous présentons trois différentes approches de reconnaissance des catégories d'objets génériques : approche basée sur l'apparence, approche basée sur les caractéristiques et classificateurs et approche basée sur les parties.

5.4.1 Approche basée l'apparence

Les premiers travaux réalisés sur la reconnaissance visuelle de catégorie d'objets se sont orientés vers une approche basée sur l'apparence des objets en terme d'intensité de pixel et de texture. Cependant, cette approche se basent sur des vecteurs caractéristiques de grande dimension qui sont sensibles au bruit et aux conditions d'éclairage. Pour palier à ce problème, le recours à des techniques de réduction de dimension comme l'analyse à composante principale(ACP) est généralement utilisé .

L'approche basée sur l'apparence comprend deux étapes. Dans la première étape d'apprentissage hors ligne, un ensemble de modèles d'apprentissage est obtenu. Ces images capturent généralement l'apparence d'un seul objet sous différentes orientations, directions d'éclairage ou de multiples instances d'une classe d'objets comme par exemple des visages. Les visages propres(eigenfaces en anglais) sont un ensemble de vecteurs propres utilisés dans le domaine de la vision artificielle afin de résoudre le problème de la reconnaissance de visage. Le recours à des visages propres pour la reconnaissance a été développé par Sirovich et Kirby en 1987 et utilisé par [TURK et PENTLAND, 1991] pour la classification de visages. Un ensemble de visages propres peut être dérivé en effectuant une analyse de composantes principales sur une grande collection

d'images de visage. Ces visages propres constituent une base et d'autres images de visage peuvent être représentées par un ensemble de coefficients de reconstruction. Les visages propres fournissent une représentation compacte de l'apparence du visage, tout en réduisant les dimensions de l'espace propre(eigenspace).

Dans la deuxième étape de reconnaissance en ligne, étant donné une image de test(par exemple le visage d'une personne, le système de reconnaissance projette ses parties qui correspondent à des sous images de même taille que les images d'entraînement dans l'espace propre. Les coefficients récupérés indiquent le visage spécifique de la personne.

Dans leurs travaux, [LEONARDIS et BISCHOF, 2000] ont étendu l'application de l'approche basée sur l'apparence pour la reconnaissance de visage à la reconnaissance de catégories d'objets génériques. L'algorithme qu'ils ont proposé est décrit par la figure 5.7. Le côté gauche représente l'étape d'entraînement hors ligne. L'entrée est un ensemble d'images d'apprentissage pour chaque objet. La sortie se compose des images propres et des coefficients des images d'apprentissage ou, en variante, des espaces propres paramétriques. Le côté droit de la figure 5.7 représente l'étape de reconnaissance en ligne. En entrée, il reçoit la sortie de l'étape d'apprentissage (espaces propres et coefficients pour chaque objet) et une image dans laquelle des instances d'objets d'apprentissage doivent être reconnues. A chaque emplacement de l'image, plusieurs hypothèses sont générées pour chaque espace propre. La procédure de sélection raisonne ensuite parmi différentes hypothèses, appartenant éventuellement à des objets différents, et sélectionne celles qui expliquent le mieux les données, délivrant ainsi automatiquement le nombre d'objets, les espaces propres auxquels ils appartiennent et les coefficients, via la recherche du plus proche voisin déjà effectuée à l'étape de génération d'hypothèses. Cet algorithme a été testé sur les bases d'images Columbia Object Image Library et COIL-20.

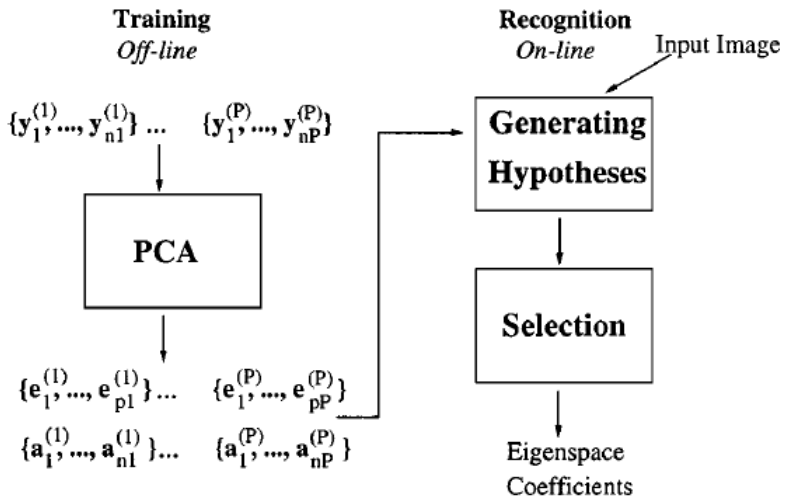


FIGURE 5.7 – Reconnaissance de catégorie d'objet selon une approche basée sur l'apparence.

5.4.2 Approche basée les caractéristiques/classificateurs

Les approches basées sur les caractéristiques et les classificateurs consistent en deux étapes principales. La première étape est l'extraction et la représentation de caractéristiques d'image, qui vise à extraire un ensemble de vecteurs de caractéristiques (descripteurs), d'une image pour décrire son contenu visuel, et à transformer les caractéristiques extraites en représentations plus compactes et informatives en appliquant certaines méthodes de modélisation d'image. La deuxième étape est la classification d'images, qui accepte les représentations d'images sur la base des caractéristiques extraites et effectue la classification finale en utilisant certains algorithmes de reconnaissance de modèle(classificateurs). De plus, étant donné que différentes caractéristiques peuvent véhiculer des informations complémentaires les unes aux autres, des stratégies de fusion sont également nécessaires pour améliorer encore plus les performances de reconnaissance.

Les approches basées sur les caractéristiques et les classificateurs sont devenues populaires pour la reconnaissance d'objets depuis la fin des années 1990, en raison du développement avancé des caractéristiques et descripteurs d'image ainsi que des algorithmes de reconnaissance de formes. Notamment, en utilisant des descripteurs locaux, SIFT [LOWE, 2004], ainsi que la représentation sac à mots visuels [CSURKA et collab., 2004] suivi de classificateurs discriminants tels que les machines à vecteurs de support(SVM) qui est devenu le paradigme dominant depuis 2004, voir la figure 5.8.

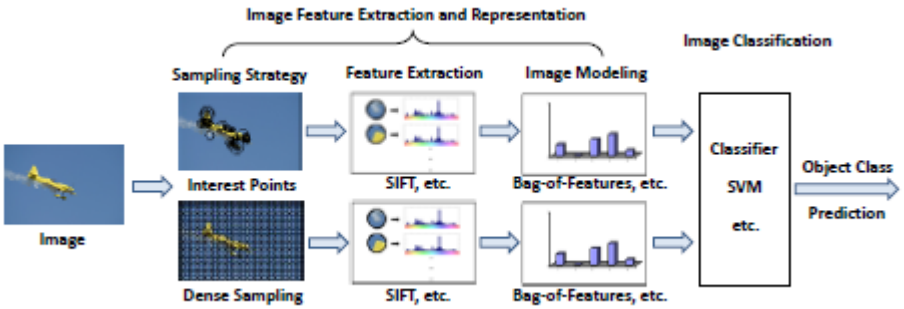


FIGURE 5.8 – Reconnaissance de catégories d'objets basée sur l'approche de caractéristiques/classificateurs, extrait de [ZHU, 2012].

5.4.3 Approche basée les parties

L'idée de base de la reconnaissance d'objets basée sur les parties a été introduite par la représentation de structure picturale de [FISCHLER et ELSCHLAGER, 1973] où un objet est modélisé par une collection de parties disposées dans une configuration déformable. Chaque partie code les propriétés visuelles locales de l'objet, et la configuration déformable est caractérisée par des connexions de type ressort (spring-like connections) entre certaines paires de parties.

Pour les visages, les parties sont des caractéristiques telles que les yeux, le nez et la bouche, et les connexions en forme de ressort permettent de varier les emplacements relatifs de ces caractéristiques. Pour les personnes, les parties sont les membres, le torse et la tête, et les connexions en forme de ressort permettent une articulation au niveau des jointures 5.9.

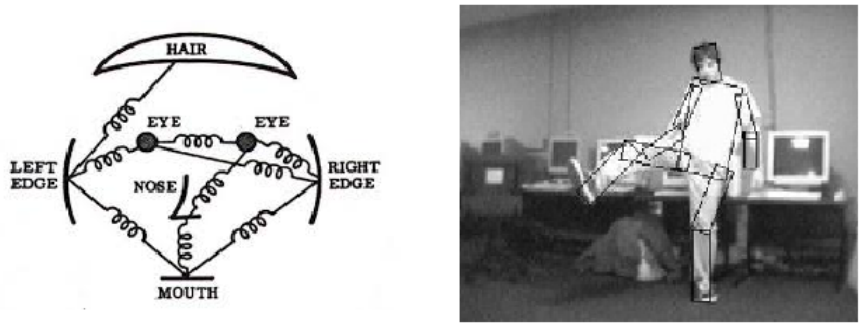


FIGURE 5.9 – Exemples de modèles d'objets basée sur les parties.

La reconnaissance est obtenue en trouvant la meilleure correspondance entre un tel modèle basé sur les parties et une image de test. La meilleure correspondance peut être trouvée en minimisant une fonction d'énergie qui mesure à la fois un coût de correspondance pour chaque partie et un coût de déformation pour chaque paire de parties connectées.

Dans leurs travaux, [FELZENSZWALB et HUTTENLOCHER, 2004] ont proposé un algorithme efficace pour le problème classique de minimisation d'énergie de la structure picturale décrit dans [FISCHLER et ELSCHLAGER, 1973], pour le cas où les connexions entre les parties ne forment aucun cycle et sont d'un type général. De nombreux objets, y compris des visages, des personnes et des animaux peuvent être représentés par de tels modèles en plusieurs parties acycliques. [FELZENSZWALB et HUTTENLOCHER, 2004] ont présentent une méthode d'apprentissage de ces modèles à partir d'exemples d'apprentissage. Cette méthode apprend tous les paramètres du modèle, y compris la structure des connexions entre les parties.

L'idée des parties et des approches basées sur la structure vient de l'observation que la plupart des objets se composent généralement de plusieurs parties individuelles qui sont disposées dans certaines structures géométriques. Par exemple, un visage se compose de deux yeux, un nez et une bouche, tandis qu'un avion se compose de deux ailes, un fuselage et une queue.

Les modèles déformables basés sur les pièces ont donc été proposés pour exploiter cette observation en décomposant un objet en parties connectées. Pour un objet, chaque pièce code ses propriétés visuelles locales, tandis que la configuration déformable est représentée par des connexions entre certaines paires de pièces pour définir sa structure géométrique globale.

La reconnaissance est obtenue en trouvant la meilleure correspondance entre un tel modèle basé sur les pièces et une image d'entrée. La meilleure correspondance peut être trouvée en minimisant une fonction d'énergie qui mesure à la fois un coût de correspondance pour chaque pièce et un coût de déformation pour chaque paire de pièces connectées.

5.5 Principales bases d'images

Dans cette section, nous présentons une liste non exhaustive des bases d'images disponibles pour l'apprentissage et le test des algorithmes de reconnaissance de catégories d'objets.

5.5.1 Caltech101 Database

La base d'images Caltech101 [1] contient en total 9146 images, réparties en 101 classes d'objets différentes (y compris les avions, les animaux, les visages, les véhicules, les chaises, les fleurs, les pianos, ⋯.) et une catégorie d'arrière-plan supplémentaire. Le nombre d'images dans chaque catégorie varie de 31 à 800, et la plupart des catégories ont environ 50 images. L'ensemble de données n'est pas divisé en un ensemble d'entraînement et un ensemble de tests prédéfinis, et la stratégie commune pour les expérimentations consiste à sélectionner au hasard (5,10,15,20,25,30) comme le nombre d'images de chaque classe pour l'ensemble d'apprentissage et le reste des images pour l'ensemble de test. La précision de classification moyenne dans toutes les classes est utilisée comme critère d'évaluation.

5.5.2 Caltech256 Database

La base d'image Caltech256 [GRIFFIN et collab., 2007] [2] est le successeur de la base d'image Caltech-101 créé en 2006. Elle contient 30 607 images avec un minimum de 80 images par catégorie qui définissent une variation intra-classe améliorée sans alignement artificiel des objets. Elle est répartie en 256 catégories. Une nouvelle catégorie d'arrière-plan plus large appelée clutter a été introduite pour tester le rejet d'arrière-plan.

5.5.3 PASCAL VOC Database

la base de données PASCAL VOC [EVERINGHAM et collab., 2010] est devenu une référence standard pour évaluer les algorithmes de reconnaissance et de détection d'objets, car toutes les annotations ont été mises à disposition en 2007 par les organisateurs mais depuis lors, les annotations ne sont plus rendu accessibles au public. La base de données PASCAL VOC 2007 contient près de 10 000 images de 20 classes d'objets, qui contiennent un nombre différent d'images, des centaines à des milliers. L'ensemble de données est divisé en un ensemble d'apprentissage prédéfini (2501 images), un ensemble de validation (2510 images) et un ensemble de test (4952 images). La précision moyenne(Mean Average Precisionen anglais, MAP) dans toutes les classes est utilisée comme critère d'évaluation. La précision moyenne(Average precision en anglais,AP) mesure l'aire sous la courbe précision-rappel pour chaque classe, et une bonne valeur de AP nécessite à la fois des valeurs de rappel et de précision élevées.

5.5.4 ImageNet Database

La base d'images ImageNet [DENG et collab., 2009] est une large base d'image organisée selon la hiérarchie de WordNet. Chaque concept significatif dans WordNet, éventuellement décrit par plusieurs mots ou phrases de mots, est appelé un ensemble de synonymes ou synset. Il existe plus de 100 000 synsets dans WordNet, et la majorité d'entre eux sont des noms. Le but de la base ImageNet est de fournir en moyenne 1000 images pour illustrer chaque synset. Les images de chaque concept sont soumises à un contrôle de qualité et des annotations humaine. Actuellement, ImageNet contient environ 15 millions d'images pour plus de 20 000 synsets, et le nombre d'images avec des

1. www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html
2. http://www.vision.caltech.edu/Image_Datasets/Caltech256/

annotations de boîte englobante est supérieur à 1 million. ImageNet peut offrir des dizaines de millions d'images bien triées pour la plupart des concepts de la hiérarchie WordNet.

5.6 Conclusion

Dans ce chapitre, une revue des principales approches proposées dans la littérature pour la reconnaissance des catégories d'objet est présentée. Un intérêt particulier sera accordée à l'approche basée sur les caractéristiques et les classificateurs vu qu'elle est devenue le framework le plus populaire pour la tâche de reconnaissance d'objets. Cette approche dépend de la création de mots visuels constituant un dictionnaire visuel, et qui sert à représenter le contenu des images. Nous évoquerons ces deux concepts en détails dans le chapitre suivant.

Références

CSURKA, G., C. DANCE, L. FAN, J. WILLAMOWSKI et C. BRAY. 2004, «Visual categorization with bags of keypoints», dans *Workshop on statistical learning in computer vision (ECCV'04)*, p. 1–22. 133

DENG, J., W. DONG, R. SOCHER, L. LI, K. LI et L. FEI-FEI. 2009, «Imagenet : A large-scale hierarchical image database», dans *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*. 135

EVERINGHAM, M., L. V. GOOL, C. K. I. WILLIAMS, J. M. WINN et A. ZISSERMAN. 2010, «The pascal visual object classes (voc) challenge», *International Journal of Computer Vision*, vol. 88, n° 2, p. 303–338. 135

FELZENSZWALB, P. F. et D. HUTTENLOCHER. 2004, «Pictorial structures for object recognition», *International Journal of Computer Vision*, vol. 61, p. 55–79. 134

FISCHLER, M. et R. A. ELSCHLAGER. 1973, «The representation and matching of pictorial structures», *IEEE Transactions on Computers*, vol. C-22, p. 67–92. 133, 134

GRIFFIN, G., A. HOLUB et P. PERONA. 2007, «Caltech-256 object category dataset», cahier de recherche, California Institute of Technology. 135

HARRIS, C. et M. STEPHENS. 1988, «A combined corner and edge detector», dans *Proceedings of the Fourth Alvey Vision Conference*, p. 147–151. 130

LEONARDIS, A. et H. BISCHOF. 2000, «Robust recognition using eigenimages», *Comput. Vis. Image Underst.*, vol. 78, p. 99–118. 132

LOWE, D. 2004, «Distinctive image features from scale-invariant keypoints», dans *International Journal of Computer Vision*, p. 91—110. 130, 133

MORAVEC, H. 1977, «Towards automatic visual obstacle avoidance», dans *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, p. 584–594. 130

TURK, M. et A. PENTLAND. 1991, «Eigenfaces for recognition», *Journal of Cognitive Neuroscience*, vol. 3, p. 71–86. 131

ZHU, C. 2012, *Effective and efficient visual description based on local binary patterns and gradient distribution for object recognition*, thèse de doctorat, Ecole Centrale de Lyon. xi, 133

Chapitre 6

Dictionnaires Visuels et Modèles de Dictionnaire

« *Le succès est la somme de petits efforts, répétés jour après jours.* »

Robert Collier

Sommaire

6.1 Introduction

L'approche de sac à mots visuels(BoVW) est un modèle de catégorisation d'objet très populaire en raison de sa simplicité et de ses performances remarquables O'HARA et DRAPER [2011]. Il est inspiré de la représentation mot-document (modèle de sac de mots) qui est utilisée dans l'exploration de texte [JOACHIMS, 1998] et la recherche d'informations textuelles BAEZA-YATES et RIBEIRO-NETO [1999] où un document est représenté par un ensemble de mots non ordonné. Sa première utilisation sur des images pour la reconnaissance de texture remonte à 2001 [LEUNG et MALIK, 2001]. Plus tard, [SIVIC et ZISSERMAN, 2003],[CSURKA et collab., 2004] ont été les premiers à introduire le modèle BoVW dans la communauté de la vision par ordinateur. Selon ce modèle de dictionnaire, une image est représentée sous forme d'histogramme de vecteur de caractéristiques qui indique le taux d'occurrence de mots visuels dans une image obtenu via une quantification vectorielle de l'espace des descripteurs d'images locaux (généralement des descripteurs SIFT)LOWE [2004]. Plus précisément, un framework de catégorisation d'objets basé sur un codebook commun qui implique les principales étapes d'extraction des caractéristiques, de construction du dictionnaire visuel, de codage des caractéristiques, d'agrégation des caractéristiques(feature pooling) et de classification, est partagé par plusieurs auteurs AVILA et collab. [2011]; BOUREAU et collab. [2010]; GOH et collab. [2014]; WANG et collab. [2014]. Au cours des ces dernières décennies, de nombreux travaux ont été proposés pour améliorer le modèle de dictionnaire et ceci à travers des méthodes avancées de génération de dictionnaire visuel ainsi que de nouvelles méthodes de codage et de mise en commun des caractéristiques visuelles. Dans ce qui suit, nous passons en revue les travaux connexes relatifs à de ces deux catégories.

6.2 Dictionnaire visuel et concepts

Les mots visuels constituant un dictionnaire visuel peuvent être obtenus selon deux approches différentes : la quantification vectorielle(Vector Quantization, VQ) ou la représentation parcimonieuse(Sparse Representation).

6.2.1 Dictionnaire basé sur VQ/Parcimonie

Dictionnaire basé sur VQ. En adoptant l'approche de quantification vectorielle, les mots visuels qui constituent le dictionnaire visuel sont obtenus à travers une classification non supervisé(clustering) dans laquelle les relations spatiales et sémantiques entre les vecteurs caractéristiques d'apprentissage des images sont ignorées durant leur regroupement. Soit une matrice $X \in R^{n\times p}$ représentant N points décrits selon M caractéristiques, l'algorithme des k-moyennes vise à partitionner les N points en K groupes (clusters) et représenter chaque cluster par son centre. Soit $C \in R^{p\times k}$ une matrice dont les colonnes contiennent les centres des k clusters, $C = [c_1, c_2, \ldots, c_k]$. L'objectif des k-moyennes est de minimiser une fonction objective, la fonction d'erreur quadratique, qui minimise la somme des erreurs quadratiques intra-groupe comme suit :

$$\ell_{k-moyennes}(C) = \sum_{x\in X} \min_i \|x - c_i\|_2^2 \tag{6.1}$$

$$\ell_{k-moyennes}(C) = \sum_{x\in X} \min_{b\in H_{1/k}} \|x - Cb\|_2^2 \tag{6.2}$$

où $H_{1/k} \equiv \{b | b \in \{0,1\}^k$ et $\|b\| = 1\}$, ce qui signifie, b est un vecteur binaire comprenant un codage de 1-à-k.

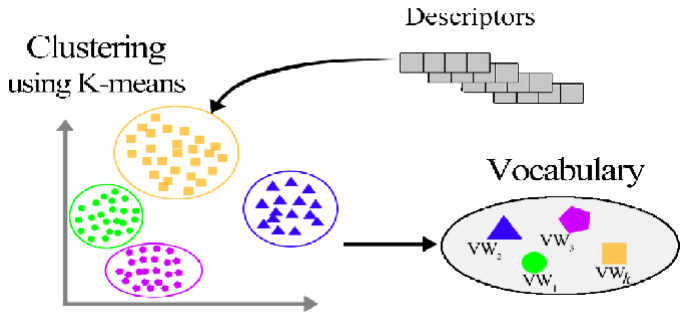


FIGURE 6.1 – Construction de dictionnaire visuel en se basant sur la quantification vectorielle, l'algorithme des k-moyennes).

Dictionnaire basé sur la parcimonie. On considère un ensemble de N patchs d'images $\{y_i\}_{i\in\{1,N\}}$. Chaque y_i est un vecteur de R^n, $n = s^2$ où s est la taille du patch carré considéré. Matriciellement, les données d'entrée seront représentées par la matrice Y, de taille $n \times N$, ou les $\{y_i\}$ sont les vecteurs colonnes. D est une matrice de taille $n \times K$ dont les vecteurs colonnes sont les K éléments du dictionnaire qu'on cherche à apprendre $\{d_j\}_{j\in\{1,K\}}$. Le dictionnaire D est composé de K colonnes $d_k, k = 1, \cdots, K$ appelées atomes, chacune d'elles supposée normalisée, c'est à dire de norme l_2 unitaire tel que $\|d_k\|^2 = 1, \forall k = 1, \ldots, K$.

Le dictionnaire est en général sur-complet, ce qui signifie qu'il comporte davantage d'atomes que la dimension de chaque atome qu'il contient (K > n). Si K < n, on dit que le dictionnaire est sous-complet, et complet si K = n. [BARTHÉLEMY, 2013]

X est la matrice de taille K × N contenant l'ensemble des projections de Y sur D. Les vecteurs colonne de X, $\{x_k\}_{k\in\{1,N\}}$ contiennent l'ensemble des projections de y_k sur D tel que $y_k \approx Dx_k$. On va chercher à représenter Y par une combinaison linéaire X d'éléments de base, notés D tel que $Y \approx DX$. La figure 6.2 illustre cette approximation pour un seul patch y.

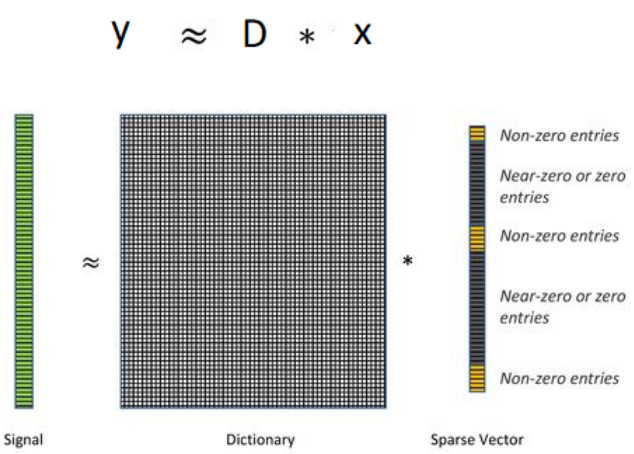


FIGURE 6.2 – Construction de dictionnaire visuel en se basant sur la représentation parcimonieuse : y ≈ D x avec x un vecteur parcimonieux.

Afin de représenter au mieux les données, images, le problème consiste donc à trouver D et X minimisant le problème d'optimisation suivant [MAZAHERI, 2015] :

$$\min_{D,X} = \{\|Y - DX\|_F^2\} \tag{6.3}$$

où $\|X\|_F$ est la norme de Frobenius de la matrice X.

L'apprentissage de dictionnaire consiste à trouver le dictionnaire optimal permettant de représenter efficacement l'ensemble des vecteurs d'apprentissage, ou vecteurs d'entrainement, de façon parcimonieuse, de telle sorte que $Y \approx DX$, avec $X \in R^{K \times N}$ une matrice dont chaque colonne $x_i, i = 1, \ldots, N$ est parcimonieuse. Le problème est donc le suivant :

$$\min_{D,X} \|Y - DX\|_F^2, \|x_i\|_0 \leq L \forall i \, et \, \|d_k\|_2 = 1 \forall k \tag{6.4}$$

avec $L > 0$ la contrainte de parcimonie de chaque colonne de X, c'est à dire le nombre maximal de valeurs non nulles dans chaque colonne $x_i, i = 1, \ldots, N$, et $\|.\|_F$ la norme de Frobenius : $\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}$

La résolution de ce problème fait appel aux méthodes d'apprentissage de dictionnaire qui se décomposent en deux étapes majeures : une étape de représentation parcimonieuse dite de codage parcimonieux et une étape de mise à jour du dictionnaire, qui sont itérées jusqu'à atteindre un critère d'arrêt comme par exemple, l'erreur de représentation des vecteurs d'apprentissage à atteindre, le test sur la convergence de cette erreur, ou le nombre maximum d'itérations fixe. L'étape de codage parcimonieux cherche à trouver la matrice de coefficients X correspondant à la représentation parcimonieuse des vecteurs d'apprentissage Y sur le dictionnaire D, fixe à cette étape. Le problème peut ainsi s'écrire :

$$\min_{x_i} \|y_i - Dx_i\|_2^2, \|x_i\|_0 \leqslant L \forall i \tag{6.5}$$

avec L la contrainte de parcimonie sur chaque vecteur x_i. Cela revient à un problème de décomposition parcimonieuse qui peut être résolu par les algorithmes de décomposition tels que MP, OMP, ou BP si la parcimonie est contrôlée par la norme l_1.

La seconde étape correspond à la mise à jour du dictionnaire D. Ayant calculé la matrice de coefficients X précédemment, cette étape a pour but de trouver le dictionnaire optimal permettant de minimiser l'erreur de représentation des vecteurs d'apprentissage. Le problème s'écrit :

$$\min_{\mathrm{D}} \|\mathrm{Y}-\mathrm{DX}\|_{\mathrm{F}}^{2}, \|d_k\|_2 = 1 \forall k \tag{6.6}$$

Il existe différentes méthodes de mise à jour du dictionnaire comme par exemple les méthodes de directions optimales(Method of Optimal Directions,MOD), K-SVD , Sparse K-SVD.

6.2.2 Dictionnaire basé sur les données/annotations

Les dictionnaires visuels se distinguent selon deux grandes catégories. La première catégorie considère les dictionnaires visuels qui suivent une approche basée sur les données seulement lors de la création des mots visuels les constituant. Ils se basent sur des méthodes d'apprentissage non supervisé en faisant appel à la quantification vectorielle dans le cas de la représentation BoVW tels que la méthode des k-moyennes or radius-based clustering et Locality-constrained Linear Coding(LLC) dans le cas de la représentation parcimonieuse. Cependant, ces méthodes n'ont aucun mécanisme pour conserver des informations discriminantes (par exemple, des catégories d'objets ou de scènes) dans les dictionnaires visuels.

La deuxième catégorie considère les dictionnaires visuels qui suivent une approche basée sur les annotations. Ils se basent sur un apprentissage supervisé qui prend en considération les étiquettes des catégories lors de la création des mots visuels constituant les dictionnaires visuels. Dans l'approche basée sur les annotations, le dictionnaire visuel est obtenu en assignant des étiquettes significatives aux patchs d'images, par exemple ciel, eau, ou végétation. Cependant, l'annotation manuelle des étiquettes de classe est une lourde tache, qui sont difficiles, voire impossibles, à obtenir dans de nombreux problèmes du monde réel.

6.2.3 Dictionnaire Global/Spécifique

Les dictionnaires visuels créés selon une quantification vectorielle se distinguent aussi selon qu'ils soient globales ou spécifiques à une catégorie. On dit qu'un dictionnaire visuel est global si l'ensemble des mots visuels qui le compose a été obtenu à partir de l'espace des caractéristiques des images de toutes les catégories. Un dictionnaire global basé sur la quantification vectorielle est généralement construit en regroupant(clustering)les descripteurs visuels qui sont choisit aléatoirement à partir de chaque classe de l'ensemble d'apprentissage. Par la suite, chaque image est représentée comme un vecteur caractéristique en calculant les histogrammes de fréquence avec les clusters appris. Cette mise en correspondance (mapping) produit la représentation de sac à mots visuels [RAMANAN et NIRANJAN, 2012].

En revanche, un dictionnaire visuel est spécifique à une catégorie si l'ensemble des mots visuels le constituant a été obtenu à partir de l'espace des caractéristiques extraites des images d'une seule catégorie/classe. Le processus de construction de ce type de dictionnaire est identique à celui d'un dictionnaire global, sauf qu'il est réalisé séparément pour chacune des catégories.

6.3 Dictionnaire visuel et synthèse des travaux

6.3.1 Selon la quantification vectorielle

Dans ce qui suit, nous présentons quelques travaux qui utilisent la quantification vectorielle dans la création de dictionnaire visuel. Suivant une approche basée sur les données, plusieurs auteurs [CSURKA et collab., 2004; LAZEBNIK et collab., 2006; NOWAK et collab., 2006; SIVIC et ZISSERMAN, 2003; ZHANG et collab., 2007] utilisent l'algorithme de clustering k-moyennes (KM) pour apprendre un ensemble de centres de cluster à partir de l'espace des caractéristiques. Chacun d'eux est considéré comme un mot visuel. Par ailleurs d'autres auteurs [FEI-FEI et PERONA, 2005; PERRONNIN et collab., 2006; QUELHAS et collab., 2007; SUDDERTH et collab., 2008; WINN et collab., 2005] utilisent le clustering à travers l'algorithme des k-moyennes dans la création de dictionnaire visuels mais en adoptant une approche basée sur les annotations. Dans leur travaux, [JURIE et TRIGGS, 2005] ont adopté une méthode de clustering basée sur le rayon plutôt qu'un clustering basé sur l'algorithme des k-moyennes pour la création du dictionnaire visuelle. De cette façon, les mots visuels sont mieux représentés, vu que toutes les caractéristiques se trouvant à l'intérieur du rayon de similarité sont associées à un seul cluster. [DORKO et SCHMID, 2005; PERRONNIN, 2008] effectuent un clustering(regroupement) des descripteurs d'images dans le but de caractériser l'apparence des classes en utilisant le modèle de mélange gaussien (GMM). [AGARWAL et collab., 2004; LEIBE et SCHIELE, 2003] utilisent la classification hiérarchique ascendante (HAC) qui a l'avantage de ne pas dépendre de l'initialisation des centres des cluster, pour la création du dictionnaire visuel. Ensuite, les deux groupes les plus similaires sont fusionnés, paire par paire, sous la contrainte que leur mesure de similarité dépasse la valeur d'un seuil prédéfini t, ce qui réduit la taille du dictionnaire appris. Dans leur travaux, [CHANG et collab., 2012] utilisent également la classification hiérarchique ascendante (HAC) pour créer des clusters sur les descripteurs SIFT. Cependant, ce sont les deux clusters les plus cohérents qui sont fusionnés. [MIKOLAJCZYK et collab., 2006; NISTER et STEWENIUS, 2006] adoptent un clustering en utilisant les k-moyennes hiérarchique (HKM) pour construire un arbre de dictionnaire visuel. Tout d'abord, les caractéristiques sont regroupées à l'aide de l'algorithme KM, puis une classification hiérarchique ascendante est effectuée pour obtenir des clusters de caractéristiques compacts au sein de chaque partition.[LIU et collab., 2007] utilisent l'approche de Maximization of Mutual Information (MMI)co-clustering pour générer un dictionnaire efficace en regroupant les clusters qui possèdent des concepts sémantiques similaires. [LARLUS et JURIE, 2006] ont proposé un modèle génératif qui intègre la construction du vocabulaire visuel avec l'apprentissage de classificateur.[MOOSMANN et collab., 2007] ont construit un dictionnaire visuel discriminant en utilisant the extremely random classification forest algorithm noté ERCF. [WANG, 2007] ont construit un dictionnaire visuel discriminant en adoptant une procédure de sélection de mots visuel au travers des dictionnaires de résolution résolution. [ZHANG et collab., 2009] ont proposé d'apprendre plusieurs codebook non redondants en exploitant un descripteur de caractéristiques,SIFT, et un algorithme de boosting. [LOPEZ-SASTRE et collab., 2013] ont proposé une nouvelle approche, Visual Word Aggregation(VWA), pour combiner des dictionnaires visuels hétérogènes basés sur le clustering via un consensus de clustering. [ALTINTAKAN et YAZICI, 2015] ont adopté les cartes auto-organisatrice (SOM) comme une méthode alternative pour générer un dictionnaire visuel au lieu de la méthode des k-moyennes.

6.3.2 Selon la représentation parcimonieuse

Dans ce qui suit, nous présentons les méthodes d'apprentissage de dictionnaire visuel couramment utilisées pour la création de dictionnaire visuels parcimonieux. Dans leur travaux, [AHARON et collab., 2006] ont proposé une approche basée sur la décomposition en valeurs singulières(K-SVD)qui permet d'apprendre un dictionnaire sur-complet adapté à un ensemble de vecteurs d'apprentissage. Cet algorithme est une généralisation de KM qui cherche la table de codes C pour représenter les données Y en résolvant :

$$\min_{C,X} \|Y - CX\|_2^F, \forall i, x_i = e_k \, pour \, un \, certain \, k \tag{6.7}$$

avec e_k un vecteur nul à l'exception d'une valeur 1 à la ligne k. L'algorithme permet ainsi de représenter chaque vecteur de Y par un seul vecteur de C avec un coefficient associé unitaire. L'algorithme K-SVD généralise le problème de l'algorithme KM 6.7 en traitant le problème décrit par l'équation 6.4, de telle sorte que chaque vecteur de Y soit représenté par une combinaison linéaire d'atomes du dictionnaire D avec des coefficients associés non unitaires. C'est une méthode qui itère entre l'étape de codage parcimonieux des vecteurs d'entrainements sur le dictionnaire courant, pouvant être réalisée par tout algorithme de poursuite tel que les algorithmes MP ou OMP, et l'étape de mise à jour du dictionnaire permettant d'améliorer la représentation des données. Cette mise 'a jour est réalisée atome par atome, en considérant les autres atomes fixes, grace à l'algorithme de Décomposition en Valeurs Singulières (Singular Value Decomposition (SVD)) [KL80]. Ainsi, à chaque itération, chaque colonne d^k du dictionnaire est mise à jour dans le but de réduire l'erreur de reconstruction.

[ZHANG et LI, 2010] ont proposé l'algorithme D-KSVD qui est une extension de l'algorithme K-SVD. Ils incorporent dans l'apprentissage du dictionnaire des informations discriminatives et des paramètres de classificateur linéaire dans la fonction objective. Dans leur travaux [JIANG et collab., 2013] ont introduit l'algorithme label consistent K-SVD(LC-KSVD) qui apprend un dictionnaire discriminant efficace pour la classification des images. LC-KSVD exploite les informations d'étiquette de classe pour apprendre le dictionnaire et incorpore le processus de construction du dictionnaire et du classificateur linéaire optimal dans une fonction objective reconstructive et discriminante mixte, puis obtient conjointement le dictionnaire appris et un classificateur efficace. [YANG et collab., 2011] ont proposé l'algorithme d'apprentissage du dictionnaire de discrimination de Fisher(FDDL) qui intègre les informations d'étiquette de classe et le message de discrimination de Fisher dans la fonction objective d'apprentissage d'un dictionnaire discriminant structuré, qui est utilisé pour la classification des modèles. [THIAGARAJAN et SPANIAS, 2011] ont présenté un algorithme itératif pour apprendre des dictionnaires permettant de calculer des codes parcimonieux locaux de descripteurs extraits de patchs d'images à partir de la pyramide spatiale linéaire de correspondance. Il effectue une itération entre une étape de codage locale sparse et une étape de mise à jour qui recherche un meilleur dictionnaire. [QUAN et collab., 2016] ont proposé une méthode discriminante de codage parcimonieux se basant sur l'apprentissage d'ensemble de classificateur. Une meilleure discriminabilité du codage parcimonieux ainsi qu'une meilleure robustesse de la classification sont obtenues en apprenant conjointement un dictionnaire pour le codage parcimonieux et un ensemble de classificateur pour la discrimination. [WANG et collab., 2018] ont proposé un algorithme d'apprentissage de dictionnaire pour la classification d'objets, unidirectional representation dictionary learning(URDL) qui exploite les directions des coefficients pour favoriser la représentation de la discrimination.

Le tableau 6.1 récapitule les différentes méthodes de construction de dictionnaire visuel utilisés dans la littérature pour la classification d'images(objets). D'après l'étude réalisée dans ce chapitre, nous avons constaté que plusieurs aspects peuvent caractériser une méthode de construction de dictionnaire visuel comme l'approche adoptée pour la construction du dictionnaire, le type de dictionnaire utilisé et l'algorithme d'apprentissage du dictionnaire employé.

En ce qui concerne l'approche adoptée lors de la construction du dictionnaire visuel, soit elle se base uniquement sur les vecteurs caractéristiques des images d'apprentissage (data-driven approach)et où aucune connaissance sur les classes d'appartenance de ces données n'est disponible. Soit au contraire, elle se base sur les étiquettes (annotations approach) des vecteurs caractéristiques des images d'apprentissage d'où la création d'un dictionnaire visuel sémantique, on parle alors de dictionnaire discriminatif.

En ce qui concerne le type de dictionnaire utilisé, un dictionnaire visuel est global dans la mesure où les mots visuels le constituant portent sur toutes les catégories des images d'apprentissage. Comme il peut être spécifique à une seule catégorie des images d'apprentissage.

La qualité d'un dictionnaire visuelle dépend aussi de l'algorithme utilisé pour la création des mots visuels. Ces algorithmes se basent soit sur une représentation parcimonieuse et des méthodes d'apprentissage de dictionnaire comme le codage parcimonieux(sparse coding), soit sur la quantification vectorielle et des méthodes de partitionnement(clustering) : KM, HCA, MS, GMM.

TABLEAU 6.1 – Quelques travaux de littérature relatifs à la création de dictionnaire, selon l'approche de dictionnaire adoptée, le type de dictionnaire et l'algorithme d'apprentissage du dictionnaire utilisé.

Work	Codebook Approach	Codebook Type	Codebook Learning Algorithm
[SIVIC et ZISSERMAN, 2003]	data-driven	category-specific codebook	KM
[LEIBE et SCHIELE, 2003]	data-driven	category-specific codebook	HAC
[CSURKA et collab., 2004]	data-driven	global codebook	KM
[AGARWAL et collab., 2004]	data-driven	category-specific codebook	HAC
[JURIE et TRIGGS, 2005]	data-driven	global codebook	MS
[DORKO et SCHMID, 2005]	data-driven	global codebook	GMM
[WINN et collab., 2005]	annotation	global codebook	KM
[FARQUHAR et collab., 2005]	annotation	category-specific codebook	GMM
[FEI-FEI et PERONA, 2005]	annotation	global codebook	KM
[NOWAK et collab., 2006]	data-driven	global codebook	KM
[LAZEBNIK et collab., 2006]	data-driven	global codebook	KM
[MIKOLAJCZYK et collab., 2006]	data-driven	global codebook	KM + HAC
[NISTER et STEWENIUS, 2006]	data-driven	global codebook	KM + HAC
[PERRONNIN et collab., 2006]	annotation	global codebook	KM
[LARLUS et JURIE, 2006]	annotation	category-specific codebook	GMM
[ZHANG et collab., 2007]	data-driven	global and category-specific codebook	KM
[QUELHAS et collab., 2007]	annotation	global codebook	KM
[MOOSMANN et collab., 2007]	annotation	category-specific codebook	RCF
[WANG, 2007]	annotation	global codebook	HAC
[PERRONNIN, 2008]	data-driven	global codebook	GMM
[SUDDERTH et collab., 2008]	annotation	global codebook	KM
[ZHANG et collab., 2009]	annotation	global codebook	KM
[LIU et collab., 2011]	data-driven	global codebook	Co-clustering
[ALTINTAKAN et YAZICI, 2015]	data-driven	category-specific codebook	SOM
[AHARON et collab., 2006]	data-driven	global dictionary	K-SVD
[WANG et collab., 2010]	data-driven	global dictionary	LCC
[ZHANG et MAYO, 2010]	annotation	global dictionary	D-KSVD
[JIANG et collab., 2013]	annotation	global dictionary	LC-KSVD
[YANG et collab., 2011]	annotation	global dictionary	FDDL
[WANG et collab., 2018]	data-driven	global dictionary	URDL

6.4 Modèles de dictionnaire et synthèse de travaux

Selon la définition de [RAMANAN et NIRANJAN, 2012], un modèle de dictionnaire fournit une distribution des mots visuels qui modélisent la globalité d'une image, ce qui rend ce modèle bien adapté à la description du contexte. Dans cette section, nous présentons quelques modèles de dictionnaire proposés dans la littérature.

6.4.1 Travaux de Lazebnik et al,2006

L'approche par Pyramide Spatiale d'Histogramme(SPM) [LAZEBNIK et collab., 2006] a été proposé afin de rajouter une information spatiale à l'approche sac à mots visuels. Cette approche extrait de l'image des informations à la fois globales et locales. Les histogrammes sont calculés sur une grille pyramidale couvrant l'image. Le niveau supérieur contient une seule case couvrant l'ensemble de la région d'intérêt. Chaque niveau supplémentaire subdivise la grille du niveau supérieur de manière à obtenir un quadrillage de plus en plus fin. A chaque case spatiale est associé l'histogramme de mots visuels correspondant. Le descripteur final est obtenu par la concaténation de l'ensemble des histogrammes, pondérés de façon adaptée.

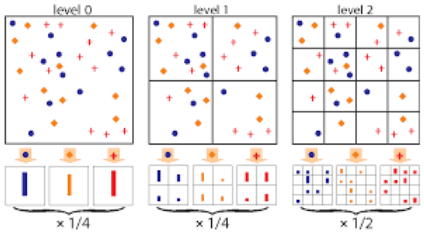


FIGURE 6.3 – Pyramide Spatiale d'Histogramme

Soit $b_i \in \mathrm{R}d$ un mot visuel ou un vecteur de base, où d est la dimension du mot visuel. Soit $\mathrm{D}_{d\times n} = (b_1, b_2, \ldots, b_n)$ la matrice qui représente un dictionnaire visuel ou un ensemble de vecteurs de base, tel que le nombre total de mots visuels est n. Soit $x_i \in \mathrm{R}^d$ la ième caractéristique locale dans une image. Soit $\mu_i \in \mathrm{R}^n$ le vecteur de coefficient de codage de x_i, tel que μ_i est le coefficient relatif au mot b_j. L'approche par Pyramide Spatiale d'Histogramme(SPM) utilise un codage dur basé sur la quantification vectorielle(VQ) qui résout le problème d'ajustement des moindres carrés contraints suivant

$$\mu_{ij} = arg\min_{\mathrm{C}} \sum_{i=1}^{\mathrm{N}} \|x_i - \mathrm{D}c_i\|^2, \|c_i\|_{\ell^0} = 1, \|c_i\|_{\ell^1} = 1, c_i \geq 0, \forall i \tag{6.8}$$

où $\mathrm{C} = [c_1, c_2, \cdots, c_{\mathrm{N}}]$ est l'ensemble des codes de X. La contrainte de cardinalité $\|c_i\|_{\ell^0} = 1$ signifie qu'il n'y aura qu'un seul élément non nul dans chaque code c_i, correspondant à l'indice de quantification de x_i. La contrainte non négative, ℓ_1 $\|c_i\|_{\ell^1} = 1, c_i \geq 0$ signifie que le poids de codage pour x est 1. En pratique, l'élément unique non nul est trouvé en recherchant le voisin le plus proche.

6.4.2 Travaux de Van Gemert et al,2008

Dans leur travaux, [VANGEMERT et collab., 2008] ont présenté une amélioration fondamentale du modèle de dictionnaire, modèle sac à mots visuels, pour la catégorisation

des images de scènes. Le modèle de dictionnaire classique utilise un codage dur pour représenter les caractéristiques de l'image avec des mots visuels. [VANGEMERT et collab., 2008] ont substitué ce type de codage par la modélisation de l'incertitude, ce qui est approprié car les vecteurs de caractéristiques ne sont capables de capturer qu'une partie de la variation intrinsèque de l'apparence visuelle. Cette incertitude est obtenue avec des techniques basées sur l'estimation de la densité du noyau. Par conséquent, [VANGEMERT et collab., 2008] ont introduit le codage souple en attribuant une caractéristique locale à tous les mots visuels. Le coefficient de codage représente l'appartenance d'une caractéristique locale à différents mots visuels. En considérant le jème coefficient de codage représentant le degré d'appartenance d'une caractéristique locale x_i au jème mot visuel, le codage souple proposé par [VANGEMERT et collab., 2008] revient à résoudre le problème suivant

$$\mu_{ij} = \frac{\exp(-\beta \| x_i - b_j \|_2^2}{\sum_{k=1}^{n} \exp(-\beta \| x_i - b_k \|_2^2} \tag{6.9}$$

où β est le facteur de lissage contrôlant la souplesse(softness) de l'affectation. Tous les n mots visuels sont utilisés dans le calcul de μ_{ij}.

6.4.3 Travaux de Yang et al,2009

Afin d'améliorer la perte en quantification du codage dur, la contrainte de cardinalité restrictive $\| c_i \|_{l^0} = 1$ dans l'équation 6.8 peut être atténuée en utilisant un terme de régularisation de la parcimonie. Dans ScSPM [YANG et collab., 2009], le codage parcimonieux représente une caractéristique locale x_i par une combinaison linéaire d'un ensemble de vecteurs de base parcimonieux. Le vecteur de coefficient μ_i est obtenu en résolvant un problème d'approximation régularisée $\ell_1 - norm$,

$$\mu_i = arg \min_{\mathrm{C}} \sum_{i=1}^{\mathrm{N}} \| x_i - \mathrm{D} c_i \|^2 + \lambda \| c_i \|^{\ell^1} \tag{6.10}$$

6.4.4 Travaux de Wang et al,2010

D'après [YU et collab., 2009], la localité est plus essentielle que la parcimonie vue que la localité aboutit systématiquement à la parcimonie mais pas obligatoirement l'inverse. Partant de ce constat, [WANG et collab., 2010] ont proposé le codage LLC, Locality-constrained Linear Coding dans lequel ils intègre la contrainte de localité de l'emplacement spatial du code où chaque descripteur est codé sur des bases sélectionnées localement. De ce fait, le codage LLC incorpore la contrainte de localité au lieu de la contrainte de parcimonie.

$$\mu_{ij} = arg \min_{\mathrm{C}} \sum_{i=1}^{\mathrm{N}} \| x_i - \mathrm{D} c_i \|^2 + \lambda \| d_i \odot c_i \|^2, 1^{\mathrm{T}} c_i = 1, \forall i \tag{6.11}$$

où $\odot$ représente la multiplication élément par élément $d_i \in \mathrm{R}^{\mathrm{M}}$ est l'adaptateur de localité, offrant une liberté différente pour chaque vecteur de base proportionnelle à sa similarité avec le descripteur d'entrée x_i.

Plus précisément,

$$d_i = \exp(\frac{dist(x_i, \mathrm{D})}{\sigma}) \tag{6.12}$$

où $dist(x_i, \mathrm{D}) = [dist(x_i, b_1), dist(x_i, b_2), \cdots, dist(x_i, b_\mathrm{M})]^\mathrm{T}$ et $dist(x_i, \mathrm{D})$ est la distance euclidienne entre x_i et b_j. σ est utilisée pour ajuster la vitesse de décroissance du poids pour l'adaptateur de localité. d_i est normalisée entre [0,1] en soustrayant $\max(dist(x_i, \mathrm{B}))$ de $dist(x_i, \mathrm{B})$. La contrainte $1^\mathrm{T} c_i = 1$ suit les exigences invariantes de décalage du code LLC. Le code LLC dans 6.11 n'est pas parcimonieux dans le sens de la norme ell_0 , mais est parcimonieux dans le sens où la solution n'a que peu de valeurs significatives.

6.4.5 Travaux de Gao et al,2013

[GAO et collab., 2013] ont proposé la représentation parcimonieuse de noyau, KSR, qui cherche la représentation parcimonieuse d'une caractéristique mappée sous la base mappée dans un espace de grande dimension.

Supposons qu'il existe un noyau $\kappa(.,.)$ induit par la fonction de mappage de caractéristiques $\phi : \mathrm{R}^d \longrightarrow \mathrm{R}^\mathrm{F}$ où $d \ll \mathrm{F}$ et $k(u_i, \mu_j) = \phi(u_i).\phi(mu_j)$ représente une similarité non linéaire entre deux vecteurs u_i et u_j, la fonction mappe la caractéristique d'entrée et la base à un espace de caractéristiques de grande dimension :

$$x \xrightarrow{\phi} \phi(x)\mathrm{U} = [\mu_1, \mu_2, \cdots, \mu_k] \xrightarrow{\phi} \upsilon = [\phi(\mu_1), \phi(\mu_1), \cdots, \phi(\mu_k)]. \tag{6.13}$$

Ensuite, les caractéristiques et la base mappées sont substituées dans la formule de codage parcimonieux, d'où la fonction objectice de KSR

$$\min_{\mathrm{U},\nu} \|\phi(x) - \upsilon\nu\|_\mathrm{F}^2 + \lambda\|\nu\|_1, \kappa(\mu_i, \mu_j) \leq 1. \tag{6.14}$$

Soit $\mathrm{K_{UU}}$ une matrice $k \times k$ avec $\{\mathrm{K_{UU}}\}_{ij} = \kappa(\mu_i, \mu_j)$, et soit $\mathrm{K_U}(x)$ un vecteur de dimension k avec $\{\mathrm{K_U}(x)\}_i = \kappa(\mu_i, x)$, l'équation 6.14 peut être écrite comme suit

$$\min_{\mathrm{U},\nu} \kappa(x,x) + \nu^\mathrm{T}\mathrm{K_{UU}}\nu - 2\nu^\mathrm{T}\mathrm{K_U}(x) + \lambda\|\nu\|_1, \kappa(\mu_i, \mu_j) \leq 1. \tag{6.15}$$

L'objectif de KSR dans l'équation 6.15 est le même que celui du codage parcimonieux sauf pour la définition de $\mathrm{K_{UU}}$ et $\mathrm{K_U}(x)$, qui peuvent être calculé à l'aide de n'importe quel noyau. Lorsque le noyau linéaire est utilisé dans KSR, $\mathrm{K_{UU}}$ et $\mathrm{K_U}(x)$ deviennent $\mathrm{U}'\mathrm{U}$ et $\mathrm{U}'x$ respectivement, et KSR est réduit à un codage parcimonieux.

6.4.6 Travaux de Wang et al (2011)

[WANG et WANG, 2011]ont abordé le problème de l'apprentissage de la représentation pyramidale spatiale optimale pour une image donnée. Les auteurs ont proposé un framework d'apprentissage à échelles multiples (MSL) pour apprendre les meilleurs poids pour chaque échelle de la pyramide. La figure 6.4 est une illustration de la représentation spatiale pyramidale. Comme l'illustre la figure 6.4, les points rouges représentent les vecteurs de coefficient codés des descripteurs d'image locale. Les pyramides grises représentent les opérations de mise en commun. Les points bleus, jaunes et verts représentent les vecteurs regroupés(pooled vectors) aux niveaux 0, 1 et 2, respectivement. Chaque point décrit les statistiques du patch se trouvant dans la région de la grille associée. Les vecteurs regroupés, mis en commun, sont généralement concaténés avec des poids fixes dans un seul vecteur, qui correspond à la représentation finale de l'image pyramidale spatiale. Le framework développé vise à déterminer la valeur optimale de ces poids en se basant sur l'apprentissage automatique.

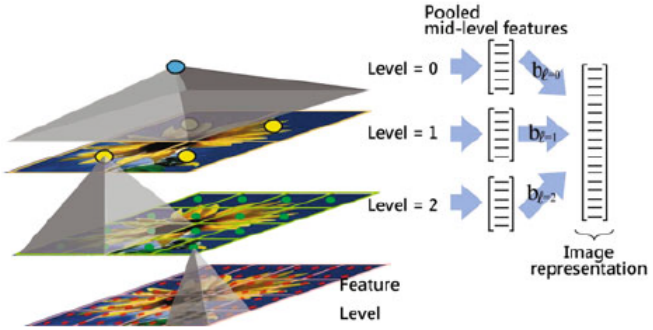


FIGURE 6.4 – Illustration du principe d'apprentissage des poids de la pyramide spatiale à différente échelles.

6.4.7 Travaux de Liu et al, 2011

[LIU et collab., 2011] ont proposé de ne considérer que les k mots visuels se situant dans le voisinage d'une caractéristique locale. Par ailleurs, les distances qui séparent cette caractéristique des autres mots visuels restants sont mises à l'infini. Cette stratégie élimine l'effet néfaste des des longues distances non fiables, même si un petite valeur de β est utilisée. Formellement, x_i représente une caractéristique locale et b_j représente le $j^{ième}$ mot visuel, le coefficient de codage j du localized soft-assignment coding s'écrit comme suit

$$\mu_{ij} = \frac{\exp(-\beta \widehat{d}(x_i, b_l))}{\sum_{l=1}^{n} \exp(-\beta \widehat{d}(x_i, b_l))} \tag{6.16}$$

où $\widehat{d}(x_i, b_l)$ est la version localisée de la distance originale $d(x_i, b_l)$. N désigne les k plus proches voisins de x_i définit par la distance $d(x_i, b_l)$, tel que $l = 1, 2, \ldots, n$.

Dans leur travaux, [LIU et collab., 2011] ont suivit le codage original d'affectation souple de [VANGEMERT et collab., 2008], [VANGEMERT et collab., 2009] pour définir $d(x_i, b_l)$ comme la distance Euclidienne carrée. Cependant d'autres distances peuvent être utilisées. La valeur β est le facteur de lissage contrôlant la douceur de l'assignement.

6.4.8 Travaux de Zhang et al,2009

Dans leur travaux, [ZHANG et collab., 2009] ont proposé d'apprendre des dictionnaires visuels multiples non redondants en exploitant le descripteur SIFT et un algorithme de boosting. L'idée de base est d'envelopper le processus de construction du dictionnaire dans une procédure de boosting. Chaque itération de boosting commence par l'apprentissage d'un dictionnaire en fonction des poids attribués par l'itération de boosting précédente. Le dictionnaire en résulte est ensuite appliqué pour coder les exemples d'apprentissage. Un nouveau classificateur est appris et de nouveaux poids sont calculés. Pour la classification des images, une réduction de 77 % des erreurs a été obtenu pour la reconnaissance de 9 classes de catégories d'objets.

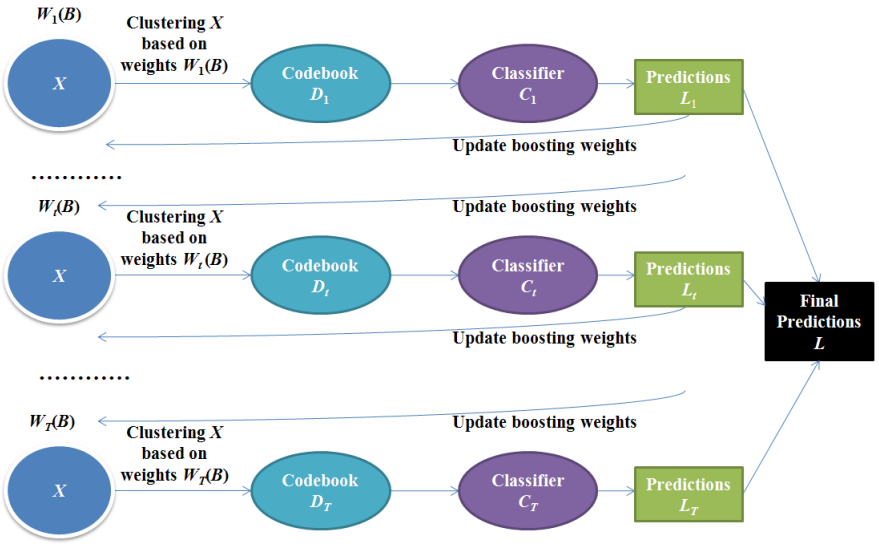


FIGURE 6.5 – Illustration du framework basé sur l'apprentissage de dictionnaires visuels multiples non redondants en utilisant un algorithme de boosting.

6.4.9 Travaux de Oliveira et al,2012

[OLIVEIRA et collab., 2012] ont développé le codage spatial parcimonieux, Sparse Spatial Coding (SSC), qui combine les avantages de l'approche SPM et implémente une représentation du codage spatail Euclidien. Dans la phase d'apprentissage, un dictionnaire est construit à partir d'un ensemble de patchs aléatoires extraits de l'ensemble d'images d'apprentissage. Ces patchs sont normalisés et transmis à un processus d'apprentissage de dictionnaire. La phase de codage consiste à l'extraction des descripteurs locaux, SIFT ou SURF, et la génération du code en se basant sur le dictionnaire et sur la quantification de chaque descripteur. Une contrainte spatiale est utilisé au lieu de la parcimonie. Les codes associés à chaque région sont mis en commun pour former une signature globale de l'image. Cette signature de l'image est soumise par la suite à une classification en ligne.

Le terme de régularisation imposé dans le codage parcimonieux peut conduire à la sélection de bases différentes pour des patchs similaires. Les auteurs proposent de surmonter ce problème en imposant une contrainte spatiale. Soit les caractéristiques d'entrée X_i et soit D un dictionnaire visuel. Comme l'illustre la figure 6.6, le codage Euclidien spatial permet de sélectionner la base la plus proche dans le dictionnaire.

6.4.10 Travaux de Zhang et al,2013

Dans leur travaux, les auteurs [ZHANG et collab., 2013] ont présenté une nouvelle méthode de classification des images en utilisant le codage spatial pyramidal robuste(SP-RSC) comme l'illustre la figure 6.7. Une image est divisée en sous-régions à différentes échelles. Un codage parcimonieux robuste est adopté pour générer le dictionnaire visuel et coder les caractéristiques locales de l'image avec une contrainte spatiale.

Différent de l'approche SPM de [LAZEBNIK et collab., 2006], le dictionnaire basé sur le codage SP-RSC est concaténé avec chaque résultat de codage des sous-régions qui

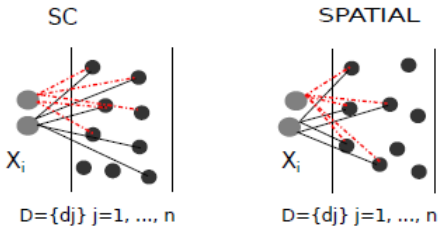


FIGURE 6.6 – Illustration du principe de codage spatial parcimonieux.

ont la même localité spatiale et la même échelle de segmentation. Pour le codage parcimonieux robuste, l'approche de l'estimation du maximum de vraisemblance (EMV) est adoptée avec une minimisation de certaines fonctions des résidus de codage. Cette fonction est associée à la distribution des résidus de codage qui codent de manière robuste la caractéristique locale donnée avec des coefficients de régression parcimonieux.

Par ailleurs, les auteurs ont étendu le codage de la pyramide spatiale en utilisant un codage parcimonieux robuste au lieu d'un codage parcimonieux, à la fois pour la construction du dictionnaire visuel ainsi que le codage des caractéristiques locales.

Le codage parcimonieux suppose que l'erreur de reconstruction suit une distribution gaussienne ou Laplacienne, alors que le codage parcimonieux robuste n'impose pas de telles contraintes, ce qui permet de coder les caractéristiques locales plus efficacement.

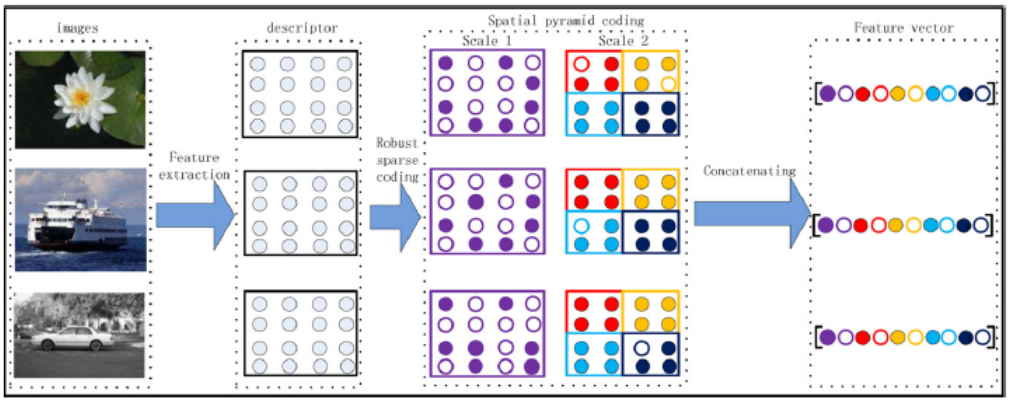


FIGURE 6.7 – Illustration du codage robuste parcimonieux de la pyramide spatiale.

6.4.11 Travaux de JinPark et al,2015

[JIN PARK et KIM, 2015] ont proposé un modèle de dictionnaire hybride qui se base sur deux dictionnaires visuels différents générés à partir de différents types de descripteurs, caractérisant les premier et arrière plans de l'image. Le premier dictionnaire est généré à partir du descripteur SIFT et le deuxième dictionnaire est généré à partir du descripteur LBP modifié. Les objets de premier plan ou les régions contenant une sil-

houette distincte sont encodés sur la base du descripteur de caractéristique SIFT et les autres régions telles que le ciel, la pelouse ou le sol sont encodées sur la base du descripteur de caractéristique LBP. Pour établir un critère de sélection des descripteurs, les auteurs ont introduit une pyramide de saillance basée sur la Transformée de Phase de Fourier(PFT).

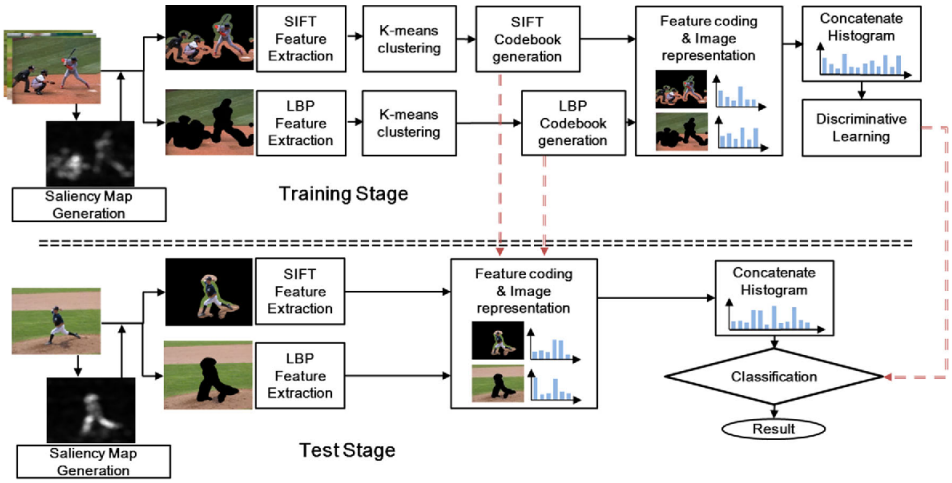


FIGURE 6.8 – Illustration du modèle de dictionnaire hybride proposé par [JIN PARK et KIM, 2015].

6.4.12 Travaux de Goh et al,2014

Dans leur travaux, [GOH et collab., 2014] ont proposé une architecture hiérarchique hybride basée sur les machines Boltzmann(RBM) restreintes pour encoder les descripteurs SIFT et fournir une représentation vectorielle pour la catégorisation des images. L'architecture hybride fusionne les forces complémentaires de l'approche sac à mots visuel et des architectures profondes. En particulier, les auteurs exploitent la puissance de modélisation des descripteurs locaux et la l'agrégation spatiale de l'approche sac à mots visuel, ainsi que la capacité d'adaptation et de représentation de l'apprentissage profond. Contrairement à d'autres méthodes d'apprentissage du dictionnaire, les dictionnaires sont régularisés pour être conjointement parcimonieux et sélectifs basé sur des distributions de loi de puissance. L'architecture de codage visuel hiérarchique, coloré en gris dans la figure 6.9 est responsable de la transformation des caractéristiques locales en codes de caractéristiques.

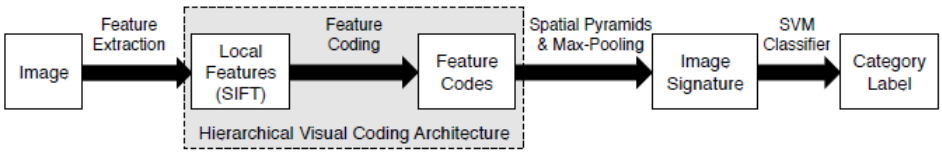


FIGURE 6.9 – Illustration du codage hiérarchique profond [GOH et collab., 2014].

6.4.13 Travaux de Lopez et al, 2013

Dans leur travaux, [LOPEZ-SASTRE et collab., 2013] ont proposé de construire un dictionnaire visuel W^* par un regroupement de consensus, agrégation de mots visuelles(VWA), en combinant m dictionnaires visuels hétérogènes $\{W_1, W_2, \ldots, W_m\}$. Étant donné un ensemble de m dictionnaires $\{W_1, W_2, \ldots, W_m\}$, l'approche de regroupement par consensus CC trouve le dictionnaire visuel W^* qui minimise le nombre total de désaccords avec les m regroupements,

$$W^* = CC(W_1, \ldots, W_m) \tag{6.17}$$

La sortie de l'algorithme de regroupement de consensus, W^{ast}, spécifie les centres qui définissent le regroupement des données. D'abord, les images sont représentées par des caractéristiques locales, comme le descripteur SIFT. Ensuite, les processus de quantification vectorielle commencent, et m algorithmes de clustering sont exécutés. L'approche VWA peut rapprocher des informations de clustering sur le même ensemble de données provenant de différents algorithmes de clustering et/ou de différentes exécutions du même algorithme. Une fois l'agrégation de clustering terminée, chaque image I_i peut être représentée en utilisant une approche de sac à mots visuels avec le nouveau dictionnaire W^*, et par conséquent calculer l'histogramme des nouveaux mots visuels $H(I_i, W^*)$.

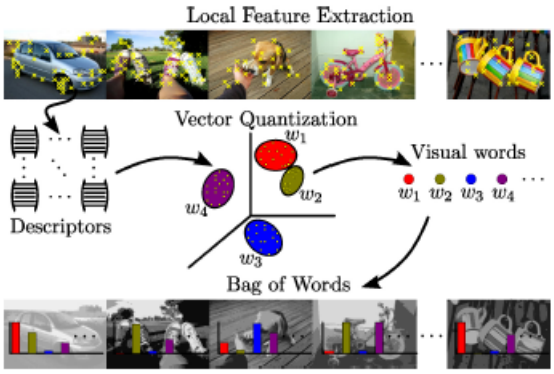


FIGURE 6.10 – Illustration de l'approche agrégation de mots visuelles(VWA)[LOPEZ-SASTRE et collab., 2013].

6.4.14 Travaux de Quan et al,2016

Dans leur travaux, les auteurs [QUAN et collab., 2016] ont développé une méthode de codage parcimonieux discriminative basée sur un ensemble de classificateur. Au lieu d'apprendre un seul classificateur linéaire, ils entrainent plusieurs classificateurs linéaires basés sur différents sous-espaces de codes parcimonieux provenant de différents sous-ensembles de signaux d'entrée pendant l'apprentissage du dictionnaire. En apprenant conjointement un dictionnaire pour le codage parcimonieux et en apprenant un ensemble de classificateur pour une tache de classification, les avantages de la méthode proposée sont doubles : une meilleure discrimination du codage parcimonieux et une meilleure robustesse de la classification.

Comme l'illustre le tableau 6.2, nous synthétisons les modèles de dictionnaire adoptés dans la littérature ainsi que l'algorithme de construction du dictionnaire visuel correspondant.

En ce qui concernent les modèles de dictionnaire visuels étudiés, les aspects qui peuvent influencer leur qualité sont la méthode de construction du dictionnaire(qui se base soit sur la quantification vectorielle ou la représentation parcimonieuse et l'apprentissage de dictionnaire), les méthodes de codage(Hard [LAZEBNIK et collab., 2006], Soft [**?**], Sparse [YANG et collab., 2009],[GAO et collab., 2013], LLC [WANG et collab., 2010], Localized Soft [LIU et collab., 2011], Sparse Spatial [OLIVEIRA et collab., 2012], SPR-SC [ZHANG et collab., 2013] Deep Hierarchical [GOH et collab., 2014]), les méthodes de mise en commun maximale(maximum pooling) ou moyenne(average pooling), l'intégration de l'information spatiale à travers la représentation de la pyramide spatiale(SPM) et d'autres représentations de pyramide spatiale avancée. [LAZEBNIK et collab., 2006], MSL [**?**], KSRSPM [GAO et collab., 2013]), l'utilisation de plusieurs dictionnaires visuels hétérogènes à résolution multiple [ZHANG et collab., 2009], la combinaison de plusieurs modèles de dictionnaires visuels issus de différents types de caractéristiques d'apprentissage(SIFT, LBP) [JIN PARK et KIM, 2015], la combinaison de plusieurs dictionnaires visuels selon un consensus de clustering[LOPEZ-SASTRE et collab., 2013].

TABLEAU 6.2 – Quelques travaux relatifs à des modèles de dictionnaire visuel proposés dans la littérature ainsi que à l'approche d'apprentissage adoptée(VQ :Vector Quantization,SC :Sparse Coding))

Work	Codebook Learning	Codebook Model
[LAZEBNIK et collab., 2006]	VQ(KM)	Hard coding + SPM(Baseline)
[VANGEMERT et collab., 2008]	VQ(MS)	Soft Coding
[YANG et collab., 2009]	VQ(KM)	Hard coding + SPM
[YANG et collab., 2009]	SC	Sparse Coding + SPM + Max-pooling(SCSPM)
[WANG et collab., 2010]	SC	Locality constrained Linear Coding + SPM (LLC)
[LIU et collab., 2011]	SC	Localized Soft Assignment coding + SPM (LSC)
[WANG et WANG, 2011]	SC	Spatial Pyramid + Robust Sparse Coding (SP-RSC)
[OLIVEIRA et collab., 2012]	SC	Sparse spatial coding + SPM (SSC)
[GAO et collab., 2013]	SC	Kernel Sparse Representation + SPM(KSRSPM)
[LOPEZ-SASTRE et collab., 2013]	VQ(KM)	Consensus Clustering
[JIN PARK et KIM, 2015]	VQ(KM)	Hybrid Bag of Feature Model + SPM
[QUAN et collab., 2016]	SC	Sparse coding + Ensemble Learning

Références

AGARWAL, S., A. AWAN et D. ROTH. 2004, «Learning to detect objects in images via a sparse, part-based representation», dans *Transactions on Pattern Analysis and Machine Intelligence (PAMI'04), IEEE*, p. 1475—-1490. 144, 147

AHARON, M., M. ELAD et A. BRUCKSTEIN. 2006, «K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation», *IEEE Transactions on Signal Processing*, vol. 54, nº 11, p. 4311–4322. 145, 147

ALTINTAKAN, U. L. et A. YAZICI. 2015, «Towards effective image classification using class-specific codebooks and distinctive local features», *IEEE Transactions on Multimedia*, vol. 17, nº 3, p. 323–332. 144, 147

AVILA, S., N. THOME, M. CORD, E. VALLE et A. DE ALBUQUERQUE ARAÚJO. 2011, «BOSSA: Extended bow formalism for image classification», dans *International Conference on Image Processing (ICIP)*, p. 2909–2912. 140

BAEZA-YATES, R. A. et B. RIBEIRO-NETO. 1999, *Modern Information Retrieval*, ACM Press/Addison-Wesley. 140

BARTHÉLEMY, Q. 2013, *Représentations parcimonieuses pour les signaux multivariés. Thèse de Doctorat.*, thèse de doctorat, Université de Grenole. 141

BOUREAU, Y., F. BACH, Y. LECUN et J. PONCE. 2010, «Learning mid-level features for recognition», dans *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, p. 2559–2566. 140

CHANG, L., M. DUARTE, L. ENRIQUE SUCAR et E. MORALES. 2012, «A Bayesian approach for object classification based on clusters of SIFT local features», dans *Journal Expert Systems with Applications*, p. 1679—1686. 144

CSURKA, G., C. DANCE, L. FAN, J. WILLAMOWSKI et C. BRAY. 2004, «Visual categorization with bags of keypoints», dans *Workshop on statistical learning in computer vision (ECCV'04)*, p. 1–22. 140, 144, 147

DORKO, G. et C. SCHMID. 2005, «Object class recognition using discriminative local features», cahier de recherche, IEEE Transactions on Pattern Analysis and Machine Intelligence. 144, 147

FARQUHAR, J., S. SZEDMAK, H. MENG et J. SHAWE-TAYLOR. 2005, «Improving bag-of-keypoints image categorisation : Generative models and pdf-kernels», cahier de recherche, LAVA report, University of Southampton, U.K. 147

FEI-FEI, L. et P. PERONA. 2005, «A bayesian hierarchical model for learning natural scene categories», dans *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05)*, p. 524–531. 144, 147

GAO, S., I. W. TSANG et L. CHIA. 2013, «Sparse Representation With Kernels», *IEEE Transactions on Image Processing*, vol. 22, nº 2, p. 423–434. 150, 156, 157

GOH, H., N. THOME, M. CORD et J.-H. LIM. 2014, «Learning Deep Hierarchical Visual Feature Coding», *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, p. 2212–2225. xi, 140, 154, 156

JIANG, Z., Z. LIN et L. DAVIS. 2013, «Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, nº 11, p. 2651–2664. 145, 147

JIN PARK, D. et C. KIM. 2015, «A Hybrid Bags-of-Feature model for Sports Scene Classification», *Journal of Signal Processing Systems for signal, image, and video technology(J Sign Process Syst)*, p. 249–263. xi, 153, 154, 156, 157

JOACHIMS, T. 1998, «Text categorization with support vector machines: Learning with many relevant features», dans *Proceedings of the European conference on machine learning(ECML'98),Springer, Berlin, Heidelberg*, p. 137–142. 140

JURIE, F. et B. TRIGGS. 2005, «Creating efficient codebooks for visual recognition», dans *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05),Beijing, China*, p. 604–610. 144, 147

LARLUS, D. et F. JURIE. 2006, «Latent Mixture Vocabularies for Object Categorization», dans *Proceedings of the British Machine Vision Conference*, BMVA Press, p. 98.1–98.10. 144, 147

LAZEBNIK, S., C. CORDELIA SCHMID et J. PONCE. 2006, «Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories», dans *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2006),17-22 June 2006, New York, NY, USA*, p. 2169–2178. 144, 147, 148, 152, 156, 157

LEIBE, B. et B. SCHIELE. 2003, «Interleaved object categorization and segmentation», dans *Proceedings of British Machine Vision Conference (BMVC'03), BMVA Press, Norwich, U.K*, p. 759–768. 144, 147

LEUNG, T. et J. MALIK. 2001, «Representing and Recognizing the Visual Appearance of Materials using Three-dimensional Textons», *International Journal of Computer Vision*, vol. 43, nº 1, p. 29–44. 140

LIU, L., L. WANG et X. LIU. 2011, «In defense of soft-assignment coding», dans *2011 International Conference on Computer Vision*, p. 2486–2493. 147, 151, 156, 157

LIU, T., J. SUN, N. ZHENG, X. TANG et H. SHUM. 2007, «Learning to detect a salient object», dans *2007 IEEE Conference on Computer Vision and Pattern Recognition*, p. 1–8. 144

LOPEZ-SASTRE, R. J., J. RENES-OLALLA, P. GIL-JIMENEZ, S. MALDONADO-BASCON et S. LAFUENTE-ARROYO. 2013, «Heterogeneous Visual Codebook Integration via Consensus Clustering for Visual Categorization», *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 23, p. 1358–1368. xi, 144, 155, 156, 157

LOWE, D. 2004, «Distinctive image features from scale-invariant keypoints», dans *International Journal of Computer Vision*, p. 91—110. 140

MAZAHERI, J. A. 2015, *Représentations parcimonieuses et apprentissage de dictionnaires pour la compression et la classification d'images satellites. Thèse de Doctorat. Traitement du signal et de l'image*, thèse de doctorat, Université de Rennes 1. 142

MIKOLAJCZYK, K., B. LEIBE et B. SCHIELE. 2006, «Multiple object class detection with a generative model», dans *IEEE conference on computer vision and pattern recognition*, p. 26—36. 144, 147

MOOSMANN, F., B. TRIGGS et F. JURIE. 2007, «Fast discriminative visual codebooks using randomized clustering forests», *In Neural information processing systems (NIPS'07)*, p. 985—-992. 144, 147

NISTER, D. et H. STEWENIUS. 2006, «Scalable Recognition with a Vocabulary Tree», dans *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR'06), New York, USA*, p. 2161–2168. 144, 147

NOWAK, E., F. JURIE et B. TRIGGS. 2006, «Sampling Strategies for Bag-of-Features Image Classification», dans *Proceedings of Ninth European Conference of Computer Vision ECCV 2006, Springer*, p. 490–503. 144, 147

O'HARA, S. et B. DRAPER. 2011, «Introduction to the Bag of Features Paradigm for Image Classification and Retrieval», *CoRR*. 140

OLIVEIRA, G. L., E. NASCIMENTO, A. VIEIRA et M. CAMPOS. 2012, «Sparse Spatial Coding: A novel approach for efficient and accurate object recognition», dans *2012 IEEE International Conference on Robotics and Automation*, p. 2592–2598. 152, 156, 157

PERRONNIN, F. 2008, «Universal and adapted vocabularies for generic visual categorization», dans *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI'08)*, p. 1243–1256. 144, 147

PERRONNIN, F., C. DANCE, G. CSURKA et M. BRESSAN. 2006, «Adapted vocabularies for generic visual categorization», dans *European Conference of Computer Vision ECCV 2006, Springer*, p. 464–475. 144, 147

QUAN, Y., Y. XU, Y. SUN, Y. HUANG et H. JI. 2016, «Sparse Coding for Classification via Discrimination Ensemble», dans *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 5839–5847. 145, 155, 157

QUELHAS, P., F. MONAY, J.-M. ODOBEZ, D. GATICA-PEREZ et T. TUYTELAARS. 2007, «A thousand words in a scene», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, nº 9, p. 1575–1589. 144, 147

RAMANAN, A. et M. NIRANJAN. 2012, «A Review of Codebook Models in Patch-Based Visual Object Recognition», *Journal of Signal Processing System,Elsevier North-Holland*, p. 333–352. 143, 148

SIVIC, J. et A. ZISSERMAN. 2003, «Video Google: a text retrieval approach to object matching in videos», dans *Proceedings of the IEEE International Conference on Computer Vision (ICCV'03)*, p. 1470–1477. 140, 144, 147

SUDDERTH, E. B., A. TORRALBA, W. T. FREEMAN et A. S. WILLSKY. 2008, «Describing visual scenes using transformed objects and parts», *International Journal of Computer Vision*, vol. 77, nº 1–3, p. 291–330. 144, 147

THIAGARAJAN, J. et A. SPANIAS. 2011, «Learning dictionaries for local sparse coding in image classification», dans *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, p. 2014–2018. 145

VANGEMERT, J., J. GEUSEBROEK, C. VEENMAN et A. SMEULDERS. 2008, «Kernel codebooks for scene categorization», dans *European Conference on Computer Vision (ECCV'2008), PART III. LNCS, Springer*, p. 696–709. 148, 149, 151, 157

VANGEMERT, J., C. VEENMAN, A. SMEULDERS et J. GEUSEBROEK. 2009, «Visual word ambiguity», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1271–1283. 151

WANG, J., J. YANG, K. YU, F. LU, T. HUANG et Y. GONG. 2010, «Locality-constrained Linear Coding for image classification», dans *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, p. 3360–3367. 147, 149, 156, 157

WANG, L. 2007, «Toward A Discriminative Codebook: Codeword Selection across Multiresolution», dans *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR'07)*, p. 1—-8. 144, 147

WANG, L., L. LIU, L. ZHOU et K. CHAN. 2014, «Application of SVMs to the Bag-of-Features Model: A Kernel Perspective», dans *Support Vector Machines Applications*, édité par Y. Ma et G. Guo, Springer, Germany, p. 155–189. 140

WANG, S. et Y. WANG. 2011, «A Multi-Scale Learning Framework for Visual Categorization», dans *ACCV 2010, Part I, LNCS, Springer*, p. 310—-322. 150, 157

WANG, X., Y. LID, S. YOU, H. LI et S. WANG. 2018, «Unidirectional Representation Based Efficient Dictionary Learning», *IEEE Transactions on Circuits and Systems for Video Technology*, p. 1–16. 145, 147

WINN, J., A. CRIMINISI et T. MINKA. 2005, «Object categorization by learned universal visual dictionary», dans *Proceedings of the 10th IEEE International Conference on Computer Vision(ICCV'05)*, p. 1800–1807. 144, 147

YANG, J., K. YU, Y. GONG et T. HUANG. 2009, «Linear spatial pyramid matching using sparse coding for image classification», dans *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR'09), Miami, FL, USA*, p. 1794–1801. 149, 156, 157

YANG, M., L. ZHANG, X. FENG et D. ZHANG. 2011, «Fisher Discrimination Dictionary Learning for Sparse Representation», dans *Proceedings of the 2011 International Conference on Computer Vision*, p. 543–550. 145, 147

YU, K., T. ZHANG et Y. GONG. 2009, «Nonlinear learning using local coordinate coding», dans *NIPS*. 149

ZHANG, E. et M. MAYO. 2010, «Improving Bag-of-Words model with spatial information», *International Conference Image and Vision Computing*, p. 1–8. 147

ZHANG, J., M. MARSZALEK, S. LAZEBNIK et C. SCHMID. 2007, «Local features and kernels for classification of texture and object categories: A comprehensive study», dans *International Journal of Computer Vision, IEEE*, p. 213–238. 144, 147

ZHANG, L., Z. GU et H. LI. 2013, «Sdsp : A novel saliency detection method by combining simple priors», dans *Proceedings of the 20th IEEE International Conference on Image Processing(ICIP'13)*, IEEE, p. 171–175. 152, 156

ZHANG, Q. et B. LI. 2010, «Discriminative K-SVD for dictionary learning in face recognition», dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 2691–2698. 145

ZHANG, W., A. SURVE, X. FERN et T. DIETTERICH. 2009, «Learning non-redundant codebooks for classifying complex objects», dans *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09), ACM, Montreal, Quebec, Canada*, p. 1241–1248. 144, 147, 151, 156

Chapitre 7

Une nouvelle approche de catégorisation d'objets

« *Je pense 99 fois et ne découvre rien. Je cesse de penser, je me plonge dans le silence et la réalité apparait.* »

Albert Einstein

Sommaire

7.1 Introduction

La catégorisation d'objets à partir d'images est l'un des problèmes les plus actifs et les plus difficiles de la vision par ordinateur. Il s'agit de prédire la classe/catégorie d'objet sémantique à partir de pixels d'image qui contiennent des variations élevées causées par des changements d'éclairage, des occlusions partielles, un encombrement d'arrière-plan, des déformations géométriques, une mise à l'échelle, des changements de perspective et différents points de vue. Dans le chapitre précédent, nous avons constaté que de nombreux travaux de recherche ont été consacré à l'amélioration le modèle de dictionnaire original et ceci à travers des méthodes avancées de génération de dictionnaire visuel ainsi que de nouvelles méthodes de codage des descripteurs d'images et de leur mise en commun(agrégation).

Une approche classique adoptée pour construire un dictionnaire visuel, et déterminer les mots visuels le constituant et ceci indépendamment du fait qu'il soit global, spécifique à une catégorie, ou sémantique est généralement obtenue par une quantification vectorielle qui regroupent les vecteurs caractéristiques d'apprentissage d'images. Cette approche fait appel à des méthodes de classification non supervisée telles que les méthodes de k-moyennes, hiérarchique k-moyennes, décalage moyen qui tentent de regrouper les caractéristiques d'apprentissage de bas niveau et représenter le centre de chaque cluster par un seul et unique mot visuel.

Cependant, une image peut être représentée de différentes manières [ZHANG et collab., 2009] et souvent un seul et unique dictionnaire visuel, qu'il soit basé sur les données ou les annotations, ne suffit pas pour décrire complètement le contenu de l'image. Pour cette raison, plusieurs auteurs se sont concentrés sur la construction de plusieurs dictionnaires visuels soit à partir d'un ensemble de vecteurs d'apprentissage homogène [ZHANG et collab., 2009] caractérisé par des descripteurs locaux de l'image de même type, comme le descripteur SIFT uniquement, soit à partir d'un ensemble de vecteur d'apprentissage hétérogène [JIN PARK et KIM, 2015] caractérisé par des descripteurs locaux d'images de différents types(desripteurs SIFT et LBP).

Dans ce chapitre, nous proposons une nouvelle approche de catégorisation d'objet. Nous présentons nos principales sources d'inspiration. Ensuite, nous présentons le framework de catégorisation d'objet proposé. Nous décrivons la méthode de génération de dictionnaire visuel proposée. Nous décrivons ensuite le modèle hybride de dictionnaire proposé. Enfin, nous présentons l'ensemble de données utilisé et les résultats expérimentaux.

7.2 Clustering/Biclustering

Le clustering est une technique d'exploration de données les plus populaires pour la découverte de connaissances dans des ensembles de données. Il consiste à regrouper les données en clusters, de telle sorte que les données à l'intérieur d'un cluster soient plus similaires les uns aux autres, qu'ils ne le sont avec les données à l'extérieur du cluster. Les techniques de clustering attribuent une donnée à un cluster sur la base de similarité globales, c'est à dire des mesures de similarité calculées pour tous les attributs. Ces mesures de similarité sont généralement basées sur le calcul de distance, comme les distances Euclidienne, Manhattan.

Contrairement au clustering qui vise à regrouper dans une matrice de données, les variables qui appartiennent à un certain modèle global dans les données, le biclustering, connu aussi sous le nom de simultaneous clustering, est une méthode d'analyse

de données conçue pour détecter les modèles locaux dans les données. Le biclustering opère simultanément sur l'ensemble des objets et des attributs d'une matrice de données, à la recherche de sous matrices constituées de sous ensembles d'objets qui ont un modèle très cohérent sur un sous ensemble d'attributs. Plus récemment, ce terme a été utilisé dans l'analyse d'expression génétique des données.

Le classification simultanée a suscité beaucoup d'attention en tant qu'une technique importante et puissante dans l'analyse bidirectionnelle des données, permettant de contourner certaines limites de l'approche de classification non supervisée [CHARRAD et BENAHMED, 2011]. Elle a été d'abord exploité pour l'analyse des données biologiques des matrices d'expression génétiques [CHENG et CHURCH, 2000; DHILLON, 2001; HARTIGAN, 1972]. Un aperçu des algorithmes de classification simultanée est présenté par [MADEIRA et OLIVEIRA, 2004]. L'approche de classification simultanée a été utilisée dans de nombreux autres domaines tels que l'exploration du Web [CHARRAD et collab., 2009] et l'exploration de texte [BICHOT, 2010]. [CHARRAD et BENAHMED, 2011] ont catégorisé les méthodes de classification simultanée selon cinq catégories principales : les méthodes de graphe bipartite, les méthodes de minimisation de la variance, les méthodes de regroupement bidirectionnelle, les méthodes de reconnaissance des motifs et les méthodes probabilistes et génératives.

Les méthodes de classification bidirectionnel utilisent un regroupement unidirectionnel afin de produire des partitions sur les deux dimensions de la matrice de données séparément. Elles font généralement appel aux méthodes classiques de partitionnement de données : k-moyennes, cartes auto-organisatrices, classification hiérarchique. Les résultats obtenus sur chaque dimension sont combinés pour produire des sous partitions de lignes et de colonnes appelées bi-partition. Ces méthodes identifient les partitions sur les lignes et les colonnes constituant la matrice de données mais pas directement les bi-partitions. Contrairement à l'approche de clustering adoptée dans la génération des mots visuels, nous pensons qu'une approche de clustering bidirectionnelle pourrait conduire à la création d'un dictionnaire visuel plus efficace pour la représentation des objets dans les images.

7.3 Description générale de l'approche proposée

Ce chapitre présente une nouvelle approche de catégorisation d'objets qui se base sur une quantification vectorielle bi-directionnelle. En considérant un ensemble de m échantillons d'apprentissage représenté par le descripteur SIFT, cet ensemble peut être représenté par la matrice de caractéristiques F($m \times n$). Par conséquent, un algorithme de regroupement bi-hiérarchique est appliqué sur les deux dimensions de la matrice de caractéristique F séparément, résultant en deux dictionnaires visuels : un dictionnaire basé sur les patchs et un dictionnaire basé sur les caractéristiques. Par conséquent, le modèle de dictionnaire basé sur les patchs et le modèle de dictionnaire basé sur les caractéristiques utilisent chacun un codage dur pour la représentation d'une image avec les mots de code correspondants. Chaque dictionnaire est utilisé, séparément, pour coder et représenter une image via le modèle de dictionnaire basé sur les patchs et le modèle de dictionnaire basé sur les caractéristiques, respectivement. Enfin, un modèle de dictionnaire hybride est obtenu, menant à une meilleure représentation de l'image pour une tâche de catégorisation d'objets.

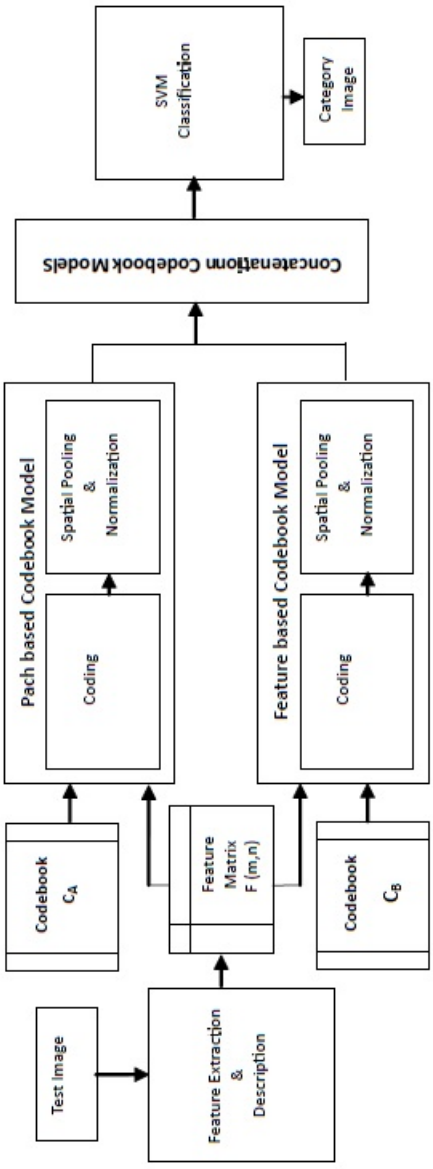


FIGURE 7.1 – Un framework de catégorisation d'objet basé sur un modèle de dictionnaire hybride

7.4 Détection et description de caractéristique

Étant donné une image, un ensemble de patchs locaux de l'image peut être détecté sur la base d'une représentation d'échantillonnage dense ou parcimonieuse (détection de point d'intérêt), pour former un sous ensemble représentatif de l'image. Une méthode d'échantillonnage dense où des patchs de taille fixe sont placés sur une grille régulière est considérée comme un meilleur moyen pour extraire de nombreux patchs locaux des images que la détection de points d'intérêt [JURIE et TRIGGS, 2005; NOWAK et collab., 2006]. Cette méthode d'échantillonnage a l'avantage d'éviter de d'omettre des informations visuelles importantes qui conduisent à de meilleures performances de classification. Une fois l'extraction de patchs locaux des images, plusieurs descripteurs de caractéristiques locales avec différents degrés d'invariance géométrique et photométrique peuvent être utilisés. Ces descripteurs sont conçus pour obtenir une description fiable et robuste des patchs locaux.

Nous exploitons la transformée de caractéristique invariante à l'échelle SIFT [LOWE, 2004] comme descripteur de caractéristique. Le calcul du descripteur SIFT se résume comme suit : premièrement, l'orientation et la magnitude du gradient sont calculées à chaque point d'échantillonnage dans le patch d'image de taille 16×16. Le patch orienté résultant est divisé en un certain nombre de sous-régions (4×4 sous-régions) et un histogramme d'orientation de gradient (bord) de 8 cases est calculé pour chaque sous-région. Un vecteur descripteur (caractéristique) de dimension 128 pour le patch d'image est formé en concaténant les histogrammes d'orientation de gradient de chaque sous-région. Par conséquent, une image I peut être représentée par un ensemble X de descripteurs SIFT x_j à n emplacements identifiés avec leurs indices $j = 1, \ldots, n$, tel que $\mathrm{X} = (x_1, x_2, \ldots, x_j, \ldots, x_n)$.

7.5 Méthode de génération de codebook proposée

Dans cet section, nous décrivons le processus de génération du dictionnaire visuel proposé, illustré dans la figure 7.2.

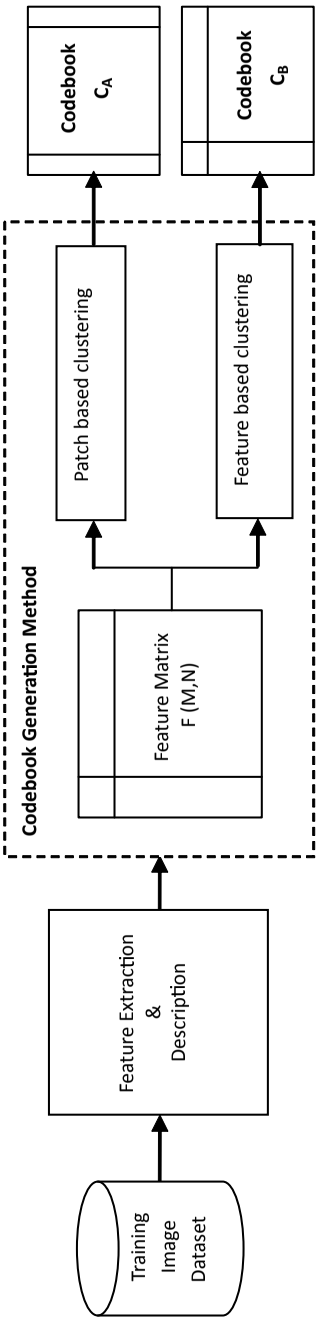


FIGURE 7.2 – Schéma de génération du dictionnaire visuel

Étant donné un ensemble d'images d'apprentissage M I = $\{I_1, I_2, \ldots, I_M\}$ appartenant à C classes, et représentées par un vecteur SIFT de dimension 128. Un ensemble de données d'images d'apprentissage peut être représenté comme une matrice de caractéristiques F(M × N) composée de M lignes et N colonnes comme illustrée dans la figure 7.3. Les étiquettes de classe correspondantes sont représentées par le vecteur d'étiquette $F = [F_1, F_2, \ldots, F_i, \ldots, F_M]^T$ tel que $F_i \in \{1, 2, \ldots, C\}$. Soit $A = \{a_1, a_2, \ldots, a_i, \ldots, a_M\}$ l'ensemble des vecteurs ligne de la matrice F . Soit $B = \{b_1, b_2, \ldots, b_{j'}, \ldots, b_N\}$ l'ensemble des vecteurs colonnes de la matrice F. Notre méthode de construction de dictionnaire est basée sur un processus de clustering bidirectionnel sur l'ensemble A et l'ensemble B séparément. Ainsi, le clustering basé sur les patchs est effectué sur l'ensemble A pour créer le dictionnaire visuel C_A, tandis que le clustering basé sur les caractéristiques est effectué sur l'ensemble B pour créer le dictionnaire visuel C_B.

Patch Set A = $\{a_{i=1,M}\}$	Feature Set B = $\{b_{j'=1,N}\}$						
	F(M,N)	b_1	b_2	...	$b_{j'}$	...	b_N
	a_1	w_{11}	w_{12}	...	$w_{1j'}$	...	w_{1N}
	a_2	w_{21}	w_{22}	...	$w_{2j'}$	...	w_{2N}
	.	.	.	...	.	...	.
	a_i	w_{i1}	w_{i2}	...	$w_{ij'}$	...	.
	.	.	.	...	.	...	.
	.	.	.	...	.	...	.
	a_M	w_{M1}	w_{M2}	...	$w_{Mj'}$	...	w_{MN}

FIGURE 7.3 – Représentation de la matrice de caractéristique F.

7.5.1 Classification basée patch

Les patchs peuvent être regroupés en clusters en fonction de leur similarité. En commençant par chaque patch $a_i \in A$ en tant que cluster distinct, une classification hiérarchique ascendante est effectué : les deux clusters de patchs les plus similaires a_i de l'ensemble A, sont fusionnés tant que la similarité moyenne entre leurs caractéristiques atteint un certain niveau ℓ. Une structure arborescente est générée et découpée en k clusters. Le centre de chaque cluster est calculé. Ensuite, une classification des k-moyennes est effectuée à en utilisant l'ensemble des centres de cluster précédemment défini comme centre de cluster initial pour obtenir des clusters de caractéristiques compactes dans chaque partition. À partir de chaque cluster résultant, nous calculons le centre du cluster et le sauvegardons dans le dictionnaire basé sur les patch, que nous notonsC_A.

7.5.2 Classification basée caractéristiques

D'autre part, l'ensemble des caractéristiques peut être partitionné en groupes homogènes. Dans un tel regroupement basé sur les caractéristiques, les caractéristiques sont considérées comme les objets et les patchs comme les caractéristiques. Initialement, l'ensemble des caractéristiques $B = \{b_1, b_2, \ldots, b_{j'}, \ldots, b_N\}$ est partitionné en utilisant une classification hiérarchique ascendante. Chaque caractéristique $b_{j'} \in B$ est affectée à un cluster distinct. Les deux clusters de caractéristiques les plus similaires $b_{j'}$ de l'ensemble B sont fusionnés tant que la similarité moyenne entre leurs patchs constitutifs atteint un certain niveau ℓ. Une arborescence est générée et découpée en k clusters. Le centre de chaque cluster est calculé. Ensuite, une classification des k-moyennes est effectuée en utilisant l'ensemble des centres de cluster précédemment définis comme

centres de cluster initiaux pour obtenir des clusters de caractéristiques compactes dans chaque partition. À partir de chaque cluster résultant, nous calculons le centre du cluster et le sauvegardons dans le dictionnaire basé sur les caractéristiques, que nous notons C_B.

7.6 Modèle de codebook proposé

Chacun des dictionnaires visuels C_A et C_B obtenu dans la section 7.5 est utilisé séparément pour représenter le contenu visuel d'une image donnée.

7.6.1 Modèle de codebook basé patch

Étant donné une image, soit $A = \{a_1, a_2, \ldots, a_i, \ldots, a_m\}$ l'ensemble des vecteurs ligne de la matrice $F(m,n)$ et soit $C_A = [C_{A1}, C_{A2}, \cdots, C_{Ai}, \ldots, C_{Ak}]$ une matrice qui désigne un dictionnaire visuel avec k mots visuels. Ainsi, chaque ligne a_i est codée par un code μ_{A_i} en utilisant un codage dur qui résout le problème d'ajustement du moindre carré contraint suivant [WANG et collab., 2012]

$$\arg\min_{\mu_A} \sum \| a_i - C_A \mu_{A_i} \|^2 \tag{7.1}$$

$\|\mu_{A_i}\|_{\ell_0} = 1, \|\mu_{A_i}\|_{\ell_1} = 1, \mu_{A_i} \geqslant 0, \forall i$
tel que le vecteur de dimension k μ_{A_i} indique le code pour a_i

La contrainte de cardinalité $\|u_{A_i}\|_{\ell_0} = 1$ et $\|mu_{A_i}\|_{\ell_0} = 1$ garantit que seul un composant des vecteurs de codage μ_{A_i} pour a_i est égal à 1 et tous les autres composants sont égaux à 0. Ainsi, la composante non nulle correspond au mot visuel le plus proche pour chaque objet patch soumis à la distance Euclidienne comme le montre l'équation 7.1. Nous introduisons l'information spatiale dans le modèle de codebook en suivant la représentation de la pyramide spatiale proposée par [LAZEBNIK et collab., 2006]. Par conséquent, une représentation pyramidale de regroupement spatial d'une image est obtenue en intégrant les réponses sur chaque code $\mu_{A_i,j}$ en une valeur à chaque niveau de pyramide ℓ en utilisant une procédure de regroupement de sommes où le i^{th} élément du vecteur caractéristique mise en commun Z_A est calculé en utilisant eq. (7.2)

$$Z_{A_i} = \sum_{j=1}^{m} \mu_{A_i,j} \tag{7.2}$$

Par conséquent, une image est décrite par l'histogramme $h_A^T = [z_{A1}^T z_{A2}^T \cdots z_{Ai}^T \cdots z_{Ak}^T]$ comme l'illustre fig. 7.4 [AVILA et collab., 2011; BOUREAU et collab., 2010].

La représentation pyramidale spatiale finale H_A d'une image est produite en concaténant tous les vecteurs caractéristiques mis en commun $H_A = h_A^{\ell_0}, h_A^{\ell_1}, h_A^{\ell_2}, \cdots h_A^{\ell_L}$ de chaque cellule de la pyramide. L'histogramme H_A est ensuite normalisé en utilisant $\ell_1 - norm$ eq. (7.3)

$$H_A = \frac{H_A}{\sum_{l=0}^{\ell} |H_A^l|} \tag{7.3}$$

7.6.2 Modèle de codebook basé caractéristiques

Étant donné la même image de test ci-dessus, considérons $B = \{b_1, b_2, \ldots, b_{j'}, \ldots, b_n\}$ comme un ensemble de vecteurs de colonne de la matrice $F(m,n)$ et que $C_B = \{C_{B1}, C_{B2}, \ldots, C_{Bi}, \cdots, C_{Bk'}\}$

		a_1	a_2	...	a_j	...	a_m
z_{A1} =	C_{A1}	$\mu_{A1,1}$	$\mu_{A1,2}$	...	$\mu_{A1,j}$	...	$\mu_{A1,m}$
z_{A2}=	C_{A2}	$\mu_{A2,1}$	$\mu_{A2,2}$	...	$\mu_{A2,j}$	...	$\mu_{A2,1}$
.	.	.	.	...	.	...	.
z_{Ai} =	C_{Ai}	$\mu_{Ai,1}$	$\mu_{Ai,2}$	...	$\mu_{Ai,j}$	...	$\mu_{Ai,m}$
.	.	.	.	...	.	...	.
.	.	.	.	...	.	...	.
z_{Ak}	C_{Ak}	$\mu_{Ak,1}$	$\mu_{Ak,2}$	...	$\mu_{Ak,j}$	...	$\mu_{Ak,m}$

FIGURE 7.4 – Représentation de la matrice d'histogramme h_A^T

désigne le dictionnaire visuel avec k' mots visuels. Chaque colonne $b_{j'}$ est encodée par un code μ_{B_i} en utilisant un codage dur qui résout le problème d'ajustement des moindres carré contraint suivant [WANG et collab., 2012]

$$\arg\min_{\mu_B} \sum \| b_{j'} - C_B \mu_{B_i} \|^2 \tag{7.4}$$

$\| \mu_{B_i} \|_{\ell_0} = 1, \| \mu_{B_i} \|_{\ell_1} = 1, \mu_{B_i} \geqslant 0, \forall i$
où le vecteur de dimension k' μ_{B_i} indique le code de $b_{j'}$. La cardinalité de contrainte $\| \mu_{B_j} \|_{\ell_0} = 1$ et $\| \mu_{B_i} \|_{ell_0} = 1$ garantit qu'une seule composante des vecteurs de codage u_{B_i} pour $b_{j'}$ est égale à 1 et toutes les autres composantes égales à 0. Ainsi, la composante non nulle correspond à la plus proche mot visuel pour chaque objet objet soumis à la distance euclidienne est donnée par eq. (7.4). Similaire, au modèle de dictionnaire basé sur les patchs, une représentation pyramidale de regroupement spatial d'une image est obtenue en intégrant les réponses sur chaque code $\mu_{B_{i,j'}}$ en une valeur à chaque niveau de pyramide ℓ en utilisant la procédure de mise en commun par la somme où le i^{th} élément du vecteur de caractéristiques mis en commun Z_B est calculé en utilisant eq. (7.5)

$$Z_{B_i} = \sum_{j'=1}^{m} \mu_{B_{i,j'}} \tag{7.5}$$

Par conséquent, la même image est décrite par l'histogramme $h_B^T = [z_{B1}^T z_{B2}^T \dots z_{Bi}^T \dots z_{Bk}^T]$ comme l'illustre la fig. 7.5

La représentation pyramidale spatiale finale H_B d'une image est produite en concaténant tous les vecteurs de caractéristiques mis en commun $H_B = h_B^{\ell_0}, h_B^{\ell_1}, h_B^{\ell_2}, \dots h_B^{\ell_L}$ de chaque cellule de la pyramide. L'histogramme H_B est ensuite normalisé en utilisant $\ell_1 - norm$ eq. (7.6)

$$H_B = \frac{H_B}{\sum_{l=0}^{\ell} |H_B^l|} \tag{7.6}$$

Enfin, un histogramme hybride $H_{hybride}$ est produit en concaténant les deux histogrammes normalisés H_B et H_B en utilisant eq. (7.7)

$$H_{Hybrid} = [H_A, H_B] \tag{7.7}$$

		b_1	b_2	...	$b_{j'}$	...	b_n
Z_{B1} =	C_{B1}	$\mu_{B1,1}$	$\mu_{B1,2}$	...	$\mu_{B1,j'}$	...	$\mu_{B1,n}$
Z_{B2}=	C_{B2}	$\mu_{B2,1}$	$\mu_{B2,2}$	...	$\mu_{B2,j'}$	...	$\mu_{B2,n}$
.	.	.	.	...	.	...	.
Z_{Bi} =	C_{Bi}	$\mu_{Bi,1}$	$\mu_{Bi,2}$	...	$\mu_{Bi,j'}$	...	$\mu_{Bi,n}$
.	.	.	.	...	.	...	.
.	.	.	.	...	.	...	.
$Z_{Bk'}$	$C_{Bk'}$	$\mu_{Bk',1}$	$\mu_{Bk',2}$	...	$\mu_{Bk',j'}$	...	$\mu_{Bk',n}$

FIGURE 7.5 – Représentation de la matrice d'histogramme h_B^T

7.7 Classification

Pour la classification, nous utilisons des machines à vecteurs de support(SVM). Dans la classification binaire, la fonction de décision qu'un SVM vise à apprendre pour une image de test a la forme suivante [ZHANG et collab., 2009]

$$g(x) = \sum_{i=1}^{n} \alpha_i y_i \kappa(\mathrm{H}_{\mathrm{H}ybrid}, \mathrm{H}'_{\mathrm{H}ybrid}) - b \tag{7.8}$$

où **K** (H,H') est la valeur d'une fonction de noyau pour la représentation d'histogramme **H** d'une image d'entraînement et la représentation d'histogramme $\mathbf{H}'$ d'une image de test. $\mathbf{y}_i \in \{+1, -1\}$ est l'étiquette de la classe de α_i. La valeur optimale du paramètre de régularisation d'apprentissage C est déterminée en effectuant une validation croisée. Pour la classification multi-classes, un-contre-reste et un-contre-un sont des techniques qui permettent d'étendre les classificateurs SVM binaires dans la pratique. Selon [ZHANG et collab., 2007], les techniques one-vs-rest et one-vs-one fournissent des résultats souvent comparables. De plus, le premier a une complexité moindre dans le nombre de catégories. Par conséquent, nous utilisons la technique one-vs-rest où pour une image de test, tous les classificateurs binaires M sont exécutés et la classe avec la réponse maximale est affectée à l'image. Le choix de la fonction du noyau $\kappa(.,.)$ a un impact important sur les performances et la vitesse de classification. Il peut s'agir de n'importe quelle fonction du noyau Mercer raisonnable. En pratique, le noyau d'intersection [BARLA et collab., 2003] s'est avéré le plus approprié pour la représentation d'histogramme. Nous comparons les performances du noyau d'intersection d'histogramme avec le noyau RBF gaussien classique. Soit $\mathrm{H}_{hybride}$ et $\mathrm{H}'_{hybride}$ la représentation d'histogramme hybride pour une image d'apprentissage et une image de test respectivement. Un SVM avec un noyau d'intersection d'histogramme et un noyau RBF gaussien sont définit en utilisant eq. (7.9) et eq. (7.10)

$$\kappa(\mathrm{H}_{\mathrm{H}ybrid}, \mathrm{H}'_{\mathrm{H}ybrid}) = \sum_{i=1}^{k} (\min(\mathrm{H}_{\mathrm{H}ybrid}{}^{i}, \mathrm{H}'_{\mathrm{H}ybrid}{}^{i})) \tag{7.9}$$

$$\kappa(\mathrm{H}_{\mathrm{H}ybrid}, \mathrm{H}'_{\mathrm{H}ybrid}) = e^{-\alpha \|\mathrm{H}_{\mathrm{H}ybrid} - \mathrm{H}_{\mathrm{H}ybrid}\|_2^2} \tag{7.10}$$

tel que α est un paramètre de normalisation.

7.8 Résultats expérimentaux

Dans ce travail, nous évaluons notre approche sur la base d'images Caltech-101 collectée par [FEI-FEI et collab., 2004] à l'aide de Google Image Search. L'ensemble de données Caltech-101 contient 9144 images répartit selon 101 différentes catégories d'objets : les animaux, les véhicules, les fleurs, ..., etc. et la classe d'arrière-plan avec une variabilité intra-classe élevée. Le nombre d'images par catégorie varie de 31 à 800. La plupart des images sont de résolution moyenne (environ 300×250 pixels). Pour comparer nos résultats à ceux rapportés dans des travaux antérieurs, nous avons suivi la même configuration expérimentale que celle suggérée par la base de données d'origine [FEI-FEI et collab., 2004] et de nombreux autres chercheurs [JURIE et TRIGGS, 2005; YANG et collab., 2009] avec 15 images d'entraînement pour chaque catégorie d'objet, y compris la classe d'arrière-plan, et nous avons utilisé jusqu'à 50 images de test par catégorie. Nous avons mesuré les performances en utilisant une précision moyenne sur 102 classes.

Le tableau 7.1 montre les performances de classification moyennes obtenues par plusieurs méthodes et la méthode proposée sur l'ensemble de données Caltech-101. Comme on peut le constater, notre approche donne la plus grande précision pour l'ensemble de données Caltech-101, à savoir 69,15 %. La méthode proposée surpasse la méthode de codage parcimonieux qui atteint une précision de 67% [WANG et collab., 2009], la SPM linéaire(LSPM) [WANG et collab., 2009] avec une précision de 53,23 % et même la SPM non linéaire [LAZEBNIK et collab., 2006] avec une précision de 56,41 %.

Expérimentalement, il a été démontré que les performances de classification augmentent avec la taille du dictionnaire visuelVANGEMERT et collab. [2010]. Une manière courante pour définir la taille du dictionnaire est de la définir empiriquement. Les méthodes de l'état de l'art utilisent des milliers de mots de visuels [MARSZALEK et collab., 2007; VANGEMERT et collab., 2009]. Dans leur travaux, [LAZEBNIK et collab., 2006] ont rapporté qu'il y avait peu de différence entre une taille de dictionnaire visuel de 200 et 400. Dans notre travail, nous avons testé notre méthode avec une taille de dictionnaire de 300. Selon les résultats rapportés par [YANG et collab., 2009], avec 15 images d'entraînement par classe et une taille de dictionnaire de 256, ScSPM et LSPM donnent une précision de 61,97% et 51,84% respectivement, ce qui est inférieur à la notre 69,15% avec une taille de dictionnaire de seulement 300. De plus, LSPM atteint une haute précision de 53,23 % avec une taille de dictionnaire de 512. Cependant, il ne dépasse pas notre méthode. ScSPM atteint une précision de $67,0 \pm 0,5$ avec une plus grande taille de dictionnaire constitué de 1024 bases par rapport à seulement 300 codes utilisés dans notre travail. Ainsi, avec un dictionnaire relativement petit, nous avons obtenu de meilleurs résultats de performance. Selon les résultats rapportés par [WANG et collab., 2018], notre méthode proposée est capable d'obtenir des résultats comparables à la méthode URDL-LE qui atteint une précision de 69,15% avec une taille de dictionnaire de 1530. Par rapport à la méthode de codage profond non supervisée proposée par [GOH et collab., 2014] qui atteint une précision de $62,5 \pm 1,4$% avec une taille de dictionnaire de 2048, notre méthode présente des performances plus élevées.

TABLEAU 7.1 – Comparaison des taux de classification sur la base d'images Caltech-101 dataset

Authors	Method	Codebook Size	15 training
[BERG et collab., 2005]		—	48.00
[AHARON et collab., 2006]	K-SVD	1530	65.2
[ZHANG et collab., 2006]	SVM-KNN	—	59.10
[LAZEBNIK et collab., 2006]	SPM	200	56.41
[GRIFFIN et collab., 2007]		—	59.00
[BOIMAN et collab., 2008]	NBNN	—	65.0 ± 1.1
[JAIN et collab., 2008]		—	61.00
[YANG et collab., 2009]	ScSPM	1024	67.0 ± 0.5
[YANG et collab., 2009]	LSPM	512	53,23
[WANG et collab., 2010]	LLC	1530	65.43
[ZHANG et LI, 2010]	D-KSVD	1530	65.10
[YANG et collab., 2011]	FDDL	1530	66.8
[OLIVEIRA et collab., 2012]	SSC	4096	69.0 ± 0.7
[JIANG et collab., 2013]	LC-KSVD	1530	67.70
[GOH et collab., 2014]	Unsupervised Deep Coding	2048	$62,5 \pm 1.4$
[QUAN et collab., 2016]	EasyDL	1530	68.4
[WANG et collab., 2018]	URDL-LE	1530	69.15
[CHEBBOUT et MEROUANI, 2020]		300	69.15

TABLEAU 7.2 – Influence du noyau des SVM sur le taux de la classification

SVM Kernel	Accuracy
Gaussien	69.15 %
Histogramme d'intersection	56.80 %

Nous avons testé notre approche avec deux noyaux non linéaires. Comme le montre la table 7.2, la classification basée sur le noyau non linéaire gaussien permet d'obtenir de bien meilleures performances que le noyau non linéaire d'histogramme d'intersection sur l'ensemble de données Caltech-101. Comme on peut le voir dans la figure 7.6, le modèle de dictionnaire basé sur les patchs donne une précision de 64 %, tandis que le modèle de dictionnaire basé sur les caractéristiques donne une précision de 66 % par rapport au modèle de dictionnaire hybride proposé qui donne une précision de 69,15 %. Par conséquent, la combinaison des modèles de dictionnaire améliore la précision de la classification comparé aux modèles de dictionnaire individuels.

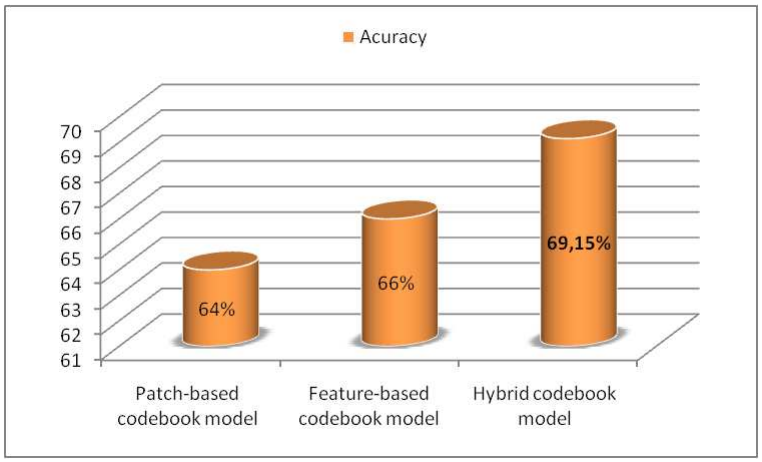


FIGURE 7.6 – Effet des modèle de dictionnaire basé sur les patchs, basé sur les caractéristiques et hybride sur le taux de classification de base d'images Caltech-101.

La principale limitation de ce travail est le manque de capacité mémoire suffisante. Alors que la majorité des travaux ont été effectués sur un poste de travail avec 16 Go de mémoire, nos expériences ont été réalisées sur un Intel Core i7 à 2,13 GHz avec 10 Go de mémoire. Ainsi, nous sommes confrontés à un problème de mémoire insuffisante lorsque nous évaluons les performances du modèle proposé sur d'autres bases d'image telle que Caltech-256, PASCAL VOC. Le même problème apparaît lorsque d'un côté, nous évaluons le modèle proposé avec d'autres méthodes de codage telles que les méthodes de codage souple ou parcimonieuse et lorsque nous augmentons le nombre de mots visuels. Cependant, avec seulement une simple méthode de codage dur et un ensemble de mots visuels limité à 300, nous avons pu obtenir un résultat très satisfaisant.

7.9 Conclusion

Dans ce chapitre , nous nous sommes intéressé au problème de la catégorisation objet/image. Dans un premier temps, nous avons proposé un framework de catégorisation d'objets basé sur un modèle de dictionnaire hybride. Par conséquent, une image est modélisée via un modèle de dictionnaire basé sur des patchs et un autre basé sur des caractéristiques séparément. Dans un deuxième temps, nous avons proposons une nouvelle méthode de génération de dictionnaire visuel qui se basée sur une approche de classification simultanée et plus précisément un algorithme de classification bidirectionnelle à travers lequel deux dictionnaire visuels différents sont construits. Malgré que notre modèle de dictionnaire hybride adopte une méthode de codage dure pour attribuer chaque descripteur d'image au mot visuel le plus proche, et que nous utilisons une taille de dictionnaire relativement petite, nos expérimentations reçoivent les résultats de classification les plus élevés sur l'ensemble de données Caltech-101 comparé à d'autres modèles de dictionnaire de l'état de l'art.

Références

AHARON, M., M. ELAD et A. BRUCKSTEIN. 2006, «K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation», *IEEE Transactions on Signal Processing*, vol. 54, nº 11, p. 4311–4322. 174

AVILA, S., N. THOME, M. CORD, E. VALLE et A. DE ALBUQUERQUE ARAÚJO. 2011, «BOSSA: Extended bow formalism for image classification», dans *International Conference on Image Processing (ICIP)*, p. 2909–2912. 170

BARLA, A., F. ODONE et A. VERRI. 2003, «Histogram intersection kernel for image classification», dans *Proceedings of IEEE International Conference on Image Processing(ICIP'03), Barcelona, Spain*, p. III–513–16. 172

BERG, A., T. BERG et J. MALIK. 2005, «Shape matching and object recognition using low distortion correspondences», dans *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR'05), IEEE*, p. 26–33. 174

BICHOT, C. E. 2010, «Co-clustering Documents and Words by Minimizing the Normalized Cut Objective Function», *Journal of Mathematical Modelling and Algorithms*, vol. 9, nº 2, p. 131–147. 165

BOIMAN, O., E. SHECHTMAN et M. IRANI. 2008, «In defense of Nearest-Neighbor based image classification», dans *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2008)*, p. 1–8. 174

BOUREAU, Y., F. BACH, Y. LECUN et J. PONCE. 2010, «Learning mid-level features for recognition», dans *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, p. 2559–2566. 170

CHARRAD, M. et M. BENAHMED. 2011, «Simultaneous Clustering: A Survey», dans *Proceedings of 4th International Conference on Pattern Recognition and Machine Intelligence(PReMI'11), Springer, Moscow, Russia*, p. 370—-375. 165

CHARRAD, M., Y. LECHEVALLIER, M. B. AHMED et G. SAPORTA. 2009, «Block Clustering for Web Pages Categorization», dans *IDEAL, Lecture Notes in Computer Science*, vol. 5788, Springer, p. 260–267. 165

CHEBBOUT, S. et H. F. MEROUANI. 2020, «A hybrid codebook model for object categorization using two-way clustering based codebook generation method», *International Journal of Computers and Applications*. 174

CHENG, Y. et G. M. CHURCH. 2000, «Biclustering of Expression Data», dans *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMBOO)*, AAAI Press, p. 93–103. 165

DHILLON, I. S. 2001, «Co-clustering documents and words using bipartite spectral graph partitioning», dans *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, p. 269–274. 165

FEI-FEI, L., R. FERGUS et P. PERONA. 2004, «Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories», dans *Proceedings of IEEE Conference on CVPR Workshop of Generative Model Based Vision (WGMBV)*, p. 59–70. 173

GOH, H., N. THOME, M. CORD et J.-H. LIM. 2014, «Learning Deep Hierarchical Visual Feature Coding», *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, p. 2212–2225. 173, 174

GRIFFIN, G., A. HOLUB et P. PERONA. 2007, «Caltech-256 object category dataset», cahier de recherche, California Institute of Technology. 174

HARTIGAN, J. A. 1972, «Direct Clustering of a Data Matrix», *Journal of the American Statistical Association*, vol. 67, nº 337, p. 123–129. 165

JAIN, P., B. KULIS et K. GRAUMAN. 2008, «Fast Image Search for Learned Metrics», dans *IEEE Conference on Computer Vision and Pattern Recognition(CVPR'08)*, p. 1–8. 174

JIANG, Z., Z. LIN et L. DAVIS. 2013, «Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, nº 11, p. 2651–2664. 174

JIN PARK, D. et C. KIM. 2015, «A Hybrid Bags-of-Feature model for Sports Scene Classification», *Journal of Signal Processing Systems for signal, image, and video technology(J Sign Process Syst)*, p. 249–263. 164

JURIE, F. et B. TRIGGS. 2005, «Creating efficient codebooks for visual recognition», dans *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05),Beijing, China*, p. 604–610. 167, 173

LAZEBNIK, S., C. CORDELIA SCHMID et J. PONCE. 2006, «Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories», dans *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2006),17-22 June 2006, New York, NY, USA*, p. 2169–2178. 170, 173, 174

LOWE, D. 2004, «Distinctive image features from scale-invariant keypoints», dans *International Journal of Computer Vision*, p. 91—110. 167

MADEIRA, S. C. et A. L. OLIVEIRA. 2004, «Biclustering Algorithms for Biological Data Analysis: A Survey», *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, p. 24–45. 165

MARSZALEK, M., C. SCHMID, H. HARZALLAH et J. VAN DE WEIJER. 2007, «Learning Object Representations for Visual Object Class Recognition», Visual Recognition Challange workshop, in conjunction with ICCV. 173

NOWAK, E., F. JURIE et B. TRIGGS. 2006, «Sampling Strategies for Bag-of-Features Image Classification», dans *Proceedings of Ninth European Conference of Computer Vision ECCV 2006, Springer*, p. 490–503. 167

OLIVEIRA, G. L., E. NASCIMENTO, A. VIEIRA et M. CAMPOS. 2012, «Sparse Spatial Coding: A novel approach for efficient and accurate object recognition», dans *2012 IEEE International Conference on Robotics and Automation*, p. 2592–2598. 174

QUAN, Y., Y. XU, Y. SUN, Y. HUANG et H. JI. 2016, «Sparse Coding for Classification via Discrimination Ensemble», dans *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 5839–5847. 174

VANGEMERT, J., C. SNOEK, C. VEENMAN, A. SMEULDERS et J. GEUSEBROEK. 2010, «Comparing Compact Codebooks for Visual Categorization», *Journal Computer Vision and Image Understanding*, p. 450–462. 173

VANGEMERT, J., C. VEENMAN, A. SMEULDERS et J. GEUSEBROEK. 2009, «Visual word ambiguity», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1271–1283. 173

WANG, C., D. BLEI et L. FEI-FEI. 2009, «Simultaneous image classification and annotation», dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*, IEEE, p. 1903–1910. 173

WANG, J., J. YANG, K. YU, F. LU, T. HUANG et Y. GONG. 2010, «Locality-constrained Linear Coding for image classification», dans *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, p. 3360–3367. 174

WANG, X., Y. LID, S. YOU, H. LI et S. WANG. 2018, «Unidirectional Representation Based Efficient Dictionary Learning», *IEEE Transactions on Circuits and Systems for Video Technology*, p. 1–16. 173, 174

WANG, X., L. WANG et Y. QIAO. 2012, «A Comparative Study of Encoding, Pooling and Normalization Methods for Action Recognition», dans *Proceeding of 11th Asian Conference on Computer Vision (ACCV'12),Springer, Daejeon, Korea*, p. 572–585. 170, 171

YANG, J., K. YU, Y. GONG et T. HUANG. 2009, «Linear spatial pyramid matching using sparse coding for image classification», dans *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR'09), Miami, FL, USA*, p. 1794–1801. 173, 174

YANG, M., L. ZHANG, X. FENG et D. ZHANG. 2011, «Fisher Discrimination Dictionary Learning for Sparse Representation», dans *Proceedings of the 2011 International Conference on Computer Vision*, p. 543–550. 174

ZHANG, H., A. C. BERG, M. MAIRE et J. MALIK. 2006, «SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition», dans *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, p. 2126–2136. 174

ZHANG, J., M. MARSZALEK, S. LAZEBNIK et C. SCHMID. 2007, «Local features and kernels for classification of texture and object categories: A comprehensive study», dans *International Journal of Computer Vision, IEEE*, p. 213–238. 172

ZHANG, Q. et B. LI. 2010, «Discriminative K-SVD for dictionary learning in face recognition», dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 2691–2698. 174

ZHANG, W., A. SURVE, X. FERN et T. DIETTERICH. 2009, «Learning non-redundant codebooks for classifying complex objects», dans *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09), ACM, Montreal, Quebec, Canada*, p. 1241–1248. 164, 172

Conclusion et Perspectives

« *Celui qui veut réussir trouve un moyen. Celui qui ne veut rien faire trouve une excuse.* »

Proverbe Français

NOUS concluons nos travaux en soulignant les principales contributions et leurs limites. Enfin, nous discutons les perspectives d'avenir, en indiquant des orientations intéressantes pour la recherche future dans ce domaine.

Rappel des contributions

Le travail de recherche mis en évidence dans ce mémoire est axé sur trois contributions principales qui se placent dans le cadre de l'analyse et l'interprétation des images.

Dans notre première contribution, nous nous sommes intéressé à la détection de la saillance dans les images. Ce problème a connu un regain d'intérêt, au cours des dernières années, de la part de la communauté des chercheurs de la vision artificielle. Ceci peut être justifié par ses applications diverses telles que la segmentation d'images, la reconnaissance d'objets, la compression des images, la recherche et l'indexation d'image basée sur le contenu.

Contrairement aux modèles de saillance étudiés dans le chapitre de l'état de l'art, nous proposons un modèle de saillance visuelle qui se base sur une approche connexioniste. Nous utilisons le réseau de neurone Self Organizng Tree dans la construction de la carte de saillance en exploitant les caractéristiques de couleur et de texture. Le réseau de neurone procède au partitionnement des vecteurs caractéristiques des pixels de l'image en différents clusters. Pour chaque cluster, nous calculons la mesure de saillance basée sur l'indice spatial, spatial cue. En supposant que la vraisemblance de la saillance d'un pixel appartenant à un cluster satisfait une distribution gaussienne. La carte de saillance finale est obtenue en calculant la probabilité marginale de la saillance. Le modèle de saillance proposé a été testé sur la base d'images MSRA-1000. Les résultats obtenus sont très encourageants.

Dans notre deuxième contribution, nous avons exploité le modèle de saillance proposé précédemment pour résoudre un problème de segmentation d'objet, connu aussi sous le nom de séparation figure/fond. Nous proposons une nouvelle méthode de segmentation d'objet qui se base sur le regroupement spectral des pixels. Cette méthode intègre la valeur de saillance des pixels, calculée à partir du modèle de saillance proposé, dans le calcul du graphe de similarité. En comparaison avec certaines méthodes typiques de segmentation d'objet, notre méthode offre de meilleurs résultats de segmentation.

Dans notre troisième contribution, nous nous sommes intéressé à l'un des problèmes de la vision par ordinateur, la reconnaissance d'objet dans les images, et plus particulièrement à la catégorisation d'objets. Nous proposons un framework de catégorisation d'objets basé sur un modèle de dictionnaire hybride. Par conséquent, une image est modélisée à travers un modèle de dictionnaire basé sur des patchs et un autre modèle de dictionnaire basé sur les caractéristiques, séparément.

Nous proposons également une nouvelle méthode de génération de dictionnaire visuel qui se base sur une approche de classification simultanée et plus précisément sur un algorithme de classification bidirectionnelle à travers lequel deux dictionnaires visuels différents sont construits. Bien que que notre modèle de dictionnaire hybride adopte une méthode de codage dur pour attribuer chaque descripteur d'image au mot visuel le plus proche, et que nous utilisons des dictionnaires visuels de taille réduite, les résultats expérimentaux obtenus sont très satisfaisants en comparaison avec des modèles de dictionnaire de l'état de l'art.

Limitations

La principale limitation de ce travail est le manque de capacité mémoire suffisante. Alors que la majorité des travaux existants dans la littérature ont été effectué sur un poste de travail avec 16 Go de mémoire, nos expériences ont été réalisées sur un Intel Core i7 à 2,13 GHz avec 10 Go de mémoire. De ce fait, nous sommes confrontés à un problème de mémoire insuffisante (Out of memory) lorsque

— nous évaluons le modèle de dictionnaire proposé sur d'autres bases d'images telle que Caltech-256 et PASCAL VOC,

— nous évaluons le modèle de dictionnaire proposé avec d'autres méthodes de codage telles que les méthodes de codage parcimonieux,

— nous augmentons le nombre de mots visuels du dictionnaire visuel,

— nous ajoutons d'autres descripteurs visuelles comme LPB, SURF,

Toutefois, avec seulement une simple méthode de codage dur et un ensemble de mots visuels limité à 300 , nous avons pu obtenir des résultats satisfaisants.

Perspectives futures

Comme perspectives pour nos futures directions de recherche, nous proposons ce qui suit :

Une amélioration significative des modèles de détection de saillance a été observée ces dernières années avec la renaissance des techniques d'apprentissage en profondeur, grâce aux puissantes méthodes d'apprentissage de la représentation. Depuis la première introduction en 2015, les méthodes de détection de saillance basées sur l'apprentissage profond, ont rapidement montré des performances supérieures aux solutions traditionnelles. Nous citons quelques travaux intéressants et motivants.

Dans les travaux de [LI et YU, 2015],les auteurs proposent un nouveau modèle de calcul pour la saillance visuelle utilisant des caractéristiques profondes à échelles multiples et calculées par des réseaux de neurones profond à convolution(CNN).

Dans les travaux de [Kruthiventi et collab., 2016], les auteurs proposent également un réseau de neurone profond à convolution (CNN) capable de prédire les fixations oculaires et de segmenter les objets saillants et ceci à partir d'un framework unifié.

Nous avons constaté que la majorité des travaux publiés dans le domaine de la modélisation de la saillance avant 2016 ont été consacré à la détection et segmentation d'un seul objet saillant. Récemment, d'autres bases d'images ont été construite, elles sont plus complexes et contiennent plusieurs objets saillants. D'autres bases ont été construites afin de pouvoir entrainer et tester des modèles de saillance se basant sur un apprentissage profond, la base MSRA10K est un exemple.

Il serait intéressant d'exploiter le paradigme de la co-classification dans le calcul des cartes de saillance. On a vu qu'il existe des travaux qui estime les mesures de saillance à partir de cluster. Pourquoi ne pas envisager d'estimer ces mesures à partir de bi-cluster ?

Avec la création de nouvelles bases d'images mixtes , dédiées à la prédiction des fixations des yeux ainsi qu'à la détection des objets saillants dans les images, il serait intéressant de penser à un modèle de saillance unifié.

Avec le développement de la technologie d'acquisition, des informations plus complètes, telles que la profondeur, la correspondance entre images ou la relation temporelle, sont disponibles pour étendre la détection de saillance d'image à la détection de saillance RGBD, la détection de co-saillance ou la détection de saillance vidéo.

La plupart des modèles informatiques d'attention visuelle sont développés dans le contexte de scènes naturelles, et leur rôle avec les images médicales n'est pas bien étudié. Les radiologues interprètent un grand nombre d'images cliniques en un temps limité, une stratégie efficace pour déployer leur attention visuelle est nécessaire. Les cartes de saillance visuelle, mettant en évidence des régions d'image qui diffèrent considérablement de leur environnement, devraient prédire où les radiologues fixent leur regard.

Toujours dans le cadre de tirer profil du paradigme de co-classifiaction, nous envisageons d'étendre la méthode de segmentation proposée en remplaçant l'algorithme de classification spectrale par l'algorithme de co-classification spectrale. Une éventuelle comparaison serait souhaitable afin d'étudier l'apport apporté.

Avec l'arrivée des réseaux profonds et l'apport de l'apprentissage profond, les chercheurs s'intéressent moins à la segmentation d'objet dans le but de de séparer un objet de son fond. Un intérêt est plutôt donné à la segmentation des catégories/instances d'objet dans les images.

Liste des Publications

Publication internationale

- Samira Chebbout, Hayet Farida Merouani, A hybrid codebook model for object categorization using two-way clustering based codebook generation method, International Journal of Computers and Applications ISSN : 1206-212X(Print) 1925-7074(Online).DOI : 10.1080/1206212X.2020.1712775

Communications internationales

- Samira Chebbout, Hayet Farida Merouani, A Novel Saliency Detection Model based on Self Organizing Tree Algorithm, IPAC 2015 :International conference on Intelligent Information Processing, Security and Advanced Communication,23–25 November 2015,Batna,Algeria.DOI : 10.1145/2816839.2816883.

- Samira Chebbout and Hayet Farida Merouani, An object segmentation method based on saliency map and spectral clustering, WCITCA 2015 : World Congress on Information Technology and Computer Applications, Hammamet,2015,pp.1-9.DOI : 10.1109/WCITCA.2015.7367036.

- Samira Chebbout,Hayet Farida Merouani,Comparative Study of Clustering Based Colour Image Segmentation Techniques.SITIS 2012 :IEEE Eighth International Conference on Signal Image Technology and Internet Based Systems,pp.839-844, Sorrento,Naples,Italy,November 25-29,2012. DOI : 10.1109/SITIS.2012.126