UNIVERSITE NATIONALE DU VIETNAM, HANOI
INSTITUT FRANCOPHONE INTERNATIONAL

ĐÀO THỦY NGÂN

CONCEPTION D'UNE HIÉRARCHIE SÉMANTIQUE ET SPATIALE DE DESCRIPTEURS LOCAUX VISUELS

THIẾT KẾ MỘT HỆ THỐNG PHÂN CẤP NGỮ NGHĨA VÀ KHÔNG GIAN CỦA CÁC CHỈ SỐ MÔ TẢ TRỰC QUAN ĐỊA PHƯƠNG

Spécialité: Réseaux et Systèmes Communicants

Code: Programme pilote

RESUME DU MEMOIRE DE FIN D'ETUDES DU MASTER INFORMATIQUE

HANOI – 2016

Travail réalisé à l'Institut Francophone International, Université Nationale du Vietnam, Hanoi

Sous la direction de: Professeur Muriel Visani

Laboratoire L3i, Université de La Rochelle

Rapporteur 1:……

Rapporteur 2:……

Le mémoire est soutenu devant le jury à l'Institut Francophone International le ……..…… 2016 à …… h…….

Le mémoire est accessible:

- au Centre d'Informations et de Bibliothèque, Université Nationale du Vietnam, Hanoi

- à l'Institut Francophone International, Université Nationale du Vietnam, Hanoi

Chapitre 1

Introduction

1.1 Contexte et motivation

Ces dernières années, le volume de données multimédia a augmenté de manière exponentielle, en parallèle avec le développement des appareils multimédia et aussi des techniques de stockage. La disponibilité d'une vaste quantité de données multimédia, notamment des images et vidéos, fournit de grandes ressources pour beaucoup de domaines d'application : journalisme, médecine, robotique... En revanche, l'explosion de données fait émerger de nouvelles questions sur les techniques de gestion automatique des images telles que : la classification des images, la recherche d'image à partir du contenu ou la reconnaissance des objets dans des images. Ce contexte conduit au développement des études sur l'analyse et sur la description du contenu des images.

L'analyse des images par le contenu est donc un sujet de recherche très étudié récemment. Appartenant au domaine de la vision artificielle, une branche de l'intelligence artificielle, il s'agit d'un domaine séduisant, pratique et dynamique avec des possibilités d'applications multiples. Dans l'ordinateur, les images sont représentées simplement par des chiffres. Cependant, au niveau des objets, les images peuvent avoir plusieurs caractéristiques spéciales. Par exemple, les documents textuels sont constitués des mots définis par une langue qui va alors en limiter leur sens, alors que pour les images, le contenu visuel peut être très varié (une plage, une montagne ou bien même de l'abstrait). La variété du contenu des images reflète la variété dans le monde réel. Dans le monde visuel, un objet peut

avoir plusieurs formes, plusieurs états et plusieurs couleurs. Par exemple, un poisson peut être grand, petit, long ou rond... La couleur d'un même objet dans les différentes images peut varier selon les conditions de capture, et notamment l'illumination. Par contre, certains objets peuvent avoir la même couleur et la même forme. Il est difficile déjà pour l'être humain de distinguer, par exemple, un chien et un loup. L'analyse des images par le contenu présente donc plusieurs défis.

1.2 Problématique

Modèle de sacs de mots visuels classique

Après une vingtaine d'années d'étude, plusieurs méthodes ont été proposées pour l'analyse des images par le contenu visuel. Parmi celles-ci, la méthode utilisant les sacs de mots visuels semble être particulièrement populaire et étudiée ces dernières années. Les étapes principales de ce modèle peuvent être résumées comme suit :

— **Détection des régions d'intérêt :** il y a deux types de régions utilisées dans cette méthode. L'une s'apple *Shape Adapted Region (SA)*, c'est une région construite en adjustant une forme elliptique selon un point d'intérêt qui est détecté par le détecteur de Harris. L'autre s'apple *Maximally Stable Region (MS)*, qui est construite à partir de la segmentation de l'image du bassin versant (*watershed image*). Selon les auteurs, ces deux types de régions capturent deux caractéristiques différentes d'une image, il faut donc utiliser deux dictionnaires séparés pour chaque type de région.

— **Extraction des descripteurs :** c'est l'étape de représentation des régions détectées par un vecteur de valeurs. Le descripteur SIFT [1] est utilisé dont chaque point d'intérêt est décrit par un vecteur à 128 dimensions. L'ensemble des points d'une image est souvent exprimé sous la forme d'une matrice à 128 colonnes où chaque ligne est un point et chaque colonne est une des 128 dimensions du descripteur.

— **Construction du dictionnaire :** pour obtenir un vocabulaire de mots visuels, les descripteurs sont regroupés en cluster. Chaque cluster correspond à un mot visuel représenté par le centre du cluster. Un algorithme de quantification vectorielle quelconque est appliqué pour faire le clustering. Dans [2], l'algorithme k-means est choisi. Grâce à une implémentation simple

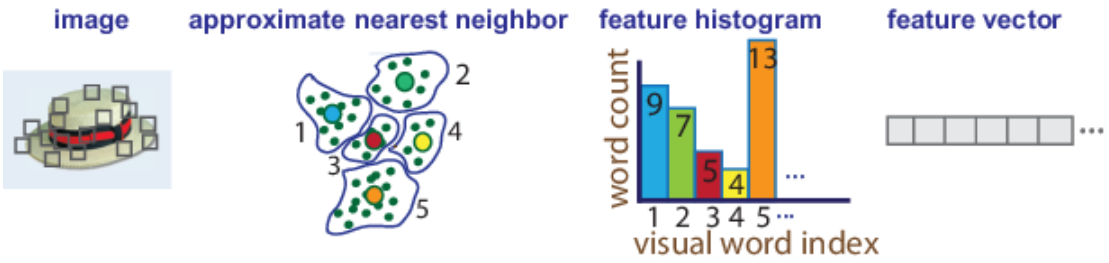

FIGURE 1.1: Illustration de l'étape d'indexation des mots visuels [1]

et une précision acceptable, cet algorithme est très utilisé en pratique et considéré comme efficace pour la construction du dictionnaire bien que ne garantissant ni l'optimalité, ni un temps de calcul polynomial).

— **Représentation des images :** pour chaque image, les descripteurs sont d'abord quantifiés en assignant chaque descripteur au mot visuel le plus proche dans le dictionnaire créé. L'image d'entrée est donc représentée par un vecteur de fréquence de mots visuels (voir la Figure 1.1). Les valeurs dans ce vecteur ne sont pas simplement le nombre d'occurrences de chaque mot, elles sont pondérées par une stratégie de pondération standardisée qui est connue sous le nom *Term Frequency - Inverse Document Frequency (TF-IDF)*.

La représentation d'une image sous la forme d'un vecteur des mots visuels, appelé aussi signature de l'image, via un dictionnaire, permet d'hériter beaucoup de techniques efficaces dans le domaine de recherche d'informations textuelles. Avec cette représentation, pour comparer deux images, on doit seulement comparer les deux histogrammes de fréquences qui les représentent. La performance de cette méthode en termes de temps de calcul dans la tâche de recherche d'objet est très remarquable car on obtient le résultat presque sans délai. Cette méthode ouvre une perspective de développement des moteurs de recherche d'images par le contenu en temps réel.

Malgré des résultats prometteurs, ce modèle présente encore quelques limites. Une limite majeure qui affecte fortement les résultats, est l'ambiguïté dans la description des images. Traditionnellement, un dictionnaire est généré en regroupant les descripteurs locaux visuels par une méthode de clustering comme *k-moyennes*, *BIRCH*,.... Dans la pratique, on s'aperçoit souvent que les dictionnaires de sacs de mots visuels construits de cette manière contiennent naturellement de nombreux synonymes et polysèmes. Dans de nombreux cas, un objet peut avoir plusieurs

1. $http://fr.mathworks.com/help/vision/ug/bagoffeatures_encodeoverview.png$

formes ou états. Pour représenter chaque forme, un ensemble différent de mots visuels est utilisé. Dans des autres cas, un mot visuel peut représenter une partie quelconque d'objets différents. Dans le monde réel, on voit aussi très souvent qu'il y a des objets qui ont des parties de forme identique. Par conséquent, deux objets différents peuvent être représentés par deux vecteurs très proches, et deux objets de même type peuvent être décrits par deux vecteurs très différents.

Techniques d'améliorations du modèle de sac de mots visuels classique

Pour réduire l'ambiguïté dans la description, de nombreuses idées d'amélioration ont été proposées. Plusieurs travaux se concentrent sur l'importance de la relation entre les caractéristiques des descripteurs bas niveau d'une image, principalement sur la position spatiale du descripteur. Dans le modèle de sacs des mots visuels original, les descripteurs dans une image sont considérés indépendants et désordonnés. Dans le monde réel, les parties d'un objet ont toujours un ordre spatial. Cet ordre peut être défini par la co-occurrence des descripteurs qui représentent les parties de l'objet. Par conséquent, la relation spatiale entre les descripteurs locaux est une information utile pour améliorer la performance du modèle de sac des mots visuels. Pour profiter la relation spatiale des descripteurs, dans des systèmes de recherche d'image par le contenu, on ajoute une étape de vérification après avoir appliqué le modèle de sac de mots classiques. Les techniques de vérification spatiale (RANSAC par exemple) peuvent améliorer la performance du modèle, mais elles sont complexes et coûteuse en temps de calcul.

Une manière plus efficace pour capturer la relation entre les descripteurs est d'enregistrer cette relation en construisant des les phrases visuelles qui sont formées en regroupant certains mots visuels selon des contraintes spécifiques. À partir des phrases construites, un nouveau dictionnaire plus descriptif peut être généré. Puis, au lieu de décrire une image comme un sac des mots visuels, on la décrit comme un sac des phrases visuelles. Similaire au modèle de sacs des mots visuels, l'idée de construire les phrases visuelles est inspirée par la notion de phrases dans le domaine d'analyse des documents textuels. Le modèle de sacs de phrases visuelles est une amélioration du modèle de sacs des mots visuels qui séduit fortement les scientifiques. Ce mémoire est une étude sur les différentes méthodes des sacs de phrases visuelles.

1.3 Objectifs du travail et principales contributions

L'objectif principal de ce travail est d'étudier les différentes méthodes existantes de sacs de phrases visuelles. Il s'agit d'une étude systématique avec un regroupement (typologie) des approches de sacs de phrases visuelles selon la méthode de construction d'une phrase.

La performance de certaines méthodes de sacs de phrases visuelles est aussi examinée. Le modèle de sac des mots visuels original [2] est considéré comme la méthode de base. Parmi les méthodes de sacs des phrases visuelles, deux méthodes appartenant à des types différents sont choisies et re-implémentées, puis elles sont comparées l'une avec l'autre ainsi qu'avec la méthode de base.

Les contributions de ce mémoire comportent deux volets :

1. Dans les années récentes, les méthodes de phrases visuelles sont devenues très populaires. De nombreuses méthodes ont été proposées mais il n'y a pas, à notre connaissance de document qui réalise une synthèse de ces méthodes. Ce mémoire serait donc le premier document qui présente une étude relativement complète et systématique des différentes méthodes existantes des phrases visuelles.
2. Malgré l'existence de plusieurs méthodes aux résultats prometteurs, chacune d'entre elles est implémentée et testée avec des dictionnaires de mots visuels et des bases d'images différentes. Cela constitue une difficulté dans la réalisation d'une comparaison entre les différentes méthodes sur la base des chiffres annoncées dans les publications associées à chacune d'elles. Il est nécessaire de réexaminer chaque méthode en les expérimentant dans les mêmes conditions. Cependant, la ré-implémentation des méthodes pour les expériences provoque beaucoup de problèmes. Dans le cadre ce mémoire, seulement deux méthodes de sac des phrases visuelles sont choisies et re-implémentées. Leurs performances sont comparées avec la méthode des sacs de mots visuels classique dans les même conditions : le même dictionnaire et la même base d'images. Pour avoir des comparaisons objectives, ces méthodes sont expérimentées avec trois bases d'images différentes.

Chapitre 2

État de l'art des méthodes de phrases visuelles

2.1 Phrases visuelles construites par fenêtres coulissantes

Dans ce contexte, "une fenêtre" est une borne dont la taille est fixée par un attribut, par exemple l'échelle, le rayon ou la longueur des axes des régions elliptiques [3, 4], parfois par une valeur constante [5]. Cette fenêtre est utilisée alternativement pour déterminer les voisins de chaque point d'intérêt dans une image, elle est donc appelée fenêtre coulissante.

Appliquant la fenêtre coulissante, la méthode de Bhatti and Hanbury [4] est une simple amélioration de la méthode de base. Le rayon de la fenêtre coulissante dans cette méthode est défini en fonction des longueurs des deux axes de la région elliptique correspondant à chaque point d'intérêt. Une phrase visuelle est simplement une paire de mots construite à partir d'un point central et d'un des voisins contenu dans la fenêtre coulissante. Un nouveau dictionnaire est créé contenant toutes les paires distinctes de mots visuels, nommé "Relational Features Codebook". Pour décrire une image, un histogramme de phrases visuelles est généré en codant les phrases avec ce dictionnaire. Bien que les résultats reportés par les auteurs ne soient pas meilleurs que ceux de la méthode de base, cette méthode est assez simple et capable de capturer la relation entre les caractéristiques des images.

Après la génération des candidats (les paires ou groupes des mots visuels qui peuvent être choisis comme phrases), les phrases visuelles sont choisies selon des critères. Ces critères sont différents selon chaque méthode. Dans [4], aucun critère n'est appliqué. Mais dans l'approche de Chen et al. [6], le critère est que les phrases visuelles doivent être discriminantes. Cette méthode est plus complexe que celle dans [4]. Tout d'abord, le voisinage spatial de chaque point est déterminé par une fenêtre coulissante ronde. Ensuite, les k voisins les plus proches du point central sont choisis pour générer les paires des mots visuels. Les paires de mots visuels les plus discriminantes sont choisies comme phrases visuelles. Selon les auteurs, cette méthode est capable de garder les phrases visuelles descriptives qui ont des basses fréquences. Les résultats expérimentaux présentés dans leur publication sont intéressants. Cependant, cette méthode est expérimentée seulement sur des images des monuments.

Dans une autre approche [7], la fréquence est utilisée comme un critère pour choisir les phrases visuelles à partir des paires des mots visuels. Pour qu'une paire de mots visuels soit une phrase visuelle, le nombre d'images dans la base d'image qui contiennent chaque mot visuel doit être supérieur à un seuil θ. Dans cette méthode, la fenêtre coulissante qui détermine les voisins des points d'intérêt est dynamique. Le rayon de la fenêtre ne dépend pas seulement du point central, mais aussi des voisins. Pour considérer si un point est le voisin d'un autre point, on compare leur distance euclidienne avec le rayon de la fenêtre coulissante (la somme des rayons des régions locales correspondant à ces points). Cette méthode est une des premières méthodes de phrases visuelles. Elle améliore la méthode de sacs de mots visuels classique sur l'efficacité (le temps d'exécution) et la performance (la qualité des résultats). Cependant, cette méthode ne fonctionne pas bien si les images ont peu de texture. Cette méthode est aussi peu performante pour les images dont l'arrière plan est complexe.

En résumé, avec la fenêtre coulissante, la relation examinée entre les mots visuels est la co-occurrence dans une région donnée, plutôt que leur distance. Dans la plupart de ces méthodes, une phrase est définie comme une paire des mots visuels. Le nombre de phrases construites à partir d'un point d'intérêt est varié, car le rayon de sa fenêtre coulissante dépend des caractéristiques du point central. Ces méthodes ne sont pas trop complexes par rapport à la méthode de base car les phrases visuelles ont une taille limitée. Par ailleurs, ces méthodes ne sont pas trop coûteuses en terme de temps de calcul. Mais, les résultats reportés montrent que ces

méthodes n'améliorent pas beaucoup la méthode de sacs de mots visuels classique. Comme le rayon de la fenêtre coulissante dépend seulement des caractéristiques du point central, les phrases visuelles ne sont pas très robustes au changement de point de vue. Dans de tels cas, les méthodes des "phrases visuelles comme groupes de k plus proches voisins" qui sont présentées dans la partie suivante semblent être plus efficaces.

2.2 Groupes de plus proches voisins

L'algorithme des k-plus proches voisins est un des algorithmes de classification les plus simples et populaires. Étant donné un point x et un ensemble de points A, cet algorithme trouve un sous ensemble de A contenant les k points les plus proches de x en utilisant une distance métrique (la distance de Mahalanobis ou Euclidienne, ou une autre distance définie par l'utilisateur).

Parmi les méthodes de sacs de phrases visuelles, beaucoup de méthodes appliquent l'algorithme des k-plus proches voisins pour déterminer le voisinage des points d'intérêt. La première proposition de sacs de phrases visuelles comme groupes de k-plus proches voisins a été publiée juste après la publication de la méthode de sac de mots visuels classique, et par les même auteurs [8]. Dans cette proposition, une phrase est décrite comme un groupe de $k+1$ points : un point central et ses k voisins les plus proches spatialement dans l'image. Pour comparer la similarité entre deux phrases visuelles, on compare d'abord les deux mots correspondants aux deux points centraux. Puis, on compte le nombre de mots communs entre ces phrases. Une paire de phrases visuelles est dite "match" (c.à.d elles décrivent la même configuration spatiale) si elles ont au moins m voisins similaires ($m \leq k$), où m est un seuil fixé heuristiquement. Dans l'expérimentation, cette méthode est appliquée pour extraire les objets, les personnages et les scènes principaux d'une vidéo. Selon les auteurs, cette méthode est assez efficace, mais elle n'est pas invariante au changement d'échelle.

Dans une autre approche [9], l'algorithme des k-plus proches voisins est combiné avec l'algorithme de triangulation de Delaunay pour former les phrases visuelles. Pour une image, d'abord les points SURF sont extraits. Les points qui correspondent aux détails les plus saillants dans l'image sont choisis pour former un ensemble de points-graines (seeds). Les "graph features" sont formés en groupant

chaque "point-graine" avec ses k voisins les plus proches spatialement. Ces points deviennent les sommets du "graph feature". Les arêtes du graphe sont déterminées en appliquant l'algorithme de triangulation de Delaunay sur les sommets. Dans cette méthode, les auteurs utilisent une structure hiérarchique : à partir d'un "point-graine", quatre graphes de taille croissante sont construits. Les tailles des graphes sont déterminées par le nombre de sommets. Le graphe de la première couche ne contient que le "point-graine". Le nombre de sommets augmente de 3 pour chaque couche suivante, la dernière couche contient 10 points : un "point-graine" et ses 9 plus proches voisins spatialement. À partir des "graph features", un dictionnaire comprenant les "graph words" est formé. Les "graph words" sont formés à l'aide des "points-graines", représentant les médianes des "graph features" dans chaque cluster. Ces "graph words" sont considérés comme étant les phrases visuelles.

Cette méthode n'utilise pas directement les descripteurs d'images pour construire les phrases visuelles. L'ensemble de "features" sont plutôt utilisés en empruntant l'idée des k-plus proches voisins pour représenter plus d'informations qu'une seul caractéristique. Les relations entre les nœuds dans un graphe sont déterminées par l'algorithme de triangulation de Delaunay, qui est invariante aux changements affines des objets dans les images comme la rotation, la translation ou le changement d'échelle. Les "graph features" sont plus robustes que les descripteurs standards. En plus, selon les auteurs, l'utilisation de la structure hiérarchique peut contribuer à l'amélioration des résultats de la recherche et de la reconnaissance. Cependant, l'inconvénient principal de cette méthode est l'absence d'une structure d'indexation pour les "graph words" qui peut provoquer plus de charge en termes de temps de calcul dans l'étape de recherche et de reconnaissance.

Différemment des voisinages déterminés par les fenêtres coulissantes, ceux déterminés par l'algorithme des k-plus proches voisins ont des formes variées. Le voisinage d'un point d'intérêt peut être défini comme l'enveloppe convexe de l'ensemble de ses k-plus proches voisins. Donc, le rayon du voisinage d'un point dépend de la distance entre lui et ses voisins. Grâce à cette caractéristique, les phrases visuelles sont plus robustes au changement de point de vue que celles construites par les fenêtres coulissantes. Cependant, les méthodes dans ce groupe ne sont pas robustes au changement d'échelle. Prenons comme exemple deux images d'un même objet ayant une échelle différente. Dans l'image ayant une échelle plus grande, on peut détecter plus de points d'intérêt que dans celle ayant une échelle plus petite. Il

peut donc y avoir beaucoup de points d'intérêt qui existent dans seulement une des deux images. Deux même objets peuvent ne pas être décrits par les mêmes phrases visuelles avec ce type de méthodes.

2.3 Chaînes des mots visuels

Il s'agit du groupe le plus spécial : les phrases visuelles des méthodes dans ce groupe sont les plus proches des phrases retrouvées dans le domaine d'analyse des documents textuels. Chaque phrase visuelle est un ensemble ordonné d'éléments (ces éléments peuvent être les mots visuels ou les histogrammes de fréquence des mots visuels). Pour les distinguer des autres groupes, on appelle ces phrases visuelles "les chaînes des mots visuels".

Dans la méthode de Nguyen et al. [10], chaque image est divisée en plusieurs régions selon deux axes principaux (l'axe vertical et l'axe horizontal). Un histogramme qui décrit la fréquence des mots visuels est généré pour chaque région dans l'image. Une phrase visuelle est une chaîne d'histogrammes selon l'axe majeur de l'image. Une image est représentée par une phrase visuelle. Cette méthode est une amélioration de la méthode *Spatial Pyramid Matching (SPM)* [11], elle hérite de l'étape de répartition de l'image et de la structure hiérarchique de SPM. Pour mesurer la similarité entre deux images, les auteurs utilisent un algorithme qui applique des coûts de suppression et d'insertion, et calcule une distance d'édition entre les deux chaînes d'histogrammes. Les résultats présentés dans [10] montrent que cette méthode donne de meilleures performances que la méthode *Spatial Pyramid Matching* et plusieurs variantes de celle-ci.

Tirilly et al. [12] a présenté une méthode qui permet d'examiner les mots visuels dans une phrase visuelle selon un certain ordre. Le processus de construction d'une phrase visuelle se présente comme suit : tout d'abord, on doit construire efficacement un dictionnaire de mots visuels et représenter chaque image par un ensemble de mots visuels en assignant un point d'intérêt au mot visuel le plus proche. Ensuite, on doit définir un axe qui est représentatif de la position de l'objet dans l'image. On projette l'ensemble des mots visuels de façon ordonnée sur cet axe principal (c'est une projection orthogonale) pour obtenir une représentation finale

de l'image entrée. Pour choisir un bon axe, il faut respecter des critères d'orientation et de direction de l'axe pour que ces projections correspondent à celles de l'objet dans l'image d'entrée. Plusieurs stratégies pour trouver un axe sont possible : utiliser simplement l'axe horizontal ou choisir aléatoirement deux axes perpendiculaires ou bien, dans le cas particulier où il y a un seul objet dans l'image, l'ACP (Analyse en composantes principales) est un bon choix. Cette méthode convient particulièrement au cas où l'image ne contient qu'un seul objet car s'il y en a plusieurs, le résultat de l'ACP est biaisé par les positions relatives de ces objets. En raison de l'utilisation simple d'un axe principal pour représenter l'image, cette méthode ne s'adapte pas aux images ayant un arrière plan complexe.

2.4 Phrases visuelles construites par régions

Les approches de ce groupe ont été plus récemment proposées. Ce groupe a des caractéristiques principales très différentes des autres groupes présentés. Les méthodes de ce groupe découpent dans une première étape les images en morceaux. Une phrase visuelle peut être définie comme l'ensemble des mots visuels ou l'histogramme des mots visuels dans une région.

Il y a différentes façons de diviser une image en régions. Dans la méthode de Jiang et al. [13], une image est simplement divisée en certaines partitions en la couplant aléatoirement avec les lignes horizontales et verticales. L'approche de sacs de sacs de mots visuels de Ren et al. [14] est plus complexe : pour morceler une image, un graphe initial est construit. Les sommets du graphe sont les points d'intérêt (détectés de manière dense) de l'image originale. Le poids de chaque arc est déterminé en fonction des couleurs dans l'espace YUV des deux terminaisons de l'arc. L'image est ensuite découpée en appliquant l'algorithme de *Normalized Graph Cut* sur le graphe initial. Dans une autre proposition [15], un graphe initial est aussi construit sur l'image mais par l'algorithme de triangulation de Delaunay. Ce graphe est ensuite couplé en plusieurs sous-graphes en utilisant l'algorithme de graph-cuts.

Le morcellement des images est une étape importante car cela améliore les résultats des méthodes dans ce groupe. En général, le morcellement permet d'accumuler plus d'informations spatiales des descripteurs. Le nombre de régions créées par le morcellement des images ou du graphe influence fortement l'efficacité des phrases

visuelles. Pour avoir des bons résultats, les approches sont souvent implémentées de manière hiérarchique. Dans [13], les images sont morcelées aléatoirement plusieurs fois. Le nombre de régions générées est le même pour toutes les couches. Les morceaux dans une même couche sont disjoints, mais les morceaux dans les différents couches peuvent se chevaucher. Dans [15] et [14], le nombre de régions augmente en fonction de la hauteur de la couche. Pour la couche 0, il y a seulement une région, l'image n'est pas divisée. Le nombre de régions augmente à 4 pour la couche 1, 16 pour la couche 2 et 64 pour la couche 3... La technique de morcellement dans ces méthodes est inspirée par la méthode *Spatial Pyramid Matching (SPM)* de Lazebnik et al. [11].

Après l'étape de découpage, c'est l'étape de construction des phrases visuelles. Dans [13], une phrase est l'ensemble des mots visuels correspondant aux descripteurs dans un morceau de l'image. Cette méthode est utilisée pour faire la recherche et la localisation des objets dans les images. Elle applique un mécanisme de vote sur tous les pixels dans l'image pour avoir les scores de confiance. En détail, le score d'un pixel est la valeur d'espérance des scores de similarité entre le requête et les régions contenant le pixel. Le score de similarité est calculé en utilisant l'histogramme d'intersection. Enfin, un seuil est utilisé pour comparer avec les scores de confiance pour déterminer la région contenant l'objet.

Différemment des phrases visuelles dans [13], celles dans [15] et [14] sont définies comme les histogrammes des mots visuels des sous-graphes (régions). Comme le graphe initial est morcelé spatialement, on peut considérer que chaque phrase représente une région dans l'image originale. Donc le nombre de phrases visuelles pour représenter une image égale le nombre de régions divisées dans l'étape de morcellement. Après que les images soient décrites par les phrases visuelles, dans l'étape de recherche, elles sont comparées avec l'image de requête qui est aussi représentée par un ensemble de phrases visuelles.

Chapitre 3

Mise en œuvre de quelques méthodes

3.1 Protocole expérimental

Pour comparer les méthodes, il faut que ses méthodes soient exécutées selon un même protocole expérimental. Nous utilisons donc un même dictionnaire pour toutes les méthodes et pour toutes les bases d'images. Les descripteurs SIFT [1] (avec l'avantage d'être peu sensibles au changement d'intensité, d'échelle et de rotation,...) sont utilisés partout dans notre expérimentation.

Le modèle de sacs de mots visuels classique [2] est choisi comme la méthode de base pour les comparaisons. Nous sommes conscient que différents améliorations sont possibles [16] mais pour des contraints de temps liés à la durée du stage, nous aurons considéré cette approche "de base". Pour assurer la cohérence entre toutes les méthodes implémentées, un seul dictionnaire est utilisé dans ce travail. Donc, pour cette méthode, le détecteur choisi est basé sur les différences de gaussiennes (DoG) pour adapter le dictionnaire.

3.2 Méthode de sacs de phrases visuelles descriptives

C'est une méthode où les phrases visuelles sont construites par les fenêtres coulissantes (groupe 1, voir section 2.1). Dans cette méthode, une phrase visuelle est construite à partir des mots visuels co-occurrents. Pour construire un sac de phrases visuelles, les étapes principales suivantes sont effectuées successivement :

— **Génération des phrases visuelles descriptives candidates :** Les DVPs candidates sont les paires de mots visuels qui se trouvent ensemble dans un histogramme spatial [17]. Un avantage de cet histogramme est l'invariance à la rotation, qui aide garder la relation spatiale entre les mots visuels qui se trouvent dans un même histogramme si le contenu de l'image est changé par permutation circulaire. Pour préparer l'étape suivante, les fréquences d'occurrences moyennes des candidates dans chaque image (noté $VPf^{(C)}$) et dans chaque catégorie (noté VPF) sont calculées.

— **Choix des phrases visuelles descriptives à partir des candidates** : Une phrase visuelle est dite descriptive si elle peut caractériser une catégorie particulière. Selon l'auteur, la fréquence d'occurrence d'une telle phrase est haute pour la catégorie qu'elle caractérise et basse pour les autres catégories. Donc, pour choisir les phrases visuelles descriptives, on calcule le score d'importance des candidates. Les scores d'importance calculés sont triés. Les candidats qui ont les scores les plus hauts sont choisis comme phrases visuelles descriptives.

— **Construction du dictionnaire des phrases visuelles descriptives** : après le choix des phrases visuelles descriptives pour chaque catégorie, les ensembles des phrases visuelles sont unis. Le nouvel ensemble formé est le dictionnaire des phrases visuelles descriptives.

— **Représentation des images** : pour chaque image, d'abord les DVPs candidates sont extraits. Les candidates qui existent dans le dictionnaire des phrases visuelles sont choisis. Ils sont ensuite utilisés pour générer un histogramme des DVPs qui décrit l'image.

Trois expériences différentes : recherche d'image, reconnaissance d'objet et reclassement d'image (*Image re-ranking*) sont effectuées dans sa méthode. Selon les résultats expérimentaux dans [3], la performance de la méthode de DVP est

meilleure que celle du modèle de sacs de mots visuels classique de 19.5% en recherche et 80% en reconnaissance d'objets, c'est un résultat impressionnant.

L'amélioration de cette méthode est due à la manière de représenter les images via un dictionnaire des phrases visuelles en remplaçant les mots visuels traditionnels. Les phrases visuelles sont compactes avec la taille fixe de deux, et descriptives car elles sont basées sur les fréquences des candidates qui sont caractérisées par chaque catégorie, donc elles sont capables de capturer certains objets et scènes. Cependant, cette méthode peut reconnaître seulement les images qui sont similaires visuellement à la requête. En outre, pour obtenir des bons résultats, la base d'images qui est utilisée pour construire le dictionnaire des phrases visuelles doit être représentative. Les phrases visuelles dans le dictionnaire sont seulement descriptives pour les catégories existantes dans la base d'images utilisée car le dictionnaire est construit en unissant les ensembles des candidates sélectionnés pour chaque catégorie.

3.3 Sacs de sacs de mots visuels

Introduite par Ren et al. [14], elle hérite de la méthode *Spatial Pyramid Matching* (SPM) [11] avec l'idée d'empiler plusieurs niveaux de morcellement d'une image pour la représenter. Dans cette méthode, une image est représentée par un ensemble de plusieurs sacs des mots visuels. Une phrase visuelle est un sac de mots visuels généré par une région de l'image. Voici le détail des 4 principales étapes :

— **Construction du graphe pondéré connexe** : Pour une image, le graphe pondéré connexe est formé par les points d'intérêt et leurs connections. Les points d'intérêt sont déterminés par la stratégie d'échantillonnage dense (dense sampling). Ils sont les sommets du graphe. Les arrêtes du graphe sont obtenues par les liens entre chaque point et ses 8 voisins sur 8 directions : est, ouest, sud nord, sud-est, nord-est, sud-ouest et nord-ouest. Cette méthode utilise l'espace de couleur YUV car les canaux de couleur y sont indépendants. Cet espace permet de mieux traiter les changements d'illumination.

— **Morcellement du graphe** : dans cette étape, le graphe créé à l'étape précédente est morcelé en plusieurs sous-graphes. L'algorithme de découpage

N-Cut est appliqué pour optimiser deux critères : le score total de dissimilarité entre les sous-graphes différents et le score de similarité dans chaque sous-graphe sont maximisés. Le nombre de régions morcelées est configuré par l'utilisateur, ce sont 1, 4, 16 ou 64 régions selon l'article original [14].

— **Représentation des régions** : après la répartition, chaque région est représentée séparément par un sac de mots visuels. Les descripteurs SIFT sont extraits aux sommets du graphe. Ils sont ensuite quantifiés en utilisant le dictionnaire des mots visuels. Un histogramme de fréquence des mots visuels est ensuite généré pour chaque région. On appelle cet histogramme une phrase visuelle.

— **Représentation des images** : finalement, une image est représentée par une pyramide de vecteurs des histogrammes des mots visuels (ou un ensemble des phrases visuelles). Inspirée par la méthode SPM, cette méthode prend en compte plusieurs résolutions de morcellement. À chaque résolution r, le grand graphe de l'image est divisé en $K_r = 2^{2r}$ sous-graphes. La représentation de l'image à la résolution r est donc un vecteurs des K_r histogrammes des mots visuels, où chaque histogramme représente une de K_r régions dans l'image. L'empilement des résolutions de 0 à r forme une pyramide des partitions. La représentation globale est donc la pyramide de $r + 1$ vecteurs des histogrammes des mots visuels.

Pour comparer deux images, les auteurs proposent une technique nommée *Irregular Pyramid Matching* (IPM). Avec l'utilisation de l'algorithme *N-Cut*, les images sont morcelées différemment, donc on doit faire une étape de mapping pour mapper les régions correspondantes entre deux images. Pour résoudre ce problème, l'algorithme *Hongrois* qui sert à trouver un appariement optimal entre les régions. Après l'étape de mapping, l'intersection entre les deux histogrammes de chaque paire de régions correspondantes est déterminée. La similarité entre deux images est calculée en utilisant la formule de "level weighted intersection".

Pour tester cette méthode, les expériences de recherche d'images par le contenu sont effectuées sur la base d'images Caltech-101. La méthode SPM est choisie pour la comparaison. Les résultats montrent que la méthode BBW est comparable à la méthode SPM et meilleure que la méthode de base. Malgré de bons résultats, cette méthode peut consommer beaucoup de temps dans l'étape de recherche à cause de l'utilisation de l'algorithme *Hongrois* dont la complexité est $O(n^3)$.

Chapitre 4

Expérimentation et discussion

Pour comparer les méthodes, des expériences de recherche d'images par le contenu sont effectuées. Un même dictionnaire de 2000 mots visuels est utilisé pour toutes les méthodes et tous les cas de test. Les méthodes sont testées trois fois sur trois bases d'images différentes : La base d'images *Holiday*, la base d'images *Caltech-101* et la base d'images *ImageNet*.

4.1 Méthode d'évaluation

Dans ce mémoire, les méthodes sont évaluées et comparées selon deux critères :

— **Mean Average Precision (mAP)** : c'est une mesure populaire pour évaluer les systèmes de recherche, qui est la mesure standard dans la communauté des conférences de recherche textuel (TREC - Text Retrieval Conferences). Parmi les mesures d'évaluation, mAP permet d'avoir une particulièrement bonne discrimination et stabilité.
La mAP est calculée sur les résultats de recherche d'un ensemble des requêtes. Pour une requête, la précision moyenne est la moyenne des k valeurs de précision obtenues pour k premières images trouvées par la recherche (le poids des valeurs de précision diminue selon l'ordre des images trouvées). Étant donné un ensemble des requêtes, la mAP est déterminée en établissant la moyenne des valeurs de précisions moyennes obtenues pour chaque requête.

— **Le temps d'exécution pour la recherche** : c'est aussi un critère important pour évaluer l'efficacité des méthodes. Comme l'étape de création des

phrases visuelles et l'étape d'indexation des images se font hors ligne, on ne prend en compte dans l'évaluation que le temps d'exécution pour l'étape de recherche.

Pour bien comparer les méthodes sur l'aspect du temps d'exécution, les tests sont lancés sur un même environnement expérimental. C'est une station de travail HP avec un processeur Intel Xeon de 16 cœurs, 32 Go de mémoire RAM et disque dur 1To.

4.2 Les base d'images utilisées

Dans ce mémoire, quatre bases d'images sont utilisées pour faire les tests. La base d'images MIRFLICKR-25000 sert à la construction du dictionnaire et les trois bases d'images restantes sont utilisées pour l'évaluation des méthodes.

Base d'images MIRFLICKR-25000

Cette base d'images est utilisée pour la construction du dictionnaire. La base d'images contient 250000 images, fournie par le LIACS Medialab à l'université de Leiden en 2008. Elle est introduite la première fois en 2008 par la commission de ACM MIR, pour évaluer les méthodes de recherche d'images par le contenu.

Flickr est une plate-forme qui permet à l'utilisateur de chercher et de partager ses images, avec des étiquettes pour chaque image. Avec une grande base d'utilisateurs, le contenu des images dans la base d'images *MIRFLICKR-25000* est très varié. Donc en utilisant cette base d'images pour la construction du dictionnaire, les mots visuels formés peuvent être assez représentatifs pour représenter différents détails dans les images.

Base d'images *Holiday*

Il y a 1491 images dans 500 catégories. En général, une catégorie contient 2, 3 ou 4 images représentant une scène distincte ou un objet différent. En utilisant cette base d'images, on ne prend pas en compte la variété visuelle des objets d'un même type. La différence des images dans une catégorie nous permet de tester la

robustesse des méthodes de recherche aux changements de rotation, d'illumination, de point de vue ou au flou.

Pour faire les expériences, la base d'images est divisée en deux parties. La première partie contenant 500 images est l'ensemble des requêtes. Les images sont choisies au hasard, une image par catégorie. La deuxième partie est l'ensemble des images restantes qui forme un pool pour la recherche. Pour une requête, une image retournée est considérée comme correcte (un bon résultat) si elle est dans la même catégorie que la requête.

Base d'images *Caltech-101*

Caltech-101 (Fei-Fei et al. [18]) est une base d'images numérisées qui contient un total de 9146 images collectées, classées en 101 types d'objets (par exemple visages, pianos, moto, ordinateur portable, etc). Les images sont très uniformes dans leur présentation, alignées à gauche ou à droite. En effet, la plupart des images dans chaque catégorie ont une taille similaire d'environ 300x200 pixels. Les objets d'intérêt ont tendance à être centrés dans les images et se présentent dans une pose stéréotypée. Les arrières-plans des images sont très hétérogènes, mais pas aussi complexe que dans les autres bases d'images utilisées dans ce travail. Le nombre d'images dans les catégories est différent, de 31 à 800 images. En fait, plusieurs catégories ne contiennent que peu d'images, c'est insuffisant pour construire le pool de recherche. Donc, seulement les 26 catégories qui ont les plus d'images sont utilisées pour faire les tests. À partir de chaque catégorie, 10 images sont choisies pour un total de 260 images comme requêtes. Le pool de recherche contient 1820 images (70 images par catégorie). Toutes les images sont choisies aléatoirement.

Base d'images *ImageNet*

ImageNet [19] est une base d'images grande et complexe qui est construite en se basant sur la hiérarchie de *WordNet*. Chaque concept du *WordNet* est éventuellement décrit par plusieurs mots ou groupes de mots, appelés *"synset"* (*"synonym set"*), et est représenté par des centaines ou des milliers d'images. Les images dans cette base sont de qualité contrôlée, elles sont 2t2 annotées sous la supervision d'humains.

Parmi plus de cent mille *synsets* disponibles, on a choisi à la main 15 *synsets* familiers comme *chien, poisson, aéroplane, vélo, maison, etc* . Chaque *synset* forme une catégorie de plus de 800 images. Pour les tests, 100 images sont choisies aléatoirement par catégorie comme requêtes, les 700 autres images sont sélectionnées pour former le pool de recherche.

4.3 Analyse des résultats

Les chiffres dans le tableau 4.1 montrent la différence de performance entre les méthodes. La méthode des sacs de sacs de mots visuels (BBW) prouve sa performance sur les bases d'image *Holiday* et *Caltech-101*. Par contre, la méthode de phrases visuelles descriptives (DVP) ne peut pas prouver son amélioration. Parmi les approches, la méthode DVP donne les moins bonnes mAPs dans tous les cas. La différence entre ses résultats et ceux des autres méthodes est assez grande. Pour la base d'images *ImageNet*, la méthode de base donne le meilleur résultat et la méthode DVP donne le plus mauvais résultat.

TABLE 4.1: mAP des méthodes sur les bases d'images différentes

	Classique	BBW-2lv	BBW-3lv	BBW-4lv	DVP
Holiday	0.524	0.564	0.554	0.51	0.388
Caltech-101	0.210	0.251	0.271	0.321	0.173
ImageNet	0.164	0.158	0.145	0.147	0.078

Les mAPs sont aussi très différentes entre les bases d'images. Cette disparité est causé par la différence entre les caractéristiques des bases d'images. Dans la base *Holiday*, les images dans une catégorie capturent seulement une scène ou un objet unique. Donc les images ne sont pas très différentes les unes des autres. Même s'il y a une rotation, transition, ou changement de luminance, les images se chevauchent souvent en partie. Ce chevauchement facilite la recherche des phrases visuelles communes. Pour cette raison, on obtient les mAPs les plus élevées sur la base *Holiday*. Pour la base d'images *ImageNet*, les objets dans une catégorie peuvent être variés en taille, forme, couleur, etc. D'ailleurs, les images peuvent contenir plusieurs objets de même type ou de différentes type. En outre, l'arrière-plan dans les images est parfois texturé ou l'arrière-plan d'une image peut être l'objet des autres images dans une autre catégorie... Donc, on peut facilement comprendre que les mAPs soient en baisse sur la base *ImageNet*. Avec la base d'images *Caltech-101*, les mAPs obtenues sont aussi moins bonnes que sur la base d'images *Holiday*.

Peut-être que la variété visuelle des objets dans une catégorie provoque des difficultés qui diminuent la performance des méthodes. En comparaison avec la base *ImageNet*, les images dans la base *Caltech-101* ne contiennent souvent qu'un seul objet. L'arrière-plan dans une image est souvent simple et l'objet est souvent au centre de l'image. Grâce à ces caractéristiques, les mAPs sur cette base d'images sont meilleures que celles sur la base *ImageNet*.

À partir des résultats dans le tableau 4.1, on peut non seulement comparer les méthodes, mais aussi vérifier l'influence de la structure hiérarchique sur la performance de la méthode des sacs de sacs de mots visuels. Dans [14], cette méthode est testée avec la base d'image *Caltech-101*. Les résultats ont montré que la structure hiérarchique du morcellement des images influence la performance de la méthode : plus on met de couches, plus haute est la performance. Ce phénomène apparaît aussi dans le tableau 4.1 avec la base d'image *Caltech-101* (la même base d'images que les auteurs ont utilisé). Cependant, il n'existe pas dans le cas des bases *Holiday* et *ImageNet*. Tandis que les mAPs sur la base *Caltech-101* augmentent progressivement et le cas BBW-4lv (4 couches de partition) donne le meilleur mAP (0.321), sur la base *Holiday*, le cas BBW-2lv nous donne le meilleur mAP (0.564), puis les mAPs diminuent. Pour la base d'images *ImageNet*, le cas de 1 couche (la méthode de base) donne le meilleur mAP (0.164) et l'ajout de plus de morcellements diminue légèrement ce résultat. Pour conclure, la structure hiérarchique proposée dans la méthode BBW semble appropriée seulement pour les images simples qui ne contiennent pas beaucoup d'objets ou d'arrière-plan texturé.

Le deuxième aspect pour évaluer les méthodes est le temps d'exécution. Le tableau 4.2 affiche la complexité théorique et les mesures pratiques (en minute) de toutes les méthodes sur les trois bases d'images. La complexité de la recherche avec la méthode classique et celle de DVP sont $O(n)$, où n est la taille du vecteur qui représente l'image. Pour la méthode BBW, une image est représentée par un vecteur des vecteurs de phrases visuelles. Pour la recherche, on doit ajouter une étape pour mettre en correspondance les régions correspondantes qui utilise l'algorithme Hongrois (Hungarian Algorithm). La complexité de cette étape est $O(m^3)$ où m est le nombre de morceaux dans l'image. Les deux images sont ensuite comparées en comparant les paires de vecteurs correspondants. La complexité de la comparaison est $O(m*n)$, où n est la taille d'un vecteur de mots visuels (c'est aussi la taille du dictionnaire des mots visuels). La complexité globale de la recherche est donc $O(m^3) + O(mn)$ pour BBW. En ce qui concerne les mesures pratiques, la

méthode DVP consomme presque le même temps, parfois même moins de temps que la méthode de base. La raison est l'utilisation du dictionnaire dans lequel les phrases visuelles sont représentées par les indexes des paires de mots visuels. La complexité de la recherche est donc $O(n)$ où n est la taille du dictionnaire des phrases visuelles.

TABLE 4.2: Temps d'exécution des méthodes sur les bases d'images différentes

	Classique	BBW-2lv	BBW-3lv	BBW-4lv	DVP
Complexité théorique	$O(n)$	$O(m^3) + O(mn)$			$O(n)$
Holiday	3m34.18s	4m49.14s	13m49.26s	127m2.45s	3m45.71s
Caltech-101	0m34.95s	1m42.23s	9m31.88s	103m7.62s	1m0.52s
ImageNet	20m49.14s	56m37.98s	306m11.76s	3563m39.53s	10m50.15s

À partir de ce tableau, la méthode BBW a exprimé sa complexité. À la différence de la méthode DVP, la méthode BBW consomme plus du temps. Quand le nombre de couches augmente, le temps d'exécution augmente de manière exponentielle. C'est à cause de l'utilisation de l'algorithme Hongrois (*Hungarian Algorithm*) duquel la complexité est $O(n^3)$ pour mettre en correspondance les partitions de deux images. On doit refaire la mise en correspondance plusieurs fois dans toutes les recherches pour comparer la similarité entre la requête et chaque image dans le pool. Pour un niveau plus haut dans la structure hiérarchique (plus de couches), le temps d'exécution augmente très vite car plus haute est la couche, plus on a de partitions, donc plus de temps consommé. Par ailleurs, on doit faire la mise en correspondance séparément pour chaque couche. Par exemple, dans le cas de 4 couches, on doit utiliser l'algorithme Hongrois 4 fois pour la couche 1 de 1 partition, la couche 2 de 4 partitions, la couche 3 de 16 partitions et la couche 4 de 64 partitions. En résumé, quand le nombre de couches utilisées dans la méthode BBW augmente, le temps d'exécution augmente de manière exponentielle dans tous les cas, mais la performance (mAP) de la méthode change de manière instable, dans certain cas la performance n'augmente pas mais diminue.

Chapitre 5

Conclusion

Ce mémoire est une bibliographie sur les méthodes existantes de sacs des phrases visuelles, qui sont les améliorations du modèle de sacs des mots visuels. Les méthodes sont recensées et groupées en 4 catégories : phrases visuelles construites par fenêtres coulissantes, groupes de plus proches voisins, chaînes des mots visuels et phrases visuelles construites par régions. Cette bibliographie est espérée comme une référence pour avoir une vue générale sur la représentation des images par le sac des phrases visuelles.

En outre dans ce travail, deux méthodes de sacs des phrases visuelles sont re-examinées : la méthode de sacs des sacs des mots visuels (BBW) et la méthode des phrases visuelles descriptives (DVP). Elles sont choisies à partir de deux groupes : le groupe des phrases visuelles construites par régions (BBW) et le groupe des phrases visuelles construites par fenêtres coulissantes (DVP). Les expériences de recherche des images par le contenu sont effectuées en utilisant un dictionnaire commun qui est généré à partir de la base d'image *MIRFLICKR-25000*. Ces méthodes sont testées séparément sur trois bases d'images différentes : *Holiday*, *ImageNet* et *Caltech-101*. Selon les résultats expérimentaux, la méthode BBW donne une assez bonne performance sur la base *Holiday* et *Caltech-101*. Parmi les méthodes, la méthode DVP est rapide mais ses performances ne sont pas à la hauteur des autres méthodes. Pour la base d'images *ImageNet*, les deux méthodes de sacs des phrases visuelles ne peuvent pas prouver leur amélioration, la méthode de base donne la meilleure performance dans ce cas. À partir des résultats expérimentaux, l'influence de la structure hiérarchique de la méthode BBW est aussi vérifiée.

Les chiffres montrent que l'augmentation du nombre de couches dans la structure hiérarchique n'améliore pas toujours la performance, mais augmente le temps d'exécution de manière exponentielle. En utilisant les bases d'images différentes, nos tests montrent aussi que la performance des méthodes dépend fortement des caractéristiques de la base d'images.

Dans le futur, pour avoir une comparaison plus complète, l'implémentation d'autres méthodes, au moins deux appartenant aux deux autres groupes, sera nécessaire. Avoir une évaluation complète des méthodes des sacs des phrases visuelles permettrait de choisir efficacement la méthode la plus adaptée à l'application souhaitée. Cela nous permettra dans le même temps de mieux comprendre le fonctionnement des méthodes étudiées et d'en proposer une amélioration pertinente.