
DÉTECTION D'OBJETS DANS LES IMAGES DE DOCUMENTS

Spécialité : Master en Système de Traitement de l'Information
Multimedia

Editer par :

Achraf BEN SALAH

Encadré par :

Laurent HEUTTE



UFR des sciences et techniques site du Madrillet
Université de Rouen

2013–2014

Remerciements

J'exprime mes profonds remerciements à *M. Laurent HEUTTE*, pour m'avoir encadré et pour m'avoir donné l'occasion de travailler sur ce sujet riche, actuel et passionnant. Sa grande disponibilité, son sens des relations, sa rigueur, son expérience, sa pédagogie et ses critiques constructives m'ont été précieux.

Table des matières

Remerciements	iii
Table des matières	v
Table des figures	vii
Liste des tableaux	ix
Résumé et Introduction générale	1
1 Résumé	1
2 Introduction générale	1
1 Cadre de TER	3
1 Introduction	3
2 Contexte	3
3 Contexte de travail	4
4 Objectif	4
5 Organisation et plan	5
2 Etat de l'art et contribution	7
1 Introduction	7
2 Notion de base	8
2.1 Mots visuels	8
2.2 Vocabulaire	8
3 Modèle vectoriel	9
4 Recherche des images par contenu	9
5 Caractérisation des régions de l'image	11
5.1 Les regions invariantes	11
5.2 Les descriptions invariantes	13
6 Méthodologie	13
6.1 Processus de travail	13
6.2 Réalisation	15

3	Expérimentation	21
1	Introduction	21
2	Base de document de validation	22
3	Résultats d'évaluation	23
4	Conclusion générale	25
	Bibliographie	27

Table des figures

1.1	Exemple des mots visuels	4
2.1	Exemple : Vocabulaire ou dictionnaire des mots visuels.	8
2.2	L'architecture d'un système d'indexation et recherche d'images par le contenu	10
2.3	Exemple : Distance cosinus	11
2.4	Construction de l'histogramme des orientations.	12
2.5	Construction d'un descripteur SIFT.	13
2.6	Processus général de recherche des objets dans les images de documents	14
2.7	Schéma de cas d'utilisation de notre application	15
2.8	Sélection de l'image requête et de la base d'image de recherche . . .	17
2.9	l'interface principale de notre application	17
2.10	Formation de l'image requête	18
2.11	Application de descripteur <i>SIFT</i>	18
2.12	Formation des mots visuels	18
2.13	Calcul de la distance entre l'image requête et l'image de la base . . .	19
2.14	Résultats de l'opération de recherche	19
3.1	Principe Rappel-Préssion	21
3.2	List des images de test	22

Liste des tableaux

2.1	Exemple : Fichier des étiquettes	14
3.1	Tableau de correspondance Rappel-précision et le temps d'exécution avec $k = 50$	23
3.2	Tableau de correspondance Rappel-précision et le temps d'exécution avec $k = 250$	24
3.3	Tableau de correspondance Rappel-précision et le temps d'exécution avec $k = 500$	24
3.4	Tableau de correspondance Rappel-précision et le temps d'exécution avec $k = 1000$	24

Résumé et Introduction générale

1 Résumé

Ce rapport présente une étude sur les performances des algorithmes de **détection d'objets dans les images de documents**. Ce sujet est inclus dans le domaine d'indexation et recherche d'images par **l'utilisation des mots visuels** ou recherche et indexation d'images **par contenu**.

Le but de mon travail est d'étudier l'approche **mots visuels** dans la recherche d'images. Pour atteindre le but, je vais utiliser le modèle vectoriel en image proposé par [J.M]. En effet chaque image (requête et les systèmes candidats) est représentée par un vecteur fréquentiel. On doit tout d'abord créer un dictionnaire des mots visuels selon un détecteur, un descripteur et un algorithme de regroupement des mots. Une fonction de correspondance entre la base d'images et la requête est associée pour obtenir la similarité entre le candidat et la requête. Le résultat sera trié selon la priorité des images les plus similaires.

Mots clés : Sift, Mots visuels, dictionnaire de mots visuels, vocabulaire de mots visuels, contenu, indexation, recherche d'images, recherche d'informations, détection des objets.

2 Introduction générale

Aujourd'hui, il existe de nombreux contenus multimédias qui sont constitués et stockés dans des environnements différents et hétérogènes tel que ils sont diffusés géographiquement et de capacités diverses. Cette masse de données est généralement indexée par des algorithmes qui vont extraire des métadonnées permettant à un utilisateur de rechercher de l'information.

Toutefois, nous éprouvons aujourd'hui une grande diversité de ces algorithmes d'indexation en terme de sources de données à manipuler en entrée, d'informations extraites en sortie, de contraintes d'exécution, de exploits, etc.

En ce qui concerne les contenus textuels, des approches d'indexation et la recherche de documents en texte, se sont suggérées comme des techniques de recherche d'informations classique en utilisant des caractéristiques hypertexte, comme des liens entre les pages et les balises HTML. Par contre cette méthode n'est pas optimale parce que le contenu des bases de données textuelles ne peut pas être abusé de façon satisfaisante pour les utilisateurs de plus en plus nombreux et diversifiés avec les méthodes traditionnelles de chaînes de caractères des systèmes bibliographiques ou de mots-clés.

Concernant l'indexation d'images, il existe des travaux dans ce domaine qui sont très nombreux. La recherche d'images sur une grande masse d'images nécessite des outils adaptés pour extraire efficacement le contenu en se basant sur des descripteurs significatifs et de trouver des images pertinentes. Les systèmes actuels permettent des recherches par des mots-clés sur les textes associés aux images.

Depuis quelques années, une nouvelle approche se développe, c'est l'indexation et la recherche des objets dans les images de documents par des mots visuels. Elle est plus optimale que les anciennes méthodes car un texte ne formule pas toujours exactement le contenu d'une image. Elle consiste à utiliser une image de requête ressemblant aux images que l'on recherche. Autrement dit en allant plus dans la pratique les systèmes d'indexation et de recherche possèdent des requêtes avec des mots-clés qui correspondent à des objets et des caractéristiques visuels.

Ce domaine d'étude entre dans les travaux de recherche qui se focalisent sur la reconnaissance d'objets et dans le domaine d'analyse et de segmentation de contenus visuels. Cette méthode a dopé les recherches des objets par le contenu des images et en particulier celui de l'apparence **visuelle**. On peut l'appliquer dans plusieurs domaines tel que la *robotique*, l'*astronomie*, l'*identification*, la *pharmacologie*, etc.

Dans la première partie de rapport, nous allons présenter le **cadre de notre travail** dans lequel nous présentons l'organisme d'accueil et l'objectif du TER. Dans la deuxième partie **l'Etat de l'Art** nous présentons les notions de ce thème de recherche et le processus de l'indexation des objets dans les images des documents. Ensuite, nous consacrons la troisième partie pour évaluer les résultats de notre travail en utilisant plusieurs exemples de bases d'images. Et finalement nous concluons ce rapport par une conclusion générale et les perspectives d'amélioration de notre travail.

Chapitre 1

Cadre de TER

1 Introduction

Il est difficile de parler d'un projet avant d'avoir fait une analyse du travail à réaliser. C'est la raison pour laquelle nous allons présenter dans cette partie le cadre de notre travail ainsi que les objectifs que nous voulons atteindre.

2 Contexte

Dépuis quelque années, avec l'explosion des informations sur internet et le développement à une grande échelle de la photographie numérique, les bases d'images contenant plus de dizaines de milliers d'images à cause de domaines d'activité professionnelle (voyage, éducation, tourisme, etc ...).

Pour manipuler ces informations contenues dans les bases d'images, un système d'indexation et de recherche d'images est mis en place. C'est pour cela qu'on peut dire que le domaine d'indexation d'images est très actif depuis une dizaine d'années. Dans ce contexte, nous proposons d'étudier un modèle efficace et générique de description d'algorithmes d'indexation. Assurément, il existe plusieurs modèles différents qui dépendent de chaque algorithme choisi. Pour cela, nous avons choisi une méthode flexible de détection d'objets dans les images de documents répondant à des besoins ou préférences d'utilisateurs et satisfaisant un contexte d'exécution donné.

Cet modèle se base sur l'approche mots visuels, tel que chaque image requête ou candidat est représentée par un vecteur des clusters ou ensembles des mots visuels (Ex : Figure 1.1). Pour avoir cet ensemble, on doit tout d'abord créer le vocabulaire de ces mots ou un dictionnaire des contenus visuels. Cet vocabulaire se construit à l'aide d'un ensemble des descriptions de base des images par un détecteur, un descripteur et un algorithme de regroupement. Un fichier des étiquettes (ou fichier

inverse) sera créée à la fin de chaque opération de regroupement pour indiquer à un candidat les clusters correspondants. Après la création de ce fichier, une fonction de correspondance se déclenche pour calculer la similarité entre la requête et tous les candidats. À la fin de ce processus les résultats seront triés selon l'ordre de la priorité des images les plus similaires.

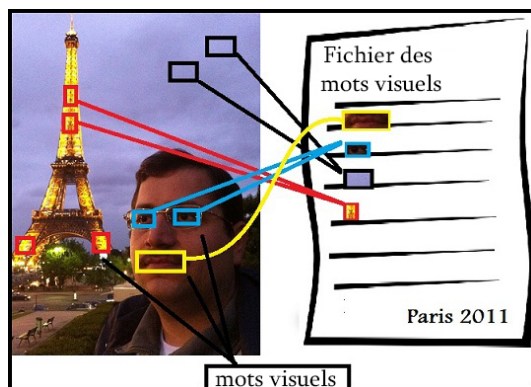


FIGURE 1.1 – Exemple des mots visuels

3 Contexte de travail

Cette étude a été effectuée dans l'équipe "Document et Apprentissage (DocApp)" de laboratoire "LITIS" de Rouen dont un axe de recherche concerne le domaine d'indexation des images. Le but est d'étudier le processus d'indexation dans les documents multimedia par leurs contenus en se basant sur le principe des mots visuels et d'évaluer les résultats attendus.

Dans le cadre de garder le maximum de la réactivité et de performance de notre travail nous avons utilisé les méthodes agiles pour adapter à un symténique de travail robuste. Notre projet a démarré par la détermination des besoins de la TER à travers les réunions que nous avons faites avec notre encadreur. Ensuite, nous avons validé la conception de l'application et les scénarios de l'utilisation de l'application. La réalisation de notre application est faite selon plusieurs versions, chaque version est validée par notre encadreur qui décrit les erreurs et les modifications nécessaires. La première version par exemple permet d'afficher l'image de test et la liste des images dans la base.

4 Objectif

L'objectif de ce TER est d'analyser le processus d'indexation des images par contenu pour comprendre ses modes de fonctionnement. Dans un deuxième temps,

nous essayons de déterminer les performances de chaque étape afin de déterminer leurs champs d'application. Pour atteindre notre objectifs, nous avons réalisé les étapes suivantes :

- Recherche approfondi sur le domaine d'indexation des images par leurs contenus.
- Recherche approfondi sur la bibliotheque des interfaces graphiques (Guii) sous MATLAB.
- Utilisation de l'UML pour construire les scénarios de l'application a l'aide de la méthode de conception **MVC** (Modèle-Vue-Controleur).
- Création d'une interface graphique avec Matlab.
- Implémentation des algorithmes de détection d'objets dans les images.
- Tester des algorithmes avec differentes méthodes d'évaluation.

5 Organisation et plan

Ce rapport suit un schéma respectant dans les grandes lignes l'évolution chronologique de notre travail. Dans un premier temps, nous présenterons les notions de travail dans le thème d'indexation des images par leurs contenus. Dans le chapitre 2, nous exposerons la procédure d'évaluation de ce procesuus ainsi que ses performances. En conclusion, nous finissons ce rapport par une conclusion générale dans laquelle nous essayons de présenter les avantages et les inconvénients de notre algorithme ainsi que les perspectives d'amélioration.

Chapitre 2

Etat de l'art et contribution

1 Introduction

L'expansion de la capacité de stockage des données numériques et l'évolution de l'internet a fait émerger des nouveaux besoins de consultation numérique. La recherche de l'information en utilisant des données textuelles est presque parfaite avec des moteurs de recherche tels que Google, Yahoo... Ces méthodes se basent sur des combinaisons de mots clés pour indexer les données numériques. Récemment, certaines applications emploient des méthodes issues du domaine de traitements automatiques du langage pour analyser sémantiquement les contenus de ces documents et affiner les résultats de recherche.

Toutefois, ces méthodes ne sont pas adaptées à l'indexation des images des bases de données numériques. En effet, un texte n'exprime pas exactement le contenu d'une image. Ceci limite beaucoup leurs performances de ces systèmes sur des documents visuels. Pour remédier à cette difficulté, plusieurs applications sur la toile ne se contentent pas par l'utilisation des métadonnées textuelles. Dans la littérature, plusieurs approches ont recouru à des données images pour décrire les images, ces caractéristiques vont former des signatures uniques pour indexer les données de la base des documents numériques. Dans cette partie, nous allons exposer quelques notions liées à l'analyse multidimensionnelle des images qui se base sur des méthodes de reconnaissance syntaxiques et structurelles. [BYRN99] modélise une telle problème de recherche d'information en utilisant les notions $(d, q, f, RSV(d, q))$ suivantes :

- d : est le système candidats ou la base des images.
- q : est la requête qui est sélectionnée par l'utilisateur.
- f : est le formalisme de représentation dans lequel chaque cluster est représenté par un vecteur des informations.
- $RSV(d, q)$: est la mesure de similarité entre la requête et le système candidats.

D'autres modèles de recherche d'information ont été également proposés comme le modèle vectoriel, modèle booléen, modèle logique, etc. Dans la suite de ce chapitre, nous commençons par la présentation des notions clés qui ont été initiées dans la littérature et adoptées dans notre travail. Ensuite, nous présentons le modèle vectoriel que nous avons adopté pour réaliser notre procédure de recherche d'image.

2 Notion de base

2.1 Mots visuels

Un mot visuel est un groupe des description invariantes dans une image. Chaque mot sera représenté par un vecteur de description représentative. Selon [JDS10], on peut obtenir ces mots visuels en quantifiant la fréquence des indices de descripteurs locaux calculés sur les zones d'intérêt de l'image. Ces mots vont être utilisés par la suite afin de réaliser une approche d'indexation d'images.

2.2 Vocabulaire

Un vocabulaire est composé par un ensemble de mots visuels qui apparaissent dans l'ensemble des documents d'une base. Par analogie aux systèmes de recherche d'information textuels, ce mode de représentation d'image est employé pour réaliser un système de recherche d'images. Selon [JDS10], on peut calculer un vocabulaire visuel en utilisant deux étapes de traitements. Dans un premier temps, on procède à l'extraction des zones d'intérêt et à la description de ces zones d'intérêts en utilisant des descripteurs images. Dans un deuxième temps, on passe à la quantification des mots visuels en utilisant une méthode de clustering dont le nombre des clusters est connu *a priori*. Ce quantificateur vectoriel est usuellement dénommé *vocabulaire visuel* (cf. figure 2.1), et un élément de son dictionnaire est appelé mot visuel.

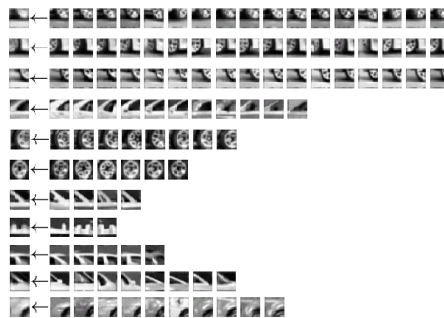


FIGURE 2.1 – Exemple : Vocabulaire ou dictionnaire des mots visuels.

3 Modèle vectoriel

Initier pour la première fois en 1979 par Salton [SB88] pour indexer des documents textuels, ce modèle est une représentation algébrique du contenu d'un document visant à rendre compte de la sémantique. En recherche d'information textuelle, l'ensemble de représentation des documents est un vocabulaire qui comprend dans son format simple les mots les plus significatifs du corpus (mots clés) comme les noms propres, les adjectifs, etc. Éventuellement, dans son format élaborées, il peut regrouper des expressions ou des entités sémantiques.

En recherche d'image, l'ensemble de représentation des images est aussi un vocabulaire composé par des mots visuels. La représentation de la requête (d) ainsi que les images candidates (q) est réalisée selon des vecteurs dont la dimension correspond à la taille du vocabulaire. Le $i^{\text{ème}}$ élément du vecteur v est employé pour référencer le poids du $i^{\text{ème}}$ élément dans le document. Etant donnée de la représentation vectorielle des images, la fonction de proximité la plus utilisée dans le modèle vectoriel est la similarité cosinus. Cette distance consiste au calcul de cosinus de l'angle entre deux vecteurs

$$\cos \alpha = \frac{v_d \cdot v_q}{\|v_d\| \|v_q\|} \quad (2.1)$$

La modélisation vectorielle du problème de mise en correspondance est généralement caractérisée par sa rapidité en temps de traitement et par sa faible consommation d'espace mémoire. Compte tenu de l'objectif de notre travail, ces caractéristiques répondent parfaitement à nos besoins ce qui nous incite à utiliser ce modèle.

4 Recherche des images par contenu

Le recherche d'image par le contenu est une technique qui se base sur les caractéristiques visuelles de l'image pour trouver des images similaires dans une base de données. Ces caractéristiques, encore appelées caractéristiques de bas-niveaux, sont généralement liées aux textures, aux couleurs et aux formes des éléments visuels de l'image. Selon [EN12] [J.M], l'architecture du système de recherche d'images est composée par les composantes suivantes :

- Requête : c'est l'image de test quand on va utilisé pour extraire les images similaires.
- Systeme condidats : c'est la base des images de documents.
- Formalisme interne des representation : selon [EN12] [J.M] le systeme condidats et la requête sont représenté par des vecteurs de fréquence.

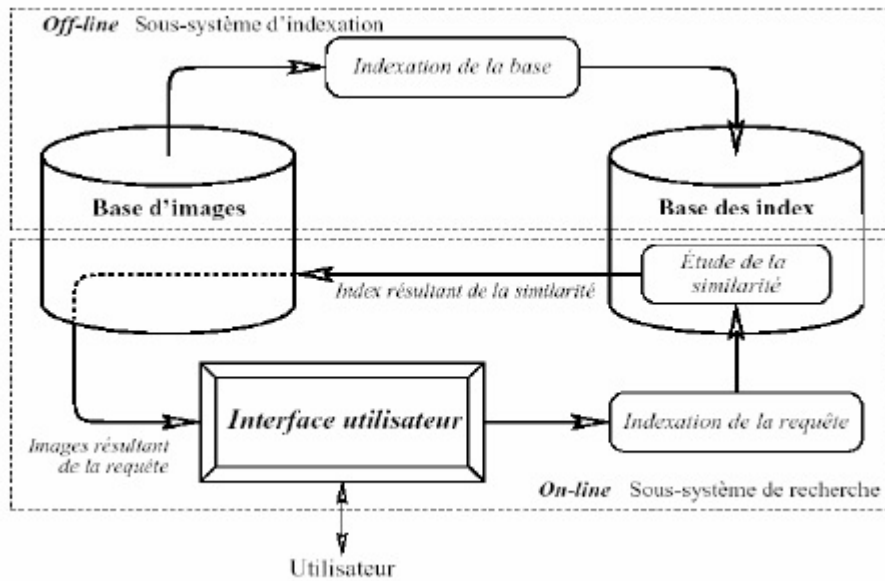


FIGURE 2.2 – L'architecture d'un système d'indexation et recherche d'images par le contenu

Le principe générale de ce système comporte deux étapes : étape en amont d'indexation d'image et étape de recherche d'image. La figure 2.2 présente cette architecture.

D'après cette figure, nous remarquons que dans l'étape d'indexation d'image, des caractéristiques sont automatiquement extraites à partir de l'image et stockées dans un vecteur numérique appelé descripteur visuel. Par analogie avec les systèmes d'indexation textuelle, on peut stocker ces caractéristiques dans des fichiers d'index et les récupérer rapidement et efficacement. La réalisation de cette approche est effectuée hors lignes pendant l'étape des prétraitements.

Dans l'étape de recherche, le système prend une ou des requêtes à l'utilisateur et lui donne le résultat correspond à une liste d'images ordonnées en fonction de la similarité entre leur descripteur visuel et celui de l'image requête en utilisant une mesure de distance. Cette étape est réalisée en ligne.

Dans notre TER, nous avons adopté cette architecture pour réaliser notre système d'indexation. Dans l'étape d'indexation, nous supposons qu'il y a k mots visuels (k -clusteur) dans le dictionnaire. Alors chaque document est représenté par la structure suivante :

$$V_{document} = (f_1, f_2, f_3, \dots, f_k)^T \quad (2.2)$$

Avec :

$$f_i = \frac{n_{OMD}}{n_d} \log \frac{N}{n_i} \quad (2.3)$$

Dont :

- n_{OMD} est le nombre d'occurrences du mot i dans le document d
- n_d est le nombre de mots dans le document d
- n_i est le nombre de documents consistant terme i
- N est le nombre de documents dans la base de donnée.

Une fois toutes les images de la base sont représentées par des vecteurs d'occurrences de mots visuels (Bag of Words, i.e. histogramme d'occurrences de mots visuels), nous procédons à la recherche des images. Dans cette étape, on va chercher de trouver une similarité entre la requête choisi et les images de la base. On dit que deux images sont similaires, si la distance entre ses vecteurs soit minimale. Dans ce contexte, plusieurs fonctions de distance ont été proposées. Les deux distances les plus populaires sont :

- **La distance euclidienne** représente la norme entre deux vecteurs dans une même espace. On peut l'utiliser pour calculer la similarité entre deux vecteurs.
- **La distance cosinus** permet de calculer la similarité entre deux vecteurs à n dimensions en déterminant l'angle entre eux.

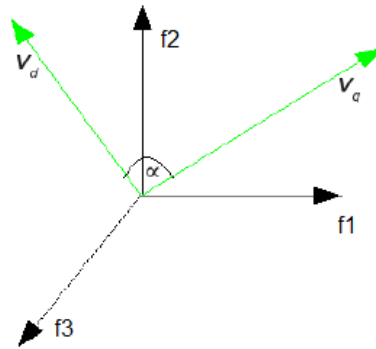


FIGURE 2.3 – Exemple : Distance cosinus

5 Caractérisation des régions de l'image

5.1 Les regions invariantes

Les regions invariantes (ou d'intérêt) sont les zones stable de l'image. Cela signifie que même en transforme l'image avec un chagement d'échelle, une rotation ou une translation, les description de ces régions restent toujours invariantes. La détection des régions d'intérêt d'une image numérique consiste donc à mettre en évidence des zones jugées « intéressantes » pour l'analyse. Dans la littérature, plusieurs catégories de détecteur ont été proposées pour sélectionner les régions de l'image :

- Détecteurs basés sur l'information de contours

- Détecteurs basés sur des points d'intérêt
- Détecteurs basés sur des informations de couleurs et de textures

Etant donnée de la performance et de la rapidité de leurs traitements, nous avons choisi d'étudier les méthodes de détection des points d'intérêt pour sélectionner les régions d'intérêt de l'image. Dans la littérature plusieurs méthodes ont été proposées dans ce contexte. Nous pouvons décomposer ces méthodes en deux familles :

- les méthodes basées sur l'analyse des coins (*Corner Detection*) : Les coins des formes de l'image peuvent être un bon moyen pour sélectionner les régions d'intérêt de l'image. En fait, ils sont les points de l'image où le contour (de dimension 1) change brutalement de direction. La méthode la plus répandue pour les détecter est probablement le **détecteur de Harris**[HS88a].
- les méthodes basées sur l'analyse des formes (*Blob Detection*) : ce sont des techniques de détection de points d'intérêt basée sur une analyse locale de l'image à l'ordre 2. Ce qui les différencie est l'opérateur de dérivation utilisé. Nous pouvons par exemple citer les méthodes basées sur l'analyse des DoG (Difference of Gaussians)[MH80], des LoG (Laplacian of Gaussian)[MS04] ou des DoH (Difference of Hessians)[MS04].

Dans notre TER nous avons utilisé le détecteur **SIFT** (Scale-invariant feature transform)[Lin12], que l'on peut traduire par **transformation de caractéristiques visuelles invariante à l'échelle**. Cette algorithme est très efficace lors de la détection des zones d'intérêt puisque il se base sur l'algorithme de harris [HS88b]. Un point d'intérêt (x,y,σ) est défini d'une part par ses coordonnées sur l'image (x et y) et d'autre part par son facteur d'échelle caractéristique (σ). L'utilisation de l'algorithme de Harris donne plusieurs candidats des points d'intérêts dans l'image. [Tay12] propose un algorithme d'amélioration de la précision des points d'intérêt en éliminant les points situés sur les arrêts et caractérisés par un faible contraste. L'algorithme SIFT finalise l'opération de détection des zones invariantes par l'assignation d'orientation. Cette étape consiste à attribuer à chaque point-clé une ou plusieurs orientations pour garantir l'invariance de ceux-ci à la rotation. Un histogramme des orientations sur le voisinage est réalisé avec 36 intervalle, couvrants chacun 10 degrés d'angle. La figure 2.4 présente la construction des l'histogramme des orientations en utilisant le descripteur de SIFT.

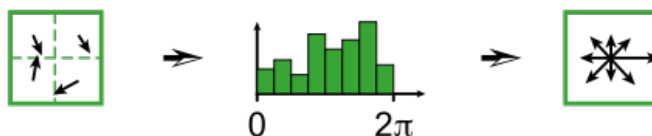


FIGURE 2.4 – Construction de l'histogramme des orientations.

5.2 Les descriptions invariantes

Les descriptions invariantes se sont des vecteurs qui présentent les facteurs des regions invariantes. Dans notre cas nous avons utilisé le descripteur **SIFT**. Cette descripteur permet de construire autour de chaque point clé une région de 16×16 pixels, subdivisée en 4×4 zones de 4×4 pixels chacune pour trouver la dépendance entre le point clé et ses voisinages. Le descripteur SIFT fourni pour chaque point clé un vecteur de 128 valeurs. La création du vecteur se base sur les facteurs de direction de la région.

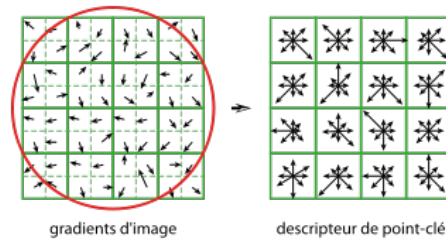


FIGURE 2.5 – Construction d'un descripteur SIFT.

6 Méthodologie

6.1 Processus de travail

L'indéterminisme du processus d'interprétation d'images et les grandes quantités de connaissances mise en œuvre nécessitent un processus de travail bien organisées. Ce processus de vision sous la conduite de stratégies est caractérisé par des étapes de traitement qui ont des rôles d'orchestrer les différentes traitement à mettre en œuvre pour atteindre un but et une meilleure resultat de recherche. Le processus de traitement se deroule en deux états : Etat en-ligne et Etat hors-ligne.

- **Etat hors-ligne** : Dans cette état le rôle principale est de construire le dictionnaire des mots visuels et un fichier des étiquettes pour chaque document dans la base.
- **Etat en-ligne** : Dans cet état on cherche a trouver des images similaires à la requête dans la base d'image.

Ces deux phases de traitement reposent sur le principe de fonctionnement des systèmes de recherche d'image par leur contenu. Dans un premier temps, des régions d'intérêt sont extraites par un détecteur Hessien affine et des descripteurs locaux SIFT sont calculés sur les zones identifiées. Dans un deuxième temps, ces descripteurs sont alors quantifiés par un quantificateur *k-moyennes*, dont le nombre des clusters

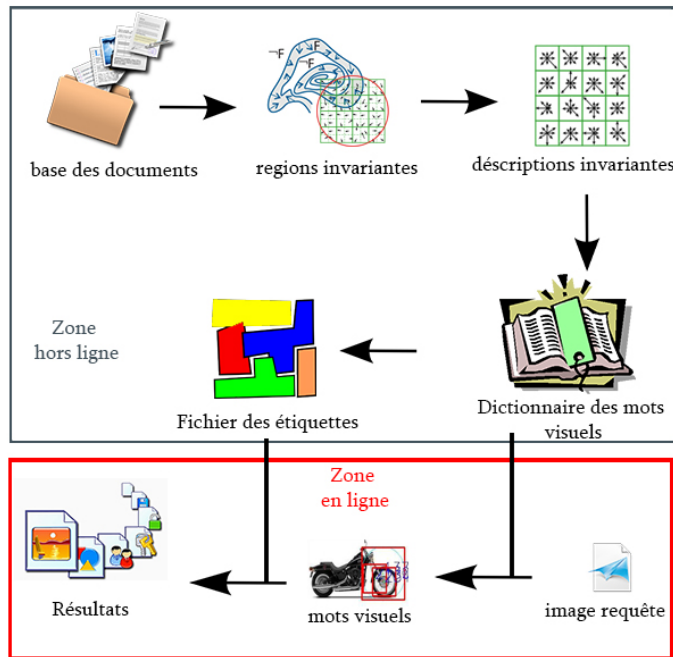


FIGURE 2.6 – Processus général de recherche des objets dans les images de documents

(mots visuels) est préalablement connu sur l'ensemble des images de la base. La quantification consiste donc à assigner chaque descripteur SIFT la classe de son k plus proches voisins euclidien, ce qui revient à l'assimiler au mot visuel le plus proche. La représentation vectorielle de l'image est calculée comme l'histogramme des fréquences d'apparition des mots visuels. Ses composantes sont pondérées et normalisées comme indiqués dans, ce qui, pour l'image i , produit un vecteur f_i à k composantes. Enfin un fichier des étiquettes est généré pour la base d'image dans laquelle on peut connaître pour chaque image les mots correspondantes et on termine par calculer le vecteur de fréquence de chaque document. Un exemple de fichier des étiquettes est présenté dans le tableau 2.1. D'après ce table, on peut trouver le premier mot visuel dans le premier document à la position 4. On peut également trouver le deuxième mot dans le premier document à la position 15.

Terme	ID ;position
1	1 ;4
2	1 ;15
5	121 ;4
3	30 ;5

TABLE 2.1 – Exemple : Fichier des étiquettes

Ce procédure de traitement est appliqué d'une part pour indexer les documents de la base et d'autre part pour décrire l'image requête. En conclusion, le mode de fonctionnement en ligne suit le schéma suivant :

1. Sélection de l'image requête
2. Détection des régions d'intérêt
3. Description des région d'intérêt
4. Formation des mots visuels
5. Création du fichier d'étiquette
6. Calcul des distances euclidiennes entre le vecteur de l'image requête et les vecteurs des de la base
7. Tries des images de la base selon la degré de ressemblance avec l'image requête

6.2 Réalisation

Pour réaliser cette approche nous avons développé une interface graphique adaptée à la procédure de recherche des images par contenu. Cette interface a été développée en utilisant le langage de programmation MATLAB et en utilisant une architecture MVC (modèle, vue, contrôleur). Le choix du langage de programmation a pris en compte la disponibilité de certains bibliothèques d'outil (*SIFT* et *k-moyennes*) nécessaires pour réaliser notre application.

L'application proposée dans le cadre de ce travail a été modélisée en utilisant le langage de modélisation graphique *UML*. Selon le schéma de cas d'utilisation de la figure 2.7, l'utilisateur de notre application a la possibilité de chercher des objets à partir l'interface d'ouverture de projet ou l'interface de création du projet. Il peut aussi suivre l'aide de l'application et enregistrer le projet crée.

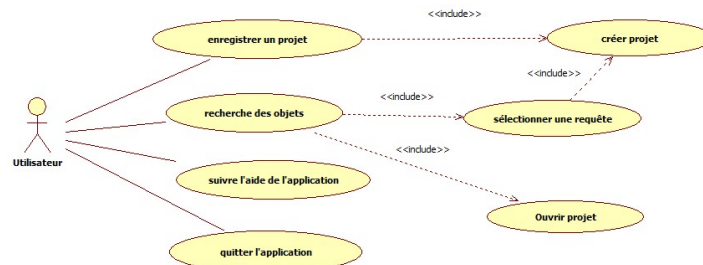


FIGURE 2.7 – Schéma de cas d'utilisation de notre application

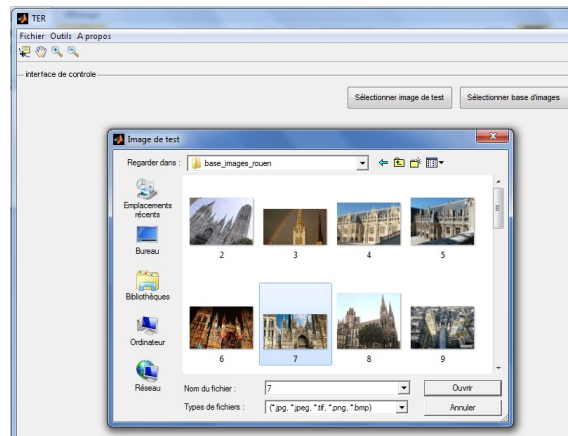
Les différentes interfaces présentées dans les figures 2.8a, 2.8b, 2.9, 2.10, 2.11, 2.12 et 2.14 sont développées selon le règle ergonomique de trois cliques qui facilite

l'utilisation de nos interfaces. Dans la suite de cette partie, on présente une démonstration de l'opération de recherche de l'image de la cathédrale de Rouen réalisée par notre application.

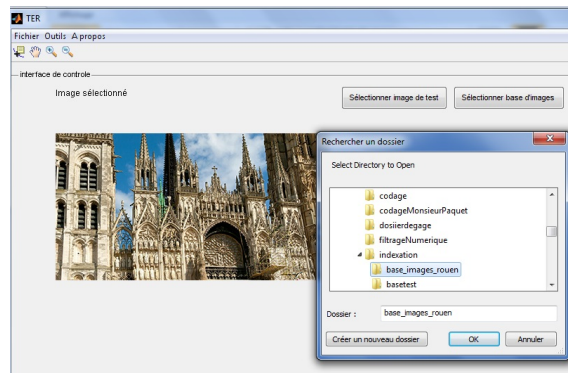
La figure 2.8 représente les interfaces de sélection de l'image de requête et de la base des images. Une fois ces deux opérations sont effectuées, on obtient l'interface de la figure 2.9. Selon cette figure, on constate que l'interface principale de notre application est décomposée en deux parties. La partie supérieure contient l'image à partir de laquelle nous allons sélectionner notre requête. La partie inférieure présente la liste des images de la base d'images dans laquelle on va rechercher nos images. En sélectionnant, les noms des images dans cette liste, nous obtenons l'affichage des images de cette base dans la partie inférieure droite de notre interface.

A l'aide de souris, nous procédons à la sélection de la partie d'image que nous voulons l'utilisée comme requête. Comme on a mentionné dans les sections précédentes, l'opération de recherche de l'image commence par la détection des régions d'intérêt dans la requête (cf. figure 2.10). Une fois cette opération effectuée, un algorithme de description des régions est appliqué (cf. figure 2.11). Ensuite, on procède à la formation des mots visuels en utilisant l'algorithme de regroupement *k-moyennes* (cf. figure 2.12). Enfin pour calculer la distance entre l'image requête et les images de la base, notre application propose à l'utilisateur deux distances qui sont : la distance euclidienne et la distance cosinus (cf. figure 2.13).

L'interface des résultats de recherche est présentée dans la figure 2.14. La liste des résultats de notre recherche est triées de manière croissante en utilisant les distances entre l'image requête et les images de notre base. Le slider dans cette interface permet de fixer le seuil de similarité que nous avons utilisé pour afficher nos images dans la liste des résultats. Pour des raisons d'évaluation, nous pouvons modifier la valeur de ce seuil. En sélectionnant, la première image dans la liste des résultats, nous obtenons une image de la cathédrale de Rouen. Ceci correspond bien à l'image requête ce qui signifie que notre procédure de recherche est assez performante. Pour s'assurer de ce constat, nous présentons dans le chapitre suivant une évaluation quantitative des résultats de notre approche.



(a)



(b)

FIGURE 2.8 – Sélection de l'image requête et de la base d'image de recherche

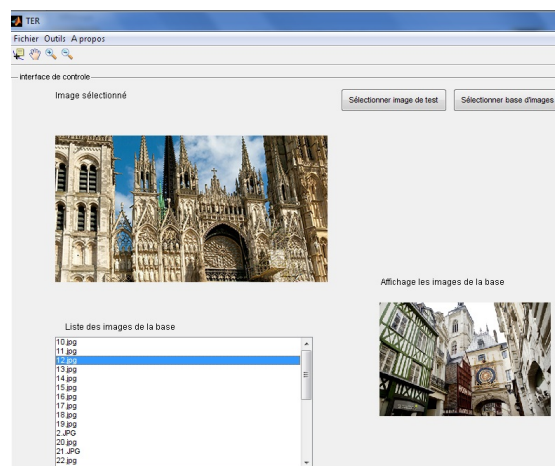


FIGURE 2.9 – l'interface principale de notre application

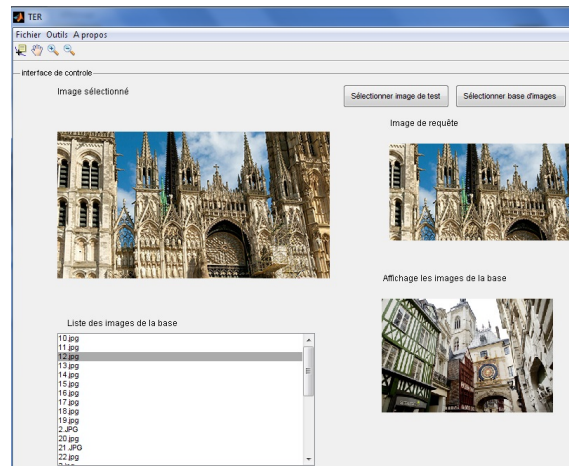


FIGURE 2.10 – Formation de l'image requête

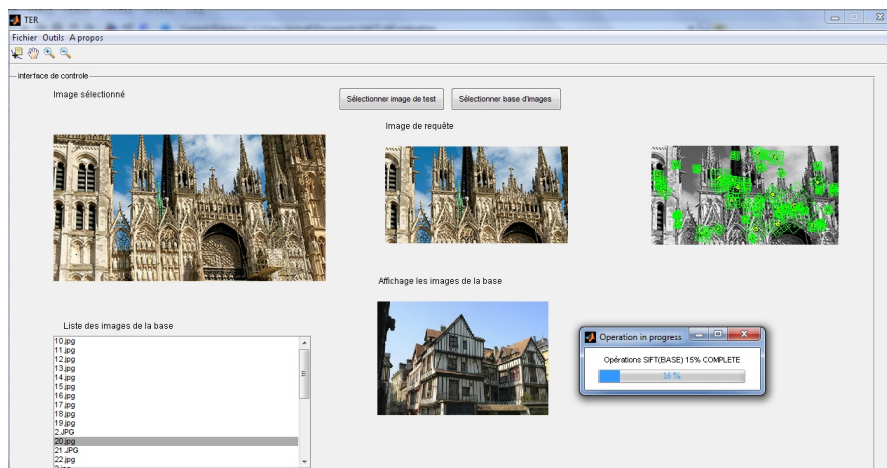


FIGURE 2.11 – Application de descripteur *SIFT*

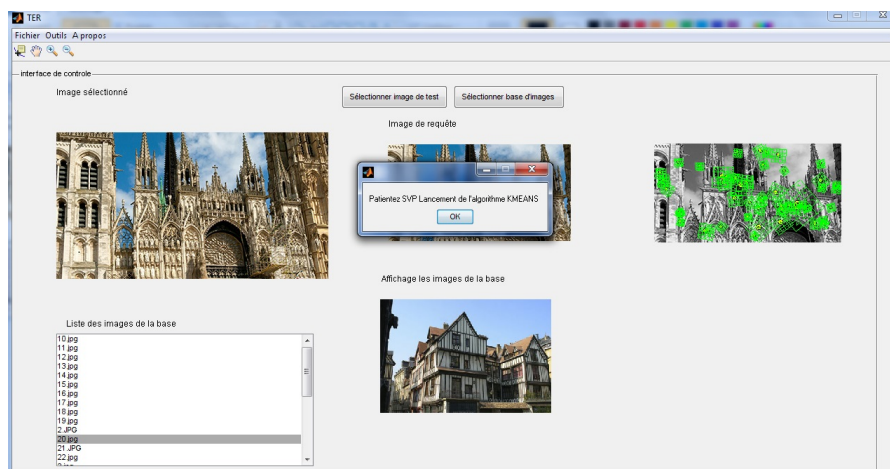


FIGURE 2.12 – Formation des mots visuels

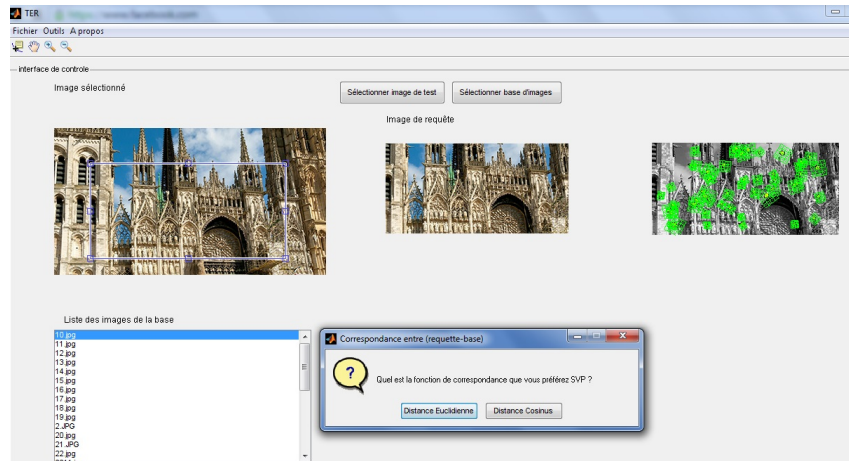


FIGURE 2.13 – Calcul de la distance entre l'image requête et l'image de la base

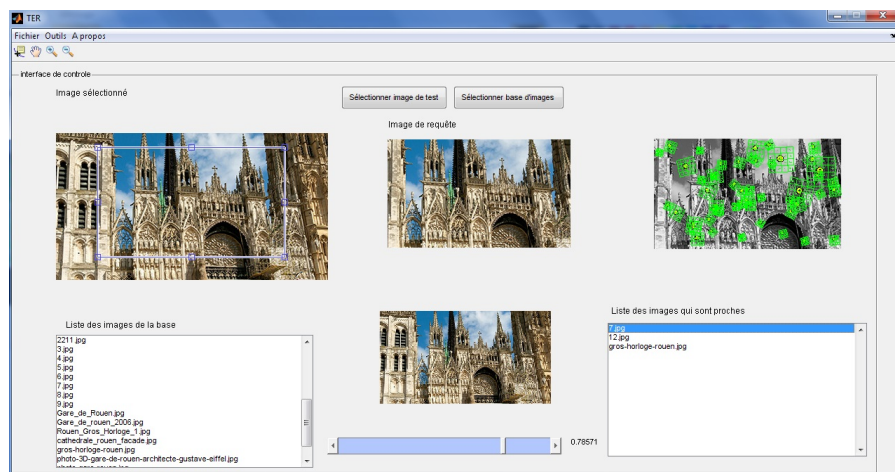


FIGURE 2.14 – Résultats de l'opération de recherche

Chapitre 3

Expérimentation

1 Introduction

Dans cette partie, nous avons réalisé un ensemble d'expérimentations qui nous a permis de déterminer les performances de notre algorithme de détection. Pour cela, nous avons employé une base des documents selon différentes catégorisation d'objets en fonction de leurs similarités ou de critères communs. Dans ce cas nous avons évalué les performances des algorithmes que nous avons implémenté en quantifiant les résultats obtenus par le principe de rappel - précision. Ce principe est une méthode classique d'évaluation des résultats qui suivent les équations suivantes :

$$Precision = \frac{\text{nombre_des_images_pertinentes_retrouvées}}{\text{total_des_images_retournées}} \quad (3.1)$$

$$Rappel = \frac{\text{nombre_des_images_pertinentes_retrouvées}}{\text{nombre_total_des_images_pertinentes}} \quad (3.2)$$

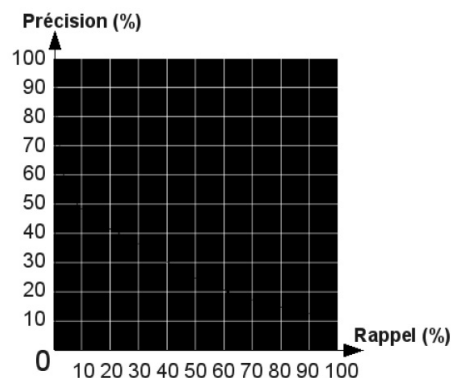


FIGURE 3.1 – Principe Rappel-Précision

Nous commençons donc cette partie par la présentation des caractéristiques d'images d'évaluation ainsi du protocole d'évaluation que nous avons adopté dans

cette étude. Ensuite, nous présentons les résultats qualitatifs et quantitatifs obtenus lors de la réalisation de ces expériences. Nous finissons cette partie par une conclusion.

2 Base de document de validation

Dans notre processus d'évaluation nous avons utilisé cinq document sélectionnées de manière aléatoire à partir différents bases de documents. La figure 3.2 présente les images d'évaluation que nous avons utilisées. D'après cette figure, nous constatons la présence des différentes catégorisation des objets sur ses exemples.

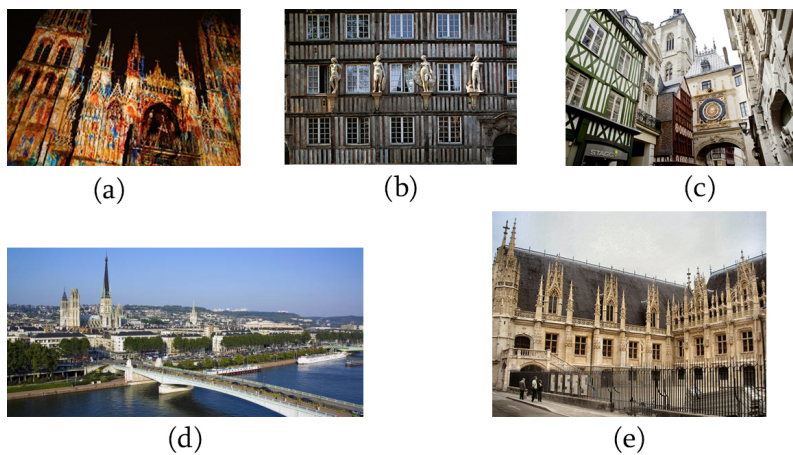


FIGURE 3.2 – List des images de test

Certaines images sont caractérisées par des propriétés physiques particulières. Par exemple, l'image 3.2.a représente des chevauchements entre les couleurs et les contours des régions. L'image 3.2.b représente présente des défauts de transparence.

D'autres images sont caractérisées par la présence simultanée des éléments graphiques qui sont caractérisés par des emplacements différentes. L'exemple 3.2 représente le cathedral qui est loin avec la présence de beaucoup de ciel ce qui permet une mauvais détection de cette image. L'exemple 3.2 illustre des objets qui sont proche de cathedral selon la caractéristique de la forme et le couleur.

A partir ces images de test on va selectionner différentes requêtes pour évaluer le processus que nous avons implimenter. Ensuite les requêtes sont représentée par un vecteur des mots visuels. Les valeurs de chaque mot visuel sont aussi calculés pour le algorithme de regroupement K-means. Grâce au fichier inverse et des mots on a une liste des mots qui construisent un dictionnaire des mots indexés. Chaque dictionnaire dans la liste est un candidat pour vérifierla similarité avec la requête.

Finalement il y a un seuil qui suit la loi de decision bayésiennes pour afficher les images qui sont les plus proches à la requette sélectionner.

3 Résultats d'évaluation

Comme nous avons évoqué dans le premier chapitre que l'objectif de ce TER est de détecter des objets d'une requête dans une base des documents caractérisés par des différentes illustrations des éléments graphiques. Dans notre procédure d'évaluation, l'information utile est représentée par les éléments graphiques des images c'est-à-dire la nature de construction des mots visuels. Rappelons que le but de mots visuels est de réduire l'information, c'est-à-dire réduire le nombre de descripteurs obtenus par le détecteur et le descripteur SIFT. La fonction **KmeansClustering-Descrip** prend en paramètres d'entrée les données suivantes :

- la liste des descripteurs, dans laquelle un descripteur est représenté comme exposé précédemment.
- K, le nombre de classes souhaitées.

Dans de cette étape de construction des mots visuels nous avons cherché le bon K-classe sur différentes jeu de données. La quantité de groupes utilisée pour tester est 50,100, 250, 500, 750 et 1000. A chaque changement de K on cherche le valeur du rappel-précision et le temps d'exécution. Voici les résultats de notre algorithme sur les images de test représentés en 3.2 :

	a	b	c	d	e
Rappel	$\frac{3}{7}$	$\frac{2}{10}$	$\frac{3}{5}$	$\frac{1}{15}$	$\frac{1}{7}$
Précision	$\frac{3}{7}$	$\frac{2}{3}$	$\frac{3}{9}$	$\frac{1}{7}$	$\frac{1}{4}$
Temps d'exécution	42 min	36 min	47 min	51 min	28 min

TABLE 3.1 – Tableau de correspondance Rappel-précision et le temps d'exécution avec $k = 50$

	a	b	c	d	e
Rappel	$\frac{5}{9}$	$\frac{3}{13}$	$\frac{3}{11}$	$\frac{2}{10}$	$\frac{3}{7}$
Précision	$\frac{5}{7}$	$\frac{3}{3}$	$\frac{3}{9}$	$\frac{2}{7}$	$\frac{3}{4}$
Temps d'exécution	1h29m	1h15m	1h37m	2h12m	1h05m

TABLE 3.2 – Tableau de correspondance Rappel-précision et le temps d'exécution avec $k = 250$

	a	b	c	d	e
Rappel	$\frac{7}{7}$	$\frac{3}{5}$	$\frac{7}{12}$	$\frac{4}{5}$	$\frac{4}{9}$
Précision	$\frac{7}{7}$	$\frac{3}{3}$	$\frac{7}{9}$	$\frac{4}{7}$	$\frac{4}{4}$
Temps d'exécution	5h 42 min	5h 39 min	5h 17 min	6h 13 min	5h 42 min

TABLE 3.3 – Tableau de correspondance Rappel-précision et le temps d'exécution avec $k = 500$

	a	b	c	d	e
Rappel	$\frac{7}{7}$	$\frac{3}{5}$	$\frac{9}{12}$	$\frac{5}{5}$	$\frac{4}{9}$
Précision	$\frac{7}{7}$	$\frac{3}{3}$	$\frac{9}{9}$	$\frac{5}{7}$	$\frac{4}{4}$
Temps d'exécution	10h 11 min	9h 45 min	10h 23min	15h 12min	10h 45min

TABLE 3.4 – Tableau de correspondance Rappel-précision et le temps d'exécution avec $k = 1000$

D'après les résultats du tableau 3.1 et tableau 3.2 on peut dire que le système manque des documents c'est-à-dire que il y a donc des **silences**. Le silence est les réponses pertinentes qui ne sont pas proposées par le système alors qu'elles existent. Il y a l'existence de silence puisque l'indexation est insuffisante. Donc on peut dire que plus le silence est grand, plus le rappel est faible. Par contre lors de calcul de la précision il y a l'existence de bruit c'est-à-dire que il y a des réponses non-pertinentes proposées par un système à cause de manque du terme dans le dictionnaire. Donc plus le bruit est grand, plus la précision est faible.

Un autre point intéressant est que le résultat avec $k = 500$ et le résultat avec $k = 1000$ sont (presque) pareils. Cela veut dire que c'est suffisant avec 500 mots et que si on augmente la quantité de mots, le résultat ne change pas ou change un peu.

Chapitre 4

Conclusion générale

Notre objectif principal concernait la détection des objets dans une base des documents. Dans un premier temps, avant d'entreprendre la mise en correspondance des images, nous avons considéré comme prioritaire de mettre au point des outils de caractérisation des images couleur. Pour cela, nous avons présenté un processus d'indexation permettant de caractériser différentes images en utilisant les mots visuels ainsi que les invariants différentiels couleur permet à la fois une détection et une description de ces mots visuels.

Dans un second temps, nous sommes intéressé au problème d'indexation d'images, cela revient à la mise en correspondance d'images. Dans ce cadre de TER, nous avons présenté une combinaison des algorithmes de mise en correspondance de détection et description des points d'intérêt et des algorithmes de mise en correspondance de regroupement de ces descripteurs pour arriver à la notion des mots visuels. La structure hiérarchique de ce processus d'indexation que nous avons suivie pour détecter des objets dans un système de candidats offre une robustesse de recherche pour arriver à des résultats parfaites. Les résultats d'évaluation ont montré que plus la valeur K-classes augmente on peut réaliser des résultats de performances plus meilleures.

D'autre part, nous avons remarqué aussi que le choix de K-classe donne une influence sur le système d'indexation. En effet, les meilleurs résultats sont obtenus sur $k = 560$ c'est-à-dire que c'est le K idéal.

Le principal inconvénient de cette approche est que le temps de création d'une base est trop long. Dans le cas où on ajoute une nouvelle image à la base, on doit recalculer tous les documents existants dans le système candidats. Cela nécessite alors un temps trop long de calcul. La recherche des objets dans les documents prend en considération tous les conditions d'images concernant l'existence des mots bruits lors de regroupement cela permet une mauvaise résultat de la recherche. Dans les perspectives de ce travail, nous pouvons essayer de développer une autre méthode de regroupement de

mots visuels comme la machine à vecteurs de support (SVM) qui permet de réduire le temps de calcul de l'algorithme de regroupement des descripteurs invariants et des mots visuels.

Bibliographie

- [BYRN99] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.
- [EN12] Sovann EN. The state of the art of content-based image retrieval system. 2012.
- [HS88a] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Manchester, UK, 1988.
- [HS88b] Chris Harris and Mike Stephens. A combined corner and edge detector. In *In Proc. of Fourth Alvey Vision Conference*, pages 147–151, 1988.
- [JDS10] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Représentation compacte des sacs de mots pour l’indexation d’images. In *RFIA 2010 - Reconnaissance des Formes et Intelligence Artificielle*, Caen, France, January 2010. Université de Caen Basse-Normandie.
- [J.M] J.Martinet. Un modèle vectoriel relationnel de recherche d’information adapté aux images. Technical report, Thèse préparée au sein de l’équipe MRIM du Laboratoire CLIPS-IMAG : Université Joseph Fourier Grenoble I, France 2004.
- [Lin12] Tony Lindeberg. Scale invariant feature transform. *Scholarpedia*, 7(5) :10491, 2012.
- [MH80] David Marr and Ellen Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167) :187–217, 1980.
- [MS04] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International journal of computer vision*, 60(1) :63–86, 2004.

- [SB88] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. In *INFORMATION PROCESSING AND MANAGEMENT*, pages 513–523, 1988.
- [Tay12] Brook Taylor. Taylor’s theorem. 1712.