ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

THÈSE PRÉSENTÉE À
L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

COMME EXIGENCE PARTIELLE
À L'OBTENTION DU
DOCTORAT EN GÉNIE
Ph.D.

PAR
YOUSSEF FATAICHA

RECHERCHE D'INFORMATION DANS LES IMAGES DE DOCUMENTS

MONTRÉAL, LE 27 DÉCEMBRE 2005

CETTE THÈSE A ÉTÉ ÉVALUÉE

PAR UN JURY COMPOSÉ DE :

Dr. Mohamed Cheriet, directeur de recherche
Département de Génie de la Production Automatisée, École de technologie supérieure

Dr. Jian Yun Nie, codirecteur
Département d'informatique et de recherche opérationnelle, Université de Montréal

Dr. Robert Sabourin, président du jury
Département de Génie de la Production Automatisée, École de technologie supérieure

Dr. Jean Meunier, examinateur externe
Departement d'informatique et recherche opérationnelle, Univeristé de Montréal

Dr. Ching Y. Suen, examinateur
Centre for Pattern Recognition and Machine Intelligence, Concordia University

ELLE A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 20 DÉCEMBRE 2005

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

INFORMATION RETRIEVAL ON DOCUMENT IMAGES

Youssef Fataicha

ABSTRACT

This thesis presents three approaches and an hybrid Information Retrieval (IR) system to locate and retrieve the informational areas in the document images.

Nowadays, scale space has been widely adopted as the most promising multi-scale image document analysis method. We propose a new approach using Separable Kernel Compact Support (SKCS) in order to analyse the composite document images. In the proposed method, SKCS transform is used to decompose an image into different scaled objects where the scale value is used for detecting progressively finer objects.

In a second work, we present a statistical and an expanded queries method of the retrieval process and the recognition errors obtained from electronic documents produced by scanning and OCR software. It uses an automatic approach to select additional words for possible erroneous terms for query expansion. The confused characters in erroneous words are located to create a collection of erroneous error-grams used to generate additional query terms. A vector space IR model is used to identify appropriate matching terms, and determine the degree of relevance of retrieved document images to the user's query. The proposed approach has been trained and tested on a thousand of different document images qualities and the performance of our method is evaluated experimentally by determining retrieval effectiveness with respect to recall and precision. The results obtained show its effectiveness and indicate an improvement as compared to standard methods such as vector space IR systems without expanded query and 3-gram overlapping.

In a third work, we present a unsupervised hierarchical method and an hybrid system for the retrieval of non textual areas. This method uses the hierarchical regions extracted at different scales. The first, a rough geometric structure on which admissible decompositions are defined with the preliminary segmented objects, is used to ensure a crude registration of the non textual areas. An accurate reconstruction is then performed for each detected area by a fusion process on the extracted objects. The features are then determined and we use a statistical model based on K-means and Multi-space Karhunen-Loève Analysers (MKL) to classify the extracted objects. The separate use of these methods did not give better results, an alternative using a mixture of k-means and MKL gives interesting results. The document image retrieval containing a given nontextual areas is largely improved by the use of an hybrid system integrating the text and the characteristics of nontext areas.

RECHERCHE D'INFORMATION DANS LES IMAGES DE DOCUMENTS

Youssef FATAICHA

RÉSUMÉ

Cette thèse présente trois modèles pour la recherche d'images de documents pertinentes à la requête d'un utilisateur.

Le premier modèle est basé sur l'approche multi-échelle pour localiser les régions informationnelles. Chaque région extraite est définie à partir de son contenu et de ses caractéristiques statistiques et géométriques. L'évaluation de l'interprétation sur des images de documents montre des résultats encourageants.

Le deuxième modèle traite de la recherche d'information reliée à la reconnaissance par OCR. Des erreurs-grammes sont collectées et utilisées pour l'expansion de la requête. Les expériences menées sur un millier d'images dégradées ont montré la fiabilité et la robustesse de notre approche.

Pour les zones graphiques, un troisième modèle vectoriel utilise des classifieurs et analyse les composantes principales des zones non textuelles pour une meilleure discrimination entre différents graphiques.

Finalement, un système hybride combine les informations textuelles et les caractéristiques des graphiques de l'image. Les requêtes portent sur des mots ou des images exemples. Les expériences ont montré une nette amélioration de la performance lorsque le processus de la recherche des images pertinentes tient compte des mots et des caractéristiques des images exemples.

RECHERCHE D'INFORMATION DANS LES IMAGES DE DOCUMENTS

Youssef Fataicha

SOMMAIRE

L'image de document est un objet intelligible qui véhicule de l'information et qui est défini en fonction de son contenu. Cette thèse présente trois modèles de repérage d'information et de recherche d'images pertinentes à la requête d'un utilisateur.

Le premier modèle de repérage des zones informationnelles est basé sur l'analyse multi échelle traduisant le contraste visuel des régions sombres par rapport au fond de l'image. Chaque région extraite est définie à partir de son contenu et ses caractéristiques statistiques et géométriques. L'algorithme de classification automatique est amélioré par l'application de règles de production déduites des formes des objets extraits. Une première évaluation de l'extraction du texte, des logos et des photographies sur les images de l'équipe Média Team de l'Université de Washington (UW-1) montre des résultats encourageants.

Le deuxième modèle est basé sur le texte obtenu par Reconnaissance Optique de Caractères (OCR). Des erreurs-grammes et des règles de production modélisant les erreurs de reconnaissance de l'OCR sont utilisées pour l'extension des mots de la requête. Le modèle vectoriel est alors appliqué pour modéliser le texte OCR des images de documents et la requête pour la recherche d'information (RI). Un apprentissage sur les images Média Team (UW-2) et des tests sur un millier d'images Web ont validé cette approche. Les résultats obtenus indiquent une nette amélioration comparés aux méthodes standards comme le modèle vectoriel sans l'expansion de la requête et la méthode de recouvrement 3-grams.

Pour les zones non textuelles, un troisième modèle vectoriel, basé sur les variations des paramètres de l'opérateur multi-échelle SKCS(Separable Kernel with Compact Support) et une combinaison de classifieurs et d'analyse de sous-espace en composantes principales MKL (Multi-espace Karhunen-Loeve) est appliqué sur une base d'apprentissage d'images de documents de Washington University et de pages Web. Les expériences ont montré une supériorité de l'interprétation et la puissance des vecteurs d'indexations déduits de la classification et représentant les zones non textuelles de l'image.

Finalement, un système hybride d'indexation combinant les modèles textuels et non-textuels a été introduit pour répondre à des requêtes plus complexes portant sur des parties de l'image de documents telles un texte, une illustration, un logo ou un graphe. Les expériences ont montré la puissance d'interrogation par des mots ou des images requêtes et ont permis d'aboutir à des résultats encourageants dans la recherche d'images pertinentes qui surpassent ceux obtenus par les méthodes traditionnelles comme révèle une évaluation des rappels vs. précision conduite sur des requêtes portant sur des images de documents.

REMERCIEMENTS

Je tiens à remercier mon directeur de thèse, M. Mohamed Cheriet, professeur à l'École de technologie supérieure, pour son suivi, pour son soutien et pour ses encouragements, ainsi que pour les conseils qu'il m'a prodigués tout au long de cette thèse.

Je remercie également mon codirecteur, M. Jian Yun NIE, professeur à l'Université de Montréal pour son suivi et pour l'aide précieuse qu'il m'a fourni au sein du RALI.

Mes remerciements vont aussi à M. Ching Y. Suen, professeur à l'Université de Concordia, pour m'avoir prodigué ses conseils et pour l'aide scientifique et la logistique qu'il m'a fournies au sein du CENPARMI, tout au long de la thèse.

Aussi, je tiens à remercier M. Jean Meunier, Professeur à l'Université de Montréal pour s'être intéressé à mon travail et avoir accepté de faire partie de mon jury et examiner mon travail.

Enfin, ma gratitude à M. Robert Sabourin, Professeur à l'École de technologie supérieure, qui a accepté de présider ce jury et organiser ma soutenance.

Tous mes remerciements vont également à tous les membres passés et présents du LIVIA, du RALI et du CENPARMI avec qui j'ai passé des moments agréables.

Enfin, je dédie ce travail à toute ma famille.

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES ALGORITHMES

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS ET SIGLES

CENPARMI	Center of Excellence on Pattern Recognition and Machine Intelligent.
UW-1	Base de 972 images à dominance textuelle de l'équipe Media-Team de l'université de Washington.
UW-2	Base de 512 images à dominance graphique de l'équipe Media-Team de l'université de Washington.
ISRI	Institut de Recherche en Sciences de l'Information de l'université de Nevada.
RALI	Recherche Appliquée en linguistique informatique.
IR	Recherche d'information.
SRI	Système de recherche d'information.
CBIR	Recherche d'Images Basée sur le Contenu.
OCR	Optical Character Recognition.
LoG	Opérateur Laplacien de la Gaussienne.
SKCS	Opérateur à noyau gaussien et à support compact.
LoSKCS	Opérateur Laplacien de SKCS.
σ	La variance de l'opérateur gaussien.
γ	l'étroitesse de l'opérateur gaussien.
2D	Bidimensionnel.
CAH	Classification Automatique Hiérarchique.
K-moyennes	Algorithme de classification automatique K-moyennes.
k	Nombre de classes de K-moyennes.
PCA	Analyse en Composantes Principales.
KL	Karhunen-Loeve.

MKL	Algorithme de classification multi-espace utilisant KL.
s	Nombre d'espace ou de classes de l'algorithme MKL.
$\Re$	Ensemble des nombres réels.
C	Matrice de covariance.
V_i	i^{me} élément du vecteur V.
ϕ	Matrice des vecteurs propres correspondants aux caractéristiques les plus signficatifs.
λ^i	i^{me} valeur propre.
P(s)	Probabilité a priori.
P(alb)	Vraisemblance.
$\overline{X}$	Vecteur moyen.
$X(t)$	Variable aléatoire de paramètre réel t.

CHAPITRE 1

INTRODUCTION

1.1 Présentation et contexte

Les bibliothèques et les centres de documentation disposent de millions de livres et de journaux sous la forme papier qui ne peuvent être saisis manuellement et qui sont difficilement indexés. La solution consiste à numériser ces documents en mode image puis à extraire de l'information à partir de ces images de façon à retrouver un document à travers son contenu. Le contexte d'application de notre étude est la recherche d'information dans les images de documents.

Dans ce contexte d'application, une requête exprime le besoin en informations d'un utilisateur et on appelle document toute unité qui peut constituer une réponse à une requête d'utilisateur. Un système de recherche d'information (RI) doit permettre de retrouver les documents pertinents à une requête d'utilisateur, à partir d'une base volumineuse de documents plein texte. C'est sur cette notion de pertinence que le système doit juger si un document doit être donné à l'utilisateur comme réponse.

Par ailleurs, une image de document peut contenir du texte, mais aussi des illustrations, des graphes et des logos. La communauté scientifique traitant de l'imagerie est très active dans le domaine de la recherche d'images basée sur le contenu (CBIR). Toutefois, notre problématique de recherche se rapporte à des images de documents alors que le CBIR se rapporte habituellement à l'extraction d'information des images non-documents. Par ailleurs, l'image de document est bien plus qu'une image à deux dimensions, elle comporte des chaînes de caractères de différentes tailles ainsi que de nombreuses illustrations, logos et graphes qui composent sans contraintes et d'une manière aléatoire le contenu du document (des exemples de documents sont à la figure 1). L'existence d'objets non tex-

tuels nécessite de nouvelles méthodes d'indexation et de recherche à intégrer au processus global de recherche d'information sur les images de documents.

(a) Liste d'adresse (b) Images publicitaires (c) Article de journal

Figure 1 Exemples d'images de documents extraits de la base d'images UW-2 (base d'images de Washington University)

Les zones textuelles de l'image peuvent être reconnues par un OCR (Optical Character Recognition) qui n'est pas exempt d'erreurs. Plusieurs travaux ont montré que la recherche d'information sur du texte obtenu par OCR dépend de la qualité de l'image et de la taille du document. Les travaux menés ces dernières années focalisent sur la correction des erreurs à l'aide de dictionnaires. Le problème est que ces derniers ne peuvent contenir tous les types d'erreurs. La motivation principale pour cette partie de la thèse est de concevoir une approche utilisant les parties des mots fréquemment erronées (erreur-grams) pour augmenter les termes de la requête et améliorer la recherche dans une collection d'images de documents de différentes qualités.

Les OCR n'ont pour seule fonction que la reconnaissance des caractères et perdent toutes les informations relatives au contenu non textuel du document. Aussi, les méthodes traditionnelles de RI sont surtout destinées aux textes, et ne sont pas directement applicables à d'autres types de données. Les objets non-textes posent problème surtout que les approches classiques dans le domaine de l'imagerie et de la reconnaissance de formes uti-

lisent de l'information a priori pour localiser et identifier la nature du contenu. Ces méthodes ne peuvent s'adapter aux images de documents dont l'information est de nature composite et structurée à l'intérieur de cadres ou en un nombre variable de formes dont la hauteur et la largeur ne sont pas constantes.

1.2 Problématiques

Comme nous venons de le préciser dans le contexte d'application, nous avons présenté les grandes lignes de la problématique générale de RI dans des images de documents et de qualités variables. Il existe d'autres problématiques techniques que nous pouvons adresser en survolant l'ensemble des problèmes. Notre travail consiste à proposer des approches pour pallier aux limitations imposées par la qualité des images de documents, et obtenir les documents pertinents à la requête d'un utilisateur.

Voici donc d'autres questions auxquelles nous croyons nécessaire de répondre afin d'avoir un système général efficace et efficient, surtout quand la performance de recherche est fonction de la qualité de l'image :

- Repérage des régions informationnelles : quels sont les facteurs de dégradation qui peuvent entâcher les images de documents ? Comment atténuer le bruit et recouvrir toute l'information de l'image ? Quelle méthode de segmentation utiliser ? Comment valider les régions détectées ?
- Interprétation des zones informationnelles : quelles sont les caractéristiques à extraire ? Comment analyser et interpréter les formes obtenues en fonction du niveau de lissage du bruit ? Comment valider les interprétations ?
- Apport de l'OCR aux textes : comment convertir les parties textuelles de l'image du document en représentation électronique ? Quels sont les facteurs qui perturbent la conversion complète ? Comment traiter les erreurs de reconnaissance ? Quelle est l'influence de la qualité du document ? Quel est l'influence des erreurs de reconnais-

sance du texte sur les approches de recherche d'information ? Comment valider les résultats obtenus ?

- Repérage des régions non-textuelles : définir les types de zones à identifier ? Quelles propriétés visuelles et de formes sont à considérer ? Quelles sont les caractéristiques pertinentes ? Comment indexer les zones non-textes ? De quelle manière les valider ?
- Représentation hybride et RI : comment combiner l'information texte et non-texte ? Comment différencier les formes des zones présentes dans l'image ? Quels sont les termes et les vecteurs d'indexation adéquats pour la représentation de l'information ? Comment mesurer les correspondances ? Quelles sont les recommandations pour améliorer le système de recherche ?
- Interrogation et requêtes : comment définir et traiter les requêtes des utilisateurs ? Comment rechercher les réponses ? Quelles sont les mesures de pertinence ? Comment définir et comparer la performance des systèmes ?

Nous avons présenté les grandes lignes de la problématique générale de RI dans les images de documents de qualité variable. Les objectifs à atteindre et les principales contributions sont développés dans la prochaine section.

1.3 Objectifs généraux et contributions

Nos objectifs sont de contribuer au développement d'un système performant de la recherche d'information basé sur les images de documents de différentes qualités. Pour cela, nous commençons par l'analyse de l'image numérisée afin d'extraire les régions informationnelles pour les utiliser dans le processus de recherche d'images pertinentes aux besoins des utilisateurs. La recherche d'information traite aisément les corpus textuelles alors que les images de documents qui contiennent aussi des zones graphiques, nous mènent à la conception d'un système de recherche hybride combinant le texte et le non-texte.

Les algorithmes fréquemment utilisés pour la détection d'objets et la reconnaissance de formes sont basés sur des méthodes regroupant les pixels en fonction d'un critère de "ressemblance". Le principal inconvénient de ces approches est la nécessité de connaître à l'avance le nombre d'objets ou de classes ainsi que leur position approximative dans l'espace de représentation. Ces techniques présentant beaucoup de restrictions, demandent des connaissances a priori et manquent de souplesse. Notre travail est motivé par l'idée de Koenderink (Koenderink, 1984) qui préconise que chaque type d'objet apparaisse sous une forme particulière à une échelle donnée. Il serait donc intéressant d'observer l'effet de l'échelle sur les formes résultantes et d'exploiter ces observations pour la séparation des zones de texte des zones non-texte. L'approche multi-échelle offre une granularité arbitrairement ajustable et intègre le concept de vues à différents niveaux et les modalités de segmentation.

La recherche d'information textuelle suppose des données numériques. La reconnaissance optique de caractères, ou l'Optical Character Recognition (OCR), est la technique qui permet de transformer un texte imprimé en pixels en un fichier numérique, composé de caractères ASCII. L'avantage théorique est indéniable. Quand une page de texte n'est qu'une image fixe, non modifiable et non lisible directement par un ordinateur, elle est aussi notablement plus volumineuse en termes de pixels. Il est donc tout à fait utile de transformer cette série de points en un fichier texte, beaucoup plus léger en nombre d'octets, et surtout indexable par n'importe quel moteur de recherche. La reconnaissance des blocs textuels à l'aide d'un OCR génère des erreurs que les méthodes existantes tentent de corriger à l'aide de dictionnaires. Ces derniers ne peuvent couvrir tous les types d'erreurs. Nous proposons donc de collecter et de modéliser, à l'aide d'un système d'apprentissage, les erreurs produites par l'OCR pour les intégrer dans un processus d'expansion de requête afin d'améliorer la recherche d'images de documents. L'algorithme est entraîné sur un millier d'images provenant de l'équipe Media-team de Washington University et testé sur

plusieurs centaines d'images Web dégradées par de multiples photocopies ou par ajout de bruit et de flou.

Pour la recherche d'information non-texte, nous nous intéressons aux objets de types logos, photographies ou graphes. Une première phase (analyse) réalise une décomposition hiérarchique de l'image et extrait pour chacune des régions un vecteur de caractéristiques composé d'une vingtaine de descripteurs (coordonnées, surface, entropie, circularité, rectangularité, différents moments, étirement, etc.). Dans une deuxième phase (décision), nous présentons des méthodes de regroupement des régions à partir des zones obtenues par la segmentation. Les zones extraites sont regroupées en familles de formes proches. L'objectif est de définir une méthode d'organisation et de déterminer l'influence des caractéristiques extraites sur la localisation et l'identification des zones non-textuelles. Les algorithmes de classification automatique comme K-moyennes et MKL (Multi-espace Karhunen Loeve) sont utilisés pour regrouper les objets de l'image en classes. Ces dernières sont raffinées pour interpréter fidèlement les types d'informations présentes. En effet, un nombre réduit de caractéristiques et des regroupements de qualité améliorent la représentation et facilitent la recherche surtout lorsqu'il s'agit d'objets images.

La mise en oeuvre de méthodes d'indexation hybrides combinant le texte et le non-texte pour une recherche d'information utilisant des mesures de correspondances entre la requête, le texte et les regroupements pré-établis des objets non-textes permet d'améliorer l'enrichissement et la puissance des requêtes et la qualité des réponses. Pour cela, nous distinguons quatres contributions principales pour répondre à ces besoins :

- l'extraction et l'indexation à plusieurs niveaux des régions informationnelles de l'image du document (Fataicha et al., 2001, 2002)
- la conversion des zones textuelles par un OCR (Optical Character Recognition) et la prise en compte des erreurs de reconnaissance dans le processus de recherche. Il s'agit ici de générer des chaînes similaires à intégrer dans l'interrogation pour améliorer les réponses à des requêtes (Fataicha et al., 2003, 2005)

- l'opérateur de segmentation que nous adoptons possède plusieurs degrés de liberté et localise une multitude d'objets informationnels. L'utilisation d'un classificateur automatique comme processus d'initialisation et d'un mélange d'analyse en composantes principales et de regroupements en sous-espaces permettent une multitude de combinaisons pour l'analyse et l'interprétation du contenu de l'image (Papier en rédaction)
- l'utilisation combinée d'informations textuelles et des caractéristiques de textures et de formes ainsi que l'élaboration de nouvelles mesures de correspondance et de performance affectent l'indexation et la recherche d'information (Papier en rédaction).

1.4 Plan de la thèse

La présentation de cette thèse est faite en huit chapitres, en plus de l'introduction. Dans le chapitre 2, nous présentons une revue de littérature pour illustrer la contribution des travaux précédents à l'élaboration de notre thèse, ainsi que l'originalité de nos principales contributions. Le troisième chapitre détaille notre méthode d'extraction d'objets et de caractéristiques. Le quatrième chapitre traite de la reconnaissance des régions textuelles et de son intégration dans la recherche d'information. Le cinquième chapitre décrit une approche améliorant la localisation et l'interprétation des régions non-textes. Le sixième chapitre présente une méthode hybride et des concepts d'indexation et de correspondance ainsi que des mesures de performance pour représenter et rechercher des informations textuelles ou non. Le chapitre sept discute les expérimentations effectuées pour valider nos différentes contributions. La conclusion de ce travail dresse le bilan de ce qui est réalisé et des recommandations pour les futurs travaux.

La Figure 2 schématise les éléments et présente les relations entre les différents chapitres de ce rapport.

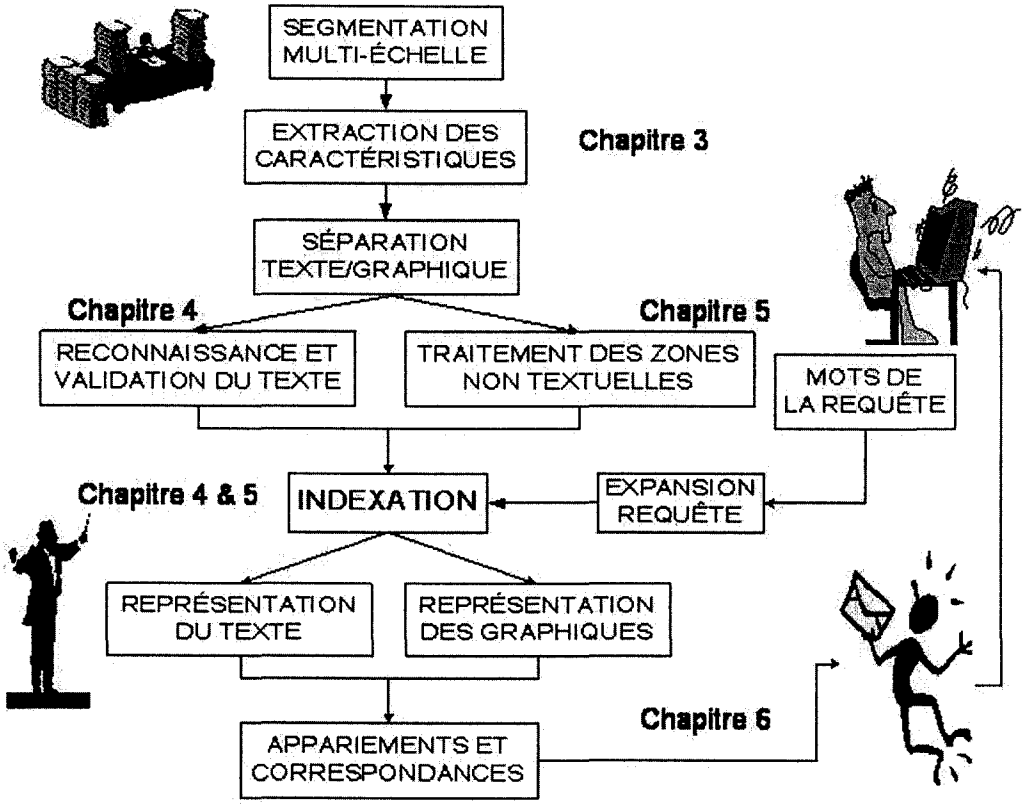


Figure 2 Schéma synoptique de notre système et éléments présentés dans cette thèse

CHAPITRE 2

ÉTAT DE L'ART

Nous présentons dans ce chapitre l'état de l'art sur la recherche d'information dans les images de documents antérieures à nos travaux. Ces travaux sont classés suivant la nature du contenu et nous nous sommes inspirés de plusieurs idées générales sous-jacentes pour développer nos méthodes de localisation, d'interprétation, d'organisation et de recherche d'information dans ce domaine.

La motivation initiale de cette thèse est la reconnaissance et la recherche rapide et fiable d'objets de l'image de document par des méthodes manipulant différents types de contenus. La première section décrit les travaux relatifs au repérage des zones informationnelles. La section 2.2 introduit les méthodes reliées à la reconnaissance du texte par OCR et sa représentation par un vecteur multidimensionnel pour la recherche d'information. La section 2.3 résume les approches existantes de reconnaissance d'objets non textuels et généralise le concept vectoriel utilisé pour représenter les mots présents dans l'image pour l'étendre aux principales caractéristiques des objets résultants de la segmentation. La section 2.4 discute des regroupements des régions extraites pour faciliter l'interprétation. La dernière section discute les méthodes de fusion texte/non-texte et l'apport de ces méthodes hybrides pour l'amélioration de la recherche d'information dans ce domaine.

2.1 Repérage fiable des zones informationnelles

La performance de l'interprétation de l'image du document dépend de la segmentation et de l'étiquettage de différentes régions informationnelles (textes, graphes, logos et illustrations). Dans la littérature, un survol du domaine de l'indexation et de la recherche d'images de documents est présenté dans (Doermann, 1998). On distingue trois approches de segmentation qui sont : méthodes descendantes (Wang et Srihari, 1989b; Tang et al., 1997; Ingold et Armangil, 1991), méthodes ascendantes (Tang et al., 1997; Jain et Yu, 1998) et

mixte (Seong-Whan et R., 2001; Cheriet, 1999; Kyong-Ho et al., 2000; Kerpedjiev, 1997). On trouve des applications de ces méthodes à des sommaires ou des articles scientifiques (Kyong-Ho et al., 2000) et aux journaux (Wang et Srihari, 1989b; Jain et Yu, 1998; Ingold et al., 2000).

L'approche descendante divise l'image en régions majeures subdivisées par la suite en sous régions. Les objets obtenus sont repérés par leurs positions, leurs tailles et leurs masses. Dans la littérature, il existe de nombreux exemples de méthodes de segmentation descendantes. La plus courante consiste à projeter récursivement l'image sur les axes des x et des y, puis analyser les histogrammes pour couper verticalement et horizontalement les différents blocs de l'image suivant des seuils évolutifs (Krishnamoorthy et al., 1993). Tang et al. (Tang et al., 1997) ont défini un langage de définition des formes et des règles décrivant les formes à reconnaître sur l'image.

L'approche ascendante extrait les composantes géométriques de base pour constituer des regroupements de plus en plus larges. Les techniques utilisées sont basées sur les méthodes d'analyse de la connexion avec les voisins pour détecter les groupes de pixels homogènes (Jain et Yu, 1998). On utilise des opérateurs comme le seuillage, la morphologie mathématique, la projection ou les opérateurs différentiels. Dans une première étape, les pixels de l'image de départ ou d'une image transformée sont regroupés en composantes connexes. Une deuxième étape consiste à extraire des caractéristiques sur ces composantes afin de pouvoir les regrouper en zones homogènes. Ces algorithmes localisent et fusionnent les composantes connexes à l'aide de seuils et de règles décrivant les formes à reconnaître. Le résultat de cette segmentation est une arborescence présentant la hiérarchie des différents blocs du document.

Quelque soit la méthode utilisée, quatre limites viendront toujours restreindre les performances :

- les problèmes d'atténuation du bruit : la question importante est de déterminer quel seuil utiliser pour lisser le bruit
- les problèmes de formes : toutes ces méthodes nécessitent la connaissance d'informations a priori sur les caractéristiques et les formes des objets. Comment définir la forme d'une zone donnée surtout que les propriétés visuelles et géométriques ainsi que la dimension varient en fonction de l'échelle et du lissage du bruit ?
- les problèmes de suivi des seuils multiples à mettre en place
- le choix des attributs pertinents : comment peut-on regrouper les attributs et les objets pour remédier au problème de discrimination entre les différents types de graphiques ?

Afin d'obtenir une segmentation robuste, il serait préférable de combiner les techniques et de choisir une stratégie basée sur une méthode mixte pour séparer le texte des graphiques.

Notre motivation est de focaliser sur les régions extraites à une échelle donnée et de suivre l'évolution des formes obtenues par l'application de différentes échelles. En effet, il s'agit de construire et d'appliquer une théorie qui permet d'effectuer une analyse de l'image à des échelles variables. Elle s'avère particulièrement utile lorsque l'on cherche par exemple à extraire des objets indépendamment de leurs tailles. Pour lisser le bruit, le principe est de noircir les plages blanches de petite taille pour obtenir des séquences continues de pixels noirs. La taille de ces plages pose problème et c'est pourquoi la représentation multi-échelle est une méthodologie largement étudiée dans la littérature (Lindeberg, 1994). L'idée principale consiste à décomposer l'image d'origine suivant une famille paramétrée de signaux réguliers, où les détails les plus fins sont éliminés au fur et à mesure que le paramètre d'échelle augmente. Notons que la plupart de ces approches peuvent être vues comme des cas particuliers de décomposition par ondelettes, qui consiste à projeter le signal sur une base de fonctions obtenues par translation et changement d'échelle d'une fonction mère appelée ondelette (Babaud et al., 1986). Dans ce travail, nous nous sommes inspirés des travaux de Cheriet (Cheriet, 1999) utilisant l'opérateur Gaussien pour filtrer le

bruit à différentes échelles et le Laplacien pour extraire les régions d'informations (parties concaves des signaux de l'image). Cette approche multi-échelle est utilisée essentiellement pour la segmentation mono-objet et il est intéressant de l'adapter et d'observer son apport dans l'extraction du contenu dans les images de documents dont les formes varient et pour lesquelles on ne dispose pas de connaissance a priori.

2.2 Méthodes de RI reliées à la reconnaissance par OCR

La recherche d'information a pour objet de répondre aux besoins des utilisateurs dans un corpus textuel donné. L'OCR transforme les images de documents en texte à l'aide de dictionnaires de correction. Dans tous les cas d'indexation automatique par OCR, il n'est guère raisonnable de s'attendre à un taux de reconnaissance supérieur à 80% (la réalité est même plus proche de 60% pour des qualités d'images moyennes). La recherche d'images de documents est difficile à cause des erreurs de reconnaissance découlant des opérations d'édition telles que la substitution, la suppression et l'insertion de caractères (Harding et al., 1997; Makinen et al., 2003; Ohta et al., 1998; Taghva et Stofsky, 2001). Beaucoup de ces études ont montré que ces trois types d'opérations -substitution, suppression et insertion de caractères - composent 80 à 90% d'erreurs.

Les systèmes proposés dans la littérature s'accordent tous pour analyser et remédier aux erreurs de reconnaissance dûes à l'OCR. Smeaton (Smeaton, 1998) emploie la forme approximative des mots dans un texte pour raffiner le processus de recherche ; mais cette approche ne peut désambiguiser les erreurs de reconnaissance. La plupart des approches à la correction des erreurs de reconnaissance se servent de dictionnaires. Les erreurs sont détectées en recherchant les mots du texte reconnus qui n'apparaissent pas dans le dictionnaire (Makinen et al., 1999; Strohmaier et al., 2003). Ce dernier ne peut couvrir tous les types d'erreurs. Durant les années 90, l'Institut de Recherche en Sciences de l'information (ISRI) de l'université du Nevada à Las Vegas a entrepris beaucoup d'expériences pour étudier la précision des OCRs et l'efficacité de la recherche à partir de textes géné-

rés par OCR (Taghva et al., 1996a; Taghva et Stofsky, 2001). Leurs recherches montrent les effets des erreurs de reconnaissance sur le classement des documents et que le retour de pertinence, processus automatique qui emploie le jugement de la pertinence de l'utilisateur pour reformuler automatiquement la requête, ne peut compenser les erreurs de reconnaissance de documents. Taghva et Stofsky (Taghva et Stofsky, 2001) ont développé un système OCRSpell, qui utilise un analyseur syntaxique, des dictionnaires spécifiques et un outil statistique de génération de mots pour remplacer les termes incorrects. OCRSpell a été utilisé dans (Taghva et al., 2002) pour traiter certaines erreurs et améliorer le rang des documents pertinents retournés lors de la recherche d'images de documents dégradées.

L'extension des mots de la requête à l'aide d'outils de génération d'erreurs tels que des "erreur-grams" pour améliorer la recherche d'informations dans un corpus de textes produits par l'OCR n'a pas été suffisamment étudiée. Les "erreurs-grams" sont les parties du mot (n-grams) susceptibles d'être erronées. L'extension automatique de requêtes est une voie de correction et d'évaluation de l'utilité et de l'impact des erreurs sur la recherche d'images de documents. Il existe plusieurs approches pour l'expansion de requête (Rocchio, 1971; Smeaton, 1998; Spink et Saracevic, 1997) et notre but est de générer des mots susceptibles d'être erronés et qui ont une relation avec les mots de la requête. Nous croyons que l'expansion de la requête améliore l'efficacité de la recherche particulièrement quand les mots incorrects sont des noms propres ou dans les documents courts en raison d'un manque de redondance. Croft et al. (Harding et al., 1997) ont augmenté les termes de la requête en utilisant les n-grams contenus dans les termes de la requête. Cette méthode a besoin de mesures restrictives pour éliminer les mots inutiles dans le processus d'expansion. Ohta et al. (Ohta et al., 1998) présentent une méthode probabiliste basée sur l'occurrence des caractères confus et les probabilités bi-grams des liens avec leurs prédécesseurs et successeurs pour approcher l'information confuse de caractères spécifiques. Les erreurs sont représentées dans des matrices de confusion dont l'utilisation augmente le taux de rappel mais diminue de manière significative la précision du système de recherche. Suen

(Suen, 1979) a montré que la fréquence de n-grams est fonction de la taille du document, du vocabulaire, de la position du mot et du corpus choisi. Les résultats obtenus par des tests sur ces méthodes ont montré une augmentation du rappel de 2 à 3 % et une baisse de la précision de 4 à 5 %. La plupart des techniques de recherche d'informations ne sont pas affectées si la précision est supérieure à 95% et sont applicables pour une reconnaissance supérieure à 80%. Pour des reconnaissances inférieures, le rappel et la précision baissent dangereusement et il est préférable de réduire le bruit de ces images à la source.

Dans des travaux récents (Fataicha et al., 2003), nous avons présenté une approche basée sur les erreur-grams et le modèle booléen. L'erreur-gram réfère à la partie du mot qui n'est pas correctement reconnue. Cette méthode a besoin d'être testée sur un large corpus d'images et validée sur des images dégradées. (Baird, 1999) survole les modèles de dégradation d'images de documents existant dans la littérature. Dans (Salton et McGill, 1983), le modèle vectoriel, contrairement au système booléen qui utilise les mots comme des variables binaires, considère l'importance des mots dans le document et leurs spécificités dans le corpus. Après avoir généré un ensemble de mots avec leurs poids à l'aide de règles de production, on a utilisé le système vectoriel de recherche d'information SMART (Salton, 1971), développé à l'université de Cornell, pour déterminer les documents pertinents et évaluer la performance de recherche. Finalement, le choix de l'OCR est basé sur les travaux récents de Souza et al. (Souza et al., 2003) qui ont étudié différents OCRs dans le but d'évaluer des critères pour le choix de filtres en utilisant des images de différentes qualités. Nous avons choisi l'OCR commercial FineReader qui a obtenu le meilleur taux de reconnaissance.

2.3 Caractéristiques des régions non textuelles

Si l'on suppose le problème de l'extraction des formes présentes dans une image et la reconnaissance du texte résolu, on peut s'intéresser à l'interprétation et à la recherche des zones non textuelles de l'image du document.

Compte tenu de l'utilisation postérieure que nous voulons donner à ce travail, il a été décidé de considérer 4 classes : textes, logos, illustrations, graphes et autres. En chaque région de l'image est calculée un vecteur de caractéristiques "signature" contenant des composantes associées aux descriptions de textures et de formes (Babaud et al., 1986). La définition d'une distance entre ces attributs permet de définir une notion de similarité entre deux régions (Venters et Cooper, 1999a). Nastar & al. (Nastar et al., 1998a; Winter et Nastar, 1999) utilisent différents types de signatures pour l'indexation et la recherche se fait sur la base de caractéristiques globales associées aux images et semblables à celles de la requête. Carson et al. (Carson et al., 1999) modélisent la distribution de ces caractéristiques comme une superposition de gaussiennes, ce qui permet de définir des classes et fournit une segmentation de l'image. Mokhtarian et al. (Mokhtarian et al., 1996a,b) s'intéressent à la mise en correspondance de formes. Les courbes subissent un lissage gaussien et sont normalisées. La mise en correspondance des courbes est réalisée par appariement entre ces représentations. Les travaux (Pentland et al., 1994; Sclaroff et Pentlab, 1995) caractérisent les formes par leurs modes de déformation et l'énergie nécessaire pour déformer une forme en une autre quantifie la qualité des appariements. L'algorithme présenté dans (Soffer et Samet, 1997, 1998) trouve des logos dans une base semblables à un logo donné. Les logos sont segmentés par composantes connexes et un vecteur de caractéristiques est extrait pour chacune des composantes. Ce vecteur est composé de quatre descripteurs globaux : moment d'inertie, circularité, rectangularité et excentricité. La recherche d'un logo semblable à un logo-requête s'appuie sur une distance (au sens des caractéristiques) entre les régions des deux logos. R.K. Srihari (Srihari, 1995) utilise la légende de l'image comme index supplémentaire à la recherche par le contenu.

La recherche d'images de documents nécessite la manipulation d'informations de différents types et formes. Le résultat de la segmentation est le repérage des zones informationnelles sous forme de blocs éparpillés sur la surface de l'image. Un travail de regroupement

des surfaces informationnelles et de recherche combinant texte et non-texte devient nécessaire pour l'interprétation et la représentation des régions informatives localisées.

2.4 Classification reliée aux régions non textuelles

Les principales faiblesses des méthodes précédentes sont la mise en jeu de beaucoup de paramètres et la représentation est statique, alors que dans les images de documents sont composites, plusieurs régions de différentes natures cohabitent dans la même image. Il est nécessaire de procéder à une interprétation de ces zones dans le but de différencier les différents types de contenu. La classification produit un regroupement de zones homogènes pour valider les formes, réduire la dimensionalité et augmenter l'efficacité de la recherche. La classification des zones générées par la segmentation d'images a été étudiée par de nombreux chercheurs. Pour améliorer le temps de recherche dans une grosse base d'images, on cherche à limiter la recherche à un certain "voisinage" par un regroupement intimement lié au stockage des vecteurs descriptifs de la base. Belaid et al (Belaid, 1994; Jain et Zhong, 1996) ont étudié un ensemble de caractéristiques et leur pouvoir discriminant. Bocchieri et Wilpon (Bocchieri et Wilpon, 1993) discutent de l'influence du nombre de caractéristiques et de la nécessité d'une sélection des caractéristiques. K-moyennes (Hartigan et Wong, 1979) est l'algorithme de classification automatique le plus populaire utilisée pour rassembler des objets en k regroupements fonction d'un critère de "ressemblance". La répartition doit minimiser l'indice de dispersion qui représente la distance par rapport aux vecteurs moyens représentant les groupes.

Le choix des caractéristiques à utiliser augmente la robustesse des algorithmes d'interprétation et de recherche d'images. Le traitement de l'image s'appuie sur des données, représentées par des mesures que l'on a accumulées sur des objets. Au sein du traitement de l'image, il existe une discipline, l'analyse statistique de l'image, qui a pour objet de décrire et d'analyser ces données. Dans la pratique, on peut être confronté à un problème de dimensionalité élevée. L'Analyse en Composantes Principales (ACP), ou la transformation

de Karhunen-Loeve (KL) permettent d'étudier les données dans un espace de dimension réduite. Une transformation KL revient à remplacer les attributs qui sont corrélés, par de nouvelles variables : les composantes principales. Ces nouvelles variables artificielles sont des combinaisons linéaires des variables initiales ou attributs, nous nous intéressons au pouvoir de réduction de la dimensionalité des données de départ. Les transformations ACP et KL sont basées sur la même opération, mais elles ont une différence qui réside dans une normalisation des données. Plus une coordonnée a un écart-type élevé, plus elle a de chiffres disparates et plus elle influera sur la comparaison.

Un choix s'impose alors : faut-il laisser intact la prépondérance de certaines coordonnées par rapport aux autres ? Ou faut-il procéder à une analyse normée en recentrant les coordonnées de l'objet par rapport à la moyenne et à la variance. L'ACP utilise cette normalisation alors que la KL laisse les variances des données intactes. Le but de la méthode est de visualiser dans un espace de dimension réduite par rapport à la dimension de départ, les proximités entre observations, et aussi les corrélations entre variables. Dans l'espace de projection, on cherche le sous-espace de dimension 1, donc un axe, qui passe au mieux à travers le nuage des données. Autrement dit, l'axe pour lequel la projection soit la moins déformée possible. On procède de même pour le sous-espace de dimension 2, puis de dimension 3, etc...

L'approche Muli-space Karhunen Loeve proposée initialement par (Cappelli et al., 2001, 1999) pourrait être qualifiée d'approche "linéaire par morceaux". Cette approche considère le modèle global non linéaire comme un mélange de modèles localement linéaires qui fournissent des groupes d'objets. Chacun de ces sous modèles peut alors être analysé par une analyse en composantes principales qui fournit une réduction de la dimensionalité pour chacun d'entre eux. Dans ce type d'approche, la méthode que nous proposons combine les résultats obtenus par le classifieur K-moyennes qui initialise les groupes et la méthode d'analyse MKL qui subdivise ces derniers tout en réduisant l'erreur de reconstruction et la dimensionalité au niveau des sous-groupes trouvés. Des travaux (Chalmond et Stéphane,

1999) ont montré l'intérêt de cette méthode par rapport à des approches classiques basées sur les KL ou l'ACP. Dans ce type d'approches, se pose notamment la question du choix du nombre de types d'objets à considérer, à laquelle il n'est pas évident de répondre, ainsi que celle du choix du nombre de caractéristiques à retenir pour la définition de l'espace de travail de dimension réduite. En plus, l'algorithme itératif est sensible à l'initialisation et risque de converger lentement.

2.5 Approche hybride pour la RI intégrant les régions non textuelles

Regrouper les images consiste à rassembler, au sein d'un même ensemble, des images ayant des caractéristiques similaires, la similarité étant calculée à l'aide d'une distance entre les descripteurs des images. Pour plus de détails, plusieurs approches de l'état de l'art sur le regroupement de données sont présentées dans (Halkidi et al., 2001; Jain et al., 1999). La performance d'un système de recherche d'images de documents dépend notamment de l'indexation qui doit permettre de retrouver toute l'information associée à ces images, du modèle de représentation qui doit être efficace et de la mesure de similarité qui doit permettre de retrouver les images de documents pertinentes.

On peut citer les logiciels Excalibur, ImageFinder, Imatch, Qbic d'IBM, Virage où le problème généralement traité est celui de la recherche d'images similaires à une image donnée. Pour des comparaisons entre ces différents logiciels, voir (Venters et Cooper, 1999b). Ces systèmes de recherche proposent d'utiliser des mots clés pour étendre automatiquement les requêtes qui portent sur des images. La correspondance mots-clés images pour la recherche d'informations est en effet un processus itératif : les utilisateurs ont des difficultés à exprimer précisément leurs besoins dès la première requête. La requête est donc modifiée itérativement [1]. Mais formuler de nouvelles requêtes n'est pas facile. Une technique [2] consiste à demander à l'utilisateur de sélectionner, parmi les documents retournés, les documents se rapprochant des plus pertinents. Certains mots de ces documents perti-

[1] Ce principe est connu sous le terme "query expansion"

[2] Connue sous le terme "relevance feedback"

nents sont alors utilisés par le système pour générer une nouvelle requête. Cette approche présente encore quelques inconvénients. Elle sollicite tout d'abord l'avis de l'utilisateur sur la pertinence des documents, ce qui demande un effort cognitif non négligeable et peut conduire à de mauvaises décisions. Les décisions peuvent être d'autant plus mauvaises que la majorité des systèmes suscitent des décisions binaires : un document est pertinent ou non. D'autres méthodes d'expansion (comme la pseudo-relevance feedback) réutilisent les documents se trouvant en tête de la liste de pertinence pour corriger la requête et augmenter la performance de la recherche.

Jing et al. (Jing et al., 2005) propose une architecture unifiée pour la recherche d'images basée sur des mots clés et des caractéristiques visuelles. La requête est formulée par des mots clés ou par une image exemple. Le jugement de pertinence des utilisateurs aux images retournées par le système est combiné aux caractéristiques visuelles des images pour représenter des concepts sémantiques à utiliser pour propager des mots-clés à d'autres images non étiquetées. Pour cela, un algorithme établi des relations entre des mots-clés et les caractéristiques visuelles des images utilisant deux modèles. Le premier utilise des probabilités de correspondance entre les mots-clés et les images de la base. Le second provient du modèle d'apprentissage utilisant les documents retournés comme réponses et jugés par l'utilisateur. Le concept d'interrogation se trouve amélioré par l'utilisation de requêtes basées sur des mots clés ou d'images exemples.

2.6 Conclusion

Le chapitre a résumé les approches de localisation et de reconnaissance en vue d'une interprétation d'objets et de recherche d'information. Quatre types d'approches ont été décrites : les techniques de segmentation multi-échelles (espaces d'échelles), la représentation des images par des vecteurs de mots pour le texte ou des descripteurs de caractéristiques pour les objets non textuels, d'autres techniques fondées sur des opérateurs de segmentation et des méthodes de classification.

Toutes les méthodes que nous avons présentées utilisent un a priori de forme ou de contenu pour reconnaître un document. Cet a priori peut être défini par l'utilisateur ou déduit du modèle de document. La solution envisagée laisse apparaître que notre problématique est de trouver un mécanisme hybride combinant texte et non texte pour faire coexister plusieurs représentations tout en évaluant la pertinence et la contribution de chacune dans l'élaboration des systèmes RI.

CHAPITRE 3

REPÉRAGE DES ZONES INFORMATIONNELLES

3.1 Introduction

Afin d'utiliser les images de documents numérisées, il est nécessaire de procéder à un traitement et une analyse approfondie de l'image de document pour mieux cerner son contenu. La segmentation est une phase essentielle du domaine de traitement et d'analyse d'images. Elle consiste à subdiviser une image en ses parties constituantes et à détecter les objets présents dans l'espace qui représente l'image. Il n'y a pas de solution générale au problème de la segmentation, mais plutôt un ensemble d'outils mathématiques combinés à l'algorithmique pour résoudre des problèmes spécifiques. Pour subdiviser l'image de document en ses parties constituantes, la littérature foisonne de descriptions de techniques de segmentation. Cependant, ces techniques s'appliquent pour la plupart à des documents dont la structure est relativement simple.

Les algorithmes fréquemment utilisés pour la détection d'objets et la reconnaissance de formes sont basés sur la méthode des K-moyennes qui regroupe les pixels en fonction d'un critère de "ressemblance". Le principal inconvénient de ces approches est la nécessité de connaître à l'avance le nombre d'objets ou de classes ainsi que leurs positions approximatives dans l'espace de représentation. Ils présentent beaucoup de restrictions, demandent des connaissances a priori et manquent de souplesse.

Le modèle d'image de document n'est pas prédéfini et peut être vu comme une opération d'agglomération des pixels, permettant de reconstituer par fusion les régions de l'image dans une approche ascendante. Dans les premières sections de ce chapitre, nous passons en revue les différentes méthodes de segmentation et d'analyse des images de documents. Puis nous aborderons notre approche pour repérer et interpréter les zones informationnelles en vue de les utiliser dans un processus de recherche d'information basé sur les

images de documents. Nous présentons dans ce chapitre les notions de base de la segmentation, les principes de recherche de caractéristiques et de formes et notre approche pour la localisation et l'interprétation des régions résultantes de la segmentation.

3.2 Méthodes de segmentation

La segmentation est l'opération qui consiste à transformer une image en primitive géométrique. Les primitives les plus employées sont les segments (contours) ou les surfaces (régions). Les méthodes de segmentation descendantes sont fondées sur le découpage de l'image en zones (ou) régions importantes. Ces zones sont à leur tour subdivisées par l'analyse des propriétés spécifiques en relation avec les objets qui composent l'image. Dans la littérature, il existe des opérateurs morphologiques, des techniques de projection et des techniques basées sur le contour ou la forme.

3.2.1 Opérateurs morphologiques

La morphologie mathématique a pour but d'analyser les formes présentes dans des images. Pour cela, on compare ces formes à un objet de référence appelé élément structurant. Les objets présents dans une image binaire sont considérés comme des ensembles de points auxquels on applique des opérations booléennes telles que l'union, l'intersection et complémentation. La morphologie mathématique nous a apporté de nouveaux concepts de transformation géométrique d'une image de notre système de vision ; notamment l'opération de dilatation et l'opération d'érosion ainsi que leurs combinaisons appelées ouverture (érosion puis dilatation) ou fermeture (dilatation puis érosion), de même ouverture suivie de fermeture.

L'érosion d'une image par un disque :

- coupe les objets au niveau de leurs étranglements,
- élimine les objets trop étroits ne contenant pas le disque,
- rétrécit les objets d'une taille correspondant au rayon du disque.

La dilatation d'une image par un disque :

- connecte les objets quand ils sont proches,
- comble les trous étroits présents dans les objets,
- élargit les objets d'une taille correspondant au rayon du disque.

L'ouverture d'une image par un disque :

- filtre les contours en éliminant les petites convexités, mais pas les concavités,
- sépare en plusieurs composantes connexes des particules présentant un étranglement assez long et étroit,
- élimine les particules trop étroites.

La fermeture d'une image par un disque :

- filtre les contours en éliminant les concavités éoites, mais pas les convexités,
- connecte les particules proches l'une de l'autre.

3.2.2 Techniques de projection

Le principe de cette technique est de projeter récursivement et alternativement l'image du document sur l'axe des x et l'axe des y pour détecter les différents blocs (Wang et Srihari, 1989a; Zen et Osawa, 1985). Le résultat de la projection est un histogramme représentant l'aspect de l'image. Les blocs sont déterminés en se basant sur des seuils évolutifs.

3.2.3 Techniques de transformée de Hough

La transformée de Hough permet de détecter les formes présentes dans l'image à l'aide de courbes représentables par une équation paramétrée. Pour détecter des lignes, on utilise l'équation d'une droite :

$$r = x \times cos(\theta) + y \times sin(\theta) \tag{3.1}$$

qui produit une courbe dans le système de coordonnées polaires (r, θ).
Cette méthode est insensible au bruit et aux inclinaisons, et nécessite beaucoup d'espaces mémoire.

3.2.4 Techniques de suivi basé contour ou texture

Un processus fiable d'extraction puis de suivi de l'information visuelle est en effet une des clés du succès, ou de l'échec, de la segmentation. Différentes techniques existent pour parvenir à cet objectif. Schématiquement, elles peuvent être divisées en deux grandes familles : celles basées sur le contour et celles basées sur la texture de l'objet. La première approche consiste essentiellement à suivre des primitives dans l'espace image ou comme des primitives géométriques (points, lignes, cercles, ...), le contour de l'objet, la projection des contours d'un objet, etc. La dernière utilise un critère de corrélation lié à l'information donnée par les niveaux de gris du motif de l'objet ou d'autres informations présentes dans ce motif (circularité, excentricité, surface, ...). Le suivi basé contour repose sur les forts gradients spatiaux délimitant le contour de l'objet ou certaines primitives géométriques présentes dans un motif (points, lignes, distances, splines, ...). En ce qui concerne le suivi dans l'espace de l'image (suivi 2D), cette approche consiste à décrire l'objet à suivre à l'aide de primitives géométriques comme des contours, des segments de droites, ou des ellipses, etc.

3.2.5 Opérateurs Laplacien

Les opérateurs Laplacien ne sont pas directionnels. En effet, si par hypothèse de départ, la fonction image est supposée continue, alors les propriétés de la dérivée seconde d'une fonction sont utilisées pour caractériser un contour par la présence d'un extremum local par le passage à zéro de la dérivée seconde.

Le Laplacien de la fonction Image est défini par :

$$\nabla^2 f = \frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial y^2} \tag{3.2}$$

Ainsi en utilisant également les propriétés de différentiation, l'opérateur Laplacien peut s'écrire sous la forme suivante :

$$\nabla^2 f(x,y) = f(x+1,y) + f(x-1,y) + f(x,y+1) + f(x,y-1) - 4f(x,y) \tag{3.3}$$

L'opérateur Laplacien est symétrique quelle que soit l'orientation choisie et il est fortement sensible aux bruits, ce qui nous conduit à essayer d'atténuer les bruits par l'emploi de filtre de lissage.

3.2.6 Technique de filtre de lissage

Le filtrage est une opération qui consiste à réduire le bruit contenu dans une image au moyen d'algorithme provenant de mathématiques par l'utilisation de méthode d'interpolation ou de la morphologie mathématique, où le pixel est considéré comme un individu cherchant son identité au sein de son voisinage. Trois types de filtrage ont été différenciés et s'appliquent à notre système de vision : le filtrage bidirectionnel, linéaire et le filtrage non linéaire.

La technique du filtrage bidirectionnel est fondée sur l'algorithme dit "RLSA" (Run Length Smoothing Algorithm). Cet algorithme est applicable à une image binaire. Le principe de cette technique est de noircir les petites plages blanches de longueur inférieure à un seuil pour obtenir des séquences continues de pixels noirs. Le lissage est appliqué d'abord horizontalement puis verticalement. L'image finale lissée est obtenue en appliquant un ET logique dans les deux images intermédiaires. Les principaux inconvénients de cette technique sont ses limites pour des petits blocs et le problème du choix des seuils à appliquer.

Le filtrage linéaire où le pixel est une combinaison linéaire de son entourage proche est un moyen pour lisser le bruit. Nous appliquons aussi à l'image une matrice utilisant la moyenne non pondérée des pixels voisins.

Le filtre médian est le plus classique des filtrages non linéaires et son efficacité est proportionnelle à la taille de la matrice et à sa forme. Ce filtre a souvent été appliqué lorsque l'image était réellement dégradée. Son efficacité est liée au nombre de pixels bruités. Par contre, ce procédé a un coût dans la restitution des informations contenues dans l'image. En effet, les contours des objets se trouvent déplacés après l'opération d'un filtre de lissage.

3.2.7 Méthodes ascendantes de segmentation

Ces méthodes consistent à repérer d'abord les objets élémentaires dans l'image, puis à construire les blocs informationnels à partir de ces objets. Cette technique consiste, dans un premier temps, à fusionner des pixels en blocs appelés aussi composantes connexes, puis à fusionner ces composantes connexes en régions homogènes.

3.3 Recherche dans les images de documents

Pour indexer des images, il faut extraire l'information pertinente, et celle-ci ne doit pas être trop volumineuse afin de permettre une recherche rapide. Deux approches sont possibles : l'utilisation de points d'intérêt ou l'utilisation de régions cohérentes.

3.3.1 Caractéristiques de points d'intérêt

Le problème est celui de la recherche d'images semblables à une image donnée. Les points d'intérêt sont les coins présents dans les images ; ils sont repérés à l'aide de détecteurs dédiés. Les caractéristiques utilisées ici prennent en compte des invariants différentiels, la géométrie de l'image, les formes des blocs etc. Un histogramme sur l'ensemble des points d'intérêt est calculé pour chaque caractéristique. Les histogrammes de l'image-requête et

ceux de la base sont ensuite comparés par corrélation. Enfin, les images de la base sont classées suivant une moyenne pondérée des corrélations. Une interaction est possible avec l'utilisateur par le choix des différents poids. Le point fort de cet algorithme réside dans le temps d'exécution, mais des problèmes subsistent à cause de la difficulté de localiser les points d'intérêts.

3.3.2 Caractéristiques des régions

L'indexation et la recherche se font selon différents types de signatures (histogrammes, transformées et ondelettes, caractéristiques discriminantes ...). Une boucle de pertinence tenant compte du jugement de l'utilisateur permet de définir une signature censée représenter l'image demandée par l'utilisateur pour la comparer ensuite aux signatures de la base. Les plus proches voisins de l'image requête sont recherchés dans la base. Une autre approche tient compte des images qui sont segmentées en régions cohérentes. Pour chaque région de chaque image, est calculé un vecteur de caractéristiques contenant des composantes associées à des descripteurs de texture et de position. La recherche d'images semblables à une requête se fait alors sur la base des signatures associées à ces régions.

3.3.3 Recherche de formes

Les images sont représentées par les formes et les textures qui appraîssent lors de la segmentation. Les images binaires sont segmentées par composantes connexes. Un vecteur caractéristiques est extrait pour chacune de ces composantes. Ce vecteur est composé de valeurs représentant des caractéristiques comme le moment d'inertie, la circularité, la rectangularité, l'excentricité, la surface etc. La recherche d'une image semblable à une image-requête s'appuie sur une distance (au sens des caractéristiques) entre les régions des deux images. L'algorithme utilisée approxime les formes de l'image par des droites et des arcs de cercle, puis les regroupe en familles de formes proches pour extraire des caractéristiques globales et dégager les familles de formes présentes.

3.4 Notre approche

Notre approche pour la détection des régions d'informations des images de documents composites a été guidée par les idées suivantes :

- l'identification de régions d'informations sans connaissance a priori sur la forme, le bruit, le nombre et la position des objets
- le lissage des dégradations qui représentent le bruit à différentes échelles
- la détection d'objets en fonction du type et du niveau de lissage
- l'utilisation combinée de différentes informations descriptives des régions de l'image
- la rapidité d'exécution.

Pour pallier aux problèmes de bruits et de contrastes présents dans l'image, notre travail focalise sur la détection de l'information contenue dans les images de documents composites par l'approche multi-échelle. Cette dernière lisse le bruit à différentes échelles et construit différentes formes relatives pour chaque région informationnelle. Notre système offre une granularité arbitrairement ajustable, permettant ainsi d'intégrer le concept de vues à différents niveaux et le choix entre plusieurs modalités de segmentation. Notre choix est motivé par l'idée de Koendreink (Koenderink, 1984) qui préconise que chaque type d'objet apparaîsse sous une forme particulière à une échelle donnée. Il serait donc intéressant d'observer l'effet de l'échelle sur les formes résultantes de la segmentation de l'image et d'exploiter ces observations pour la détection automatique de régions textuelles, de logos, de tableaux, d'illustrations, etc. Comme nous sommes concernés par la segmentation, notre représentation de l'image en espaces d'échelles est basée sur le lissage du bruit. Une image plus lisse nous permettra d'obtenir des surfaces peu influencées par le bruit, mais dont la localisation peut ne pas être très exacte, tandis qu'une image moins lisse, subira l'influence du bruit, mais la localisation des surfaces sera plus précise et validera les zones obtenues à des niveaux supérieurs. Pour cela, nous utilisons les

hypothèses développées dans (Cheriet, 1999) qui considèrent que les parties convexes de l'image convoluée correspondent aux objets informationnels et les parties concaves aux bruits et aux détails. Nous utilisons l'opérateur Gaussien pour filtrer le bruit à différentes échelles et le Laplacien pour extraire les régions d'informations (parties concaves des signaux de l'image).

Cette approche multi-échelle est utilisée essentiellement pour la segmentation mono-objet et il est intéressant de l'adapter et d'observer son apport dans l'extraction du contenu dans les images de documents composites dont les formes sont quelconques et sans connaissance a priori. Nous présentons dans ce qui suit l'approche que nous utilisons pour extraire de l'information à différentes échelles et identifier les objets d'intérêt. Nous commençons par la construction de l'espace d'échelles, suivi de l'opérateur Laplacien et les algorithmes utilisés.

3.4.1 Représentation en espaces d'échelles "scale-space"

Dans ce travail on s'intéresse particulièrement à la représentation multi-échelle où l'opérateur Gaussien est habituellement utilisé comme noyau pour les diverses propriétés dont il jouit. La gaussienne est établie comme le seul opérateur capable de changer d'échelle pour tenir compte de la spécificité du bruit et de ses propriétés statistiques. Cet opérateur utilise le paramètre d'échelle sigma (σ) d'une distribution gaussienne pour réduire le bruit à différents niveaux. Comme nous sommes concernés par le traitement d'images, on réutilise les définitions et la formulation de l'espace d'échelles en deux dimensions, développés par Cheriet (Cheriet, 1999) :

$g(x, y, \sigma)$ est le noyau Gaussien donné par la formule :

$$g(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3.4}$$

Dans la formule précédente que nous utilisons pour modéliser discrètement le bruit, g est l'opérateur gaussien, σ est l'écart type et (x, y) indique les coordonnées du pixel dans l'image. Le volume de l'objet croît et les contours sont lissés lorsque σ est grand.

3.4.2 Repérage des régions homogènes

On a vu dans la section précédente que la gaussienne permet la représentation en espace d'échelles des images plus ou moins lissées. Cependant, notre objectif est de détecter les objets perçus dans l'image en utilisant l'opérateur Laplacien de la Gaussienne (LoG).

L'opérateur (LoG) est défini par la formule mathématique :

$$LoG_\sigma = \nabla^2 g_\sigma(x, y) = (\frac{x^2 + y^2}{\sigma^4} - \frac{1}{\sigma^2})e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{3.5}$$

Cet opérateur, dont le profil est à la figure 3, est utilisé essentiellement pour la détection de formes et de contours dans une image. Dans le domaine fréquentiel, les spectres de Fourrier de l'opérateur LoG montrent que lorsque σ est grand, le spectre devient étroit et filtre de plus en plus les hautes fréquences. La figure 5 montre que pour σ petit, la bande passante du filtre s'élargit du côté des hautes fréquences, ce qui explique la suppression du bruit dans les régions détectées.

Après avoir appliqué le filtre LoG à l'image, un seuillage des valeurs positives permet de faire ressortir les différents objets de l'image. Les résultats obtenus valideront les régions détectées aux niveaux supérieurs (σ plus grand) pour ne garder que les objets informationnels. Il est clair que la performance d'utilisation et la taille du filtre dépendent de la valeur de σ.

3.4.3 Modélisation du système

Notre modèle est composée de deux étapes permettant de définir et de valider les informations véhiculées par l'image et les caractéristiques des régions. La phase ascendante

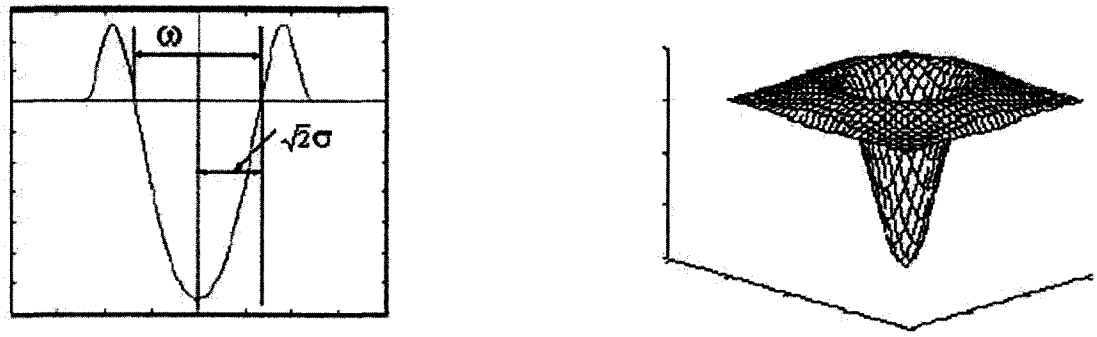


Figure 3 Profil 1D et 2D de *LoG*

convolue l'image avec le filtre, localise les régions d'informations et extrait les caractéristiques à différentes échelles. La technique descendante identifie et valide les objets obtenus. La figure 4 présente les traitements relatifs au modèle "espace d'échelles" pour l'analyse de document qui consiste en la construction du filtre, les interractions entre les méthodes ascendantes et descendantes ainsi que les variations de l'échelle. L'algorithme 1 détaille le processus du multi-échelle pour la localisation des régions informationnelles de l'image alors que l'algorithme 2 présente le processus de la validation descendante des objets extraits. Toute région extraite à grande échelle dont tous les pixels disparaissent ou il ne reste que de minusules taches à petite échelle est considéréé comme bruit et ne doit pas être considérée comme région informationnelle. Un exemple d'une image synthétique et de bonne qualité est à la figure 5. Les résultats montrent que la valeur de l'échelle σ influe beaucoup sur le nombre et la forme des taches obtenues.

Pour détecter les objets d'une image, on repère un pixel allumé et non traité auquel on affecte un numéro de séquence ou d'objet dans une table de correspondance. Le parcours de la surface de cet objet se fait à l'aide de l'algorithme des graphes connexes. Pour l'identification, on exploite les observations et les constatations déduites des expérimentations pour aider à la détection automatique des différents types de contenu.

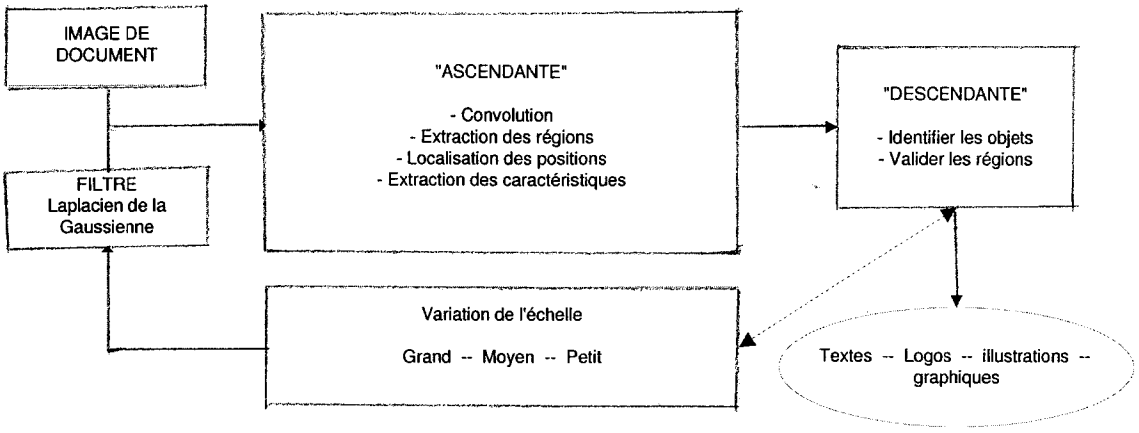


Figure 4 Modèle "espace d'échelles" pour l'analyse de document

(a) image originale

(b) Images segmentées aux échelles 30, 19 et 5

Figure 5 Exemple de segmentation multi-échelle.

Algorithme 1 : Extraction des régions informationnelles de l'image.

Entrée : Une page composite d'une image de document

Sortie : Localisation des régions informationnelles

pour $\sigma \leftarrow grand$ **à** $petit$ $(pas = 0.5)$ **faire**

- construire le filtre
- convoluer avec l'image en niveau de gris
- extraire les régions d'informations de l'image
- supprimer les objets dont tous les pixels sont devenus éteints à l'échelle courante
- extraire les caractéristiques des objets restants.

Algorithme 2 : Validation descendante des objets retenus.

Entrée : Liste des objets retenus

Sortie : Étiquetage des régions détectées

pour *chaque objet trouvé* **faire**

Reconstruire la forme

Identifier la nature de l'objet à partir des formes obtenus

Étiqueter les objets par les termes texte , logo, ligne ou illustration.

3.5 Interprétation des régions détectées

Les résultats obtenus par notre technique d'extraction est un ensemble de figures géométriques représentant les régions informationnelles de l'image. La qualité d'une identification sera fonction de l'interprétation des formes obtenues et de l'entropie de cette région. Pour identifier les régions de l'image comme texte, logo, trait ou image, nous avons observé les résultats de la segmentation des images de la base UW-2 pour avancer les critères suivants qui restent à justifier par rapport à bases plus représentatives :

- Identification de logos
  - le logo se trouve dans les coins de l'image
  - à grande échelle, les logos sont ovales et étirés en hauteur ou en largeur
  - la surface est calculée aux échelles supérieures à 20 et doit être proche de la valeur

$$\pi * (\frac{longueur}{2})^2 * (\frac{largueur}{2})^2 \tag{3.6}$$

  - la mesure de l'entropie est aux alentours de 5 d'après nos premières expériences.
- Identification de textes
  - la forme de l'objet est rectangulaire et étirée en largeur

$$surface \approx hauteur * largeur \tag{3.7}$$

  - un petit ratio longueur/largeur
  - l'entropie est petite relativement aux logos. Elle est aux alentours de 2 pour notre base d'images. La projection horizontale des pixels de l'image à petite échelle est périodique et composée de pics et de vallées
  - la densité d'une zone textuelle est plus élevée que celle contenant une équation mathématique.
- Identification de traits
  - les traits sont identifiés à petite échelle
  - le ratio largeur/hauteur est grand (petit) pour les traits horizontaux (verticaux)
  - la surface est égale à la largeur (longueur) pour les traits horizontaux (verticaux)
  - l'entropie est proche de 0.
- Identification de tables

  Le texte entre 2 traits horizontaux similaires est considéré comme table.
- Identification d'images

  Les objets restants sont considérés comme des régions constituant des illustrations.

Par souci de clarté, nous avons évité de reprendre les définitions de base conduisant à l'extraction et le calcul d'informations pertinentes comme par exemple :

- boîte englobante
- entropie
- etc....

La validation de ces heuristiques est faite dans la section 7.1.2 du chapitre "Expérimentation" où nous continuons d'appliquer des caractéristiques supplémentaires et chercher de nouvelles approches pour améliorer la qualité de notre interprétation et remédier aux problèmes de sur/sous-segmentation.

3.6 Conclusion

Nous avons présenté dans ce chapitre les différentes méthodes de segmentation pour identifier les régions informatives dans les images de documents. Le problème est celui de la recherche de l'information contenue dans une image de document. La littérature est suffisamment abondante et le sujet est certes beaucoup trop vaste mais incontournable. Un opérateur de détection de formes n'est efficace que lorsque le processus qui le précède (reconnaissance, localisation, interprétation ...) peut exploiter, de façon optimale, les données qui lui ont été transmises par le module de segmentation. Chaque opérateur peut convenir selon le type d'image à traiter. Tout de même, notre choix s'est porté pour l'opérateur Laplacien de la Gaussienne, il présente un avantage non moins considérable d'être multi-échelle pour le lissage de bruits. Il est intéressant de mettre en oeuvre ces opérateurs pour améliorer la détection des régions informationnelles. Enfin nous signalons que le choix de la variance de la Gaussienne est primordial, c'est sur elle que nous avons bâti le processus du multi-échelle et résolu le problème du lissage du bruit. L'opérateur Laplacien de la Gaussienne localise les régions cohérentes pour extraire les formes présentes dans l'image et la méthode de segmentation multi-échelle où le bruit est réduit progressivement pour valider les blocs obtenus à grande échelle, où le bruit est très présent, par les blocs extraits à petite échelle et où le bruit est totalement supprimé. Cette validation impose nécessairement l'utilisation d'une approche ascendante qui permettra de supprimer les régions créées par le phénomène de lissage de bruit introduit.

Finalement, une méthode d'interprétation utilisant des règles de production permet de mesurer l'efficacité de notre interprétation. Il s'agit de déduire la nature du contenu de la région extraite à partir des formes géométriques obtenues. Ainsi, les logos ne sont généralement pas texturés et ont une forme ellipsoidale. Ce mode d'interprétation reste insuffisant et le chapitre 5 tente de remédier au problème des zones graphiques qui restent collées aux textes. Notre approche se base sur l'analyse descendante des formats correspondants aux régions d'informations. Elle fonctionne dans le cas général mais ne prend pas en compte

tous les cas particuliers. Par ailleurs, il n'est pas raisonnable de chercher à traiter tous les cas. Nous en concluons qu'une bonne approche de détection et d'analyse doit être capable d'analyser et de synthétiser le comportement à adopter face à de nouvelles situations. Dans le futur, Il est essentiel d'éprouver notre prototype avec de gros volumes de données.

CHAPITRE 4

RECHERCHE D'INFORMATION RELIÉE À LA RECONNAISSANCE PAR OCR

4.1 Introduction

La recherche documentaire est le processus de sélectionner dans une collection les documents pertinents à la requête d'un utilisateur. Des laboratoires de recherche comme l'ISRI (Institut de Recherche en Sciences de l'Information de l'université du Nevada) ont mené des recherches sur les interactions entre OCR et recherche d'information depuis les années 80. Les résultats obtenus montrent que la fréquence des mots dans le document influe le rang des images de documents pertinents retournés et que le rappel décroit lorsque la recherche porte sur des documents courts ou que la recherche est basée sur l'appariement exacte de termes. Prenons comme exemple une société qui manipule diverses bases de documents et qui recherche un nom spécifique d'un client. Dans le cas où le client est enregistré comme "riemannian" dans une collection et reconnu comme "ricmanuinn" ou "licmamian" dans les autres, le procédé de recherche ne retourne que les images de documents pertinentes de la première collection et provoque une diminution du rappel.

Un dictionnaire ne peut modéliser les diffétrentes erreurs de l'OCR et le but de cette recherche est d'éviter son utilisation pour corriger le texte reconnu. Notre approche utilise les sous-chaînes mal reconnues par l'OCR dans le processus d'expansion de la requête de l'utilisateur. Un apprentissage utilisant les textes originaux fournis par les concepteurs de la base de documents et les textes des images reconnus par l'OCR localise les erreur-grams et déduit les règles d'inférence. Un poids dépendant de la fréquence est affecté à chaque erreur-grams et intégré dans un modèle vectoriel de recherche d'information dans lequel chaque document est représenté par un vecteur où chaque élément reflète l'importance d'un terme dans le document et dans la collection. Nous comparons les mots de la

requête de l'utilisateur à ceux reconnus sur les images de documents à travers des opérations vectorielles comme la distance ou le cosinus. Les images de documents pertinentes à la requête sont classées par ordre décroissant de la similitude et retournées à l'utilisateur.

Pour valider l'efficacité du système, nous considérons différentes formes de dégradation. Les images dégradées sont obtenues par l'application de modèle de dégradation (ajouter du bruit, lissage, flou etc.) ou par une dégradation physique résultant de multiples impressions ou photocopies. La qualité d'une image cause des problèmes pour les raisons suivantes :

- l'image est ancienne et souffre de dégradation physique
- la production de l'image par un appareil de mauvaise qualité qui génère des variations de la qualité, du contraste, et de la position
- l'image est une copie de mauvaise qualité dûe à la qualité de l'encre et à sa diffusion.

L'OCR commercial Finereader est appliqué aux différents types d'images de documets (articles, journaux, publicités, cartes de visites, manuels, formulaires etc.). Trois ensembles de données sont utilisés, le premier pour l'apprentissage et les deux autres pour les tests et la validation. Les erreur-grams et les règles de correction sont générées en premier sur la base d'apprentissage pour les incorporer au processus d'expansion de la requête. Des expériences sont menées pour constater l'amélioration de la performance de la recherche par rapport aux méthodes standards comme le modèle vectoriel sans expansion ou le recouvrement par 3-grams. Nous décrivons une approche améliorant l'efficacité de la recherche sur du texte OCR obtenues à partir des images de document de différentes qualités. La section 4.2 définit la reconnaissance optique de caractères (OCR). La section 4.3 expose la plateforme du processus de recherche basé sur les erreurs de l'OCR. La section 4.4 catégorise les différentes erreurs rencontrées et présente l'algorithme d'appariement pour construire les erreurs-grams. La section 4.5 détaille le processus de recherche et les mesures de la performance. La conclusion de ce chapitre est à la section 4.6.

4.2 Reconnaissance optique du texte de l'image de document

S'il est vrai que les nouvelles technologies permettent de prendre efficacement le relais du papier dans certains cas, celui-ci reste néanmoins un média courant bien ancré dans notre société par l'habitude, la simplicité d'utilisation et l'atmosphère qu'introduit son utilisation (livres, lettres, etc.). À tout cela, s'ajoutent tous les anciens documents qui ne sont toujours pas en version électronique et nécessitent donc d'être "informatisés". On réalise mieux maintenant tout l'avenir que l'OCR a devant elle et toute l'importance de la recherche entreprise sur le sujet.

4.2.1 Reconnaissance de mots

Deux approches s'opposent en reconnaissance des mots : globale et analytique. L'approche globale a une vision générale du mot, elle se base sur une description unique de l'image du mot, vue comme une entité indivisible. Disposant de beaucoup d'informations, elle absorbe plus facilement les variations au niveau de l'écriture. Cependant, cet aspect généraliste la limite à des vocabulaires distincts et réduits. En effet, la discrimination de mots proches est très difficile, et l'apprentissage des modèles nécessite une grande quantité d'échantillons qui est souvent difficile à réunir. Cette approche est souvent appliquée pour réduire la liste de mots candidats dans le contexte d'une reconnaissance à grands vocabulaires. Il est nécessaire d'utiliser dans ce cas des primitives très robustes pour ne pas manquer le mot réel parmi les mots candidats. Le mot reconnu est ensuite trouvé à l'aide de primitives de plus en plus précises (ou d'un classifieur de plus en plus fin).

L'approche analytique permet de s'affranchir de ces limites mais nécessite une interprétation locale basée sur un découpage (segmentation) du mot. La difficulté d'une telle approche peut être résumée par le dilemme suivant : "pour reconnaître les lettres, il faut segmenter le tracé et pour segmenter le tracé, il faut reconnaître les lettres". Il s'ensuit qu'un processus de reconnaissance selon cette approche doit nécessairement se concevoir comme un processus de relaxation alternant les phases de segmentation et d'identification

des segments. La solution communément adoptée consiste à segmenter le mot manuscrit en parties inférieures aux lettres appelés graphèmes et à retrouver les lettres puis le mot par combinaison de ces graphèmes. C'est une méthode de segmentation explicite qui s'oppose à la segmentation interne où la reconnaissance des lettres s'opère sur des hypothèses de segmentation variables (générées en fonction des observations courantes). Cette approche est la seule applicable dans le cas de grands vocabulaires. Elle peut s'adapter facilement à un changement de vocabulaire. Elle permet théoriquement une discrimination plus fine des mots car elle se base sur la reconnaissance des lettres qui la composent et il est possible de récupérer l'orthographe du mot reconnu. Son inconvénient principal demeure la nécessité de l'étape de segmentation avec les problèmes de sous- ou de sur-segmentation que cela implique.

Certaines des approches actuelles se proposent de tirer avantage des deux méthodes, réduisant la complexité de l'approche globale en l'appliquant sur des entités plus petites (lettres). L'approche analytique recherche la séquence de lettres contenues dans l'image à reconnaître. Certains modèles permettent de combiner ces deux niveaux en un seul et peuvent ainsi s'affranchir de la segmentation préalable.

4.2.2 Reconnaissance d'images de documents

La reconnaissance ou plutôt l'analyse d'images de documents concerne tout le processus de conversion de l'image. Ce processus est relatif à toutes les questions autour du langage écrit et sa transformation numérique : reconnaissance de caractères, formatage du texte, structuration du contenu et accès à l'information pour des applications d'indexation.

S'agissant souvent d'un processus de rétroconversion d'une structure existante, le processus de reconnaissance est guidé par un modèle explicite ou implicite de la classe étudiée. Le modèle décrit les éléments composant le document et leurs relations. Cette description peut être physique, relatant le format de mise en page, logique décrivant l'enchaînement des sous-structures, ou sémantique portant sur le sens affecté à certaines parties. L'OCR

est une étape importante dans la rétroconversion du document. Il encode évidemment les caractères et participe de manière très active à la reconnaissance de la structure.

Ce processus serait sans doute clair et "simple" s'il ne s'agissait que de documents textuels pour lesquels on dispose d'une structure éditoriale hiérarchique, le problème est beaucoup plus délicat pour d'autres classes de documents où l'information n'est pas très organisée et le contenu est hétérogène (comprenant un mélange d'imprimé, de manuscrit et de graphique), comme c'est le cas pour les formulaires, les documents postaux ou techniques, les magazines, etc. Dans ce cas, il n'existe pas de modèle direct pour décrire la composition du document et l'on a souvent recours à un mélange de techniques de traitement d'images et du langage pour extraire l'information. Le monde économique s'est emparé très tôt de cette technologie (le premier OCR date des années soixante). Il a finalisé les premiers travaux sur la reconnaissance optique des caractères et propose continuellement des OCR avec des performances de plus en plus élevées. Aujourd'hui, il existe au moins une vingtaine d'OCR dont les plus connus sont TextBridge (Xerox), FineReader (Abbyy), Omnipage (Caere) et Capture(Adobe).

4.2.3 Besoins actuels en analyse de documents

Les OCR sont très performants sur la recherche de la structure physique et vont aller en s'améliorant. Cependant, le principal intérêt d'un document ne se trouve pas dans sa forme physique mais dans son contenu logique. Ce contenu est encore hors de portée des OCR qui ne savent pas localiser les zones d'intérêt ni comprendre le contenu. les OCR ne sont pas adaptés à la recherche des zones d'intérêt, car ils sont incapables d'extraire l'information logique, à cela s'ajoute les limites sur la qualité des documents qui occasionne trop d'erreurs. Pour extraire l'information logique, il a été plus efficace de rechercher des points d'ancrage (par association d'images de composantes connexes) pour extraire une information utile. Dans la même idée, l'analyse de la bibliographie et des tables de matières a également montré les limites de l'OCR. En effet, il est incapable de décomposer

une citation en ses différents champs utiles (auteurs, titre, conférence, etc.). Des techniques additionnelles à la recherche d'information comme l'étiquetage, ont dû être mises en place pour retrouver ces différentes parties.

Les OCR n'atteindront jamais les 100% de bonne reconnaissance pour les images de documents dégradés. On observe par ailleurs, que différents engins d'OCR produisent des erreurs différentes. C'est cette idée qui a conduit les chercheurs ces dernières années à développer des techniques de combinaison de moteurs OCR afin d'améliorer le résultat global. L'objectif est de tirer parti des avantages de chaque OCR et d'écarter leur faiblesse. Dans les différentes études effectuées, il est montré que de l'ordre de 50% d'erreur est éliminée par la combinaison de plusieurs OCR ayant des taux de reconnaissance individuels de l'ordre de 97%. Cela étant, ce gain ne peut être atteint que dans la mesure où les erreurs proviennent des OCR et non de la qualité de l'image, et où les OCR sont de bonne qualité.

Les idées de base utilisées dans notre approche résident dans le traitement des erreurs de reconnaissance par OCR et leurs intégration dans le processus de recherche afin d'améliorer les performances da la recherche des images de documents.

4.3 Architecture de l'approche proposée

L'architecture du système est donnée à la figure 6. L'approche proposée quant à elle est décrite par les trois étapes suivantes :

- Première étape : l'équipe média-team qui a conçue la base d'images servant d'apprentissage fournit le texte originale des images que nous utilisons pour construire les erreur-grams et les règles de correction.

  Pour mesurer la concordance entre le texte original de l'image et le texte obtenu lors de sa reconnaissance par l'OCR, nous avons besoin d'une fonction de distance d'ap-

pariement. L'algorithme de programmation dynamique "distance d'édition" calcule la distance entre deux sous-chaines et détecte les segments qui différent entre les mots originaux (texte ASCII fourni avec la base) et les mots de l'OCR. Le résultat de l'appariement est le nombre de transformations rendant les deux mots identiques. On appelle erreur-gram les parties mal reconnues d'un mot donné. Un exemple est le mot "schultz" qui est reconnu comme "sehnltz", une erreur-gram est "chu" qui est reconnue par l'OCR comme "ehn". Nous appliquons la mesure de distance d'édition sur le texte OCR pour générer la collection d'erreur-grams et les textes relatifs aux trois ensembles d'images.

- Seconde étape : le système utilise les mots de la requête, les erreur-grams, le texte ASCII fourni par le concepteur et le texte OCR pour trouver les mots à rechercher, éliminer les mots inutiles, lemmatiser, identifier les termes appropriés et appliquer le modèle vectoriel pour l'indexation et la détermination des images de documents pertinents.
- Finalement, la mesure de la performance du système de recherche est comparée à différentes méthodes pour montrer les améliorations obtenues lors de la recherche des images pertinentes.

Étant donnée une image scannée, la méthode proposée consiste à :

- Localiser et extraire le texte de l'image
- Comparer le texte reconnu avec le texte original (fourni par le concepteur). Les erreurs sont modélisées dans un ensemble d'erreur-grams et de règles de correction
- Mesurer l'efficacité du système de recherche en utilisant le classement par ordre de pertinence, le rappel et la précision.

Les étapes présentées à la figure 6 sont détaillées dans les sections suivantes.

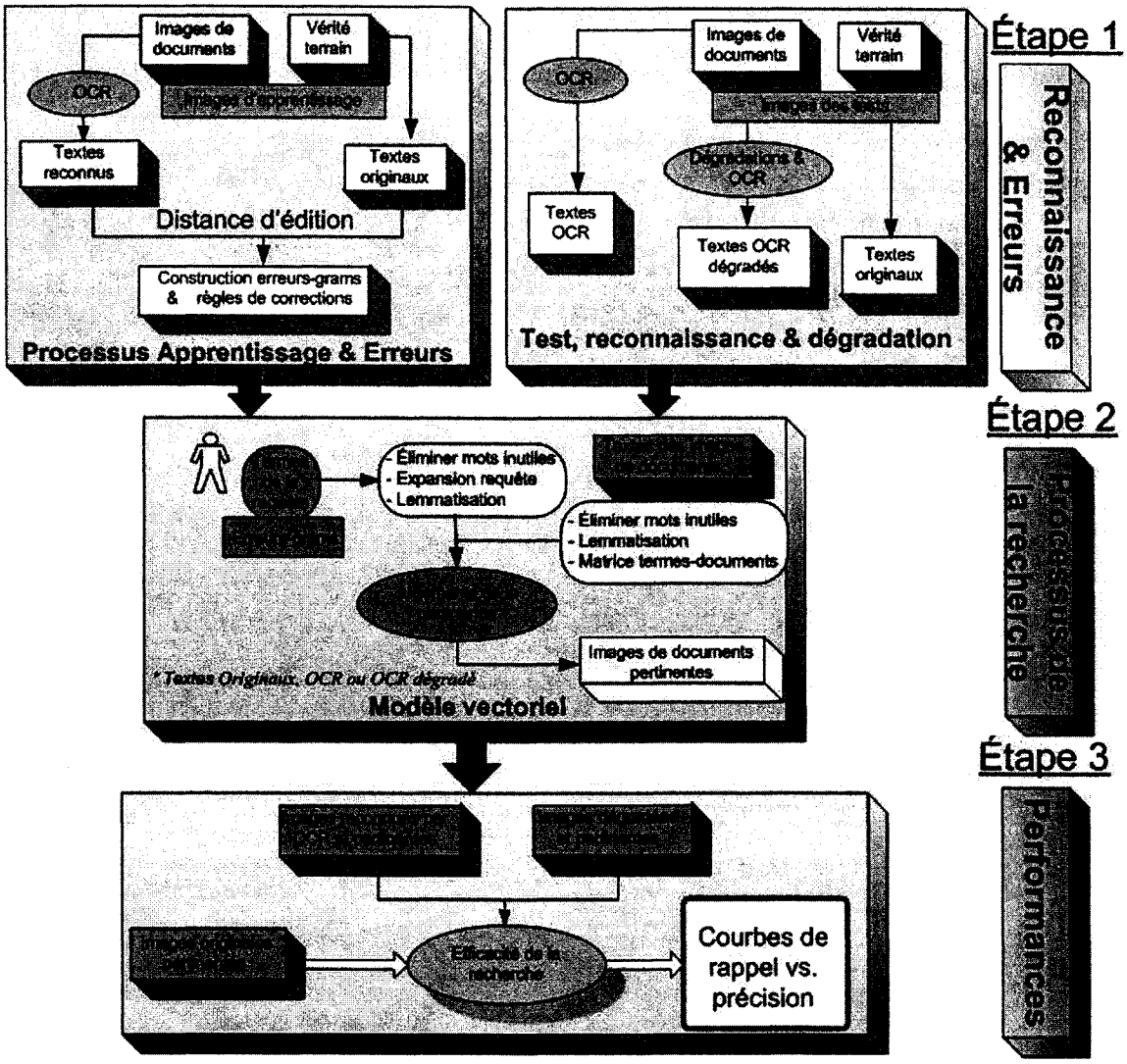


Figure 6 Recherche d'information reliée aux erreurs de l'OCR extrait de (Fataicha et al., 2005)

4.4 Appariements et erreurs de l'OCR

Un OCR commercial est utilisé pour reconnaitre les caractères contenus dans les zones textuelles de l'image du document. Notons que l'OCR produit des erreurs dont des exemples sont données au Tableau I.

Tableau I

Groupe d'erreurs et exemples

Groupe d'erreurs	mot correct	exemple d'erreur
Substitution	nuit	cuit
Suppression	Info	Nfo
Insertion	Kylie	Ikylie

Les différences entre les chaînes de caractères du texte original (le code ASCII fourni) et du texte reconnu par l'OCR permettent de localiser les sous-chaîne erronées et de calculer les probabilités correspondantes. Compter ces erreurs à la main est une opération très coûteuse, c'est pourquoi l'algorithme de programmation dynamique - distance d'édition - a été adoptée pour construire les erreurs-grams et les règles de production à intégrer dans le processus de recherche.

4.4.1 Algorithme de distance d'édition

L'algorithme de distance d'édition est basé sur la différence entre deux chaînes de caractères. Il utilise la programmation dynamique et effectue l'appariement de chaînes de caractères sans information a priori (Bunke et Csirik, 1975). La distance entre deux mots est le nombre d'opérations d'édition qui rendent les deux mots identiques.

Posons M_{ori} l'ensemble des mots originaux (fournis avec la base), M_{ocr} les mots retournés par l'OCR, et $s = e_1; e_2; ...; e_n$ la séquence des opérations d'édition transformant la chaîne x en y. Le coût $c(s)$ de cette séquence est donné par la formule c(s) = $\sum_{i=1}^{n} c(e_i)$, où $c(e_i)$ est le coût du i^{th} opération d'édition.

Étant donnée deux chaînes de caractères x et y et le coût de chaque opération d'édition pour transformer x en y, la distance entre x et y est définie par la formule :
d(x,y) = { min{c(s)} : s est la séquence d'opérations d'édition transformant x en y}

Les ensembles de mots sont ventilés entre les ensembles :

$M_{rec} = \{words \in M_{ori} \bigcap M_{ocr}\}$ pour les mots bien reconnus et

$M_{remo} = \{words \in M_{ori} - M_{rec}\}$

$M_{remr} = \{words \in M_{ocr} - M_{rec}\}$ pour les mots restants.

Nous adaptons cette mesure pour la construction d'erreur-grams et des règles de correction pour les images de documents. L'algorithme (Ukkonen, 1983) est utilisé pour calculer la distance d'édition $d()$ à l'aide de la programmation dynamique. Une matrice $D_{0..|x|,0..|y|}$ est calculée, où chaque élément $D_{i,j}$ représente le nombre minimal d'opérations pour rendre identique les sous-chaînes $x_{1..i}$ et $y_{1..j}$, x est une chaîne, $|x|$ sa longueur et x_i est le i^{me} caractère de x.

L'algorithme 3 calcule graduellement la distance entre deux chaînes x et y avec un coût des opérations d'édition fixé à 1.

Algorithme 3 : Algorithme distance d'édition.

Entrée : deux chaînes de caractères x et y

Sortie : distance entre les deux chaînes

$D_{i,0} \leftarrow 0$;

$D_{0,j} \leftarrow 0$;

pour $i \leftarrow 1$ ***à*** $|x|$ **faire**

  pour $j \leftarrow 1$ ***à*** $|y|$ **faire**

    si $x_i < y_j$ **alors**

      $D_{i,j} \leftarrow D_{i-1,j-1}$

    sinon

      $D_{i,j} \leftarrow 1 + min(D_{i-1,j}; D_{i,j-1}; D_{i-1,j-1})$

4.4.2 Erreur-grams et les règles de correction

Notre algorithme traite les mots du texte original abscents dans le texte OCR M_{remo}. Il utilise la distance d'édition pour trouver dans l'ensemble des mots du texte OCR abscents dans le texte original M_{remr} les mots les plus proches. Les sous-chaînes erronées dans le texte OCR sont appelées les erreur-grams et les couples formés de l'erreur-grams et de son correspondant dans le texte OCR constituent les règles de corrections. Les résultats sont vérifiés et les prédecesseurs et successeurs de chaque caractère confus sont utilisés pour considérer les alentours des chaînes erronées dans les statistiques classifiant les erreur-grams selon leurs occurences. L'algorithme 4 est le processus de localisation et de construction des erreurs-grams et des règles de correction. Le poids calculé par cet algorithme quantifie l'importance de l'erreur et évalue sa pertinence lors du processus de recherche. A chaque règle de correction est associé une probabilité pour qu'une chaîne de caractères A_i dans l'image du document soit reconnue comme la chaîne B_j dans le texte OCR. Cette probabilité est déterminée à l'aide de la matrice de confusion C_{ij} dont les éléments représentent l'occurence de B_j sachant A_i. Elle est calculée par la formule :

$$P(B_j) = \sum_{A_i} P(B_j|A_i)P(A_i) \tag{4.1}$$

où $P(B_j|A_i)$ dénote la probabilité conditionnelle de B_j sachant que c'est A_i qui est reconnu,

$$P(B_j|A_i) = \frac{P(A_i|B_j)P(B_j)}{P(A_i)} \tag{4.2}$$

Ces probabilités portent sur les chaînes de caractères et sont utilisées dans l'expression de la requête à la section 4.5.3.

4.5 Processus de recherche

Avec l'appariement de chaînes de caractères comme outil de base, l'appariement exact devient insuffisant à cause des erreurs de reconnaissances. Un mot mal reconnu devient

Algorithme 4 : Algorithme "erreurs-grams".

pour $x_i \in M_{remo}$ **faire**

  pour $y_j \in M_{remr}$ **faire**
    calculer $d_{ij} = d(x_i, x_j)$
      * selectionner $\{y_j \in M_{remr}$ / d_{ij} est le minimum$\}$
      * vérifier les appariements obtenus
      * extraire les chaînes de caractères mal reconnus
      * construire les erreurs-grams et les règles de correction correspondantes
      * mettre à jour la fréquence des erreurs.

calculer le poids indiquant l'importance de chaque erreur-gram.

difficile d'accès et toutes les requêtes s'y référant posent problème. Quand les données sont bruitées ou mal reconnues, comme c'est le cas avec le texte OCR, l'appariemnt exacte de chaînes de caractères devienne inapproprié et d'autres mesures sont à considérer pour faciliter l'indexation et la recherche d'information dans les textes OCR.

Le but principal de notre approche est d'ajouter de nouveaux termes à la requête pour pallier aux erreurs de reconnaissance. Le processus de recherche prépare les images de documents pour faciliter l'accès lors de la recherche. Il identifie les éléments du document potentiellement indexables, supprime les mots inutiles, lemmatise et calcule les poids des mots restants pour constituer le fichier inversé à travers lequel le processus de recherche trouve les documents pertinents à la requête de l'utilisateur.

L'appariement porte désormais sur les poids des termes de la requête étendue et la matrice "fichier inversé" MxN termes-documents où M est le nombre de documents dans la collection et N est le nombre de termes considérés lors de l'indexation. La similarité calculée pour chaque image de document permet au système de recherche de présenter les images par ordre de pertinence à la requête de l'utilisateur.

La conception de notre système de recherche est basée sur 3 modules qui sont :

- étendre la requête en ajoutant les mots générés par les erreurs-grams et les règles de correction à partir des termes de la requête initial. Assigner des poids à la liste obtenue pour une utilisation lors du processus de recherche
- sélectionner les images pertinentes à la requête de l'utilisateur. La phase d'indexation est responsable de la sélection d'index et de la structuration du fichier inversé à partir des images de la collection. Les nouveaux documents ne peuvent être considérés qu'après avoir subi cette phase d'indexation
- décider de la pertinence des images et de la manière de les présenter à l'utilisateur.

4.5.1 Expansion de la requête

Pour chaque mot de la requête, nous générons les mots en substituant les erreur-grams par leurs correspondants dans les règles de correction. Prenons le mot "light" comme exemple, il est prouvé statistiquement que l'OCR confond "i" avec "l" et "g" avec "e", etc. Les erreur-grams nous permettent de considérer les mots $"llght", "lighl", "right"$, etc. (32 mots au total) comme reliés à "light". Si la probabilité de confusion est limitée à 10^{-3}, la liste des mots à ajouter à la requête est :
$< light; llght; ligit; lighd; lieht; iight; lighl; ligbt; right >$
et les poids correspondants aux probabilités de confusion de ces mots (voir les formules de calcul à la section 4.5.3) sont :
$< 1; 0.096; 0.0092; 0.0086; 0.009; 0.0036; 0.0032; 0.004; 0.001 >$.

Certains mots, comme "right" dans l'exemple précédent peuvent prêter à des confusions dans les réponses. En effet, le mot "right" utilisé dans les mots de la requête peut nuire au besoin de l'utilisateur. Le modèle vectoriel utilisé affecte le poids 0.001 à "right" dans notre exemple pour influencer l'ordre et l'importance des documents qui le contiennent.

4.5.2 Processus d'indexation

Les documents sont reconnus à travers un ensemble de termes. La technique d'indexation repose sur la collecte de tous les termes contenus dans le document, desquels on supprime les mots inutiles comme "le", "chose" et on lemmatise pour ne considérer que les racines des mots comme par exemple "image" pour "images" et "imageries". Les mots restants constituent les termes d'indexation.

Le modèle vectoriel utilise des vecteurs pour représenter les documents de la collection et les requêtes. Un vecteur est obtenu pour chaque document et chaque requête à partir de l'ensemble des termes d'indexation avec les poids associés. Le poids est fonction de la fréquence du mot dans le document et une méthode populaire pour le calculer est le tf*idf. La notation tf*idf est très connue dans le milieu de la RI. Cela désigne un ensemble de schémas de pondération (et de sélection) de termes. tf signifie "term frequency" et idf "inverse document frequency". Par tf, on désigne une mesure qui a rapport à l'importance d'un terme pour un document. En général, cette valeur est déterminée par la fréquence du terme dans le document. Par idf, on mesure si le terme est discriminant (ou non-uniformément distribué). la fréquence du terme (tf) du terme t_i dans le document d_j est calculé par la formule :

$$tf_{ij} = \frac{frequence_{ij}}{Max_l \ frequence_{lj}}$$

où $Max_l \ frequence_{lj}$ est la fréquence maximale des termes dans le document d

$$idf_i = log\frac{N}{n_i}$$

où N est le nombre de documents dans le corpus et n ceux qui contiennent le terme t_i.

Dans le tf*idf, la valeur du poids associée au terme t_i et au document d_j est calculé par :
$d_{ij} = tf * idf = [tf_{ij}/Max[tf_{ij}]] * log(N/n)$

Une formule tf*idf combine les deux critères qu'on a vu précédemment :

- l'importance du terme pour un document (par tf)
- le pouvoir de discrimination de ce terme (par idf). Ainsi, un terme qui a une valeur de tf*idf élevée doit être à la fois important dans ce document, et aussi il doit apparaître peu dans les autres documents. C'est le cas où un terme correspond à une caractéristique importante et unique d'un document.

Avec une telle formule, on peut donc choisir de garder seulement les termes dont la valeur de tf*idf dépasse un certain seuil.

4.5.3 Calcul de la similarité

Avec les mesures précitées, la mesure du cosinus $sim(q, d_i)$ détermine le degré de similarité entre la requête $q = q_1; q_2, ; ...; q_t$ et le document $d_i = d_{i1}; d_{i2}, ; ...; d_{it}$. La similarité est alors mesuré par le cosinus et calculé par la formule suivante :

$$sim(q, d_i) = \frac{\sum_{j=1}^{t} q_j \, d_{ij}}{\sqrt{\sum_{j=1}^{t} q_j^2} \sqrt{\sum_{j=1}^{t} d_{ij}^2}} = \frac{\sum_{j=1}^{t} q_j \, d_{ij}}{||q_j|| \, ||d_{ij}||}$$

où d_{ij} est le poids du terme t_j dans le document d_i et q_j est le terme de la requête, ce qui est déterminé comme suit :

$$q_j = \begin{cases} 1 & \text{si } q_j \text{ est un terme de la requête} \\ \prod_{j=1}^{s} Pr(q_j) \, . \, idf_j & \text{si } q_j \text{ est un terme ajouté} \\ 0 & \text{autrement} \end{cases}$$

Dans la formule ci-dessus, s est le nombre d'erreur-grams utilisées lors de l'expansion et $Pr(q_j)$ est la probabilité $P(Mot_{OCR}/Mot_{ORI})$. Par exemple, si 'light' est un mot de la requête, on détermine les erreurs-grams telles que 'llght', ... 'right' (Mot_{OCR}). Donc on détermine $P(Mot_{OCR}/light)$ pour sélectionner les mots les plus probables et les intégrer dans le processus de recherche.

Les documents sont alors classés par ordre de similarité à la requête. Les documents dont la similarité est supérieure à un certain seuil sont retenus dans la liste des documents pertinents tandis que tous les autres sont considérés comme non pertinents.

4.5.4 Mesures de la performance

La performance est déterminée sur la base d'images de documents pertinentes trouvées à partir d'un ensemble de requêtes sélectionnées aléatoirement. La liste des images pertinentes trouvées à partir des textes originaux fournis par le concepteur est comparée à celles obtenues à travers les textes OCR. L'évaluation des différentes méthodes est basée sur l'efficacité de la recherche utilisant les valeurs moyennes de la précision versus le rappel, qui sont calculées à partir des équations suivantes :

i) le rappel mesure la proportion de document pertinents retrouvés parmi tous les documents pertinents dans la base. Il est calculé par la formule :

$$RAPPEL = \frac{\text{total d'images pertinentes trouvées}}{\text{total d'images pertinentes dans toute la collection}}$$

ii) la précision mesure la proportion de document pertinents retrouvés parmi tous les documents retrouvés par le système. Elle est calculée comme suit :

$$PRECISION = \frac{\text{total d'images pertinentes trouvées}}{\text{total d'images trouvées}}$$

iii) qualité-distance (QD) est utilisée pour mesurer la performance de l'approche "Recouvrement par 3-grams". Cette approche est basée sur la décomposition des chaînes de caractères T en des parties de 3 caractères successifs $((c_i, c_{i+1}, c_{i+2}))_{i\in[1,n-2]}$, où T est de longeur n et représente le terme de la requête et c_i son i^{th} caractère. La distance entre 2 chaînes x et y est mesurée par la qualité-distance QD ; $QD(x, y)$ est le nombre de 3-grams différents entre x et y. Cette mesure est considérée comme un seuil d'acceptation des mots les plus proches. Ainsi, $QD = 1$ signifie que les mots considérés comme proches ne peuvent avoir plus d'un 3-gram différent du mot de la requête.

iv) on utilise aussi la précision moyenne pour comparer les performances de différentes méthodes. La précision moyenne est une moyenne de précision sur 11 points de rappel (0, 0.1,..., 1.0).

4.6 Conclusion

Ce chapitre a présenté une nouvelle approche pour traiter l'information textelle afin d'améliorer l'efficacité de la recherche d'images de documents. Le traitement des chaines de caractères dans un corpus texuel est un domaine fertile et intéressant pour la communauté scientifique. Il y a deux grands axes en rétroconversion de documents :

- tout ce qui concerne la première couche de codage du document (reconnaissance des mots de l'image de document) : dans cette tâche, les OCR sont très performants mais la qualité de la reconnaissance se dégrade lorsque les images sont des documents scannés ou de mauvaises qualités. Ce qui nécessite des améliorations.
- le deuxième axe concerne la détection et la correction des erreurs de reconnaissance : ce point n'est pas du tout pris en compte par les OCR et reste du domaine de la recherche. Tout système de traitement et de recherche d'images de documents doit prendre conscience de ces deux parties et donc du chemin à faire pour améliorer la recherche d'information dans ce domaine.

L'approche que nous avons proposée collecte les erreurs-grammes et les règles de correction pour les utiliser dans l'expansion des requêtes afin de réduire le temps et d'améliorer la performance de la recherche. Nous avons montré que les n-grammes et leurs probabilités d'apparition influe sur la mesure de la similarité et l'ordre des images considérées comme pertinentes à la requête. En outre, l'apport de l'OCR est important pour des documents de qualité inférieure ou provenant d'archives. Nous avons discuté la nature des dégradations qui affectent les images et leurs effets sur la précision des résultats obtenus. Les tests et

les validations pour mesurer l'efficacité de la recherche par l'utilisation d'erreurs-grams et l'expansion des mots de la requête sont décrits au chapitre "Expérimentations".

CHAPITRE 5

REPÉRAGE DES ZONES NON TEXTUELLES

5.1 Introduction

Les systèmes de recherche d'information sur les images fonctionnent en deux phases distinctes. Une phase 'hors ligne' d'indexation sur la base d'une analyse statistique des attributs de formes et de textures des images et une phase 'en ligne' de recherche d'images par similarité des index. Dans les logiciels existants, le problème généralement traité est celui de la recherche d'images similaires à une image donnée. Ces techniques déterminent pour chaque image un vecteur de caractéristiques contenant des valeurs associées aux descripteurs de texture, de forme et de position. La définition d'une distance entre ces attributs permet de définir une notion de similarité entre deux images. La distribution de ces caractéristiques est alors modélisée pour définir des classes regroupant les images similaires et facilitant le processus de recherche.

L'image de document est de nature composite et sa segmentation par une technique offrant plusieurs degrés de liberté permet de s'adapter aux différents types de contenu pour extraire le maximum d'informations. Dans notre approche, l'opérateur 'SKCS', que nous présentons dans la prochaine section, permet l'extraction de l'information avec une bonne précision. Le résultat de cette segmentation est une multitude de tâches localisant les régions informationnelles. La fusion des objets dont les boîtes englobantes se chevauchent permet de réduire le nombre de régions à traiter pour le rapprocher de celui des images originales. Notons que la multitude d'objets et le nombre important de formes obtenues nous mènent à appliquer une technique de classification en vue de regrouper les objets similaires et de faciliter l'interprétation et la recherche d'information.

5.2 Passage de l'opérateur LoG à l'opérateur SKCS

Il n'existe pas de méthode universelle de segmentation pour les images de documents à cause de la grande variété d'objets, de la présence du bruit et de la variation du contraste. Ce qui rend difficile le repérage des zones informationnelles et l'extraction des caractéristiques. Il faut faire appel selon les cas à telle ou telle technique, voire même à plusieurs simultanément.

L'opérateur Laplacien de la Gaussienne $L \circ G$ utilisé au paragraphe 3.2.2 est limité à la variation du paramètre d'échelle σ, son seul degré de liberté. L'analyse des résultats et de la sensibilité de l'opérateur par rapport aux images de documents se trouvent limitées par rapport aux différents contenus. Nous avons observé que la qualité de la segmentation décroit pour les images à faible contraste. Pour σ grand, une importante quantité d'information est retournée lors de la segmentation, mais le temps de calcul reste prohibitif. Pour faire face à ces problèmes et pour préserver les propriétés de la gaussienne les plus importantes et les plus utiles, Cheriet et al (Remaki et Cheriet, 2000) ont proposé une nouvelle famille de noyaux, à support compact, qui dérivent de la gaussienne, appelés KCS (Kernel with Compact Support) pour générer l'espace d'échelles dans la représentation multi-échelle. Cette transformation permet de concentrer l'information dans une boule autour du pixel pour réduire la dimension du filtre de 11.32σ pour la gaussienne à seulement 2σ pour le KCS. Notons que le KCS est obtenu par une transformation topologique de l'espace $\Re^2$ dans une boule par un changement de variables. Cette transformation concentre toute l'information dans cette boule et rend nulle la valeur de la gaussienne en dehors du support.

5.2.1 Formulation du KCS

La famille de noyaux à support compact est obtenue par transformation topologique de l'espace $\Re^2$ dans une boule de rayon unitaire, par un changement de variables. Cette transformation a pour effet de grouper toute l'information dans la boule unitaire. Ainsi, avec

les nouvelles variables, la Gaussienne sera étendue à tout l'espace en prenant des valeurs nulles en dehors de la boule. L'expression générale du noyau KCS est donnée ci-dessous.

$$\phi_{\sigma,\gamma}(x,y) = \begin{cases} \frac{1}{C_\gamma \sigma^2} \quad e^{(\frac{\gamma\sigma^2}{x^2+y^2-\sigma^2}+\gamma)} & \text{si } x^2+y^2 < \sigma^2 \\ \\ 0 & \text{sinon} \end{cases}$$

où, σ est le rayon de la boule supportant le noyau et γ le paramètre qui contrôle la largeur du pic du noyau lui donnant ainsi un second degré de liberté. Il est important de mentionner que le paramètre γ contrôle la distance entre les zéros du KCS et l'origine des axes. Il n'affecte pas la nature des fonctions $\phi_{\sigma,\gamma}$, elles demeurent des noyaux à support compact. De plus, si $\gamma \geq 2$ (Remaki et Cheriet, 2000), les caractéristiques désirées sont garanties, ce qui signifie que la première et la seconde dérivée ont le même comportement que la Gaussienne. Cependant, la formule du KCS n'est pas séparable et une amélioration de cet opérateur est possible par le passage à une nouvelle forme d'opérateur 2D sous forme d'une somme de deux filtres 1D. Cette nouvelle formulation ouvre la porte à de nouveaux paramètres augmentant le nombre de degrés de libertés lors de la segmentation.

5.2.2 Formulation du SKCS

La version séparable du KCS est définie comme le produit de noyaux KCS monodimensionnel. Pour exprimer cette nouvelle version de noyaux, le noyau générateur de la famille du SKCS est donné par la formule suivante (Ben Braiek et al., 2005) :

$$SKCS(x,y) = \Phi(x,y) = \Phi_{\sigma_1,\sigma_2,\gamma_1,\gamma_2}(x,y) = \phi_{\sigma_1,\gamma_1}(x)\phi_{\sigma_2,\gamma_2}(y) \tag{5.1}$$

$$\Phi_{\sigma_1,\sigma_2,\gamma_1,\gamma_2}(x,y) = \begin{cases} \frac{e^{\gamma_1}}{C_{\gamma_1\sigma_1}} \quad \frac{e^{\gamma_2}}{C_{\gamma_2\cdot\sigma_2}} \quad e^{\frac{(\gamma_1\sigma_1^2)}{x^2-\sigma_1^2}} \quad e^{\frac{(\gamma_2\sigma_2^2)}{y^2-\sigma_2^2}} & \text{si } x^2 < \sigma_1^2 \;\; et \;\; y^2 < \sigma_2^2 \\ \\ 0 & \text{sinon} \end{cases}$$

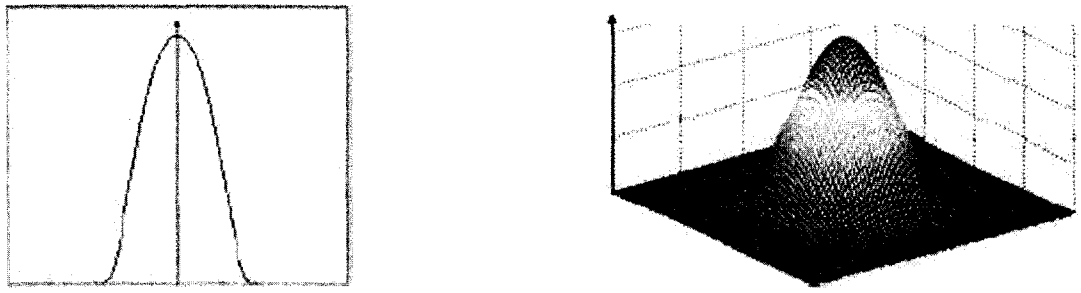


Figure 7 Le profile 1D et 2D du noyau de l'opérateur "SKCS"

Avec le SKCS (Séparable KCS), quatre degrés de liberté s'offrent à nous (σ_1 et σ_2 délimitent le support du noyau et γ_1 et γ_2 la largeur du pic sur l'axe des abscisses et des ordonnées respectivement) pour varier la nature et la qualité de la segmentation.

Comme le KCS, le SKCS présente le même comportement que le noyau Gaussien, les profils 1D et 2D du SKCS sont représentés à la figure 7. Toutefois, le Laplacien du SKCS montre, pour $\gamma = 1$, un nouveau maximum à l'origine (voir figure 8). Mais, on peut facilement montrer que si $\gamma > 2$, les dérivées première et seconde du SKCS présentent le même comportement que le noyau Gaussien, e.g. la convolution avec le SKCS n'introduit aucun nouvel extremum.

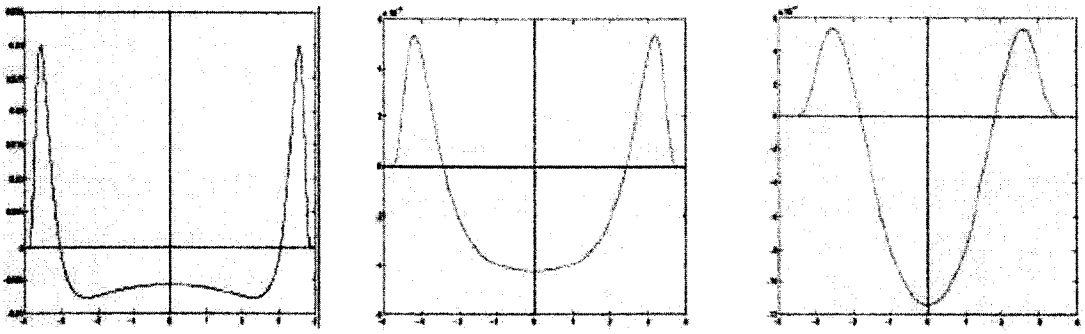
Dans notre application, on localise les régions informationnelles et en conséquence, notre intérêt va porter sur le Laplacien du SKCS ($LoSKCS$) qui est obtenu par la formule :

$$LoSKCS = \nabla^2 \phi_{\sigma,\gamma}(x,y) \tag{5.2}$$

$$\nabla^2 \phi_{\sigma_1,\sigma_2,\gamma_1,\gamma_2}(x,y) = (\tfrac{\partial^2}{\partial x^2}\phi_{\sigma_1,\gamma_1}(x))\phi_{\sigma_2,\gamma_2}(y) + (\tfrac{\partial^2}{\partial y^2}\phi_{\sigma_2,\gamma_2}(y))\phi_{\sigma_1,\gamma_1}(x)$$

où $\frac{\partial^2}{\partial s^2}\phi_{\sigma,\gamma}(s) = 2\ \zeta\gamma\sigma^2(\frac{(s^2-\sigma^2)(.\sigma^2-3s^2)+2s^2\gamma\sigma}{(s^2-\sigma^2)^4})e^{\frac{(\gamma\sigma^2)}{s^2-\sigma^2}}$

et $\zeta = \frac{e^{\gamma}}{C_{\gamma\sigma}}\ \gamma\sigma$



(a): γ=1 (b): γ=2.1 (c) γ=4

Figure 8 Influenence du paramètre γ sur la largeur du pic de l'opérateur "LoSKCS", pour $\sigma = 4$

La validation de nos résultats expérimentaux portant sur le repérage des zones non textuelles serait faite sur la base de 4 variables qui sont σ_1 et γ_1 sur l'axe des abscisses et σ_2 et γ_2 sur l'axe des ordonnées.

5.3 Architecture de l'approche proposée

L'approche pour la localisation et l'identification des zones non textuelles est guidée par les idées suivantes :

- la segmentation basée sur l'opérateur SKCS qui nous offre plusieurs degrés de liberté
- l'utilisation des positions spatiales et du multi-échelle pour la localisation et la fusion des blocs
- l'utilisation combinée de différentes informations de formes et de textures pour regrouper les objets similaires en vue de différencier les types de contenus.

La figure 9 présente l'architecture et les principaux modules de notre approche. Le module de localisation des régions non textuelles est réalisé en segmentant l'image par l'opérateur SKCS. Il extrait les régions informatives à différentes échelles de façon à intégrer progressivement le bruit pour obtenir différentes formes pour chaque objet. On déterminera la meilleure échelle pour ce processus dans le chapitre 7 relatif aux expériences menées. Les modules restants qui présentent les caractéristiques extraites, la fusion des zones détectées, la classification et la recherche des régions informatives sont détaillés dans les sections suivantes.

5.4 Fusion des objets

L'objectif de la fusion est de définir une notion de connexité élargie pour réduire le nombre d'objets à traiter et faciliter l'interprétation. Nous nous sommes interessés à cette étape de pré-traitement car c'est elle qui va reconstituer les régions et les rapprocher des formes originales. Durant le processus de fusion, l'algorithme 5 travaille sur un ensemble E, constitué des objets extraits et subdivisés en plusieurs parcelles disjointes, et sur un ensemble C constitué des objets en construction. Une fois la fusion terminée, tous les objets dont les boites englobantes se chevauchent seront regroupés pour ne former qu'un seul objet dans l'ensemble C.

La fusion se fait de gauche à droite et de haut en bas, l'objet situé à gauche et celui situé en haut et qui se chevauchent servent de base pour déterminer les coordonnées de la boite englobante du nouveau objet et dont les descripteurs seront modifiés en fonction des actions de fusion que nous avons sélectionnées.

On considère qu'il y a recouvrement entre 2 objets x et y si :

$[(x_1 + h_x) > y_1 \quad ET \quad x_1 < y_1] \quad ET \quad [(x_2 + l_x) > y_2 \quad ET \quad x_2 < y_2]$

où $x_1,\ x_2,\ y_1\ et\ y_2$ représentent les coordonnées nord-ouest de la boite englobante de l'objet x alors que h_x et l_x sont respectivement la hauteur et la largeur de l'objet x.

Les paramètres $x_1,\ x_2,\ h_x\ et\ l_x$ sont des éléments du vecteur de l'objet x.

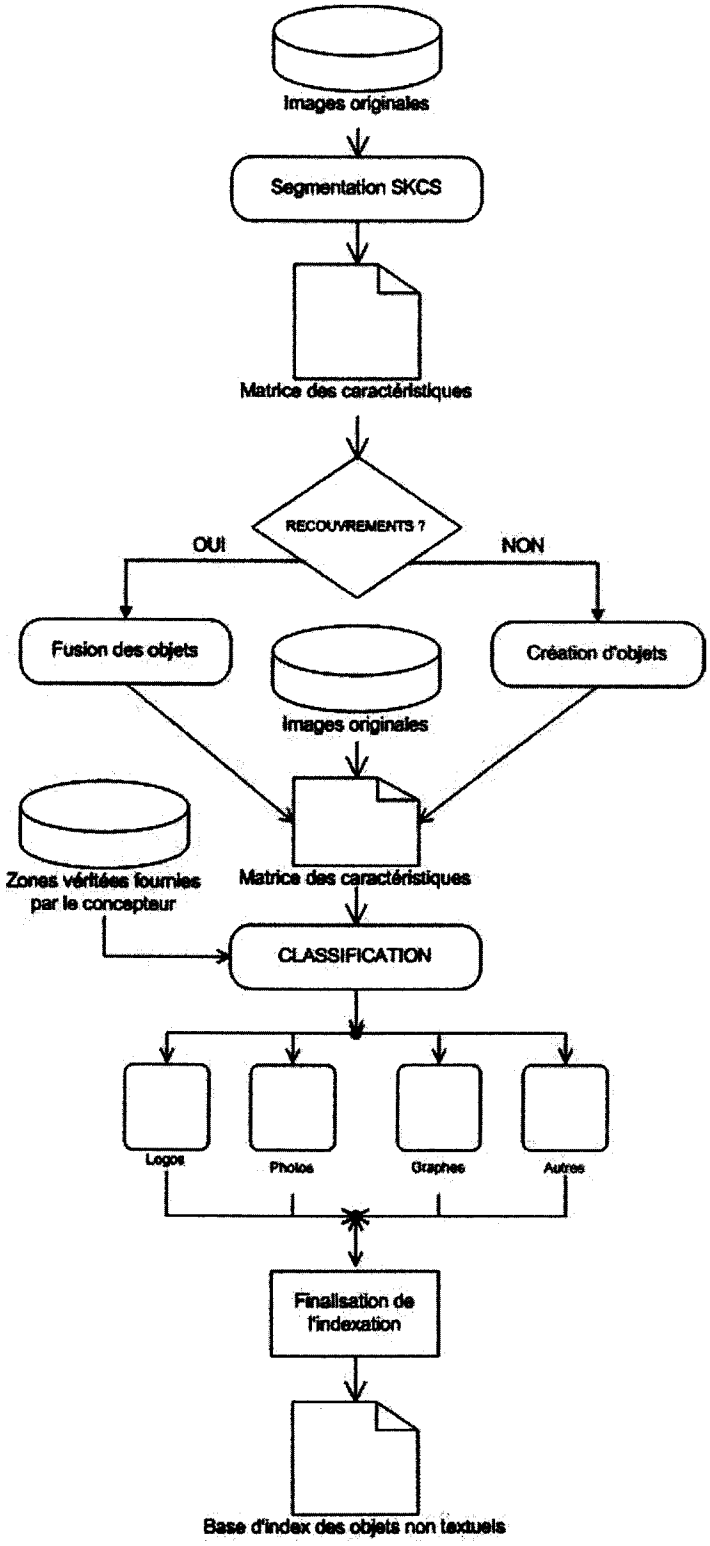


Figure 9 Architecture de traitement des zones non textuelles

Le même procédé est appliqué pour chaque coin de la boîte englobante de l'objet y (on vérifie si l'un des coins nord-est, sud-ouest ou sud-est se chevauche avec l'objet x).

L'algorithme 5 présente les actions entreprises pour la fusion d'objets. L'opération de fusion opérée sur l'image originale de la figure 10 a réussi à intégrer les vecteurs caractéristiques des différentes parcelles pour ne produire que deux vecteurs caractéristiques représentant les deux illustrations effectives.

Algorithme 5 : Fusion des objets

Fusion des objets

E Ensemble des objets à traiter

C Ensemble des objets obtenus par la fusion

debut

  pour *chaque objet* O_i *de* E **faire**

    si *Recouvrement avec un* O_j *de* C **alors**

      même objet ;

      mise à jour de O_j

    sinon

      si *Recouvrement avec plusieurs* O_j *de* C **alors**

        fusion de tous les O_j recouverts dans O_j^c

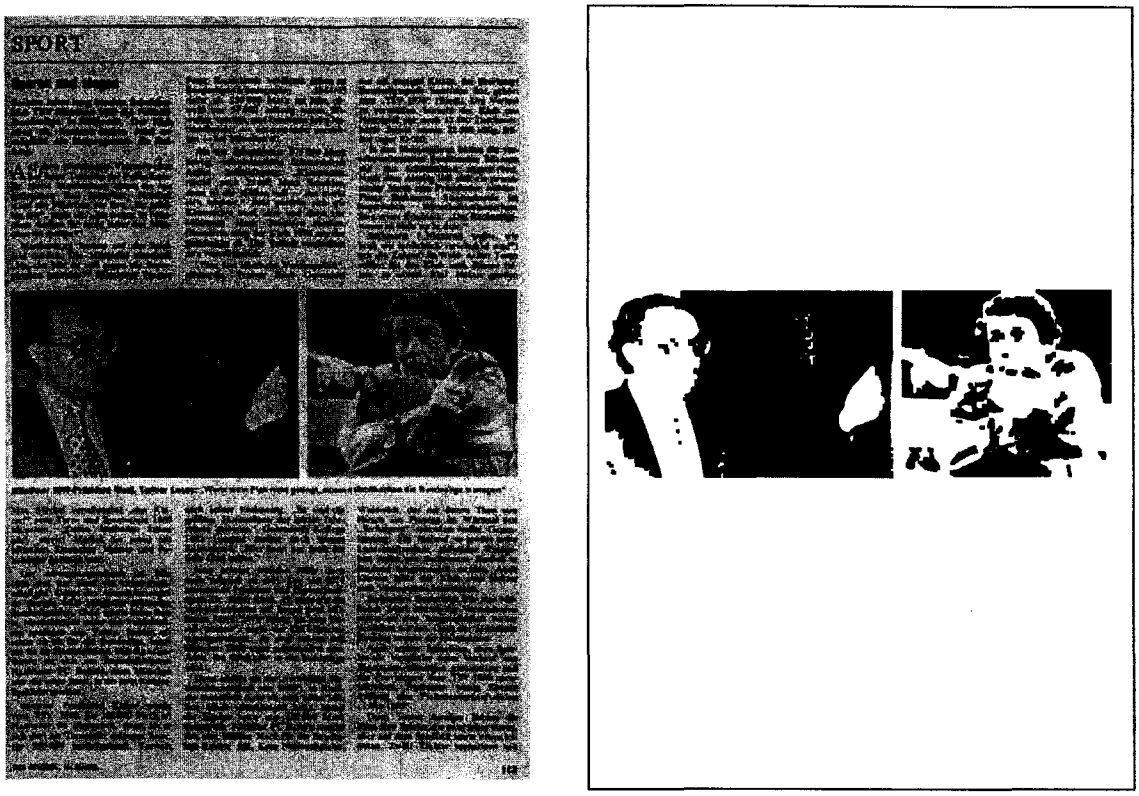
        C=C-{O_j fusionnés}+O_j^c

      sinon

        nouvel objet créé

        $C = C + O_i$

fin



(a) Image originale

(b) Extraction des zones non textuelles (2 objets seulement à considérer après la fusion)

Figure 10 Image ségmentée par le SKCS (σ et $\gamma = 5$) et fusion des tâches recouvertes après la suppression des taches minuscules

5.5 Définition de la classification

Une fois les objets informationnels localisés suite au processus de segmentation d'images, on cherche à identifier ce que représente chacune des régions. L'interprétation automatique des objets devient nécessaire à cause du nombre important d'objets extraits lors de la segmentation.

Le but est alors de classer automatiquement un nouvel objet ou de prédire son type (logo, photographie ou graphe). Il s'agit de regrouper les objets en thèmes correspondants à la vérité fournie par le concepteur de la base d'images. On procède par l'attribution des vec-

teurs de caractéristiques des régions de l'image à des classes connues à priori (c'est la classification supervisée) ou à des agrégats produits de la classification automatique (classification non supervisée). Le problème de la classification est en général de construire une procédure permettant d'associer une classe à un objet. Nous détaillons dans cette section les principes de base et le concept de la classification, détaillés à la figure 11, en vue de l'utiliser dans notre approche.

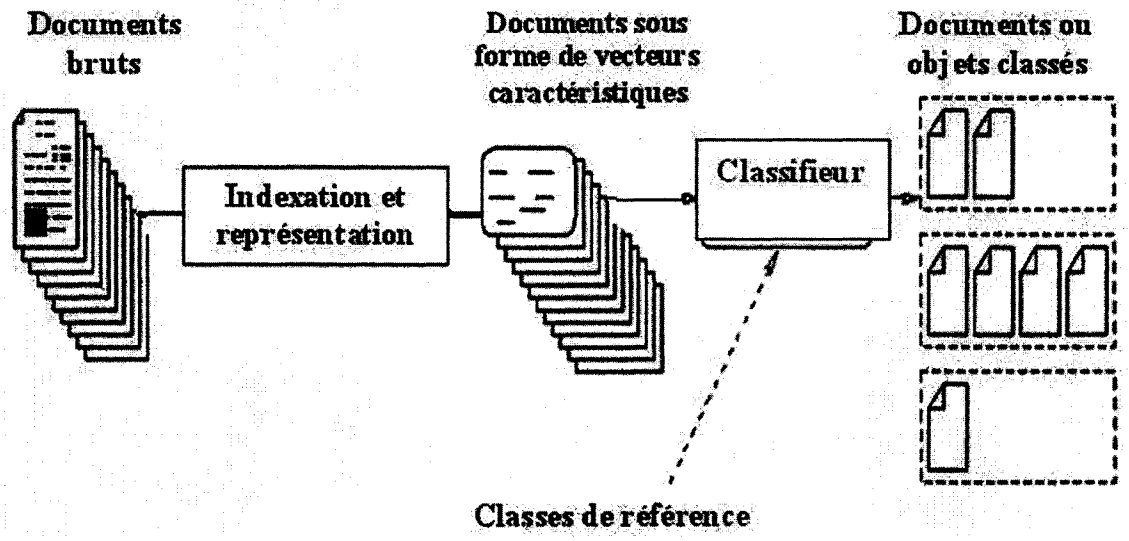


Figure 11 Concept de la classification

5.5.1 Classification supervisée

Dans la classification supervisée, on connaît les classes possibles et on dispose d'un ensemble d'objets déjà classés, servant d'ensemble d'apprentissage. Le problème est alors d'être capable d'associer à tout nouvel objet sa classe la plus adaptée, en se servant des exemples déjà étiquetés. La classification supervisée consiste à déterminer une procédure de classification :

$$f : \vec{d_{ij}} \rightarrow C_k \tag{5.3}$$

qui à partir du vecteur de description $\vec{d_{ij}}$ de l'objet j et de l'image d_i, détermine sa classe C_k avec le plus faible taux d'erreur.

La classification supervisée suppose connues deux fonctions :

- la première fait correspondre à tout vecteur $\vec{d_{ij}}$ une classe C_k. Elle est définie au moyen de couples $(\vec{d_{ij}}, C_k)$ donnés comme exemples au système,
- la deuxième fait correspondre à toute image de document d_i ses vecteurs descriptions $\vec{d_{ij}}$.

La performance de la classification dépend notamment de l'efficacité de la description $\vec{d_{ij}}$ et de la fiabilité du système d'apprentissage pour classer efficacement tout nouvel exemple (pouvoir prédictif).

5.5.2 Classification non-supervisée

Dans la classification non supervisée, les classes ne sont pas connues à l'avance, et les exemples disponibles sont non étiquetés. Le but est donc de regrouper dans un même cluster (ou groupe) les objets considérés comme similaires. Le problème est alors de définir cette similarité entre objets qui est estimée par une fonction calculant la distance entre ces objets. Une fois cette fonction distance définie, la tâche de clustering consiste à réduire au maximum la distance entre membres d'un même cluster, tout en augmentant au maximum la distance entre clusters. À la fin du processus de classification non-supervisée, les documents doivent appartenir à l'une des agrégats générées par la classification automatique.

Après avoir choisi les paramètres sur lesquels va porter la classification, de nombreuses techniques peuvent alors être envisagées et on distingue deux catégories de classifieurs : hiérarchiques et non-hiérarchiques.

Dans la classification non-hiérarchique, les individus ne sont pas structurés de manière hiérarchique. Le résultat obtenu est soit des partitions (Chaque objet ne fait partie que d'un

sous-ensemble) ou des groupes recouverts (chaque individu peut appartenir à plusieurs groupes avec une probabilité p_i).

Dans la classification automatique hiérarchique(CAH), les sous-ensembles créés sont emboîtés de manière hiérarchique les uns dans les autres. On distingue la CAH descendante (ou divisive) et la CAH ascendante.

La CAH descendante part de l'ensemble de tous les individus et les fractionne en un certain nombre de sous-ensembles, chaque sous-ensemble étant alors fractionné en un certain nombre de sous-ensembles, et ainsi de suite.

La CAH ascendante sélectionne les deux objets les plus proches (au sens de la mesure de similarité entre caractéristiques choisie) pour les regrouper. Pour déterminer quelles classes on va fusionner, on utilise un des critères d'agrégation présentés dans la section 5.5.4. On regroupe à chaque étape les deux clusters dont les moyennes de distances entre voisins sont les plus faibles. Le regroupement ascendant des objets deux à deux se traduit par un dendrogramme (figure 12) dont les feuilles sont les objets et dont les noeuds sont les clusters (groupes). Le nombre de clusters est fonction du niveau de la coupe du dendrogramme.

Malgré une complexité en $O(n^2)$, où n est le nombre d'objets à classer, la méthode de classification hiérarchique ascendante est la plus utilisée pour classer des documents. Cette méthode produit une classification hiérarchique comme montrée à la figure 12.

Par ailleurs, d'autres méthodes existent. Le principe de ces méthodes repose sur la minimisation de distance et consiste à rechercher la classe la plus proche pour chaque vecteur. La notion de proximité est liée à la distance considérée. On peut distinguer deux méthodes :

- méthodes non itératives qui consistent à parcourir les vecteurs des objets de l'image et à déterminer la classe la plus proche parmi les K possibles

– méthodes itératives qui reposent sur des algorithmes de regroupements répétitifs d'objets dans différentes classes jusqu'à stabilisation (Le contenu des classes est identique entre deux itérations). La méthode la plus populaire est connue sous le nom de K-moyennes que nous adoptons pour sa simplicité et que nous présentons dans la section 5.6.1.

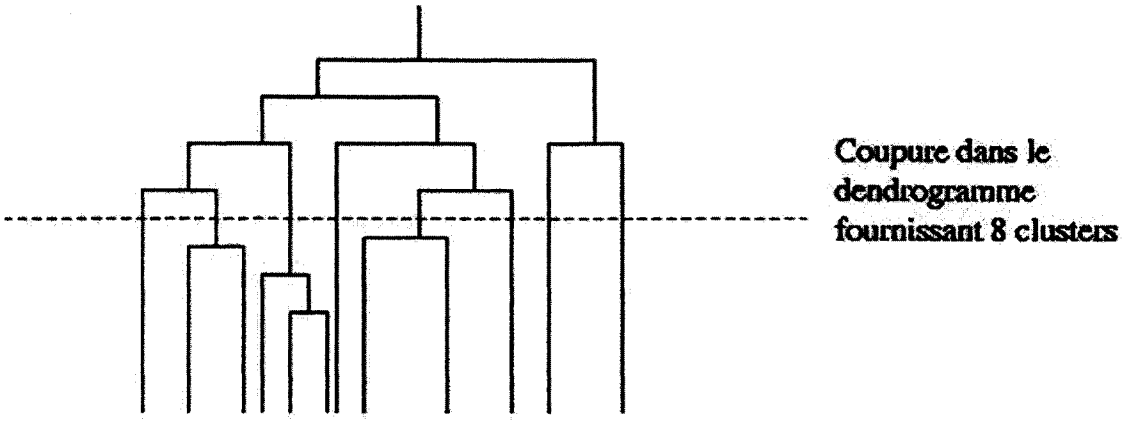


Figure 12 Exemple de hiérarchie produite par un algorithme de classification ascendante (CAH)

5.5.3 Définition de la distance

Une distance d est une application de $E \times E$ dans $\Re^+$ (E est l'espace de projection des objets à traiter) telle que les propriétés suivantes soient vérifiées (Saporta, 1990) :

$$d(a,b) = 0 \Leftrightarrow a = b \qquad \forall (a,b) \in E \times E \tag{5.4}$$

$$d(a,b) = d(b,a) \tag{5.5}$$

$$d(a,b) \leq d(a,c) + d(c,b) \qquad \forall c \in E \tag{5.6}$$

Dans notre cas, les points sont les vecteurs caractéristiques des objets et la distance entre deux vecteurs $\vec{x}$ et $\vec{q}$ se fait via la métrique de Minkowski L_p :

$$d_p(\vec{x},\vec{q}) = \left(\sum_{j=1}^{n} |x_j - q_j|^p \right)^{\frac{1}{p}} \qquad p \in [1,\infty[\tag{5.7}$$

Selon la valeur de p, on retrouve plusieurs distances connues :

- Si $p = 1$, cette distance est la distance de Manhattan définie par
  $d_m(a, b) = |a_1 - b_1| + |a_2 - b_2| + \cdots + |a_n - b_n|$
- Si $p = 2$ c'est la distance euclidienne définie par
  $d_e(a, b) = \sqrt{\sum_i (a_i - b_i)^2}$
- Si $p = \infty$ c'est la distance de Chebyshev définie par
  $d_c(a, b) = max\{|a_1 - b_1|, |a_2 - b_2|, \ldots, |a_n - b_n|\}$

La distance euclidienne favorise les vecteurs dont toutes les variables sont assez proches alors que la distance de Manhattan permet de tolérer une distance importante sur l'une des variables.

5.5.4 Critère d'agrégation

Le critère d'agrégation permet de comparer les classes deux à deux pour sélectionner les classes les plus similaires suivant un certain critère. Les critères les plus classiques sont le plus proche voisin, le diamètre maximum, la distance moyenne et la distance entre les centres de gravités.

- Plus proche voisin
  La distance entre la classe C_p et la classe C_q est la plus petite distance entre un élément de C_p et un élément de C_q ;

$$L(C_p, C_q) = \min\{d(i, j); i \in C_p, j \in C_q\}; \tag{5.8}$$

- Diamètre maximum
  La distance entre la classe C_p et la classe C_q est la plus grande distance entre un élément de C_p et un élément de C_q.

$$L(C_p, C_q) = \max\{d(i, j); i \in C_p, j \in C_q\}; \tag{5.9}$$

- Distance moyenne

  La distance entre la classe C_p et la classe C_q est la moyenne des distances entre les éléments de C_p et les éléments de C_q.

$$L(C_p, C_q) = \frac{\sum_{i,j}\{d(i,j); i \in C_p, j \in C_q\}}{Card(C_p) \times Card(C_q)}; \tag{5.10}$$

- Distance entre les centres de gravité

  Si G_p est le centre de gravité de la classe C_p et si G_q est le centre de gravité de la classe C_q alors la distance entre la classe C_p et la classe C_q est la distance entre leurs centres de gravités. $L(C_p, C_q) = d(G_p, G_q)$

5.6 Stratégie de classification proposée

Après avoir déterminé les vecteurs caractéristiques des régions non textuelles, un besoin d'interprétation des différentes régions informatives nous conduit à regrouper les objets similaires dans des partitions disjointes. Il est difficile d'opter pour une classification supervisée à cause de la taille de notre base d'images. Nous regroupons les objets similaires dans des classes pour les confronter ensuite à la base véritée (ground-truths) et juger de la qualité des regroupements. Le nombre de groupes est relativement stable, c'est pourqoui nous appliquons le classifieur k-moyennes comme processus d'initialisation et ensuite un mélange d'analyse en composantes principales et une subdivision spaciale des objets en sous-classes pour réduire la dimensionnalité des vecteurs de caractéristiques et augmenter la précision lors de la recherche d'objets similaires.

5.6.1 Classification automatique par l'algorithme k-moyennes

La classification est utilisée pour rassembler les zones extraites de l'image de document en K groupements, fonction d'un critère de "ressemblance". Parmi les algorithmes de regroupement, la méthode des k-moyennes est la plus utilisée. Il s'agit d'une approche facile à implémenter et qui consiste à répartir les objets des images en k groupes autour de k

centres appelés noyaux ou centroïdes : un objet image est dans un groupe (ou cluster) s'il est plus proche, en fonction d'une distance choisie, du centroïde de ce groupe que de n'importe quel autre centroïde des autres groupes. Le centroïde de chaque groupe, recalculé à la fin de l'exécution de la méthode, correspond au barycentre de chaque groupe.

La répartition des différents vecteurs des objets dans les différents groupements minimise l'indice de dispersion. Ce dernier représente une mesure de distance du vecteur échantillon aux vecteurs représentatifs du cluster. L'algorithme k-moyennes développé dans le cadre du traitement de l'image est décrit par l'algorithme 6.

L'algorithme k-moyennes nécessite de choisir au départ un nombre k de points dans l'espace vectoriel des objets comme représentants des k classes à trouver. Généralement, les k images sont choisies au hasard. Des méthodes ont été proposées pour l'initialisation, Cutting et al. (Cutting et al., 1992) utilisent par exemple une classification hiérarchique ascendante appliquée à un sous-ensemble des documents pour trouver les k centres de départ. P. Bellot (Bellot et El-Beze, 1999) utilise également une classification hiérarchique ascendante jusqu'à obtenir un nombre k de classes. Une fois les k centres choisis, la méthode des k-moyennes range chaque document dans la classe la plus proche. Les centres des classes sont alors recalculés jusqu'à la stabilisation de la solution (classes identiques entre 2 itérations).

L'algorithme s'arrête lorsque plus aucun document ne change de classe ou lorsqu'un nombre prédéterminé d'itérations sur l'ensemble des documents est effectué.

La figure 13 montre le résultat de la classification, dans 4 classes différentes, des 3343 objets obtenus lors de la segmentation. La mesure d'excentricité est une mesure de la forme ellipsoidale de l'objet. La valeur est entre 0 et 1. Une ellipse dont l'excentricité est 0 est réellement un cercle, alors qu'une ellipse dont l'excentricité est 1 est une ligne segment. On remarque que cette mesure n'est pas discriminante pour ce niveau de classification car

Algorithme 6 : Algorithme k-moyennes dans le cadre du traitement de l'image.

Préliminaire : d est la distance euclidienne.

$\vec{x}$ est le vecteur des caractéristiques de l'objet x.

Initialisation : Choisir k vecteurs caractéristiques (distincts) y_1 y_2 ... y_k (on les appelle centres des classes)

tant que *les nouvelles classes sont différentes des anciennes* **faire**

Reconstruire les k nouvelles classes C_i telles que

$C_i = \{\ x\ \mid d(y_i, \vec{x}) < d(y_j, \vec{x})\ \forall j \neq i\}$

Recalculer le centre des classes : $y_i = \frac{\sum_{x \in C_i} \vec{x}}{|C_i|}$

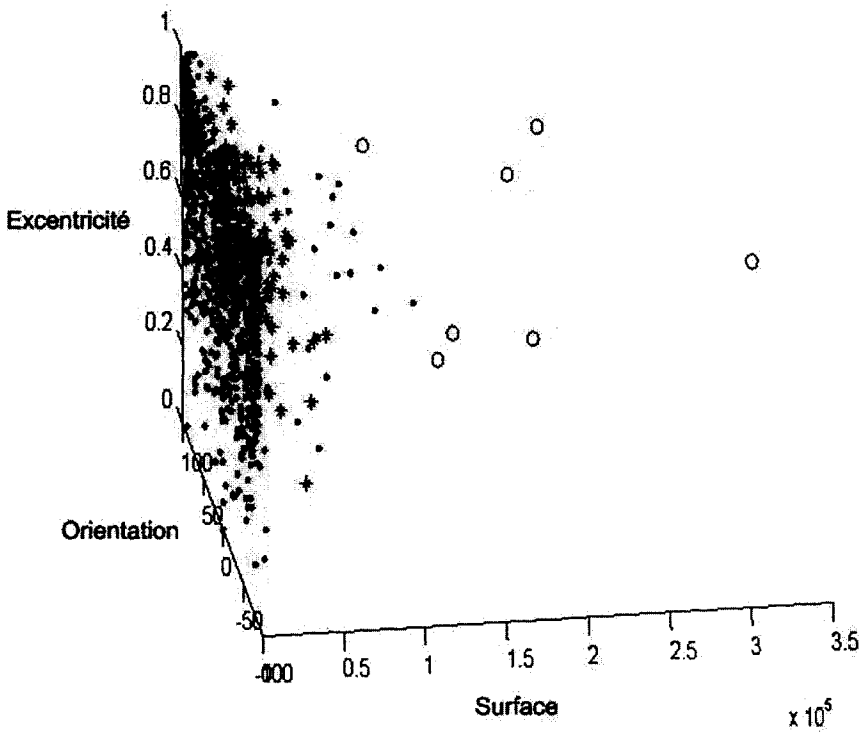


Figure 13 K-moyennes appliqué aux 1759 vecteurs des objets trouvés après fusion lors de la segmentation

sa valeur ne donne aucune indication sur la classe probable de l'objet. Par contre, la surface est une mesure discriminante car elle influe sur la classe d'appartenance de l'objet.

La qualité de la classification par l'algorithme K-moyennes se mesure par :

- une petite variance dans les classes (classes compactes)
- une large distance entre les centres des classes (classes bien isolées).

L'algorithme des k-moyennes est indépendant de l'ordre dans lequel les objets sont considérés, de plus, l'ajout d'un nouveau objet au terme de la classification ne remet pas nécessairement celle-ci en cause. Leur principal inconvénient concerne la détermination du nombre k de classes à trouver ainsi que le choix des k classes de départ.
Dans nos expériences, c'est la mesure du rappel vs. précision pour chaque classe et chaque type d'objet que nous considérons pour mesurer la qualité de la répartition et le nombre de classes.

5.6.2 Optimisation des classes par MKL

La méthode Karhunen-Loeve Multi-espaces (MKL) réduit la dimensionnalité des vecteurs caractéristiques et regroupe les objets proches dans un espace de projection. Cette section présente la transformation KL et la classification par MKL pour répartir les objets en sous groupes représentant fidèlement le nuage des vecteurs caractéristiques.

5.6.2.1 Transformation Karhunen-Loeve

La transformation de Karhunen-Loeve a pour but de transformer un tableau de données pour une meilleure interprétation. L'idée de la transformation est de réaliser un changement de base pour obtenir de nouveaux axes où l'information contenue sur chaque axe est distribuée de façon optimale. Une image est considérée comme un tableau de données ayant u objets représentées par des vecteurs caractéristiques de dimension d.

On peut directement analyser les données brutes. L'origine du nuage est définie par l'observation qui a comme coordonnées $(0, 0, 0, ...)$. Cette origine est très rarement intéressante car elle n'a aucun sens physique. Le plus intéressant est de choisir comme référence le centre de gravité $(g_1, g_2, g_3, ...)$. L'analyse devient une analyse centrée sur le nuage de points. Elle revient à remplacer chaque observation par l'écart de ses coordonnées avec les écarts avec la moyenne. La transformation de Karhunen-Loeve garde la prépondérance de certaines variables sur d'autres.

Pour maximiser l'information sur chacun des axes, il faut se mettre dans le référentiel qui permet la comparaison de ces variables. Pour cela on se place dans le référentiel dont la base est constituée des vecteurs propres de la matrice de corrélation.

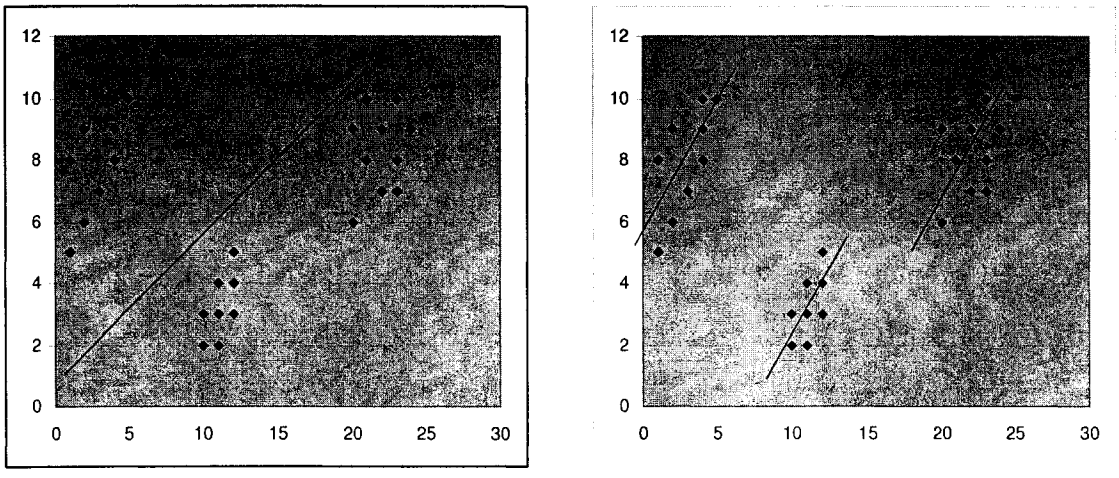
Le principe de la méthode détaillée dans l'algorithme 7 réside dans :

- l'utilisation des vecteurs caractéristiques pour calculer la moyenne de chaque vecteur
- le centrage de la matrice des caractéristiques
- le calcul de la matrice de covariance et sa décomposition pour déterminer les valeurs et les vecteurs propres.

Avec les valeurs et les vecteurs propres, nous avons la contribution des différents axes de la projection des vecteurs caractéristiques. L'énergie d'un axe est le ratio de la valeur propre de l'axe sur la somme des valeurs propres. Cette énergie est la quantité d'information expliquée par cet axe. La qualité de la transformation en gardant les axes les plus importants est simplement la somme des énergies de ces axes. Plus la qualité sera proche de 100%, plus la transformation sera sans perte d'information, ce qui est équivalent à dire que le dernier axe ne contient que très peu d'information. Pour avoir des chances d'obtenir une bonne transformation, il est souhaitable de minimiser le taux d'erreur (avoir une

qualité supérieure à 90%) et de réduire la dimensionnalité. Cette transformation permet d'effectuer :

- la réduction du nombre de variables : grâce à la quantité d'information portée par chaque variable principale, on est à même de sélectionner un nombre réduit de variables en se fixant un pourcentage minimum de l'information expliquée. Il est souhaitable d'avoir une qualité supérieur à 90%.
- l'optimisation dans le sens où l'on obtient la plus grande distribution de variances sur les axes. Dans l'espace KL, toute l'énergie est contenue dans la somme des valeurs propres λ_i. Mieux, l'énergie de chaque axe est donnée par la valeur propre associée. Comme par définition $\lambda_1 > \lambda_2 > ... > \lambda_n$, l'axe 1 est énergétiquement prépondérant sur les autres.
- la préclassification : il est alors nécessaire de décomposer le nuage de points constitué par les vecteurs caractéristiques des régions détectées dans l'espace des paramètres.



a) Projection pour la transformation KL

b) Projection multi-espace MKL

Figure 14 Projection sur un axe (K=1) pour la transformation KL à gauche et MKL à droite avec S=3 et K=1

5.6.2.2 Multi-espace KL "MKL" (Cappelli et al., 2001)

La figure 14 montre que la séparation des données par KL est linéaire. Pour remédier à ce problème, nous avons besoin de répartir les objets à traiter en sous-espaces afin de rendre notre système d'interprétation plus précis. Pour celà, la méthode MKL a l'avantage de subdiviser les classes initiales en répartissant ses objets en sous-classes. Chaque sous-classe représente un groupement d'objets dans l'espace de projection. La méthode MKL remédie au problème de la linéarité de KL par une représentation plus fidèle du nuage de points. Les étapes importantes de MKL sont exposées dans l'algorithme 8. L'initialisation des sous-espaces peut être parfaitement aléatoire de sorte que chaque sous-espace sera formé de m/s objets où m est le nombre total d'objets et s le nombre de sous-espaces à considérer. Chaque sous ensemble P_i est un sous-espace KL de dimension k_i. La répartition des objets de la classe est bâtie sur les mesures de distance et de similarité suivantes :

- la projection d'un vecteur $x \in \Re^n$ dans l'espace $S_{\overline{x},\phi_k}$ est définie par :

$$KL(x, S_{\overline{x},\phi_k}) = \phi_k^T(x - \overline{x}) \tag{5.11}$$

  ϕ_k est la matrice des vecteurs propres de l'espace S.

  Le retour à l'espace d'origine pour un vecteur $y \in \Re^k$ relativement à $S_{\overline{x},\phi_k}$ se fait à l'aide de la formule $KL^{-1}(y, S_{\overline{x},\phi_k}) = \phi_k y + \overline{x}$

- la distance entre un vecteur $x \in \Re^n$ et l'espace $S_{\overline{x},\phi_k}$ est illustrée à la figure 15. Elle est calculée par la formule :

$$d_{FS}(x, S_{\overline{x},\phi_k}) = \sqrt{\|x - \overline{x}\|_2^2 - \|y\|_2^2} \quad \text{où} \quad y = KL(x, S_{\overline{x},\phi_k}) \tag{5.12}$$

- la similarité entre le vecteur requête q de dimension n et les vecteurs du sous-espace $y_1, y_2, ...y_m$ de dimension k est déterminée par :
  - la projection r de q est calculée par la formule 5.11. Elle produit un vecteur r de dimension k
  - les distances euclidiennes entre r et $y_1, y_2, ...y_m$ sont calculées dans l'espace $\Re^k$ à l'aide de la formule 5.12.

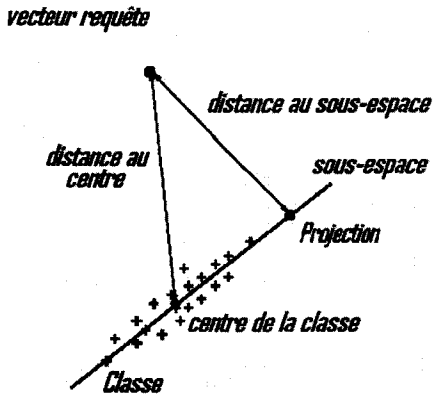


Figure 15 Projection et calcul des distances entre un vecteur et un sous-espace

Le partitionnement $\wp = \{P1, P2, ..., Ps\}$ de l'espace P et les résulats de la classification $K = \{k1, k2, ...ks\}$ doivent respecter les conditions suivantes :

- $\bigcup_{i=1...s} P_i$ =P $\quad P_i \bigcap P_j = \Phi \quad \forall i, j = 1...s, i \neq j$
- $m_i = card(P_i) \geq \lfloor \frac{m}{s+1} \rfloor \quad \forall i = 1...s$
- $k_i < m_i, \quad k_i > 0, \quad k_i < n \quad \forall i = 1...s.$

La transformée MKL est définie par l'ensemble des sous espaces $S = \left\{ s_i | s_i = s_{\overline{x}, \phi_{i,k_i}}, i = 1...s \right\}$
où $\overline{x_i} = \frac{1}{m_i} \sum_{x \in P_i} x$ est la moyenne
ϕ_{i,k_i} est la matrice des k_i vecteurs propres de C_i et
les $C_i = \frac{1}{m_i} \sum_{x \in p_i} (x - \overline{x})(x - \overline{x})^T$ correspondent aux k_i plus importantes valeurs propres.

5.6.3 Évaluation de la qualité de la classification automatique

Les techniques classiques pour mesurer, comparer et évaluer les systèmes de classifications automatiques reposent sur la qualité de partition ou de regroupement (classes obtenues, éléments par classe, moyenne et écart-type des classes obtenues).
Pour évaluer l'homogénéité des objets dans une classe, on peut utiliser la moyenne ou l'écart-type $\sigma = \sqrt{V}$ pour exprimer la dispersion dans chaque classe :

$$V = \sigma^2 = \frac{1}{c}\sum_{k=1}^{c}(card(C_k) - moy)^2 \tag{5.13}$$

où $moy = \frac{1}{c}\sum_{k=1}^{c} card(C_k)$ est le nombre moyen d'éléments par classe et c est le nombre de classes obtenues.

Comme notre problématique relève du domaine de la recherche d'information, nous favorisons la mesure du rappel vs. la précision qui est communément utilisée pour évaluer la qualité de nos solutions :

- le rappel est la portion d'objets d'une classe et d'un type donnés par rapport au nombre total d'objets de ce même type dans la base. Il exprime le pourcentage des éléments d'un type T_j dans une classe donnée. Sa définition est la suivante :

  $$Rappel(C_i, T_j) = \frac{n_{ij}}{nt_j} \in [0, 1] \tag{5.14}$$

  où n_{ij} est le nombre d'éléments de la classe C_i qui sont de type T_j
  et nt_j est le nombre total d'objets de type T_j dans la base

- la précision exprime la part d'éléments d'une classe C_i qui ont bien été regroupés à l'intérieur d'un même groupe. Sa définition est la suivante :

  $$Precision(C_i, T_j) = \frac{n_{ij}}{nc_i} \in [0, 1] \tag{5.15}$$

  où nc_i est le nombre total d'objets dans la classe c_i. Ces deux mesures permettent de caractériser à la fois la pureté d'un groupe (précision) mais également la qualité d'éléments oubliés d'une classe (1-Rappel).

Nous présentons les résultats de la classification en termes de nombre d'éléments bien interprétés par rapport au nombre total d'objets dans la base.

5.7 Conclusion

D'un point de vue général, notre concept est centré autour de la représentation vectorielle d'objets, le SKCS, à l'aide de ses degrés de libertés, s'avère un opérateur puissant quant à la détection des régions informationnelles en général et le repérage des zones non textuelles objet de cette partie. Le processus de fusion permet de réduire le nombre d'objets avant d'introduire la phase de classification pour faciliter l'interprétation. Le chapitre a introduit le concept de classes de caractéristiques en tant que cadre général de la classification d'objets. Ces classes sont définies par des similitudes des vecteurs caractéristiques des objets fondées seulement sur des mesures caractérisant les textures et les formes des objets extraits.

Notre argument est alors que le concept de classes d'objets constitue un cadre puissant pour l'indexation d'objets. Le chapitre a proposé un algorithme de regroupement selon la distance entre les objets qui détermine une répartition grossière des objets. Un deuxième algorithme basé sur l'approche MKL décompose les éléments de chaque classe en sous-espaces fidèles aux regroupements des vecteurs lors de la projection dans l'espace des caractéristiques. Cette opération augmente la précision de notre système d'interprétation et réduit la dimensionalité de chaque sous-espace pour ne tenir compte que des caractéristiques les plus importantes intrinsèques aux sous-classes.

Algorithme 7 : Algorithme de la transformation Karhunen-Loeve

transformation KL

Estimer les paramètres σ^2, ν et Φ de la transformation KL sur l'échantillon $[x_1, x_2, \ldots, x_u]^T$ assurant que l'erreur de reconstruction est $< \epsilon$.

u	Taille de $[x_1, x_2, \ldots, x_u]^T$
d	Dimension des vecteurs de départ
k	Dimension des vecteurs après réduction
ϵ	Seuil de reconstruction

debut

Calculer la moyenne empirique $\bar{x} = \frac{1}{u}\sum_{i=1}^{u} x_i$.

Calculer la matrice de covariance $C = \frac{1}{u}\sum_{i=1}^{u}(x - \bar{x})(x - \bar{x})^T$.

Décomposer la matrice C en $\phi^T C \phi = D$

- ϕ est la matrice orthonormale des vecteurs propres

- D est la matrice diagonale des valeurs propres λ_i, avec $\lambda_1 \geq \ldots \geq \lambda_r$),

$(1 \leq i \leq \min(d, u))$.

pour $j = 1$ ***à*** $d - 1$ **faire**

ϕ_j est la matrice des premiers j colonnes de ϕ.

Calculer l'erreur de reconstruction $\varrho = \frac{\sum_{j+1}^{u} \lambda_j}{\sum_{i=1}^{u} \lambda_i}$.

si $\varrho < \epsilon$ **alors**

$k = j$

$\Phi_k = \Phi_j$

$\sigma^2 = \frac{1}{d-k} \sum_{i=k+1}^{d} \lambda_i$

$j = d - 1$.

fin

Algorithme 8 : Algorithme Multi-space KL (Cappelli et al., 2001).

u nombre de vecteurs dans l'espace

k dimension des vecteurs après réduction

d_{FS} distance entre un vecteur $x \in \Re^d$ et un sous espace $S_{\bar{x},\phi k}$

begin

P1 = P % Initialisation des sous espaces %

pour $i = 2$ ***à*** s **faire**

- Appliquer KL à P1

- **if** $k_i > kmax$ **then**

  $k_{max} = k_i$

- Sélectionner dans Q_i les $\lfloor m/s \rfloor$ objets les plus proches de l'espace $S_{\bar{x}_i,\phi k_i}$

(m est le nombre de vecteurs de l'espace et s le nombre de sous-espaces à créer)

$$Q_i = \left\{ x_{i1}, x_{i2}, \ldots, x_{ih} | d_{FS}(x_{ir}, S_{\bar{x}_i,\phi_{k_i}}) \leq d_{FS}(x_{it}, S_{\bar{x}_i,\phi_{k_i}}) \forall x_{ir} \in Q, \forall x_{it} \notin Q \right\}$$

- $P_1 = P_1 - Q_i$

- $P_i = Q_i$;

$\rho = P_1, P_2, \ldots, P_s$; % Optimisation des sous espaces %

$iter = 0$;

répéter

$\rho_{old} = \rho$; $P_1 = P_2 = \cdots = P_s$;

pour $i = 1$ ***à*** u **faire**

Déterminer le sous espace S_t le plus proche de x_i :

$MKL(x_i, S) = \langle t, y \rangle$

$t = argmin_{j=1\cdots s}(d_{FS(x_i,S_j)})$ et $y = KL(x_i, S_t)$ avec $y \in \Re^{k_t}$

$P_t = P_t \bigcup x_i$

Balancer $\rho = P_1, P_2, \ldots, P_s$ pour que chaque P_i contienne $h = \lfloor m/s \rfloor$ objets ;

$iter = iter + 1$;

jusqu'à $\varrho = \varrho_{old} \vee iter = iter_{max}$;

end

CHAPITRE 6

APPROCHE HYBRIDE POUR LA RECHERCHE D'INFORMATION

L'indexation est le processus de représentation des données en vue de faciliter la recherche. La performance d'un système de recherche d'images de documents dépend étroitement de l'indexation qui doit permettre de retrouver toute l'information véhiculée dans le modèle de représentation. La mesure de la similarité permet de retrouver les images de documents pertinentes à une requête de l'utilisateur.

Ce chapitre décrit la méthodologie proposée pour mettre en oeuvre un modèle fiable et performant de recherche de différents types d'information dans les images de documents.

6.1 Définition de la recherche d'images

La plupart des systèmes de recherche proposent d'utiliser des mots clés pour étendre automatiquement les requêtes qui portent sur des images. La correspondance mots-clés images pour la recherche d'information est en effet un processus itératif : les utilisateurs ont des difficultés à exprimer précisément leurs besoins dès la première requête. La requête est donc modifiée itérativement [1]. Mais formuler de nouvelles requêtes n'est pas facile. Une technique [2] consiste à demander à l'utilisateur de sélectionner, parmi les documents retournés, les documents se rapprochant des plus pertinents. Certains mots de ces documents pertinents sont alors utilisés par le système pour générer une nouvelle requête.

C'est dans ce contexte que l'article (Jing et al., 2005) publié en juillet 2005 propose une architecture unifiée pour la recherche d'images basée sur des mots clés et des caractéristiques visuelles. La requête est formulée par des mots clés ou par une image exemple. Le jugement de pertinence des utilisateurs aux images retournées par le système est combiné

[1] Ce principe est connu sous le terme "query expansion"

[2] Connue sous le terme "relevance feedback"

aux caractéristiques visuelles des images pour représenter des concepts sémantiques à utiliser pour propager les mots-clés à d'autres images non étiquetées. L'algorithme établi des relations entre des mots-clés et les caractéristiques visuelles des images utilisant deux modèles. Le premier formule, à l'aide d'experts, des probabilités de correspondances entre les mots-clés et les images de la base. Le deuxième provient du modèle d'apprentissage utilisant les documents retournés comme réponses et jugés par l'utilisateur. Ces concepts améliorent la recherche et les réponses aux requêtes basées sur des images exemples. Toutefois, ces approches présentent des inconvénients. Elles sollicitent tout d'abord l'avis de l'utilisateur sur la pertinence des documents, ce qui demande un effort cognitif non négligeable et peut conduire à de mauvaises décisions. Les décisions peuvent être d'autant plus mauvaises que la majorité des systèmes suscitent des décisions binaires : un document est pertinent ou non.

Dans notre cas, les requêtes doivent permettre aux utilisateurs de rechercher les images contenant un type d'objet ou des caractéristiques bien définis (forme circulaire ou image exemple). Dans la littérature, les mots clés proviennent généralement d'un expert ou de la légende des graphiques. Dans notre approche, les images de documents produisent des mots provenant du texte reconnu par OCR et des objets non textuels résultant de la segmentation. Une fusion des mots-clés et des caractéristiques des objets devrait augmenter la puissance d'interrogation et répondre à des requêtes portant sur des mots-clés ou des images exemples. La structure d'indexation doit localiser un groupe d'objets similaires, prédire le type de l'image exemple et répondre en partie aux requêtes portant sur des objets non textuels. Pour cela, nous allons définir la notion d'indexation par le contenu et proposer un modèle vectoriel basé sur une structure d'indexation multi-niveau et arborescente.

6.2 Modèle d'indexation textuel

Pour les régions textuelles reconnues par OCR, nous avons présenté dans le quatrième chapitre une représentation vectorielle qui s'est avérée efficace et que nous résumons ici. Soient $D = \{d_1, d_2, \ldots, d_i, \ldots, d_m\}$ un ensemble d'images de documents (le corpus) et $T = \{t_1, t_2, \ldots, t_j, \ldots, t_n\}$ un ensemble de mot-clés indexant ces images, le modèle vectoriel pour les éléments textuels représente un document d_i et une requête q par un vecteur dans un espace à n dimensions :

$$\vec{d_i} = (\omega_{1,i},\ \omega_{2,i},\ \ldots,\ \omega_{j,i},\ \ldots,\ \omega_{n,i}) \tag{6.1}$$

$$\vec{q} = (\omega_{1,q},\ \omega_{2,q},\ \ldots,\ \omega_{j,q},\ \ldots,\ \omega_{n,q}) \tag{6.2}$$

où $\omega_{j,i}$ est le poids du terme t_j dans le document d_i et $\omega_{j,q}$ est le poids du terme t_j dans la requête q. La formule la plus classique pour calculer le poids est le tf-idf présenté dans la section 4.5.2.

La mesure de similarité présentée au paragraphe 4.5.3 et qui correspond au cosinus de l'angle formé par les vecteurs $\vec{x}$ et $\vec{q}$ retourne une liste L_0^t d'images de documents considérées comme pertinentes par le système. Le vecteur $\vec{L_0^t}$ associé à cette liste contient le degré de pertinence attaché aux textes OCR des images dans le modèle textuel.

6.3 Définition de l'indexation par le contenu

Les images sont souvent indexées manuellement en leurs affectant des mots clés. Mais l'indexation manuelle des images est une tâche fastidieuse et nécessite un temps non-négligeable. De plus, les résultats des interrogations dépendent de l'ensemble des mot-clés disponibles et de la subjectivité humaine. Contrairement aux documents textuels, l'image ne porte pas de sémantique directement accessible à la machine. C'est pourquoi depuis les années 90, (Kanungo et al., 2002; Chang, 2001; Hull, 1996; Muller et Rigoll, 1999; Wang et Srihari, 1989b; Jain et Yu, 1998) de nombreux travaux de recherche ont été menés pour développer l'indexation automatique d'images par le contenu. La difficulté principale est

d'extraire des descripteurs textuels et visuels suffisamment significatifs pour permettre de retrouver la sémantique associée à l'image. Pour qu'un système de recherche d'images soit performant, il faut que l'indexation logique soit pertinente et que l'indexation physique permette un accès rapide aux documents recherchés. Nous allons définir davantage ces deux concepts dans les deux sections suivantes.

6.3.1 Indexation logique

L'indexation logique consiste à extraire et à modéliser les caractéristiques de l'image qui sont principalement la forme, la couleur et la texture. Chacune de ces caractéristiques pouvant être considérée pour l'image entière ou pour une région de l'image (localisation spatiale et segmentation en régions d'intérêt). Les caractéristiques les plus discriminantes pour les images de documents sont :

- la forme : les techniques de modélisation sont classées en deux catégories. L'approche «contour» décrit une région au moyen des pixels situés sur son contour. L'approche «région», que nous utilisons, considère une région par rapport aux caractéristiques des pixels que cette région contient
- la texture : une texture peut être caractérisée par les attributs de contraste, de direction, de régularité et de périodicité du motif. Dans le cadre de la recherche par le contenu, elle permet de distinguer des zones de textures similaires, mais de sémantique différente (par exemple, l'ellipse du logo et l'ellipse du texte)
- les mots-clés : ces mots sont considérés comme indices textuels. Ils sont affectés aux images manuellement ou proviennent de leurs légendes.

Les principaux systèmes actuels de recherche d'images par le contenu sont QBIC (Niblack, 1993), VisualSeek (Smith et Chang, 1996), SurfImage (Nastar et al., 1998b) et NeTra (Ma et Manjunath, 1997). Ils se basent principalement sur les caractéristiques visuelles, et n'utilisent que peu ou pas les indices textuels. D'autres systèmes de recherche

(Chang, 2003; Jing et al., 2005) tentent de combiner les mots-clés et les caractéristiques visuelles pour étendre les mots-clés aux images non étiquetées.

6.3.2 Indexation physique

Nous traitons des images de documents composites et notre mode d'indexation et de recherche doit considérer une variété de types de contenus. Contrairement aux images simples traitées par les systèmes pré-cités dans le paragraphe 6.3.1, une image de document est composite et peut contenir une variété de régions informationnelles textuelles ou non. De nombreuses techniques basées sur des arbres ou des vecteurs ont été proposées, mais ces techniques souffrent de faiblesses dues notamment à la multi-dimentionnalité de l'indexation logique (recherche sur plusieurs caractéristiques à la fois : forme, texture,...) et au grand volume de données. C'est pourquoi nous regroupons itérativement les objets non textuelles proches afin de finaliser l'interprétation, réduire l'espace de recherche, accélérer la recherche et améliorer la précision.

Nous allons présenter dans les sections suivantes les différents modèles et la structure d'indexation ainsi que les algorithmes de recherche et les mesures à appliquer pour traiter les zones graphiques.

6.4 Modèle d'indexation non textuel proposé

Les travaux menés consistent à extraire des caractéristiques de formes et de textures et à regrouper les images similaires. Chaque image est représentée de façon à permettre sa comparaison avec les autres. Elle est caractérisée par un vecteur dont les éléments correspondent aux descripteurs représentés par une valeur, booléenne ou numérique, qui tente de caractériser son apparition.

6.4.1 Informations représentées

À chaque objet extrait de l'image est associée un ensemble d'attributs de bas niveau (forme, texture ...). Chaque attribut étant décrit par un ensemble de descripteurs (surface, rectangularité, circularité, moments, entropie etc). La méthode proposée pour l'adaptation de la classification automatique à la problématique de recherche d'images se distingue par son originalité et sa simplicité : elle est basée sur la sélection d'un ensemble d'apprentissage, tiré aléatoirement des bases d'images utilisées, pour effectuer une analyse statistique du comportement de tous les paramètres à définir.

Nous allons tester l'influence de certains paramètres comme le nombre de classes à considérer et les caractéristiques discriminantes pour différencier les objets présents et faciliter la recherche. Le problème pour les images de documents consiste à déterminer la manière de passer de la notion de fréquence d'apparition des mots clés, comme c'est le cas pour le texte OCR, à une représentation des descripteurs pour une éventuelle combinaison.

En définitif, pour un ensemble de descripteurs et pour une collection d'images, une matrice $Descripteurs \times Images$ est formée par la juxtaposition des vecteurs associés à tous les descripteurs. Celle-ci jouera le rôle de la matrice $Termes \times Documents$ habituellement utilisée pour la recherche d'information textuelle.

Pour mesurer la similarité entre un objet x et une requête q représentés par des vecteurs multi-dimensionels $\vec{x} = (x_1, x_2, \cdots, x_n)$ et $\vec{q} = (q_1, q_2, \cdots, q_n)$, on a coutume de prendre une des distances L_p présentées à la section 5.5.3.

6.4.2 Indexation multi-niveau

Les images de documents sont composites et la distance entre l'objet requête I_q et les objets de l'image I_i doit considérer les aspects textuels et les caractéristiques visuelles de l'information véhiculée. Nous présentons le contexte d'évolution et discutons de l'indexa-

tion multi-niveau et de son apport à la recherche d'information. Ensuite, nous proposons des mesures considérant les informations textuelles et les vecteurs de caractéristiques visuelles.

Contexte de l'étude

La performance en terme de temps d'exécution d'une requête basée sur la similarité d'objets est fortement dépendante du nombre d'objets à comparer, puisque l'algorithme de base consiste à calculer la distance de l'objet requête avec toutes les images de la base. Les approches de la classification automatique (ou clustering) et d'indexation améliorent ces performances.

Le principe est de découper l'espace et d'indexer les différents objets de l'espace obtenu, à l'aide d'une structure arborescente. L'objectif est de diminuer le nombre d'objets à comparer et par conséquent le nombre de distances à calculer entre descripteurs. La structure d'index permet l'accès à un sous-ensemble d'objets de la base les plus proches de ceux de la requête.

Pour cela, nous répartissons les objets extraits en sous espaces homogènes sur la base des caractéristiques discriminantes pour chaque espace et construisons les index sur l'espace ainsi transformé.

Concept logique de l'indexation multi-niveau

Après avoir défini la représentation et les mesures de distances entre deux objets, nous allons procéder à une indexation multi-niveaux pour représenter les résultats obtenus par les approches K-moyennes et MKL présentées au chapitre précédent.

Dans la plate forme générale de la figure 16, Les objets textuels sont traités par un modèle vectoriel basé sur le texte OCR. Quant aux parties non textuelles, les regroupements d'objets en classes homogènes produisent des classes représentées par des noeuds dans notre

structure logique. Chaque classe est localisée par son vecteur moyen visuel $\vec{C_k^v}$. La classe visuelle d'un objet I_{ij} de l'image i et de vecteur caractéristique $\vec{I_{ij}^v}$ est déterminée par la fonction objective :

$$C(I_{ij}) = \text{argmin}_{k\in\{1,2,\ldots,c\}} d(\vec{I_{ij}^v}, \vec{C_k^v}). \tag{6.3}$$

où d est la distance euclidienne telle que définie au paragraphe 5.6.1.

La classification hiérarchique possède deux propriétés intéressantes. Tout d'abord, l'usager ne doit pas définir le nombre de groupes à obtenir car l'utilisateur a du mal à évaluer le nombre de concepts ou de types de contenus présents dans l'image. Ensuite, la méthode induit naturellement une hiérarchie entre les groupes d'objets. Pour leur défense, notons que les méthodes de partitionnement peuvent être appliquées récursivement sur chaque groupe créé à l'étape précédente. Quelle que soit la méthode envisagée, une organisation hiérarchique de documents est utilisable si la hiérarchie n'est pas trop profonde. On remarque que beaucoup de travaux montrent que les résultats ne sont généralement pas entièrement satisfaisants aux yeux des utilisateurs, c'est pour cela que le jugement de ces derniers et des mesures de performance sont indispensables pour remédier aux erreurs suivantes :

- un objet a été rangé par erreur
- deux groupes auraient dû être fusionnés.

La première erreur altère tout d'abord la compréhension de la classification produite. Elle force ensuite l'utilisateur à remettre en question les résultats proposés par la machine. La seconde erreur est moins grave puisque l'utilisateur poursuit le travail de classification en fusionnant manuellement deux groupes qu'il juge similaires. Contrairement au travail impliqué par le premier type d'erreur, l'usager ne doit pas parcourir la hiérarchie en profondeur mais simplement identifier, grâce à leurs noms, deux classes similaires qui auraient dû être regroupées.

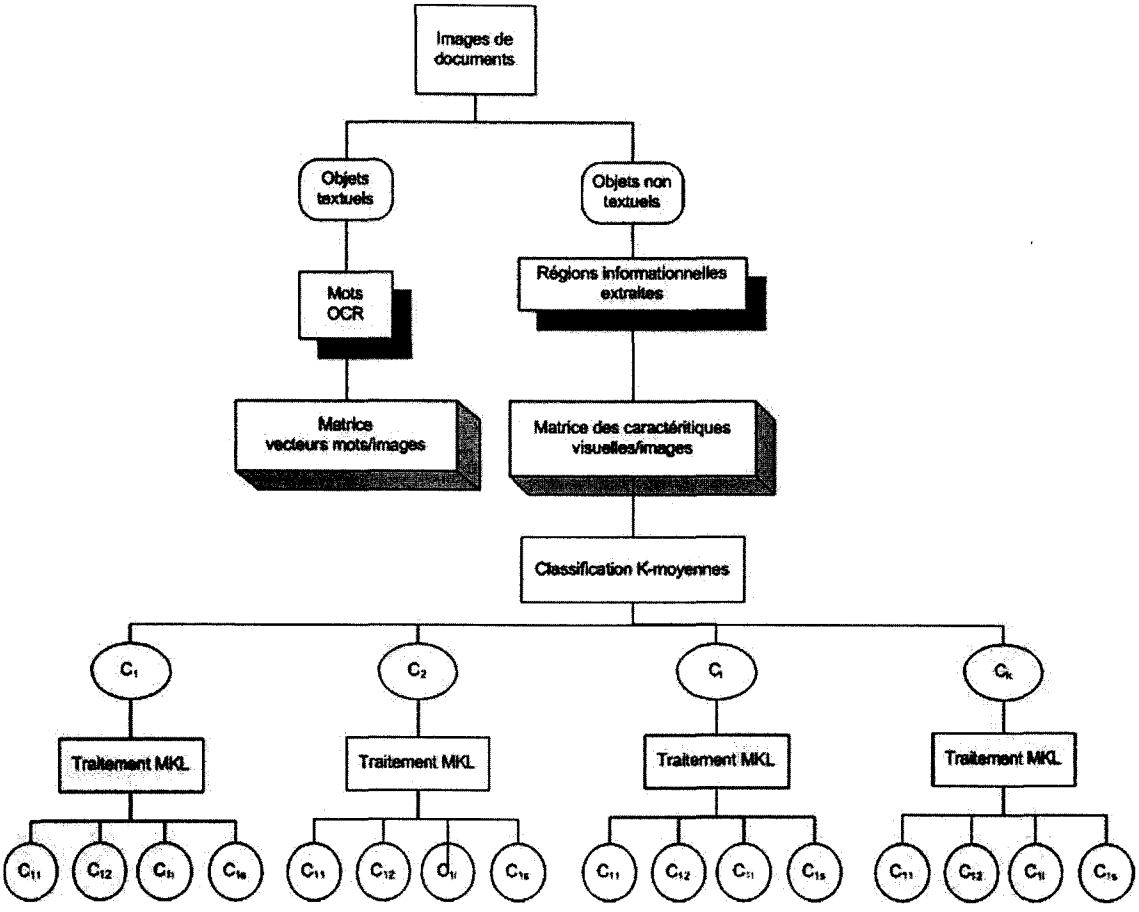


Figure 16 Concept de l'indexation multi-niveau

Structure physique de l'indexation multi-niveau

Notre indexation multi-niveau a une structure arborescente dont les feuilles représentent les objets des images de la base et les noeuds les différents regroupements. Le premier niveau donne les informations relatives aux classes obtenues par l'algorithme k-moyennes (centres des classes, rappel vs. précision pour les objets de la classe, etc.). Le deuxième niveau informe sur les résultats obtenus par la subdivision des classes du premier niveau, on y trouve les transformations dûes à la réduction de la dimensionnalité et les centres des sous-espaces.

Dans le but de simplifier la formulation, nous nous limiterons, dans la suite, à des descripteurs de trois niveaux. Le premier représente les classes obtenues par K-moyennes, le second les sous-espaces générés par MKL et le dernier les feuilles de l'arbre qui sont les objets extraits des images de documents de la base. Néanmoins, notre approche peut se généraliser à des images représentées par des vecteurs stockés dans des arbres de plus de trois niveaux.

La figure 17 montre que les noeuds de la structure sont représentés par un descripteur dans un espace à 4 classes k-moyennes et 4 sous-espaces MKL. Le premier niveau d'index de la figure 17 est représenté par 4 vecteurs et le second niveau par 16 vecteurs d'index. À chaque noeud résultant de la classification k-moyennes est associé un n-uplet [rappels/précisions, centre, distance maximale, pointeurs]. Ces noeuds sont subdivisés en quatre sous-espaces (index de deuxième niveau) ; ils sont représentés par les n-uplets [rappels/précisions, centre, distance maximale, mode de transformation, nombre d'objets de cette classe et un pointeur vers le premier objet]. Les feuilles de l'arborescence sont les vecteurs caractéristiques des objets non textuels de l'image. Chaque objet de l'image étant représenté par son vecteur de caractéristiques.

Après avoir défini les bases d'indexation de l'information graphique, nous présentons au paragraphe suivant le modèle de combinaison texte-descripteurs et au paragraphe 6.6 les algorithmes de recherche des images pertinentes.

6.5 Modèle de combinaison texte-descripteurs

De part sa nature, l'image de document est composite et son mode d'accès privilégié reste le texte. L'apport des indices visuels n'est pas à négliger surtout que la requête peut porter sur des logos, des graphes ou des illustrations. L'indexation proposée au paragraphe 6.4 est un processus de représentation sous forme d'index du contenu des images. L'opération combinant le texte et les vecteurs caractéristiques implique l'intervention d'un processus capable d'extraire automatiquement ces termes et ces objets avec une précision fortement

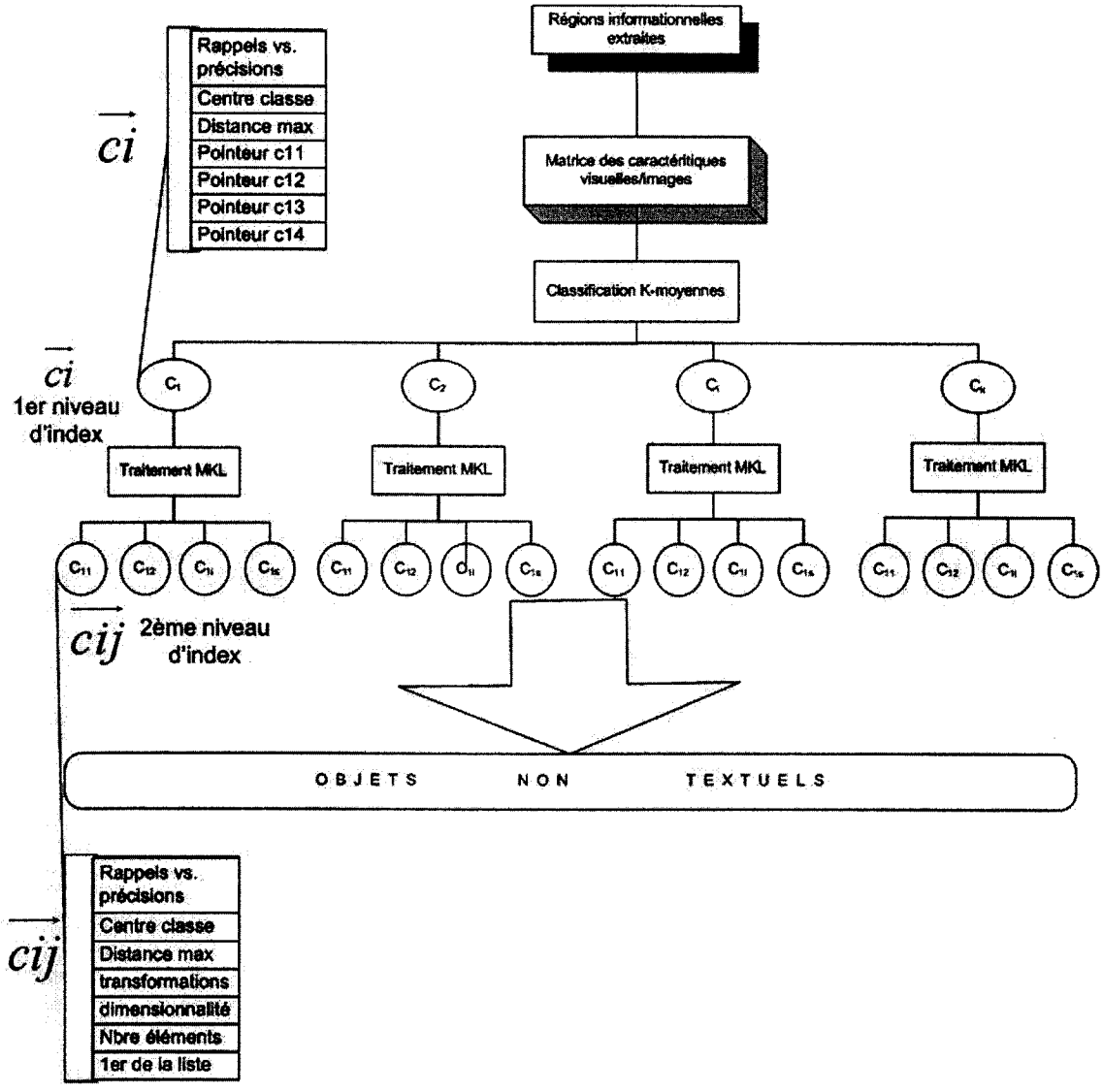


Figure 17 structure physique de l'indexation multi-niveau

liée à la méthode employée et à la finesse des informations extraites. Des méthodes d'indexation fondées sur le texte et les caractéristiques associés aux objets se développent avec la mise en oeuvre de processus d'analyse fondé sur des comparaisons de vecteurs. La recherche d'information met en évidence un contraste entre les termes employés par les utilisateurs dans le cadre de leur formulation de requête et les termes susceptibles de « répondre » à cette requête. Il y a donc une distinction réelle entre la pertinence du système

et la pertinence de l'utilisateur. L'amélioration de cette pertinence passe avant tout par une adéquation entre les termes utilisés pour questionner un système ou une base de connaissance et les termes réellement présents dans les documents capables de véhiculer les idées présentes dans la requête.

On a vu que la mesure appropriée au contenu textuel est le cosinus, alors que la distance L_p est la plus utilisée pour les descripteurs de textures et de formes. Nous adoptons comme formule la combinaison convexe des mesures textuelles et visuelles donnée par :

$$d(I_i, I_q) = \lambda sim^{mots}(I_i, I_q) + (1 - \lambda) L_p^{descripteurs}(I_i, I_q) \tag{6.4}$$

où λ détermine la contribution de chaque type de contenu et défini à partir du contenu de la requête. Ainsi $\lambda = 1$ implique une requête textuelle et une recherche basée sur les caractéristiques textuelles seulement, $\lambda = 0$ pour les requêtes par l'image exemple et $\lambda = 0.5$ pour les requêtes combinant texte et caractéristiques visuelles.

Ce modèle est à retenir pour la combinaison des indices textuels et visuels surtout que chaque image de document ou même l'image requête est composée de plusieurs objets. Chaque objet a ses propres caractéristiques et appartient à un type bien déterminé (texte, logo, illustration ou graphe).

Le paragraphe 6.5.1 présente les étapes d'initialisation et les traitements relatifs aux requêtes, le paragraphe 6.5.2 est une panoplie de mesures portant sur les informations graphiques des images et le paragraphe 6.5.3 combine les listes textuelles et graphiques des images pertinentes à la requête d'un utilisateur.

6.5.1 Initialisation et traitement de la requête

Les vecteurs représentatifs de l'information véhiculée dans notre base d'images doivent être constitués au préalable, c'est pourquoi nous construisons hors ligne deux matrices $mots \times images$ et $descripteurs \times objets$.

La requête par une image exemple doit subir la segmentation par l'opérateur SKCS pour extraire les objets à confronter à ceux de la base. Les opérations d'initialisation portant sur les objets des images de la base et les éléments de la requête sont les suivantes :

- la représentation de chaque classe $k \in \{1, 2, \ldots, c\}$ de premier niveau par son vecteur $\vec{C_k}$
- la représentation de chaque classe $l \in \{1, 2, \ldots, s\}$ de deuxième niveau par son vecteur $\vec{C}_{k,l}$
- la construction, à partir des mots de la requête, du vecteur $\vec{q_0}$
- l'extension des mots de la requête pour constituer le vecteur $\vec{q^t}$
- l'extraction des objets de l'image exemple de la requête et de leurs caractéristiques $\vec{q}^v_{ij}$.

Pour le traitement textuel, nous avons vu dans le chapitre 4 qu'à chaque image I_i est associée une distance textuelle $sim^t_{I_i}(q^t)$ qui est le cosinus de l'angle formé par les vecteurs de la requête q^t et celui relié aux mots OCR de l'image. Les images considérées comme pertinentes à la requête sont classées par ordre de similarité dans la liste L_0. La section suivante traite des mesures à prendre pour constituer la liste des réponses pertinentes à partir des zones graphiques.

6.5.2 Traitement des zones graphiques

À partir des caractéristiques de formes et de textures extraites des régions localisées dans l'image exemple de la requête, on détermine pour chaque région la classe d'appartenance de premier niveau, celle qui minimise la distance :

$$C^{kmin}(q_j) = \text{argmin}_{k\in\{1,2,\ldots,c\}} d^v(\vec{q_j}, \vec{C_k}) \tag{6.5}$$

Ensuite, on repére la classe d'appartenance de deuxième niveau à l'aide de la formule :

$$C^{lmin}(q_j) = \text{argmin}_{l\in\{1,2,\ldots,s\}} d^v(\vec{q_j}, \vec{C}_{kmin,l}) \tag{6.6}$$

On applique la transformation MKL décrite par $\phi_{kmin,lmin}$ du noeud $C_{kmin,lmin}$ au vecteur $\vec{q_j}$ pour obtenir le vecteur $\vec{q_j}'$ réduit de la requête (projection du vecteur de la requête $\vec{q_j}$ sur le sous-espace induit).

À la fin, nous estimons la nature de l'objet requête que nous traitons. Il s'agit de savoir si l'on est en présence d'un logo, d'une photographie ou d'un graphe. Chaque noeud sélectionné retourne le rappel vs. la précision pour chacun des objets. La combinaison de ces différentes mesures nous permet de calculer le score global en terme de probabilité d'appartenance de l'image requête q^v au type d'image $type$ par la formule :

$$P^v_{type}(q^v) = \alpha \frac{\sum_{i=1}^{j} R_i^{type}}{\sum_{type}\sum_i R_i^{type}} + \beta \frac{\sum_{i=1}^{j} Pn_i^{type}}{\sum_{type}\sum_i Pn_i^{type}} \tag{6.7}$$

où "$type$" correspond au logo, photographie ou graphe, j est le nombre d'objets localisés dans l'image exemple de la requête, R_i^{type} et Pn_i^{type} sont respectivement le rappel et la précision obtenus pour le type d'objet "type" aux différents objets i de l'image exemple de la requête.

α et β déterminent l'importance à donner aux rappels versus précisions. Il y a plusieurs méthodes pour calculer une moyenne d'un ensemble de rappels et de précision. La définition (et donc le calcul) des moyennes peut être synthétisée et généralisée à l'aide de la formule unique suivante :

$$\bar{x}(m) = \sqrt[m]{\frac{1}{n}\sum_{i=1}^{n} x_i^m} \tag{6.8}$$

où l'on retrouve :

* pour $m = 1$, la moyenne arithmétique

* pour $m = 2$, la moyenne quadratique

* pour $m = -1$, la moyenne harmonique

* lorsque $m \to 0$, la limite de $\bar{x}(m)$ est la moyenne géométrique

Si a et b sont deux réels positifs tels que a < b, alors on a :

$$a < M_{harmonique}(a,b) < M_{geometrique}(a,b) < M_{arithmetique}(a,b) < M_{quadratique}(a,b) < b \tag{6.9}$$

La F-mesure que nous adoptons est l'indicateur de synthèse communément utilisé pour évaluer les algorithmes de recherche d'information. La F-mesure (F) permet de combiner les deux mesures précédentes : c'est la moyenne harmonique du Rappel et de la Précision. Elle est définie par :

$$F = \frac{2RPn}{R + Pn} \tag{6.10}$$

6.5.3 Combinaison texte - non-texte

Pour formuler la fusion des résultats obtenus à partir du texte et ceux déduits des objets de l'image exemple, on affecte les numéros 1 à j aux listes L_j obtenues à partir des caractéristiques des objets de la requête et le numéro 0 à la liste textuelle L_0 obtenue à l'aide des mots de la requête. Pour uniformiser le contenu de toutes les listes, une normalisation des mesures de similarité est obtenue par les formulations suivantes :

- normaliser la similarité des images de la liste textuelle L_0 :

  $$P^t(I_i \in L_0) = \frac{sim_{L_0}(\vec{I_i^t})}{\sum_n sim_{L_0}(\vec{I_n^t})}.$$

  où n représente le nombre d'images dans la liste et I_n^t la $n^{ème}$ image de la liste textuelle L_0.

- normaliser la similarité des images des listes relatives aux objets non textuels L_j :
  Soit $interv = dmax(L_j) - dmin(L_j)$ la différence entre la distance maximale et minimale des éléments de la liste L_j ;

  $$P^v(I_i \in L_j) = \frac{\frac{dmax(L_j) - d_{L_j}(\vec{I_i^v})}{interv}}{\sum_n \frac{dmax(L_j) - d_{L_j}(\vec{I_n^v})}{interv}}.$$

où n représente le nombre d'images dans la liste.

I_i^v la $i^{ème}$ image de la liste visuelle L_j.

$d_{L_j}(\vec{I_i^v})$ est la similarité de l'image I_i (distance euclidienne par rapport au vecteur de la requête) dans la liste L_j.

La pertinence globale combinant les listes textuelles et non textuelles $P^{v \vee t}(I_i)$ d'une image I_i de la liste L_j est calculée par :

$$P^{v \vee t}(I_i|type) = P^t(I_i \in L_0) \cdot \delta(I_i \subset L_0) + \sum_{k=1}^{j} \omega(L_k|type) \cdot P^v(I_i \in L_k) \cdot \delta(I_i \subset L_k) \tag{6.11}$$

où

$$\delta(I_i \subset L_k) = \begin{cases} 1 & \text{si } I_i \subset L_k \\ 0 & \text{sinon} \end{cases}$$

et $\omega(L_k|type)$ est le taux de rappel vs. précision présent au noeud relatif à la liste L_k. Sa valeur est déterminée lors de la confrontation des résultats de la classification à la vérité terrain fournie avec la base d'images utilisée.

6.6 Algorithme de recherche hybride

La recherche d'information dans les images de documents porte sur des mots ou des images requêtes. Nous avons présenté dans le paragraphe 4.5.3 la mesure de la similarité textuelle ainsi que les mesures de performance correspondantes. Nous discutons dans cette section de l'algorithme de recherche considérant l'information textuelle et non textuelle des images.

À chaque image va correspondre un vecteur textuel et des vecteurs de caractéristiques visuelles. La fusion des réponses obtenues sur la base des mots du texte et celles des objets non textuels permet d'intégrer aux requêtes des images exemples.

Un exemple est la recherche d'un logo dans une image publicitaire partant d'un logo prototype et d'un nom de référence. Deux fonctions principales sont à développer. La première est la recherche des images contenant le nom de référence, une extension des termes de la requête et une recherche basée sur le modèle vectoriel telles que présentées dans le chapitre 4 retournent une liste d'images de documents classée par ordre de similarité.

La deuxième fonction recherche dans l'arborescence les neuds dont les vecteurs d'index sont proches des caractéristiques du logo prototype. L'algorithme 9 parcourt en profondeur l'arbrorescence en choisissant à chaque niveau d'index le noeud dont le vecteur est le plus proche. Les taux de précisions vs. rappels provenant des noeuds sélectionnés et les distances des images du noeud par rapport à la requête affecteront les mesures de pertinences globales des images sélectionnées.

6.6.1 Modalités de recherche

La plate forme générale de recherche d'images de documents pertinentes à une requête repose sur cinq étapes qui sont :

- l'indexation des données présentes dans la base
- l'extraction de l'information textuelle et visuelle de la requête
- le traitement textuel et visuel pour préparer les listes d'images pertinentes
- la normalisation et la combinaison des listes obtenues
- l'évaluation des réponses du système.

Dans notre approche, chaque image est représentée par un descripteur d'objets, chaque objet pouvant être considéré comme une image à part entière. Par conséquent, une image requête peut être comparée non seulement aux images entières, en appliquant une recherche globale, mais également en considérant chaque objet comme une image de la base. L'algorithme pour la recherche d'images de documents pertinentes est répété pour chaque

objet. Pour comparer les objets de l'image requête aux objets de la base, lorsque l'image exemple est composée de plusieurs objets, la requête est décomposée en sous-requêtes, une sous-requête par objet. La recherche accède aux noeuds de l'arbre proches des objets qui composent la région requête. L'algorithme s'applique donc à chaque objet, créant ainsi plusieurs ensembles résultats d'identificateurs d'image. Par conséquent, pour chaque image dont l'identificateur apparaît dans l'union des ensembles, une distance est calculée, en accédant aux descripteurs de chaque objet. L'ensemble résultat final contient les identificateurs des images telles que les distances sont inférieures à un seuil donné.

L'algorithme 9 est un ensemble de modules traitant des aspects textuels et des caractéristiques des images de la base et des images exemples des requêtes. Les mots de la requête sont intégrés dans un traitement textuel qui retourne une première liste des images pertinentes. Aussi, les caractéristiques des objets non textuels de la requête sont utilisées pour parcourir l'arborescence et sélectionner les noeuds proches. Ces derniers sont représentés par des vecteurs d'index qui pointent sur une liste d'images susceptibles de correspondre à celle de la requête. Les mesures de similarité dans les différentes listes sont normalisées et fusionnées pour constituer une liste de pertinence globale répondant à la requête de l'utilisateur.

6.6.2 Amélioration du processus de la recherche

Les images retournées par le système ne sont pas toutes pertinentes pour les utilisateurs à cause de la segmentation, de l'interprétation ou de l'indexation. Il faut alors faire le tri des résultats et ne conserver que ce qui répond effectivement à la requête. Il est parfois difficile d'évaluer une image de document surtout si elle a été repérée dans plusieurs listes. Voici quelques critères qui peuvent aider à faire une bonne analyse et une bonne évaluation.

- L'image est-elle bien indexée ?
- Les caractéristiques sont-elles reliées à l'image ?
- Les mots clés sont-ils représentatifs de la requête ?

Algorithme 9 : Algorithme générale de recherche - traitement du texte et du graphique - Traitement de la requête

- construire hors-ligne les vecteurs et l'arbre d'indexation des images de la base
- segmenter et extraire les objets de l'image exemple de la requête s'il ya lieu
- construire en-ligne les différents vecteurs relatifs à la requête.

Pertinence textuelle

- étendre les mots de la requête
- construire la liste L_0 des images pertinentes de la base par rapport à q.

Pertinence graphique

- **tant que** *il reste des objets* q_j *de la requête* **faire**
  - déterminer la classe d'appartenance de 1er niveau
  - appliquer la transformation MKL aux vecteurs q_j
  - déterminer la classe d'appartenance du 2ème niveau
  - calculer la distance entre les objets de la requête et les objets de la classe sélectionnée :

    pour *chaque objet de la requête* q_j **faire**

    pour *chaque objet du noeud* o_i **faire**

    Calculer la distance $d(o_i, q_j)$
  - construire la liste L_j des images de documents pertinentes à q_j.

Pertinence globale et mesure de la performance

- normaliser les mesures de pertinence dans chacune des listes
- calculer la pertinence globale des images des listes
- calculer le rappel et la précision des réponses du système.

- L'image répond-elle bien ou en partie à la requête ?
- La décomposition des images est-elle claire, précise et facile à comprendre ?

En outre, de nouvelles utilisations des SRI sont apparues. En effet, la diffusion et la gestion de l'information deviennent primordiales, que celle-ci soit recueillie par un ou plusieurs experts, ou par un groupe d'utilisateurs collaborant à une tâche commune. Les évaluations des résultats par les utilisateurs peuvent être intégrées dans la modélisation, que ce soit au niveau de l'indexation, qu'au niveau de l'interrogation. Le système et les utilisateurs collaborant à une recherche commune, interagissent, s'entraident afin d'augmenter la qualité du système. Ce processus d'amélioration est caractérisé par les modules suivants :

- à chaque image pertinente pour le système est associée une évaluation de l'utilisateur : soit (0) non pertinente, (?) pas d'avis et (1) pertinente. Si l'information concernant un critère d'évaluation n'est pas disponible pour l'utilisateur, nous inscrivons un point d'interrogation, « ?», à sa description. Il peut arriver également que l'on ajoute, à ce point d'interrogation, de l'information entre parenthèses dans le cas où l'information n'a pu être vérifiée.
- la formule que nous adoptons pour tenir compte des évaluations des utilisateurs afin de corriger notre indexation et d'améliorer la performance de notre système de recherche est la suivante :

$$d(I_i, C_k) = \alpha \ \ (1 - eval(I_i)) \ \ d(I_i, C_k) \tag{6.12}$$

où $d(I_i, C_k)$ est la distance à mettre à jour et qui est associée aux objets de l'image I_i dans la classe C_k.

α est la pénalisation appliquée aux objets sélectionnés de l'image I_i. Sa valeur dépend des jugements des utilisateurs. Le terme $eval(I_i)$ est l'évaluation de l'image par l'utilisateur.

L'effet sur la structure d'indexation est bénéfique surtout que toute l'information reliée à

l'évaluation des utilisateurs est mise à la disposition des noeuds sélectionnés. Les objets rangés par erreur ou les groupes à défusionner sont repérés par une analyse de l'historique des informations provenant des utilisateurs. Archiver l'historique des distances et les changements opérés pour les objets de l'arborescence corrige l'indexation et améliore la recherche d'images pertinentes.

6.7 Conclusion

Les logiciels capables d'automatiser le traitement d'images de documents (stockage, indexation, extraction d'information, etc.) ne permettent d'automatiser que certaines tâches du traitement de flux de données comme par exemple la lecture dans des zones bien localisées de formulaires, de chèques, etc. Notre approche consiste à utiliser une plate forme basée sur le modèle vectoriel dont les mots-clés et les caractéristiques des objets non textuels sont représentés par des vecteurs. Les résultats de la classification automatique sont une arborescence où chaque noeud est un vecteur représentant un groupe d'objets proches dans l'espace de projection des caractéristiques.

L'avantage de cette représentation est la capacité de fournir une structuration des objets facilement explorable et exploitable par un système de recherche d'information. Cette organisation permet de comparer le contenu de la requête à des représentants de groupes pour ne sélectionner que des parties des images susceptibles de satisfaire à la demande de l'utilisateur. Finalement, la fusion des listes d'images pertinentes obtenus pour chaque partie de la requête et la considération portée aux jugements des utilisateurs sont des valeurs représentant toute l'information menant à un système basé sur une multitude d'index représentant chacun un groupe fiable pour la recherche d'information dans les images de documents.

CHAPITRE 7

EXPÉRIMENTATION ET VALIDATION

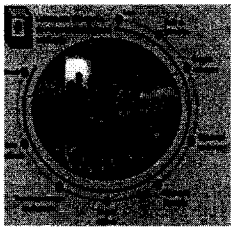
Les expériences portent sur les trois modèles présentés dans cette thèse. Le premier traite du repérage des zones informationnelles de l'image et de l'interprétation des blocs obtenus par des règles de production déduites de la texture et des formes. Le second applique le modèle vectoriel pour la recherche d'information textuelle tenant compte des erreurs de reconnaissance de l'OCR. Le dernier modèle localise, interprète et organise les zones non textuelles pour une recherche hybride d'images de documents combinant les informations collectées à partir de données textuelles et de zones graphiques provenant d'images de documents. Cette approche hybride augmente la puissance d'interrogation et améliore la performance de la recherche d'information. Nous utilisons dans nos expériences des collections d'images de pages webs et les bases UW-1 et UW-2 de l'université de Washington. UW-1 est à dominance textuelle (articles scientifiques, pages de livres, etc.) alors que UW-2 est à dominance graphique (cartes d'affaires, journaux, images publicitaires, etc.).

7.1 Repérage des zones informationnelles de l'image

La base UW-2 (Phillips, 1993) utilisée dans les expériences menées dans cette section est conçue par l'équipe Media Team de l'université de Washington. Elle contient 512 images de documents correspondant à des domaines variés (pages publicitaires, correspondances, formulaires, journaux, etc.). La taille des pages varie entre 10 mots pour les images publicitaires et 500 pour les journaux et les articles scentifiques. Les zones non textuelles sont au nombre de 1114 que les concepteurs divisent en trois catégories qui sont les logos, les photographies et les graphes.

7.1.1 Segmentation des blocs de l'image

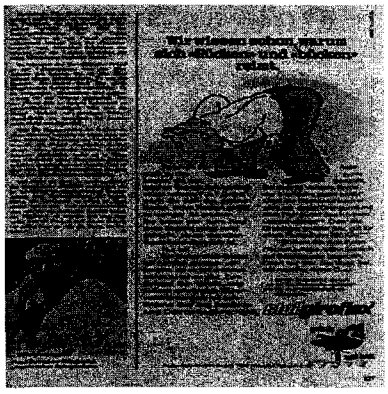
Le traitement d'images désigne en informatique l'ensemble des traitements automatisés qui permettent, à partir d'images numérisées, de produire d'autres images numériques ou d'en extraire de l'information. Des exemples de la segmentation d'une image de bonne qualité et d'une image à contraste varié sont aux figures 18 et 19. La valeur du paramètre d'échelle σ de l'opérateur Laplacien de la Gaussienne "*LoG*" influence les formes ainsi que le nombre d'objets obtenus. Les petites échelles produisent une multitude de petits objets à cause du bruit présent sur l'image et qui perturbe l'information aux alentours alors qu'une grande valeur de σ intègre le bruit dans l'information et produit des zones très homogènes.



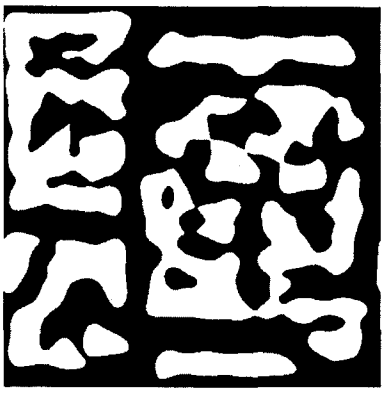
(a) image originale

(b) Images segmentées aux échelles 30, 19 et 5

Figure 18 Segmentation multi-échelle d'une image de bonne qualité



(a) Image originale



(b) $\sigma = 10$

(c) $\sigma = 5$

(d) $\sigma = 1$

Figure 19 Segmentation multi-échelle d'image de document de U-W1 à contraste varié

7.1.2 Interprétation des régions segmentées

Pour l'image synthétique de bonne qualité à la figure 18, la segmentation à grande échelle fusionne les différents blocs et intègre le bruit dans la communauté de pixels avoisinante. La segmentation à petite échelle sépare progressivement l'information du bruit et aboutit à des formes étirées horizontalement pour le texte et une dispersion en blocs non uniformes des zones graphiques.

Cette constation est difficile à généraliser pour les images de moins bonne qualité, comme c'est le cas dans notre base. L'exemple de l'image à la figure 19 montre que les blocs

informationnels représentant l'information à grande échelle intègre les parties bruitées ; en revanche, on arrive difficilement à distinguer les zones textuelles des non textuelles lorsque l'image est de mauvaise qualité. Dans la multitude d'objets de différentes formes à valider aux échelles supérieures, on remarque que le rapport hauteur / largeur de la boîte englobant des illustrations ou des logos reste supérieur à celui des zones non textuelles. Une séparation texte/non texte reste difficile à réaliser lorsque les frontières de séparation sont très bruitées ou que le texte se chevauche avec la boîte englobante de la zone non textuelle. Les tableaux II, III et IV montrent les résultats obtenus. Le premier présente la décomposition de l'image b) à l'échelle 30 de la figure 18, le deuxième est une analyse par classe de documents et le dernier tableau récapitule les formes des objets obtenus. On remarque que la détection du texte est facile aux petites échelles (σ petit) à cause de l'étirement horizontal des tâches informationnelles. Des portions de logos et d'illustrations ont des formes non homogènes qui peuvent être reconstituées aux échelles supérieures pour approcher les formes originales.

Tableau II

Résultats de la segmentation de l'image de la figure 18 pour $\sigma = 30$

Numéro Objet	Coordonnées lignes	Coordonnées Colonnes	Surface objet	Type objet
1/30	40 à 96	20 à 76	2556	Logo
2	94 à 169	206 à 356	5639	
3	206 à 393	81 à 373	23171	

Total logos : 1 Texts : 0 Lines : 0

Chaque région est identifiée par un numéro et ses coordonnées. Le premier objet de l'exemple au tableau II occupe l'espace entre les lignes 40 à 96 et les colonnes 20 à 76. Son type "logo" est défini par la règle de production du paragraphe 3.5 qui préconise que les logos sont ovales et étirés en hauteur ou en largeur.

Tableau III

Résultats obtenus : objets détectés dans les catégories de documents composites pour 4 échelles différentes

Type de contenu	Nombre d'objets localisés pour				Nombre total de zones
	$\sigma = 5$	$\sigma = 10$	$\sigma = 15$	$\sigma = 30$	
Logos	10	92	182	209	227
Photos	24	103	220	311	323
Graphes	32	98	154	250	440
Textes	5292	3225	1145	460	5670

Tableau IV

Formes des objets obtenus

Objet	Forme
Textuel	Rectangle étiré en largeur Rapport hauteur/largeur très petit Histogramme de distribution uniforme
Logo	Ovale étiré en hauteur Histogramme distribué en normal centré réduit
Graphiques	On ne différencie le texte qu'à une échelle proche de 1 Pour les formes verticales, on a un rapport largeur sur hauteur très petit
Illustrations	Formes libres éparpillées Les détails apparaissent aux petites échelles

Une caractéristique comme l'entropie distingue le texte du non-texte. La validation sur les régions obtenues à la figure 18b à l'échelle 5 a montré que les 19 zones textuelles ont une entropie moyenne de 1.9 avec un écart type de 0.24.

Les observations sur la segmentation subie par des images et dont les résultats sont présentés au tableau III montrent qu'à grande échelle, les logos et les illustrations ont une forme elliptique et la valeur de l'entropie est supérieure à 5. L'entropie des zones obte-

nues à petite échelle différencie les zones rectangulaires non textes des régions textuelles. En général, les formats textes sont étirés horizontalement, les logos ont des formes ovales étirées verticalement, alors que pour les images d'illustrations, on trouve des formes quelconques éparpillées sur une surface plus ou moins délimitée à des échelles supérieures. Ces mesures sont subjectives et exigent des seuils de tolérance pour obtenir de meilleurs résultats. Dans nos expériences, il nous a été difficile, dans nos appariements, de définir des intervalles de confiance autour des surfaces de mesures pour les rectangles, les ellipses, les cercles et les lignes. Des méthodes robustes d'apprentissage sont nécessaires pour évaluer la performance et la précision de nos résultats.

Ces observations montrent que ces mesures restent subjectives et la segmentation reste sensible à beaucoup de paramètres pour améliorer la détection et l'interprétation des régions segmentées. C'est pour cela que nous allons migrer vers l'opérateur SKCS (Separable Kernel with Compact Support) qui est à noyau gaussien. Le SKCS offre plusieurs degrés de liberté au lieu du seul σ pour l'opérateur LoG. Les sur-segmentations seront réduites par un processus de fusion de blocs. Les résultats apportés par ces deux éléments sont détaillés dans la section 7.3.

7.2 Traitement relié aux textes OCR

Les techniques de lecture optique de caractères sont généralement appliquées aux activités bureautiques. Les grandes institutions ou les entreprises sont confrontées à la transformation de grandes quantités d'informations sous forme papier vers une forme électronique, sans passer par la saisie. Pour cette raison, la reconnaissance du texte de l'image tend de plus en plus à utiliser les logiciels d'OCR.

7.2.1 Collection d'images et de données

Trois corpus d'images de documents ont été collectés.

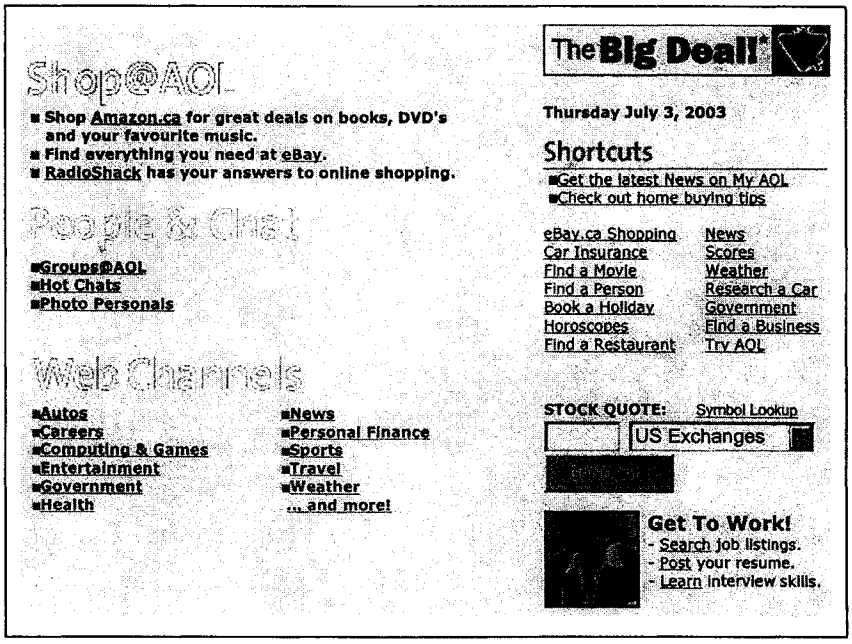
Le premier est utilisé pour l'apprentissage et la construction d'erreurs-grams. Il contient 979 pages de journaux et de documents techniques provenant de la base de l'équipe média-team de l'université de Washington (Phillips, 1993). La taille moyenne est de 510 mots par page.

Le second corpus est appelé "TEST1" et se compose de 100 images Web avec une moyenne de 410 mots par page. Il est ensuite dégradé à l'aide de modèles de traitement d'images et par des photocopies répétitives. Nous additionnons à l'image du bruit et du flou pour construire un ensemble d'images dégradées. La dégradation utilise un bruit gaussien avec trois différentes valeurs de σ alors que le flou est obtenu par la convolution de l'image avec une fenêtre remplaçant chaque pixel par la moyenne des 8 (fenêtre de taille 3×3) ou des 24 (fenêtre de taille 5×5) pixels voisins. Les photocopies ajoutent aléatoirement à l'image des tâches sel et poivre causant des coupures de caractères et des champs informationnelles. Le nouveau corpus de Test-dégradés obtenu suite à ces dégradations contient 700 images obtenues par l'application de 7 types de dégradation aux 100 pages Web du corpus de test. La figure 20 montre une page Web originale à laquelle on a appliqué 3 types de dégradations.

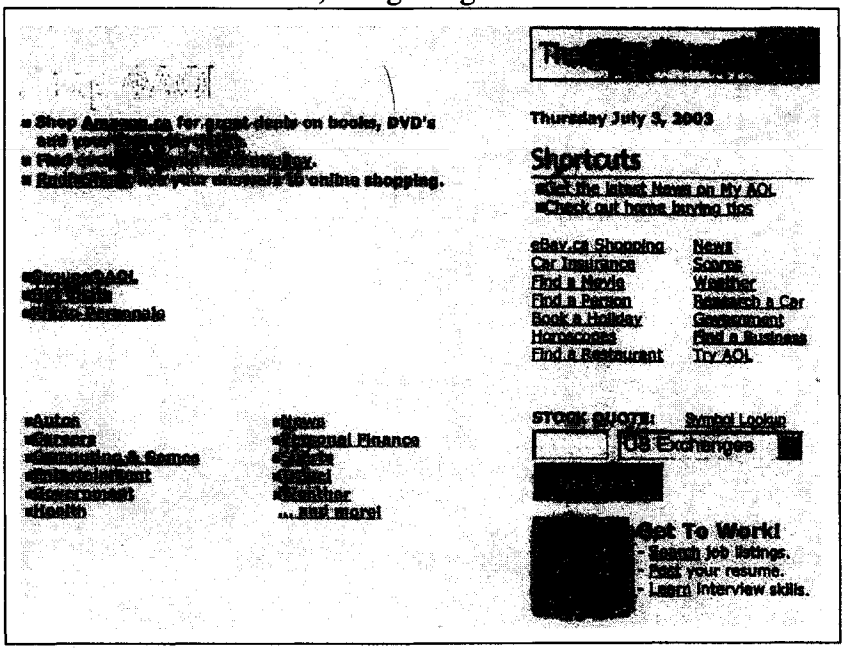
Le troisième corpus appelé "Test2" contient 200 images de la base UW-2 de l'équipe Media Team de l'université de Washington. On trouve des cartes de visites, des pages publicitaires, des manuels ou des formulaires. Le nombre moyen de mots dans chaque page est de 44 pour les images publicitaires et de 304 pour les manuels et les formulaires. Ce corpus a été considéré pour tester la robustesse et la sensibilité des systèmes sur des images présentant des textes courts et des noms de personnes ou de places.

On termine avec la phase d'interrogation où nous avons choisi aléatoirement, à partir du contenu des documents, 50 requêtes. Chaque requête contient une moyenne de 3 mots. Pour mesurer la performance, nous avons considéré comme pertinentes les réponses obtenues par SMART sur la base du texte électronique fourni par le concepteur de la base

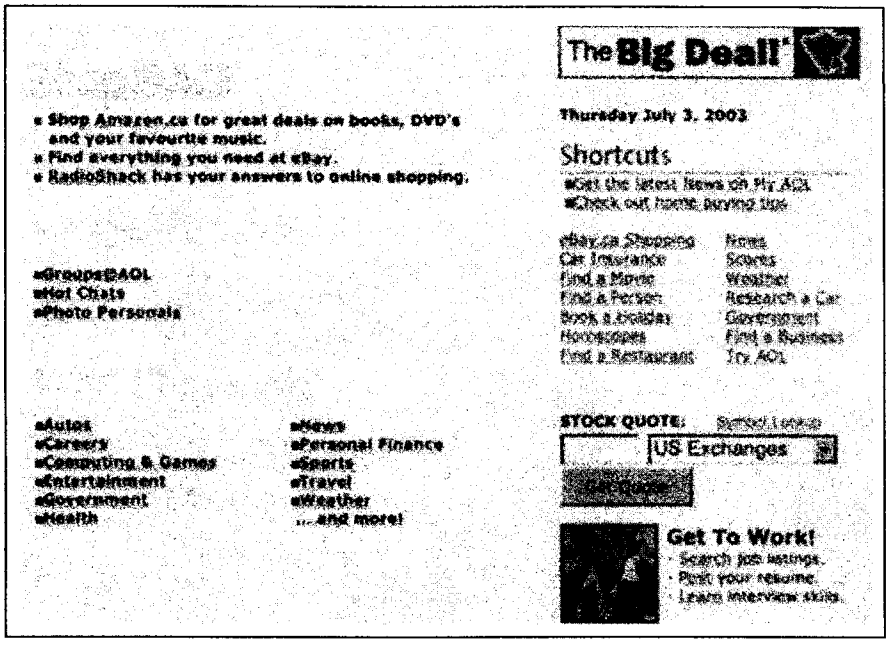
d'images UW-1. Les documents retournés par différents systèmes sont comparés à ceux considérés comme pertinents (réponses de SMART) pour déterminer leur pertinence pour chaque requête.



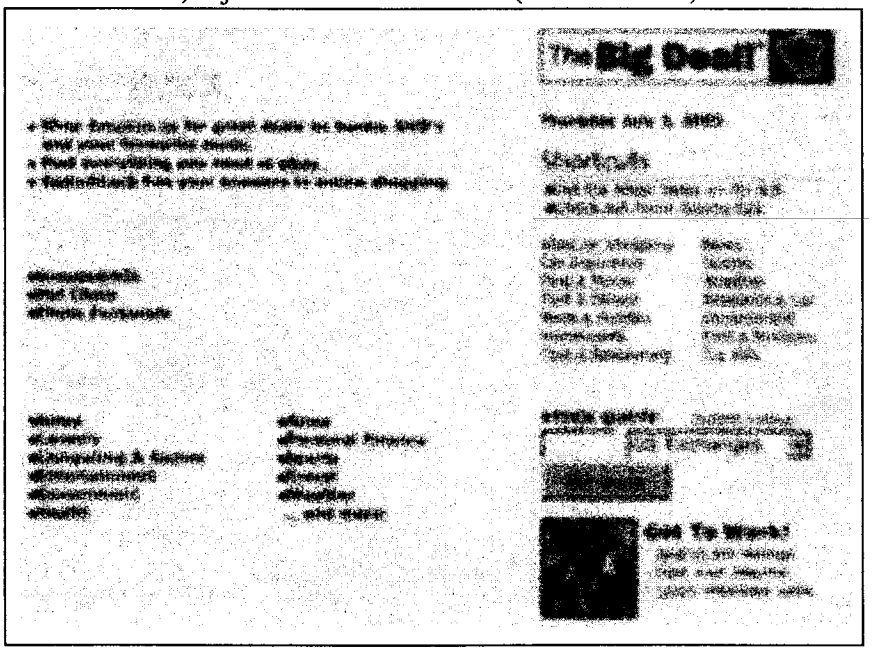
a) Image originale



b) Dégradation par des photocopies



c) Ajout de bruit et de flou (fenêtre 3×3)



d) Ajout de bruit et de flou (fenêtre 5×5)

Figure 20 Une page Web et ses correspondants dégradés

7.2.2 Reconnaissance par OCR

Nous avons utilisé la base d'images UW-1 (979 images) pour l'apprentissage et la collecte des erreurs-grammes. Ensuite, un ensemble de collections, "Test" (100 images web), "Test1" (700 images dégradées) et "Test2" (512 images de UW-2 et 200 pages web) ont testé la fiabilité et la robustesse de notre approche.

7.2.2.1 Apprentissage

L'application de l'algorithme de la distance d'édition pour localiser les erreurs de reconnaissance a produit les résultats qui sont dans le tableau V. Les images originales contiennent 614 zones non-textes qui expliquent le nombre élevé de mots dans le texte OCR par rapport aux textes originaux : 499 123 mots dans les documents originaux et 528 315 mots extraits par l'OCR. Seulement 468 619 mots des 499 123 sont correctement reconnus. Nous utilisons la programmation dynamique avec une distance inférieure à deux pour construire les erreurs-grams à partir des 5185 mots extraits. Notons que des améliorations sont possibles moyennant un prétraitement de l'image réduisant le bruit et distinguant les zones textes considérés comme bruit.

À l'issu de ces expériences, nous avons obtenu 6933 substitutions, 2216 suppressions et 2319 insertions. Les résultats obtenus par l'algorithme Edit-distance sont introduits dans le constructeur des erreur-grams et des règles de corrections. On a construit 2822 erreur-grams et règles de correction. Les 20 premières erreur-grams et leurs règles de production sont au tableau VI.

7.2.2.2 Test

Les résultats de la reconnaissance des images des collections de test et du test-dégradé sont aux tableaux VII et VIII . Notons une diminution de la performance de la reconnaissance sur les images dégradées et observons que l'ajout de bruit gaussien n'influence pas la

Tableau V

Apprentissage pour la reconnaissance du texte, 979 images scannées sont reconnues par un OCR commercial

	Nombre de mots	Nombre de caractères	Taux de reconnaissance
Image originale	499 123	2.9 Moctets	
Reconnaissance par OCR	528 315	3 Moctets	
Correctement reconnu	468 619	2.74 Moctets	93.8%
Distance d'édition$\leq$ 2	5 185	30 591	1.03%
Total de reconnaissances	473 804	2.78 Moctets	94.83%

Tableau VI

Les 20 premières erreur-grammes et la probabilité P que l'erreur-gramme A_i de l'image originale soit confondue avec B_j dans le texte OCR

A_i	B_j	$P(B_j/A_i)$	A_i	B_j	$P(B_j/A_i)$
th	di	0.12	i	1	0.064
i	l	0.096	th	dh	0.059
h	i	0.092	l	1	0.05
th	ti	0.089	l	i	0.036
t	d	0.086	y	v	0.034
r	l	0.086	z	s	0.033
th	dh	0.077	t	l	0.032
he	ie	0.066	the	die	0.029
the	tie	0.065	e	0	0.026
t	l	0.064	ize	ise	0.024

précision de la reconnaissance. La figure 21 montre que la reconnaissance par OCR résiste au bruit, mais la performance décroît à mesure que le flou augmente.

Tableau VII

Reconnaissance et erreurs de l'OCR sur "Test1". 100 pages Web images dégradées par des photocopies

	Sans dégradation	Photocopies
Nombre de mots	40 642	40 642
Mots extraits par OCR	34 318	22 600
Mots bien reconnus	29 368	15 278
% bonne reconnaissance	72.26%	37.59%

Le taux de reconnaissance sur la collection d'images "test1" (73%) est inférieur à celui de l'ensemble d'apprentissage (93%) à cause de la résolution et de la qualité de l'impression des pages Web.

Tableau VIII

Reconnaissance et erreurs de l'OCR sur "Test1". 100 pages Web images dégradées par du bruit Gaussien et du flou

Taille fenêtre du flou	3x3			5x5		
σ du bruit Gaussien	0.1	0.01	0.001	0.1	0.01	0.001
Nombre de mots	40 642			40 642		
Mots extraits par OCR	20 356	20 528	20 680	10 870	9 980	10 294
Mots bien reconnus	15 906	15 730	16 170	3 728	3 310	3 530
% de mots bien reconnus	39.14%	38.70%	39.79%	9.41	8.35%	8.91%

Les images dégradées présentent un faible taux de reconnaissance lorsque l'OCR est utilisé pour extraire le texte. Beaucoup de facteurs tels que la taille, la police, les coupures, les fusions et les tâches blanches présentes sur les caractères sont employés pour indiquer la qualité textuelle des images. Les expériences menées sur les collections "Test" et "Images

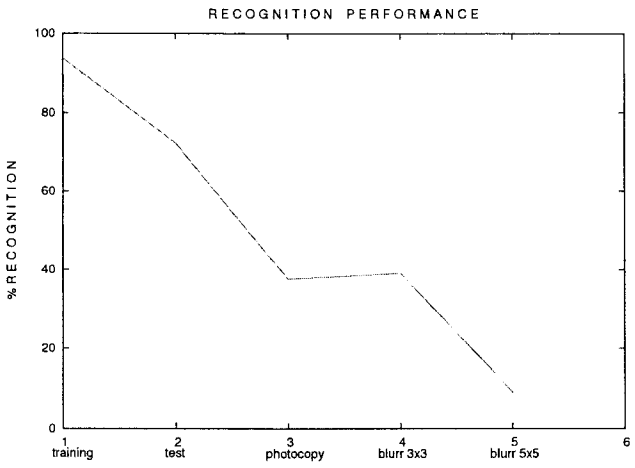


Figure 21 Dégradation de la reconnaissance sur la collection "Test" et les images dégradées

dégradées" montrent une diminution significative du taux d'identification de 93.8% sur l'ensemble d'apprentissage à 72.26% avec la collection "Test" et environ à 10% avec les images dégradées. Deux photocopies successives des images produisent un taux de reconnaissance de 38%, qui a le même effet que la dégradation avec du flou résultant des 8 pixels voisins.

7.3 Traitement des zones non textuelles

La chaîne de traitement développée autour des objets non textuels et de la recherche d'images pertinentes à une requête d'utilisateur suit trois étapes qui sont :

- les traitements de bas niveau pour réhausser les caractéristiques propres des objets, avec, par exemple, des opérateurs de fusion d'objets
- l'appariement de vecteurs utilisant des distances sur les caractéristiques de formes et de textures des objets de l'image
- les traitements de haut niveau qui utilisent des critères de voisinage et de regroupement pour améliorer la représentation de l'information, réduire l'espace mémoire utilisé et accélérer la recherche.

Le résultat d'appariement ou de recherche est une image de vraisemblance accompagnée d'une mesure de similarité comprise entre 0 et 1. Plus l'image ressemble à l'objet de référence, plus le résultat va se rapprocher de la valeur 1. Les appariements portent sur un ou plusieurs objets à cause de la nature de nos images et de la multitude de sous objets produits par la segmentation. Ces traitements successifs dont les résultats sont additifs déterminent le degré de pertinence des images de la base.

Nos principales motivations dans cette partie est l'amélioration de performances telles :

- la capacité de stockage
- la vitesse de traitement
- l'élimination de variables non discriminantes et source de bruit
- l'interprétation des caractéristiques pour identifier les zones informatives de l'image.

7.3.1 Collection d'images utilisées

La base UW-2 utilisée dans nos expériences est conçue par l'équipe Media Team de l'université de Washington. Elle contient 512 images et 1040 zones non textuelles dans la vérité terrain fournie avec la base UW-2 et que les concepteurs divisent en trois catégories que sont les logos, les illustrations et les graphiques.

Nous avons corrigé l'interprétation fournie par le concepteur pour supprimer toutes les parties textuelles considérées comme logos ou graphiques afin d'uniformiser l'étiquetage des zones non textuelles des 200 pages Web que nous intégrons à notre collection. Les corrections apportées à l'étiquetage des zones non textuelles des images de UW-2 sont décrites dans le tableau IX. Notons donc que notre méthode de segmentation n'est pas adaptée à la segmentation des dessins et des graphes de par la texture de ces derniers. Une partie des graphiques reste mal reconnue à cause des courbes et des dessins dont un exemple est à l'image (a) de la figure 22 où le SKCS n'a pu détecter toute l'information

contenue dans l'image. Nous avons profité de cet exercice pour étiqueter de façon identique et constituer la vérité terrain des régions non textuelles des pages webs de notre collection.

Tableau IX

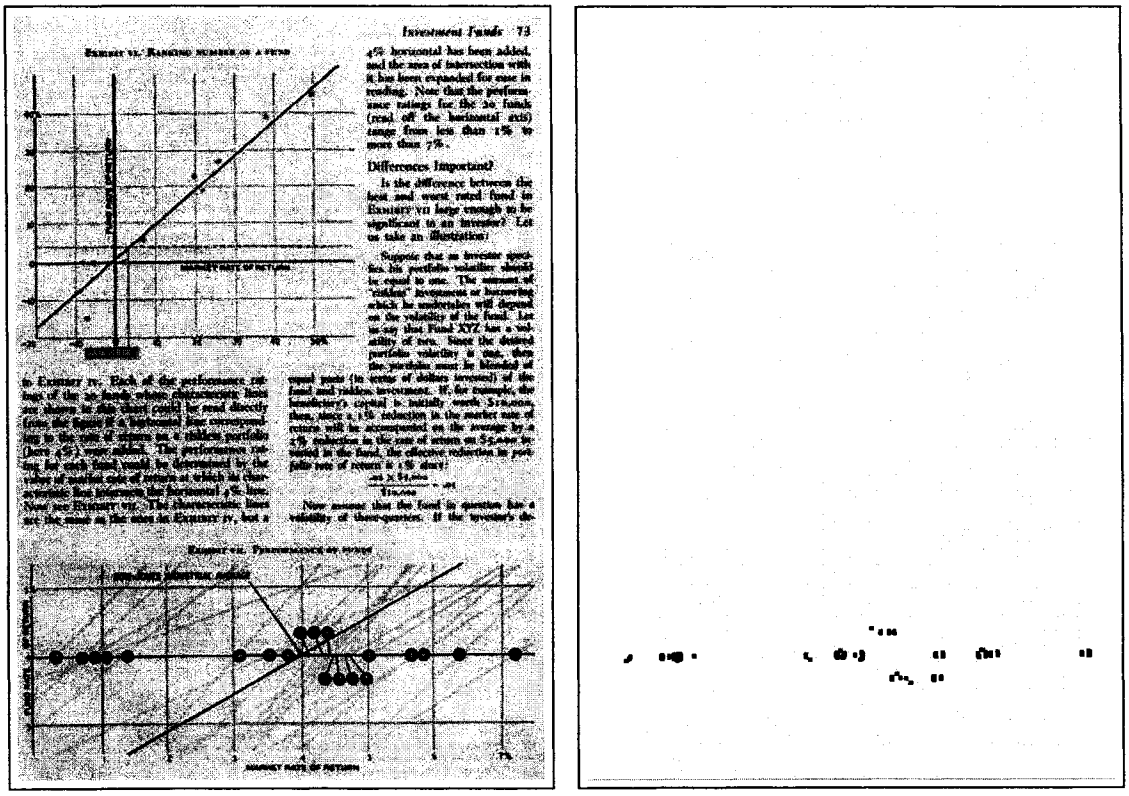
Reconnaissance des zones non textuelles après correction de l'interprétation du concepteur

	S K C S				
ORIGINAUX	LOGOS	PHOTOS	GRAPHES	AUTRES	TOTAL
LOGOS	189	3	4	31	227
PHOTOS	8	274	11	30	323
GRAPHES	4	8	333	95	440
AUTRES	21	13	16		50
TOTAL	222	298	364	156	1040

Les 200 images webs contiennent 708 zones non textuelles réparties en 220 logos, 242 photographies et 246 graphiques. Elles proviennent de pages publicitaires et de journaux. Les deux tiers des images de toute la collection, soit 425 images contenant 1114 zones non textuelles sont utilisées pour l'apprentissage. Le reste des images, soit 237 images et 561 objets non textuels, est utilisé pour tester notre approche. Ces images sont choisies aléatoirement et comprennent un échantillon de chaque type (cartes d'affaires, formulaires, pages publicitaires, journaux ...).

7.3.2 Localisation

La figure 23 montre les résultats de la segmentation par les opérateurs LoG et LoSKCS d'une image exemple de la base UW-2. La fusion des régions informatives et la suppression des taches de petites tailles réduisent le nombre d'objets pour cet exemple de 24 à 4.



(a) Image exemple de UW-2 (b) Segmentation par $LoSKCS$

Figure 22 Image de UW-2 avec deux graphiques non détectés par l'opérateur $LoSKCS$

Le nombre de régions informatives localisées par l'opérateur SKCS dépend des variables σ et γ. Un aperçu des résultats obtenus pour différentes valeurs de σ et de γ sur les images de la base d'apprentissage est à la figure 24. Le nombre de tâches informatives obtenues par la segmentation SKCS est à la figure 24(a). Toutefois, l'opération de fusion réduit ce nombre à des proportions dépassant les 30% comme montré à la figure 24(b). Le type d'application ou la nature de l'information recherchée nécessitent de faire varier les paramètres σ et γ au cours du traitement pour localiser les régions informatives souhaitées. On remarque que le paramètre σ influe sur le nombre de régions localisées. Par contre, pour le paramètre γ, on observe une diminution du nombre d'objets quand on prend γ grand à cause du lissage du bruit qui regroupe les objets proches.

Enfin, la figure 24(c) montre que le taux de localisation d'objets non textuels (1114 au total dans la base d'apprentissage) pour σ petit n'est pas suffisant et que les paramètres de la segmentation n'apportent finalement que peu d'informations à partir de $\gamma = 5$. Le nombre de régions non textuelles extraites pour $\sigma = 1$ sont moindres que pour $\sigma = 5$ et reste constant jusqu'à $\gamma = 10$.

Dans le choix des variables de la segmentation à considérer pour la suite de nos expériences, nous privilégions le rappel et la précision qui représentent la capacité de la détection et l'homogénéité des résultats obtenus par rapport aux objets non textuels existants et que nous essayons de localiser. La fusion et le nombre d'objets non textuels localisés montrent que la meilleure combinaison réduisant le nombre d'objets à traiter est pour $\sigma = 5$ et $\gamma = 10$. Ce résultat est très interessant pour la suite de nos expériences et nous permet de ne traiter que 1759 objets pour localiser 1011 objets non textuels (voir le tableau X).

Tableau X

Précision vs. Rappel de la segmentation par SKCS

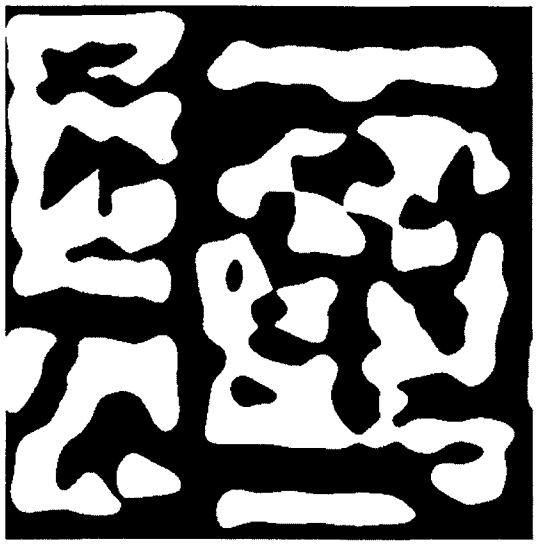
$\sigma 1$	$\sigma 2$	gamma	Nombre d'objets total	Nombre d'objets non textuels	Précision	Rappel	F-measure
1	1	2	1144	573	50.08	51.44	50.75
5	2	2	2702	1017	37.63	91.29	53.30
5	5	5	2240	1019	45.49	91.47	60.76
5	5	10	1759	1011	57.47	90.75	70.37
10	5	5	2375	1000	42.10	89.76	57.32

Nombre total d'objets non textuels = 1114

7.3.3 Classification des zones de la base d'apprentissage

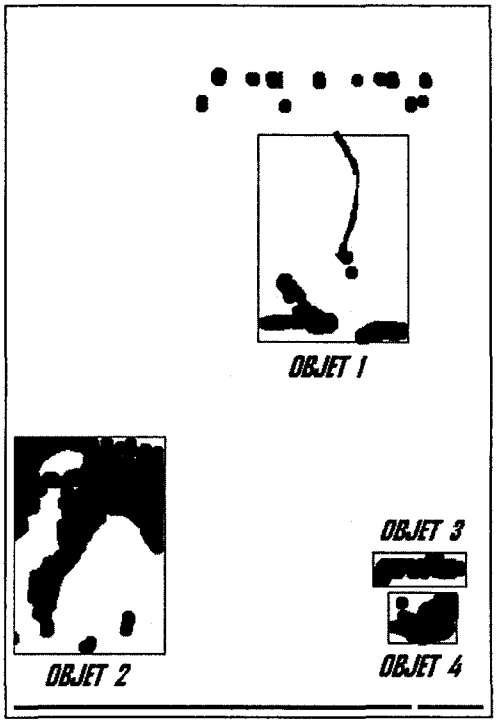
Nous allons répartir les objets obtenus lors de la segmentation dans des groupements différents pour différencier les logos des photographies et des graphes. Nous utilisons des

(a) Image originale



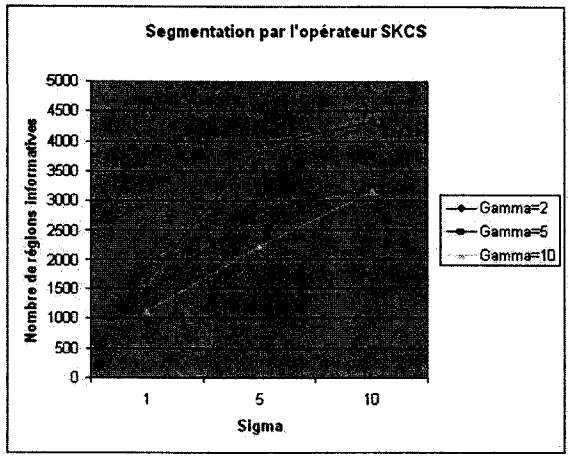
(b) Segmentation par l'opérateur LoG

(c) Segmentation par $LoSKCS$

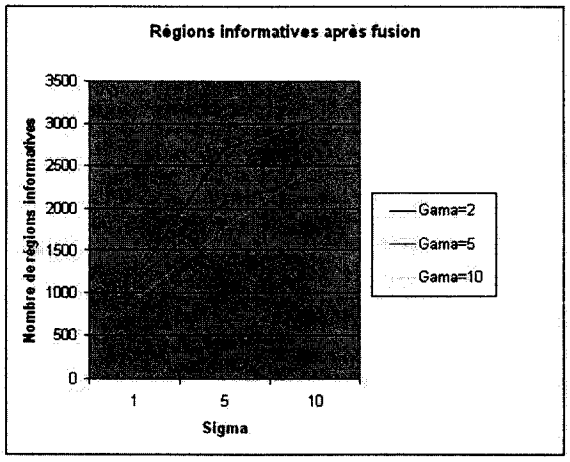


(d) Fusion des objets

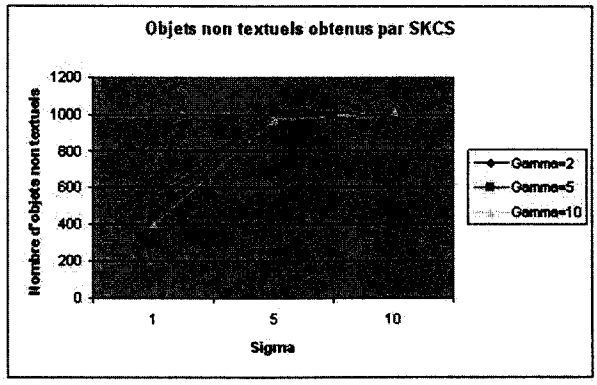
Figure 23 Segmentation par les opérateurs LoG et $LoSKCS$ et résultat de la fusion d'une image de UW-2



(a) Nombre de régions informatives avant fusion



(b) Nombre de régions informatives après fusion



(c) Nombre d'objets non textuels localisés

Figure 24 Résultats de la segmentation des images de la base d'apprentissage par l'opérateur $LoSKCS$

algorithmes de classification et présentons dans cette section les résultats obtenus par la combinaison de deux classifieurs K-moyennes et MKL. Cette classification des régions informatives repose sur l'information de texture et de forme. La liste des attributs utilisés reposent sur les mesures d'entropie, d'excentricité, diamètre, périmètre, solidité, orientation, centroides, coordonnées, hauteur, largeur, maxima, minima et des surfaces (convexe, remplie, ellipse, rectangle). Nous avons évalué la segmentation pour différents paramètres. En ce qui concerne l'extraction de régions informatives, nous avons testé notre segmentation avec un nombre de classes compris entre 1 et 8. En effet, ayant besoin, de part notre application, de reconnaître trois types d'objets qui sont les logos, les photographies et les graphes, le choix d'un nombre de classes supérieur à trois a donc été abordé de sorte à obtenir une classification plus fine des objets. La classification va diviser les groupes en sous groupes de manière à augmenter la précision du système qui reste très moyen à ce stade du traitement comme montré au tableau XI.

7.3.3.1 Apprentissage par K-moyennes

La validation de l'algorithme a été effectuée sur la base d'indice de qualité qui est l'homogénéité des classes. Chaque classe créée par l'algorithme doit être composée d'objets tous de même type (en particulier aucune classe ne doit confondre des objets de types différents). Le deuxième indice de qualité de l'analyse repose sur le nombre total de classes créées. Ces deux critères sont opposés : en effet, plus le nombre de classes créées par l'algorithme est important, plus l'homogénéité de chacune d'elle est probable. Inversement, si l'on pénalise la création de nouvelles classes, la taille des classes créées augmente, avec un risque accru d'introduire des éléments qui altèrent l'homogénéité des classes. Les pondérations choisies permettent un compromis entre ces deux critères ; le but étant de ne pas mélanger des logos avec des photographies ou des graphiques lorsque ceux-ci sont clairement distinguables visuellement.

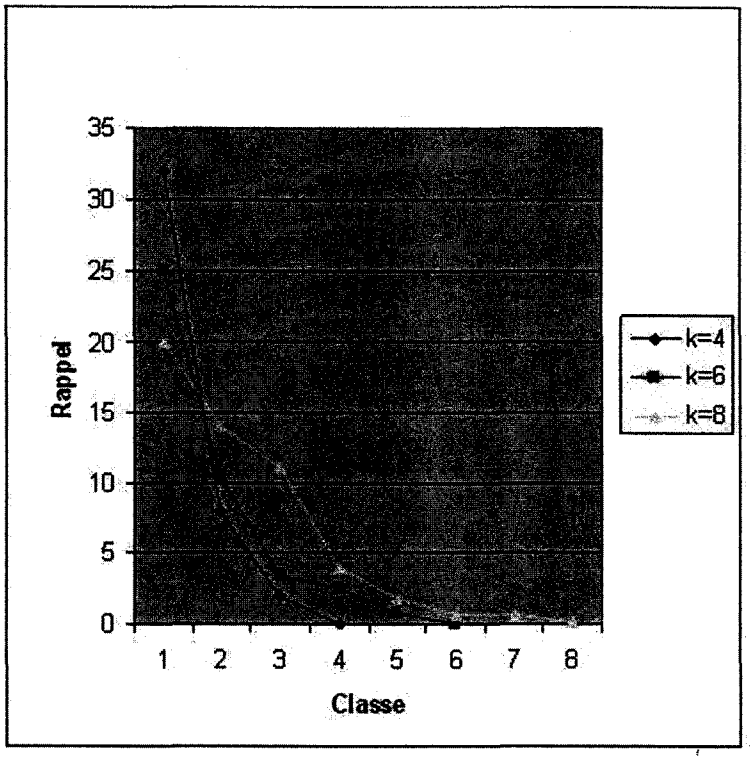


Figure 25 Rappels des zones non textuelles regroupées par K-moyennes

Homogénéité des classes

Le principal critère d'homogénéité de la classe est la bonne séparation des objets. Il faut cependant noter que les étiquettes des régions non textuelles distinguent les différents objets. Or, on remarque à la figure 26 que les classes ne sont pas homogènes. La première classe contient entre 60% et 70% de chaque type d'objet alors que les autres classes ne dépassent pas les 30%. Le détail des résultats sur l'ensemble de la base d'apprentissage est présenté dans le tableau XI. À ce niveau de l'analyse, les types d'objets peuvent être confondus dans les différentes classes de K-moyennes et notre classification n'arrive pas à distinguer chaque type d'objet dans une classe différente.

Tableau XI

Précision vs. Rappel pour l'algorithme des K-moyennes avec k=4

Classe	Nombre d'objets total	Nombre d'objets non textuels	Précision	Rappel	F-mesure
1	1531	988	64.53	88.68	74.7
2	159	21	13.20	1.88	3.30
3	51	15	29.41	1.34	2.57
4	18	2	11.11	0.17	0.35
Total	1759	1026	58.32	92.10	71.42

Nombre final de classes

La figure 25 montre que sur l'ensemble des images de la base d'apprentissage, et par rapport à l'étiquetage de celle-ci, le taux de rappel est inférieur à 5% lorsque le nombre de classes dépasse la valeur 3 ; c'est-à-dire qu'environ 5 objets sur 100 ne sont pas dans les 3 premières classes. Il est intéressant de noter que le nombre de classes ne résout pas le problème de l'homogénéité des classes. En effet, trois classes regroupent 95% des objets et l'augmentation du nombre de classes n'améliore pas l'homogénéité des classes et la répartition des objets reste identique. C'est pour cela que nous appliquons d'autres techniques de classification à chaque classe obtenue par k-moyennes pour remédier à ces problèmes.

Perspectives d'amélioration

La classification réalisée ici ne sépare pas efficacement les objets malgré la création de plusieurs classes qui s'avèrent inutiles. Ces dernières sont souvent constituées de quelques objets éloignés des classes déjà créées.

Comme nous disposons d'une vingtaine de caractéristiques, décrits au début de la section 7.3.3, pour chaque objet, une amélioration envisageable serait d'utiliser les caractéris-

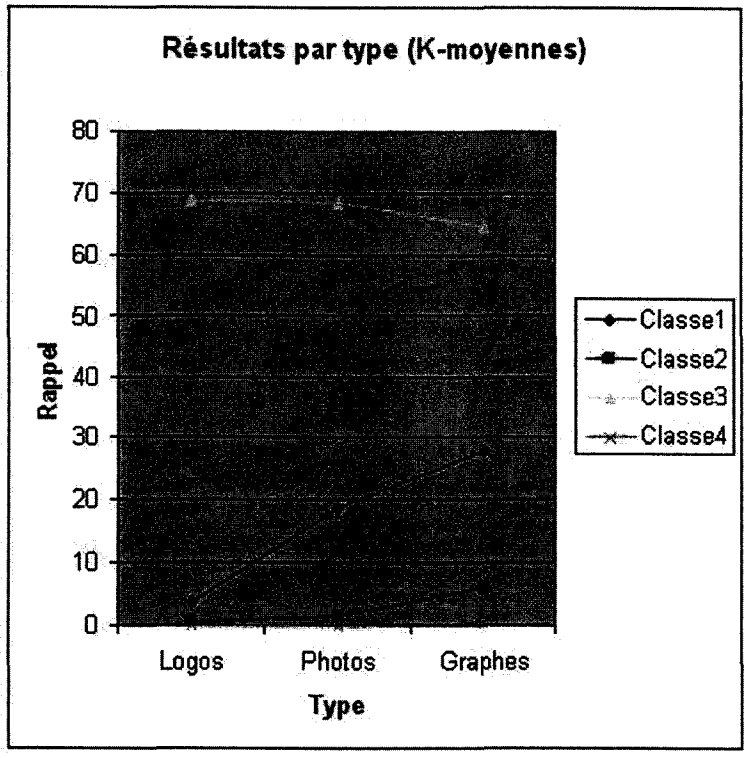
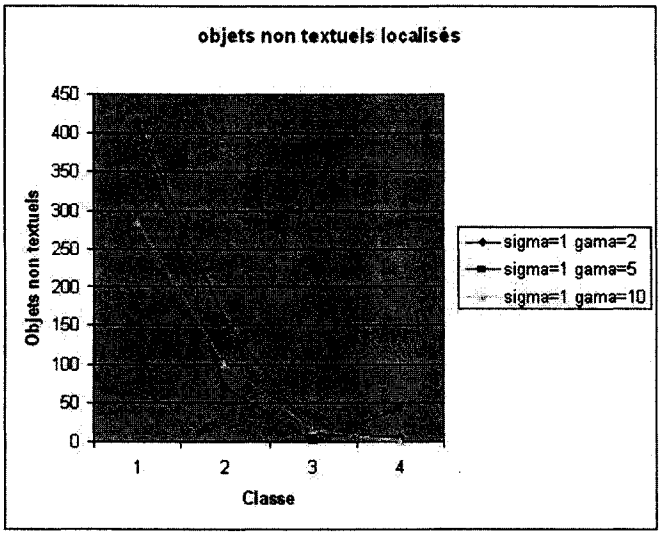


Figure 26 Rappel des types d'objets dans chaque classe de K-moyennes

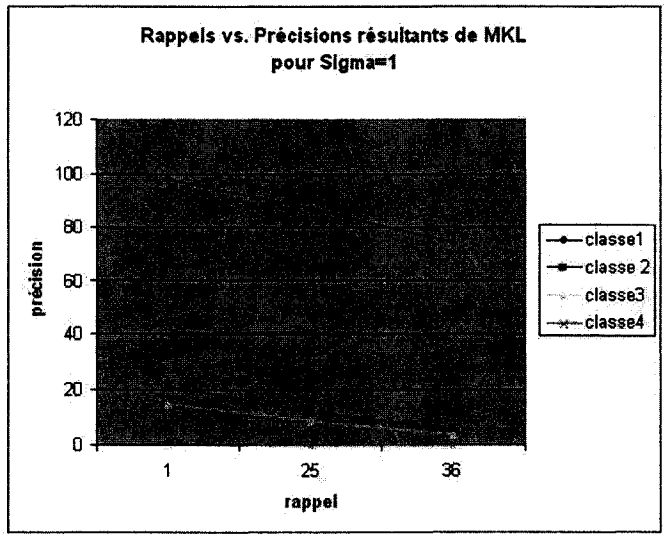
tiques les plus discriminantes pour la création d'une classe lorsque les objets sont trop éparpillés. Une alternative plus indirecte consisterait à faire varier le nombre de caractéristiques discriminantes en fonction des regroupements engendrés.

Enfin, compte tenu de l'évolution continue des puissances de calculs disponibles, il est possible que l'étape de classification définie ici devienne une définition de classes initiales ; chaque classe pourra être modélisée par les caractéristiques principales et, une fois labélisées, ces caractéristiques permettent une meilleure caractérisation de chaque objet et donc un regroupement ultérieur plus fiable pour séparer les éléments de chaque classe et augmenter la précision qui reste très moyenne à ce stade de la classification. Le MKL que nous appliquons par la suite traite les éléments de chaque classe pour définir les caracté-

ristiques principales et les regroupements réduisant l'éparpillement spatial des éléments de chaque classe.



(a) Nombre d'objets non textuels localisés



(b) Rappel vs. précision des objets non textuels

Figure 27 Résultats du classifieur MKL à 4 classes ($\sigma = 1$)

7.3.3.2 Apprentissage par MKL

Nous avons testé la classification par la méthode MKL à 4 classes pour $\sigma = 1$; la figure 27(a) montre une saturation des objets dans les deux premières classes. Un premier test du taux de rappel pour $\sigma = 1$ donne des valeurs ne dépassant pas 36% (voir la figure 27(b)). Pour les paramètres $\sigma = 5$ et $\gamma = 10$ qui représentent le meilleur compromis nombre d'objets localisés versus les objets non textuels, les résultats obtenus lors des expériences sont présentés dans le tableau XII et les figures 28 et 29.

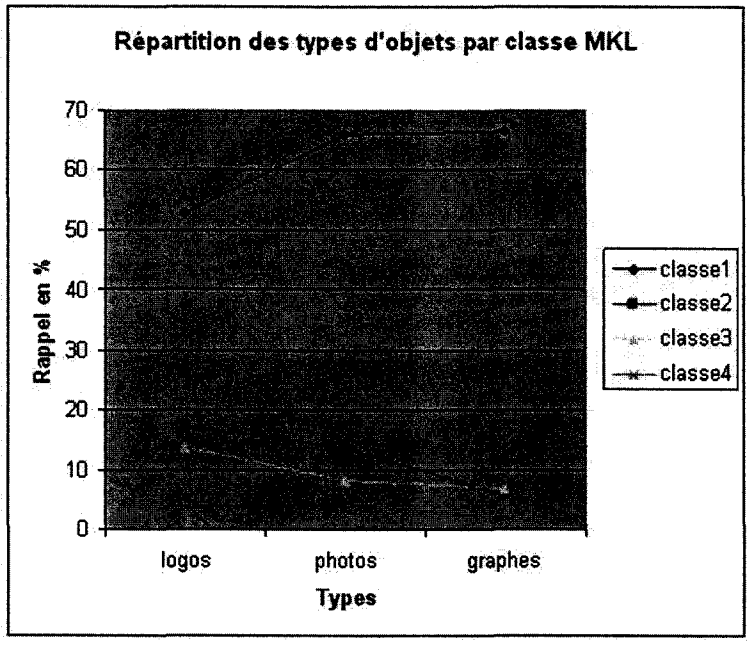


Figure 28 Rappels des types d'objets pour MKL à 4 classes

Homogénéité des classes

Dans le tableau XII, on note que le rappel vs. la précision pour le MKL à 4 classes est de 75% pour une précision égale à 74% dans la première classe. Toutefois, c'est la première classe qui contient un grand nombre d'objets non textuels au détriment du reste des classes qui ne totalisent que 17% du rappel. La répartition des logos, des photos et des graphes

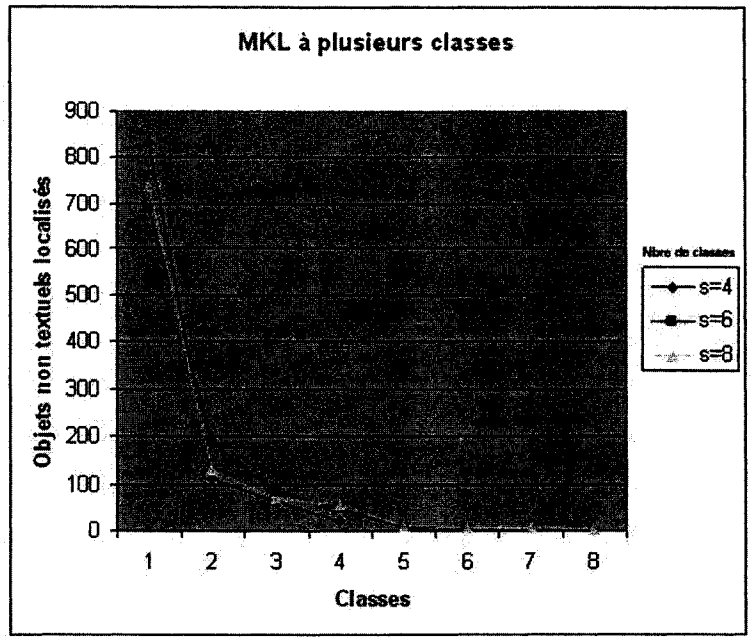


Figure 29 Localisation des objets non textuels pour MKL à 4, 6 et 8 classes

Tableau XII

Précision vs. Rappel pour l'algorithme MKL à 4 classes

Classe	Nombre d'objets total	Nombre d'objets non textuels	Précision	Rappel	F-mesure
1	1125	832	73.95	74.68	74.31
2	359	111	30.92	9.96	15.07
3	183	20	10.93	1.79	3.08
4	92	63	68.47	5.65	10.44
TOTAL	1759	1026	58.32	92.10	71.42

dans les différentes classes reste faible. En effet, la première classe regroupe entre 50% et 70% de chaque type d'objet et la différenciation reste totalement posée. On assiste au même phénomène que pour K-moyennes, un post traitement pour les classes à forte population est à considérer pour la suite des travaux.

Nombre final de classes

Le groupement des objets localisés en 4, 6 ou 8 groupes, comme montré à la figure 29, n'apporte rien de nouveau par rapport à l'algorithme K-moyennes. Comme dans la classification par K-moyennes, 80% des objets se trouvent dans les deux premières classes malgré la variation du nombre de classes entre les valeurs 4 et 8. Le nombre d'objets reste inférieur à 10% lorsque le nombre de classes dépasse la valeur 3 ; c'est-à-dire qu'environ 10 objets sur 100 ne sont pas dans les trois premières classes. Il est intéressant de noter qu'en moyenne, 3 classes regroupent plus que 95% des objets et il est donc inutile d'augmenter le nombre de classes pour améliorer l'homogénéité des groupes. On doit alors combiner différentes techniques de classification pour remédier à ces problèmes.

7.3.3.3 Apprentissage par combinaison de K-moyennes et de MKL

Les deux premières classes pour les méthodes K-moyennes et MKL cumulent environ 90% du taux de rappel ; le problème reste posé malgré l'ajout de classes supplémentaires. Pour cela, nous réutilisons les classes importantes obtenues par K-moyennes pour regrouper leurs éléments dans des sous classes à l'aide de MKL. Ce dernier définit les caractéristiques pertinentes et constitue des sous groupes minimisant l'éparpillement spatial tout en améliorant la précision de notre modèle. Cependant, nous présentons les résultats de MKL appliqué aux trois plus importantes classes de K-moyennes dans les figures 30 et 31. Le tableau XIII montre que le partage des objets d'une classe de K-moyennes en quatre sous-classes créées par MKL augmente la précision tout en préservant le rappel à des niveaux proches de ceux de K-moyennes et de MKL séparés. La figure 30 est un ensemble de courbes qui mesurent le taux de couverture des différents types d'objets dans les sous-classes de MKL. La figure 31 montre que les logos (figure 31(a)) ont un faible poids en dehors de la première classe de K-moyennes et le rappel de 70% est subdivisé entre les 4 sous-classes pour adoucir le poids et permettre ainsi une meilleure distinction des types d'objets. Pour les photos (figure 31(b)), les rappels, concentrés sur 2 classes

dans K-moyennes, sont ventilés entre 5 sous-classes avec des valeurs différentes qui vont influencer sur l'ordre de l'image concerné dans la liste de pertinence. La même remarque est portée aux graphes dont la nouvelle répartition est à la figure 31(c).

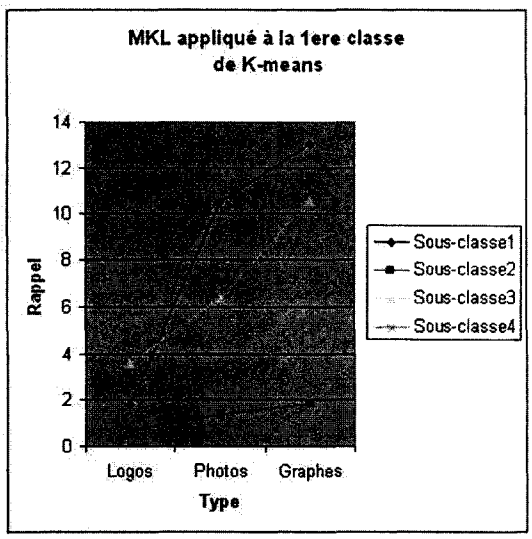
Tableau XIII

Rappels vs. Précisions de MKL à 4 sous-classes appliqué à la première classe de K-moyenne

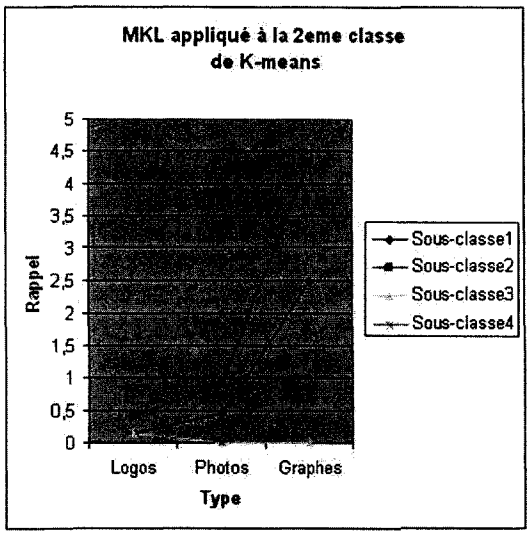
Sous-classe	Nombre d'objets total	Nombre d'objets non textuels	Précision	Rappel	F-mesure
1	615	553	89.91%	49.65%	63.96%
2	538	303	56.31%	27.20%	36.68%
3	291	100	34.36%	8.98%	14.23%
4	87	32	36.78%	2.87%	5.33%
TOTAL	1531	988			

Homogénéité des classes

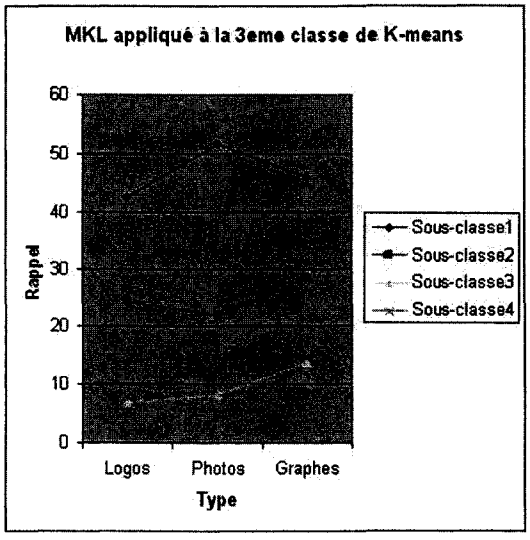
D'après les résultats présentés à la figure 34, nous pouvons affirmer que la combinaison de K-moyenne et de MKL donne de meilleurs résultats. Cependant, comme nous l'avons indiqué précédemment, il est important de prendre en compte la ventilation des types d'objets dans chaque sous classe, d'autant plus que notre application portant sur des images de documents impose une interprétation assez fiable des objets afin de fournir une réponse adéquate à l'utilisateur. Dans ce cas, il est important d'analyser le contenu des sous classes pour vérifier l'homogénéité des classes. En ce qui concerne la combinaison des algorithmes K-moyennes et MKL, nous obtenons les meilleurs résultats sur la composante précision qui est de 89.91% et de 56.31% sur la première et la deuxième sous-classes respectivement. Et comme 90% des éléments de la première classe sont non textuels, nous pouvons nous rendre compte de par les améliorations que la combinaison des deux algorithmes est un bénéfice du côté de la décentralisation des objets. La subdi-



(a) MKL appliqué à la première classe de K-moyennes



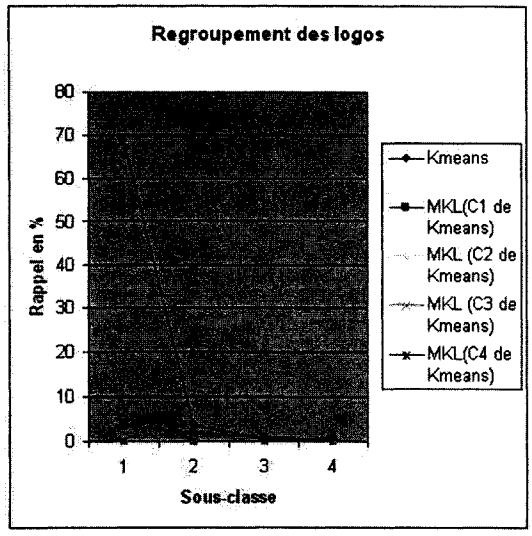
(b) MKL appliqué à la deuxième classe de K-moyennes



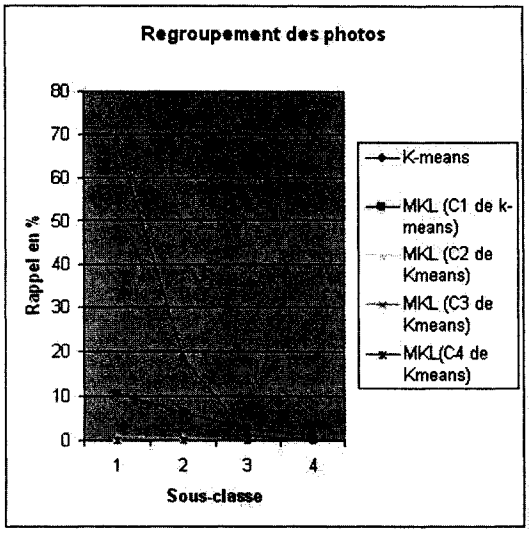
(c) MKL appliqué à la troisième classe de K-moyennes

Figure 30 Rappels obtenus après décomposition des classes de K-moyennes par MKL

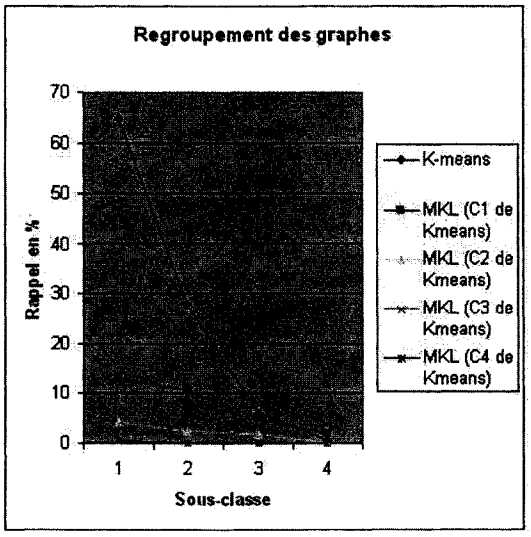
vision des classes importantes en sous classes améliore la distribution des objets d'intérêt pour augmenter l'efficacité de la recherche par la suite.



(a) Rappels des logos



(b) Rappels des photos



(c) Rappels des graphes

Figure 31 Rappels après décomposition des classes de K-moyennes par MKL

Comparaison des classes

Comme nous l'avons présenté dans la structure d'indexation à la section 6.4.2, l'arbre d'indexation est composé de quatre noeuds de premier niveau ; c'est à dire quatre groupes K-moyennes, qui sont subdivisés en quatre sous groupes MKL qui forment les noeuds

de deuxième niveau. Il est primordial de définir la fonction d'objectif à optimiser sur la base de l'évaluation de la précision versus le rappel. Dans notre cas, nous considérons une précision de 90% pour un rappel de 50% dans la première sous classe comme un résultat prometteur. Les graphiques 30 et 31 prouvent que la dominance d'une classe sur les autres, que ce soit pour les objets non textuels en général ou pour un type particulier, diminue et que le pouvoir discriminatoire croît en conséquence. On remarque que 52% et 22% des photographies, 42% et 24% des logos et 46% et 29% des graphes sont respectivement à la première et à la deuxième classe. Toutefois, le tableau XIV récapitule les meilleurs

Tableau XIV

Tableau récapitulatif des meilleurs résultats de K-moyennes et de MKL

Méthode	Nombre d'objets	Objets non textuels	Précision	Rappel	F-mesure
SKCS	1759	1026	6.56%	90.75%	12.23%
K-moyennes					
(classe 1)	1531	988	64.53	88.68	74.7%
(classe 2)	159	21	13.20	1.88	3.30%
MKL					
(souss-classe 1)	1125	832	73.95%	74.68%	74.31%
(sous-classe 2)	359	111	30.92	9.96	15.07%
K-moyennes+MKL					
classe 1					
sous-classe 1	615	553	89.91%	49.65%	63.96%
sous-classe 2	538	303	56.31%	27.20%	36.68%

résultats dans les différentes classes, la subdivision produit une meilleure ventilation des objets et différencie les différents types d'objets, mais une bonne discrimination repose sur la finesse dans la distinction des objets par les caractéristiques. Ces premiers résultats déterminent des poids à appliquer lors de la recherche et influe dans le classement des images considérées comme pertinentes.

7.3.4 Classification sur la base de test

Dans le cadre de l'analyse et de l'interprétation des résultats observés lors de l'apprentissage, des tests sont réalisés pour prouver la robustesse de notre modèle de localisation et de reconnaissance. Les expériences sont menées sur la base de test qui est composée de 237 images contenant 561 objets non textuels dont 125 sont des logos, 124 des photos et 177 des graphes. Les critères principaux d'évaluation des tests sont la qualité de la reconnaissance et de la localisation de symboles non textuels. D'autres critères seront étudiés pour compléter l'analyse : étude fine par classe, par nombre de symboles, etc. Les résultats obtenus sur la base des images tests sont comparés à ceux obtenus sur la base d'apprentissage. Ces vérifications fournissent une mesure de la couverture et de la stabilité de notre modèle de localisation et de reconnaissance.

Test des algorithmes K-moyennes et MKL séparément

La figure 32 illustre les rappels des objets non textuels sur la base de test. La comparaison avec les figures 25 et 29 relatives à la base d'apprentissage permet de constater les mêmes tendances dans l'allure des courbes. En effet, 90% des objets non textuels sont regroupés dans les quatre premières classes, malgré l'augmentation du nombre de classes. Les dernières classes ne contiennent que très peu d'objets et il est indispensable de réitérer le processus de classification sur les classes les plus importantes pour améliorer les mesures de précision qui sont très moyennes à ce niveau du traitement.

La répartition des zones non textuelles en logos, photos et graphes est présentée dans les tableaux XV à XVIII. Le tableau XV donne le résultat de la classification avec K-moyennes à quatre classes : on observe que les deux premières classes s'accaparent à elles seules au delà de 80% de logos, photos et graphes. La même tendance apparaît au tableau XVI qui ventile par type d'objet le contenu des classes de MKL à 4 classes : le nombre de logos, photos et de graphes dépasse les 80% pour les deux premières classes, alors que l'ajout de classes supplémentaires a réduit ce taux à 75% réparti sur 3 classes différentes. La figure

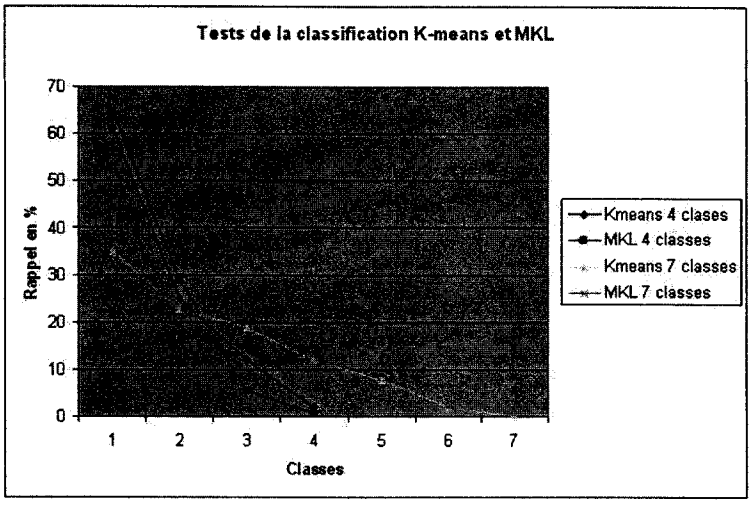


Figure 32 Rappels des zones non textuelles sur la base de test

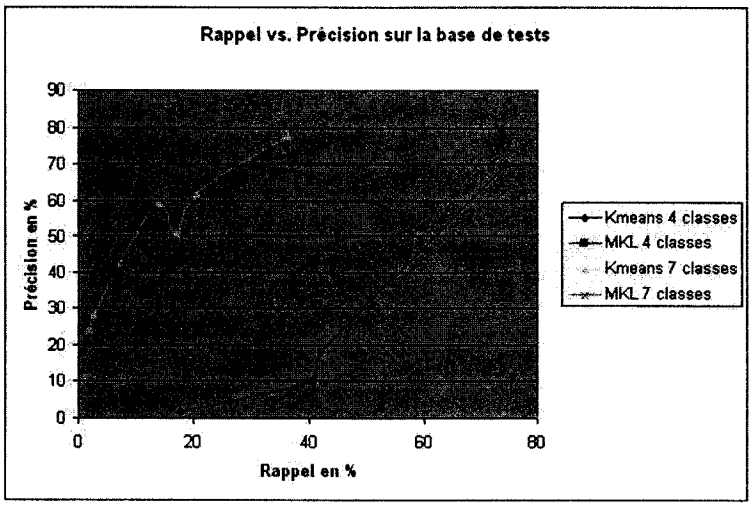


Figure 33 Rappels vs. Précision sur la base de tests pour K-moyennes et MKL séparés

33 montre que la précision augmente et le rappel diminue lorsque le nombre de classes croît. Pour la classification avec MKL, les tableaux XVII et XVIII prouvent que l'augmentation du nombre de classes n'améliore pas l'homogénéité des classes. La première classe contient au delà de 50% de chaque type d'objet, ce qui prouve que le recouvrement des logos, photos et graphes est élevé dans les classes importantes de la classification par

Tableau XV

Test par K-moyennes à 4 classes

Classe	Nombre d'objets	Objets non textuels	Précision en %	Rappel en %	Logos	Photos	Graphes
1	526	392	74.52	69.87	111	64	83
2	242	102	42.14	18.18	8	38	61
3	140	57	40.71	10.16	4	16	33
4	57	10	17.54	1.78	0	7	0
Total	965	561			123	125	177

Tableau XVI

Test par K-moyennes à 7 classes

Classe	Total objets	Objets non textuels	Précision en %	Rappel en %	Logos	Photos	Graphes
1	263	205	77.94	36.54	48	32	50
2	187	116	62.03	20.67	35	29	38
3	187	95	50.80	16.93	29	21	33
4	131	77	58.77	13.72	7	18	34
5	96	41	42.7	7.30	4	9	22
6	56	16	28.57	2.85	0	9	0
7	45	11	24.44	1.96	0	7	0
Total	965	561			123	125	177

K-moyennes ou MKL séparés. Pour palier à ces limites, la section suivante est une combinaison de K-moyennes et de MKL pour diminuer le recouvrement des types d'objets et pour augmenter la précision dans les classes importantes.

Tableau XVII

Test par MKL à 4 classes

Classe	Total objets	Objets non textuels	Précision en %	Rappel en %	Logos	Photos	Graphes
1	589	325	55.17	57.93	81	85	78
2	258	179	69.37	31.90	26	29	79
3	90	52	57.77	9.26	13	11	20
4	28	5	17.85	0.89	3	0	0
Total	965	561			123	125	177

Tableau XVIII

Test par MKL à 7 classes

Classe	Nombre d'objets	Objets non textuels	Précision en %	Rappel en %	Logos	Photos	Graphes
1	512	373	72.85	66.49	65	77	125
2	224	98	43.75	17.47	23	27	36
3	85	34	40	6.06	12	9	4
4	63	28	44.44	4.99	19	6	0
5	51	18	35.29	3.21	1	0	12
6	19	7	36.84	1.25	0	6	0
7	11	3	27.27	0.53	3	0	0
Total	965	561			123	125	177

Test de la combinaison K-moyennes et MKL

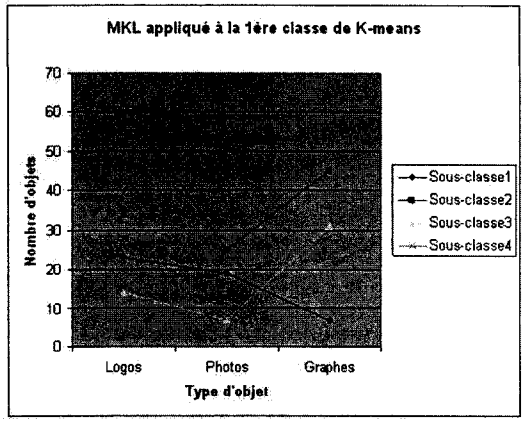
La combinaison de K-moyennes et de MKL a pour but de détecter les classes importantes de K-moyennes pour appliquer ensuite MKL dans le but d'augmenter la précision et l'efficacité dans les sous-classes de MKL. En effet, la combinaison de K-moyennes et de MKL constitue une approche intéressante pour la résolution de problèmes de discrimination ob-

servés lors de l'utilisation de ces classifieurs séparément. La méthode de K-moyennes est tout d'abord appliquée ; puis les performances obtenues sur les classes importantes sont analysées et comparées aux résultats obtenus dans les sous-classes de MKL. Les figures 34 et 35 montrent une nette amélioration de la précision et du recouvrement des types d'objets dans plusieurs sous-classes. La figure 34 montre la ventilation des d'objets dans les différentes sous-classes obtenues par la combinaison de K-moyennes et de MKL. Il apparaît que la subdivision des classes importantes de K-moyennes permette effectivement d'améliorer le recouvrement des différents types d'objets tout en augmentant la précision comme c'est le cas dans le tableau XIX où la précision dépasse les 70% dans trois des quatre sous-classes. Dans la figure 35, Les logos sont très présents dans la sous-classe 2 de la première classe de K-moyennes, les photos dans les sous-classes 3 et 4 de le première classe et la deuxième classe respectivement. Les graphes sont en majeure partie dans les sous classes 2 et 3 de la première classe et aussi dans la deuxième sous-classe de la 2ème classe. Toutes ces localisations présentent une précision meilleure que celles des classes de K-moyennes ou de MKL séparés.

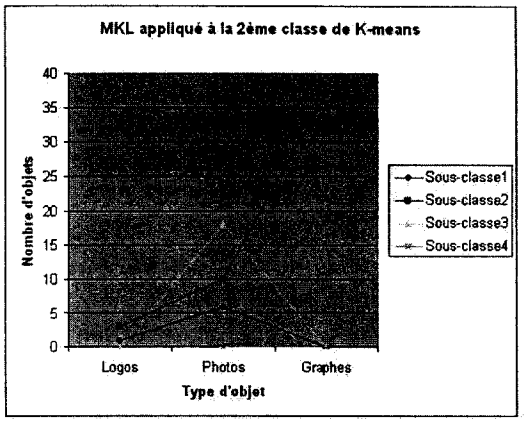
Tableau XIX

Rappels vs. Précisions de MKL à 4 sous-classes appliqué à la première classe de K-moyenne

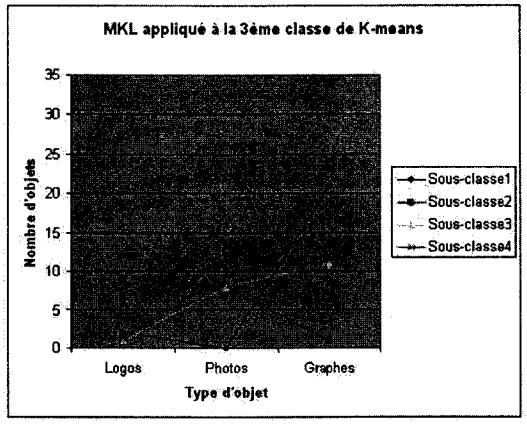
Sous-classe	Nombre d'objets total	Nombre d'objets non textuels	Précision	Rappel	F-mesure
1	262	204	77.86%	36.36%	49.57%
2	112	82	73.21%	14.61%	24.36%
3	92	70	76.08%	12.47%	21.44%
4	60	36	60%	6.41%	11.59%
TOTAL	1531	988			



(a) MKL appliqué à la première classe de K-moyennes



(b) MKL appliqué à la deuxième classe de K-moyennes

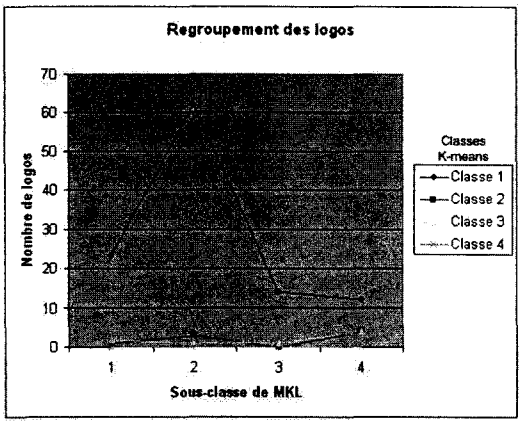


(c) MKL appliqué à la troisième classe de K-moyennes

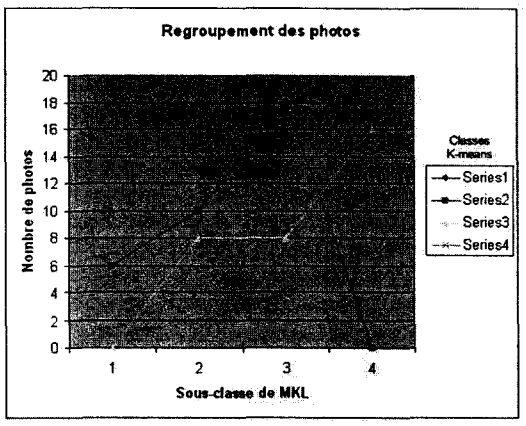
Figure 34 Ventilation des objets lors de la décomposition des classes de K-moyennes par MKL

7.4 Système d'interrogation

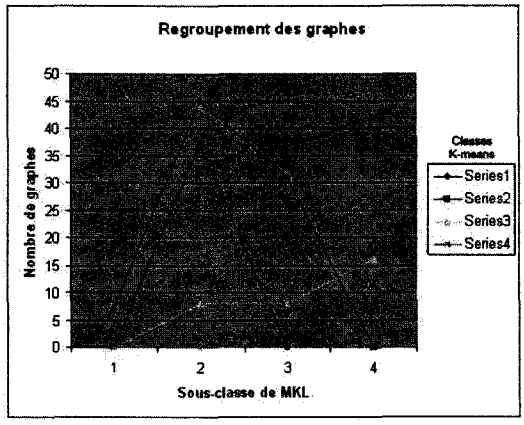
L'usager est celui qui définit une requête correspondant à son désir d'information et qui attend une liste, précise et pertinemment ordonnée, des documents. Le modèle de recherche vectoriel que nous avons développé présente différentes solutions performantes, robustes et relativement simples à mettre en oeuvre.



(a) Regroupement des logos



(b) Regroupement des photos



(c) Regroupement des graphiques

Figure 35 Test des regroupements par type d'objet obtenus par la combinaison de MKL et de K-moyennes

Le graphe rappel-précision est le plus utilisé pour comparer les résultats obtenus par des systèmes différents. Les courbes peuvent être superposées sur le même graphique pour déterminer le meilleur système. Les comparaisons sont effectuées dans trois intervalles de rappel : 0 à 0.2, 0.2 à 0.8, et 0.8 à 1. Ces trois intervalles caractérisent les performances bas-rappel, mi-rappel et haut-rappel, respectivement.

7.4.1 Requêtes sur la base d'apprentissage

Dans le processus de recherche, la performance est déterminée par l'utilisation de cinquante requêtes sélectionnées aléatoirement à partir du dictionnaire des mots de la base. Chacune porte sur 1 ou plusieurs mots. Notre méthode est comparée aux approches telles que le modèle vectoriel SMART sans expansion de requête (appelée Smart dans la suite) et le Q-gram de Ukkonen's (Ukkonen, 1983) avec des distances de 1 à 4. Un exemple de requête et des listes de pertinences obtenues est donné à la figure 36.

Requête: Information retrieval —→ Lemmatisation —→ inform + retriev

Smart Original	Smart OCR	Remarque	3-Grams	Notre approche	Remarque
654 Sim:0.37	931 Sim: 0.59	654 retrogradé	414 Sim:0.68	931 Sim:0.59	
931 Sim:0.35	394 Sim: 0.59		951 Sim:0.67	394 Sim:0.59	
394 Sim:0.35	945 Sim: 0.56		949 Sim:0.66	945 Sim:0.56	
608 Sim:0.32	408 Sim: 0.56	608 disparu	411 Sim:0.61	408 Sim:0.56	
945 Sim:0.30	951 Sim: 0.51		418 Sim:0.42	951 Sim:0.51	
147 Sim:0.30	654 Sim: 0.51	147 disparu	955 Sim:2.39	654 Sim:0.51	
408 Sim:0.29	414 Sim: 0.51		954 Sim:0.39	414 Sim:0.61	
418 Sim:0.26	948 Sim: 0.44	418 rétrogradé	417 Sim:0.39	948 Sim:0.44	
414 Sim:0.26	411 Sim: 0.43		950 Sim:0.36	411 Sim:0.43	
948 Sim:0.25	955 Sim: 0.38		953 Sim:0.36	955 Sim:0.38	
411 Sim:0.25	418 Sim: 0.38		416 Sim:0.36	418 Sim:0.38	
955 Sim:0.24	949 Sim: 0.37		949 Sim:0.34	949 Sim:0.37	
951 Sim:0.23	412 Sim: 0.36		521 Sim:0.34	412 Sim:0.36	
950 Sim:0.23	950 Sim: 0.34		412 Sim:0.34	950 Sim:0.34	
413 Sim:0.22	413 Sim: 0.28		952 Sim:0.34	413 Sim:0.28	
949 Sim:0.17	953 Sim: 0.27		656 Sim:0.34	953 Sim:0.27	
521 Sim:0.17	416 Sim: 0.27		654 Sim:0.33	416 Sim:0.27	
590 Sim:0.16	147 Sim: 0.23		590 Sim:0.33	147 Sim:0.23	
412 Sim:0.16	521 Sim: 0.22		415 Sim:0.32	521 Sim:0.22	
954 Sim:0.15	954 Sim: 0.21		945 Sim:0.32	954 Sim:0.21	
77 Sim:0.15	590 Sim: 0.21		408 Sim:0.29	590 Sim:0.21	
396 Sim:0.13	417 Sim: 0.21		933 Sim:0.27	417 Sim:0.21	
155 Sim:0.13	155 Sim: 0.21		396 Sim:0.26	155 Sim:0.21	
415 Sim:0.12	933 Sim: 0.20		455 Sim:0.26	933 Sim:0.20	
148 Sim:0.12	396 Sim: 0.20		413 Sim:0.26	396 Sim:0.20	
953 Sim:0.11	952 Sim: 0.18		941 Sim:0.26	952 Sim:0.18	
952 Sim:0.11	415 Sim: 0.17		404 Sim:0.25	415 Sim:0.17	
933 Sim:0.11	77 Sim: 0.16		457 Sim:0.23	77 Sim:0.16	
417 Sim:0.11	148 Sim: 0.16		345 Sim:0.21	148 Sim:0.16	
855 Sim:0.10			947 Sim:0.21	147 Sim:0.009	
734 Sim:0.10			410 Sim:0.18	320 Sim:0.008	
710 Sim:0.10			882 Sim:0.18	932 Sim:0.009	(i,l)
652 Sim:0.10			409 Sim:0.17	955 Sim:0.007	320 &932 intercalés
416 Sim:0.10			660 Sim:0.14	6 Sim:0.007	car (iev , ly)
318 Sim:0.10			948 Sim:0.14	608 Sim:0.006	6 & 882 car (fe ,fl)
			121 Sim:0.14	318 Sim:0.006	
			155 Sim:0.09	882 Sim:0.003	
			455 Sim:0.09	710 Sim:0.001	
			932 Sim:0.09	652 Sim:0.001	
			434 Sim:0.08	416 Sim:0.000	
			734 Sim:0.06	415 Sim:0.000	
			524 Sim:0.06	734 Sim:0.000	
			397 Sim:0.05	318 Sim:0.000	
			.....		

Figure 36 Exemple de réponses à une requête textuelle "information retrieval"

La figure 37 résume le rappel versus la précision sur la collection d'apprentissage, notre approche réalise une amélioration en termes de rappel et de précision.

Le tableau XX récapitule le rappel et la précision sur l'ensemble d'apprentissage, la précision moyenne est entre 97.81% et 99.84% pour le bas-rappel, entre 83.62% et 96.95% pour le moyen-rappel et entre 23.53% et 76.08% pour le haut-rappel. La technique "3-grammes" permet la sélection d'une large gamme de mots, et les résultats sont un rappel élevé contre une faible précision. En outre, la précision moyenne pour tous les documents pertinents (toutes requêtes confondues) est de 87.68% pour notre approche, de 86.53% pour la recherche vectorielle sans expansion et n'excède pas 65.99% pour la technique du 3-grammes. Notre approche a l'avantage de n'utiliser que les parties du mot susceptibles d'être erronées, ce qui lui confère la meilleure performance au haut-rappel.

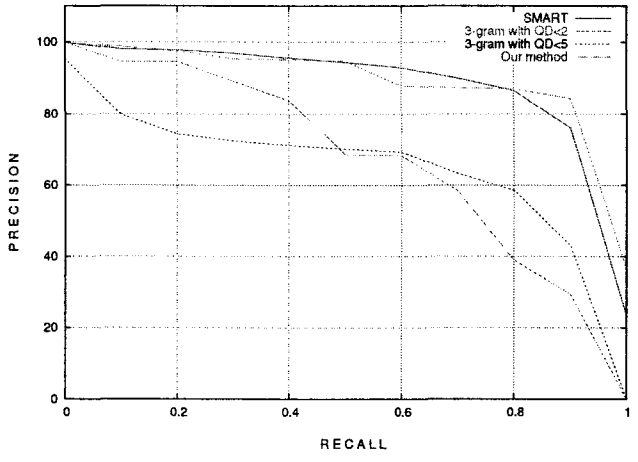


Figure 37 Rappels et précisions moyennes sur la collection d'apprentissage

Notre approche réalise une meilleure efficacité comparée aux autres méthodes. Ceci est dû à l'utilisation de critères statistiques et d'un classement des mots construits à l'aide des erreurs-grams selon leurs importances. Un exemple de traitement pour la requête "le mot schultz dans quelques journaux ?" : après avoir éliminé les mots inutiles et procéder à la lemmatisation, la requête devient "schultz journ" et l'expansion de la requête avec la technique 3-grammes ajoute des chaînes de caractères telles que "schultz journ sch chu hul ult ltz jou our urn". Si le mot "schultz" est identifié dans un document comme "sehnltz", la distance de QD entre ces mots est 4. Mais avec notre méthode, la requête sera augmentée

Tableau XX

Rappels et précisions moyennes sur l'ensemble d'apprentissage

Méthodes	bas-rappel	mi-rappel	haut-rappel	précision moyenne
Smart	97.81-99.84	83.62-96.95	23.53-76.08	86.53
Recouvrement 3-grams	94.64-99.9	38.98-89.22	0-29.35	65.99
Notre méthode	97.50-99.72	87.07-95.35	36.61-84.44	87.68

par le terme "sehnltz" qui présente une probabilité d'erreur de 0.002. L'image du document concernée serait considérée et classée dans la liste des documents pertinents. Pour apprécier la performance de notre approche, nous devrions mentionner que dans le deuxième intervalle (mi-rappel), notre système surpasse SMART de 4% et l'écart augmente jusqu'à 13.08% dans le troisième intervalle (haut-rappel). Ceci montre l'importance de modéliser les erreurs de reconnaissance pour les utiliser dans le processus de recherche.

7.4.2 Requêtes sur Test1, Test2 et Test-dégradé

Il est intéressant d'étendre nos expériences à un large éventail d'images de documents. Pour ce faire, nous avons testé notre système et comparé les résultats sur les collections d'images originales et dégradées en se basant sur la performance de la recherche. Les résultats des tests sur la collection "Test1" sont à la figure 38 et ceux de la collection "Test2" sont présentés au tableau XXI. On remarque la même tendance que pour l'ensemble d'apprentissage, excepté l'approche 3-grammes avec une petite distance, qui performe avec une précision élevée. Ceci est dû au petit nombre d'images de la collection "Test" comparé à l'ensemble d'apprentissage. Pour la collection "Test1", la précision moyenne pour tous les documents pertinents (toutes requêtes confondues) est de 87.90% pour notre approche contre 85.33% pour la recherche vectorielle sans expansion et ne dépassant pas les 65.99% pour les autres méthodes.

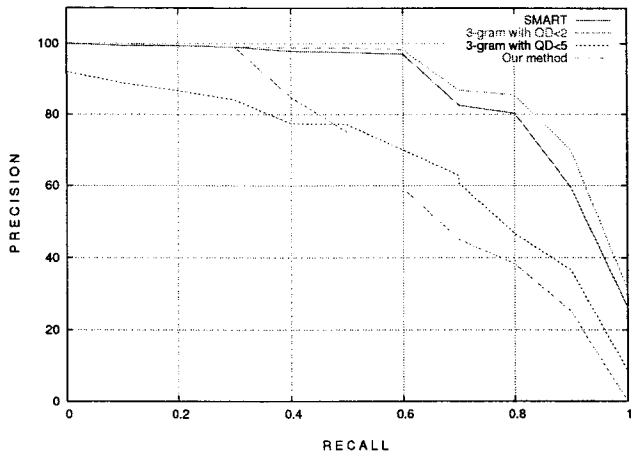


Figure 38 Rappels et précisions moyennes sur la collection "Test1"

Pour la collection "Test2", les résultats du tableau XXI prouvent que notre approche présente la meilleure précision et concourt au meilleur rappel de l'approche 3-grammes. Ceci s'explique par l'existence de mots à faible fréquence dans les documents, tels que les noms de personnes ou de lieux, et qui sont mal reconnus. L'exemple du mot "Tauberian", qui apparaît deux fois dans un document et qui est identifié comme "Tauoenan", est considéré seulement par notre méthode, avec une probabilité de 0.0009. Notons que l'expansion de requête améliore l'efficacité de la recherche, particulièrement quand les erreurs de reconnaissance portent sur des noms propres ou des documents courts (manque de redondance). L'influence des termes ajoutés grandit au fur et à mesure que le rappel augmente.

Pour les images dégradées, nous savons que le taux de reconnaissance est très faible. Nous remarquons dans les figures 39 et 40 que la précision atteint 65.54% pour des rappels inférieures à 50%. Au delà, la performance décroit et toutes les approches suivent la même tendance. La performance est meilleure avec notre expansion de requête au niveau de l'intervalle haut-rappel pour la collection dégradée par des photocopies. La précision moyenne globale est de 63.16% pour notre méthode, mais elle décroit à 60.44% pour l'approche 3-grammes à large distance, et jusqu'à 60.66% pour la recherche vectorielle sans expansion. On remarque la même tendance pour les images dégradées où la préci-

Tableau XXI

Éfficacité de la recherche sur la collection "Test2"

méthode de recherche	Rappel moyen	Précision moyenne
Notre approche	68.40%	88.82%
Smart sans expansion	41.88%	91.32%
3-gram recouvrement		
$QD \leq 1$	63.46%	70.08%
$QD \leq 4$	75.76%	60.64%

sion moyenne globale est de 59.86% pour notre méthode et de 55.62% pour le système vectoriel standard. Cependant, le taux diminue à 50.74% pour 3-grammes avec de larges distances. Ceci est dû à la capacité de l'approche 3-grammes d'extraire des parties de mots et au faible nombre d'images dans la collection de test.

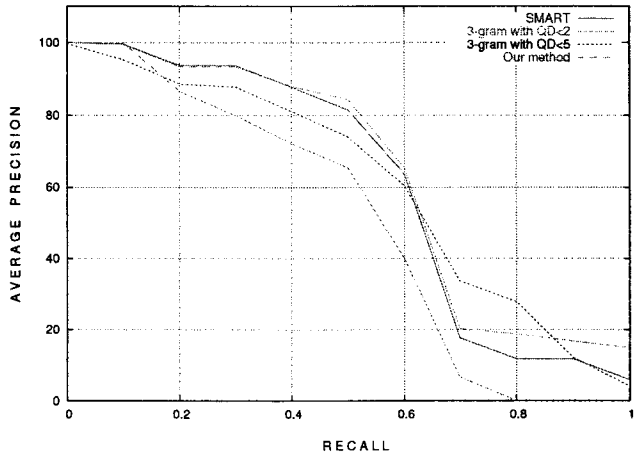


Figure 39 Rappels vs. Précisions moyennes des images dégradées par des photocopies

Les résultats obtenus pour la collection des images très dégradées, qui sont à la figure 41, montrent une baisse de la précision. Notons que les 3-grammes demeurent meilleurs dans l'intervalle mi-rappel. Le problème avec l'approche 3-grammes est la baisse de la précision au fur et à mesure que la qualité-distance (QD) augmente. Entre les intervalles

haut-rappel et haute-précision, la précision moyenne devient inférieure à 50% pour le mi-rappel et moins de 3% pour le haut-rappel ; excepté notre méthode, qui est maintenue à 6.69%. La précision moyenne globale est de 38.46% pour notre méthode, mais diminue à 33.64% pour les 3-grammes à petite QD et à 28.48% pour le modèle vectoriel standard. Nous déduisons à partir des figures reflétant la performance de la recherche que notre

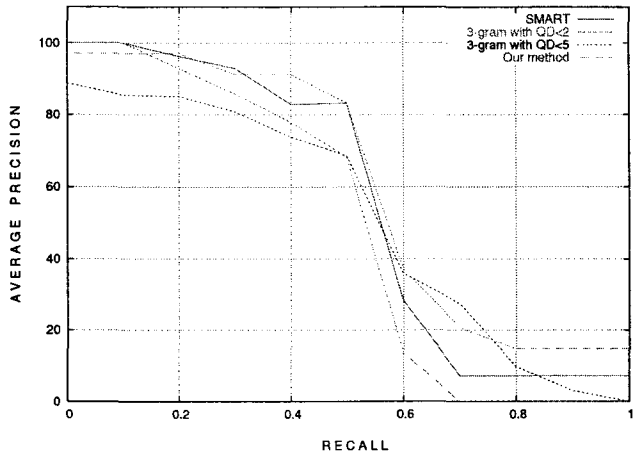


Figure 40 Rappels vs. Précision moyennes sur les images dégradés

algorithme est meilleur que les autres méthodes particulièrement sur les images dégradées. Nous avons une meilleure méthode pour ne sélectionner que les bonnes combinaisons pour l'expansion de la requête (les grammes qui sont susceptibles d'être erronés). On a amélioré ainsi la performance de la recherche dans les images de documents. Il est intéressant de voir comment la qualité d'image affecte la recherche d'information. Toutes les approches ont beaucoup de mal à améliorer la performance de la recherche lorsqu'il s'agit d'images très dégradées.

7.4.3 Requêtes portant sur les zones non textuelles

Nos expériences ont porté sur la collection d'images utilisée lors de la classification automatique des objets non textuels. L'arborescence d'indexation est donc contruite sur les images de la base UW-2 et des pages web. Nous avons appliqué des valeurs allant de 0

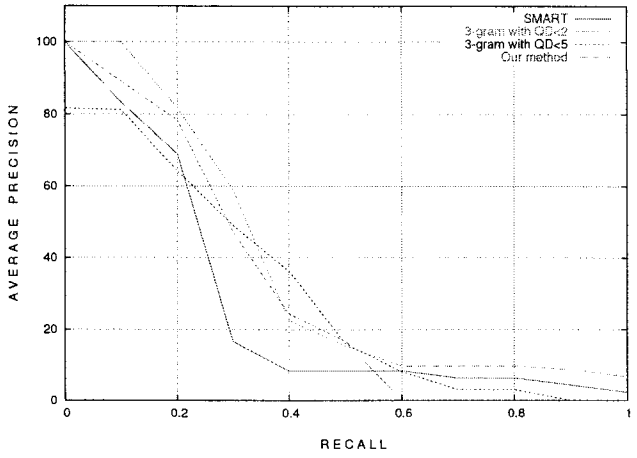


Figure 41 Rappels vs. Précisions moyennes sur les images très dégradées

(texte seulement) à 1 (graphique seulement) pour la variable λ de l'équation (6.4) du paragraphe 6.5. Les résultats des différents modes de recherche sont présentés au tableau XXII.

Tableau XXII

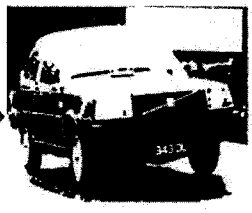
Éfficacité de la recherche des zones graphiques

Méthodes de recherche	Bas-rappel	Mi-rappel	Haut-rappel	Précision moyenne%
Texte seulement	83 à 90%	38 à 85%	15 à 30%	75%
Image seulement	79 à 85%	32 à 74%	10 à 23%	50%
Combinaison texte-non texte	93 à 95%	80 à 92%	34 à 79%	85%

L'exemple montré à la figure 42 montre que l'interrogation par une image exemple seulement retourne comme réponses pertinentes les images de documents contenant un graphique de forme similaire (les documents numéros 199 et 21 contiennent des images de voitures qui ne sont pas de marque "volvo"). Nous avons testé l'interrogation basée sur des mots seulement "volvo", le jugement de pertinence est basé cette fois-ci sur la fréquence du mot volvo sans considération du contenu graphique (les documents 522, 521 et 520

Cueillette d'une photographie de l'image n° 20 et la lancer comme image exemple

→ Segmentation et extraction →



---→ <20 1 0.16 5420 ... 0 15 242 355 ... >
---→ Sélection K-means ---→ classe 1 (988 objets)
---→ Projection ϕ ---→ <20 1 0.40 8.44 0.47 ...>
---→ Sélection MKL ---→ sous-classe1 (291 objets)

Recherche mot 'VOLVO'	Remarque	Image exemple	'volvo' + image $\lambda = 0.5$	Remarque
522 Sim: 0.79 521 Sim: 0.67 520 Sim: 0.56 20 Sim: 0.56 ...	20 à la 4ème place à cause de la fréquence du mot dans l'image 20.	20 Sim:1 199 Sim:0.8 21 Sim:0.6 ... On a 223 images contenant les 291 objets.	20 Sim:0.9 521 Sim : 0.8 522 Sim :0.25 520 Sim :0.25 523 Sim :0.1 ...	On donne 50 % aux images qui contiennent le mot 'VOLVO'. Toutes les images parasites dans l'image exemple passent en queue de liste.

Figure 42 Exemple de réponses à une requête portant sur du graphique "Voiture de marque VOLVO"

tournés comme pertinents ne contiennet pas d'images de voitures). L'utilisation combinée des index textuels et graphiques permettent de retourner les documents considérés comme pertinents dans les deux processus précédents (texte et image), c'est ainsi que le document

numéro 20 duquel on a extrait l'image exemple reprend la première place sur la liste des réponses pertinentes.

La recherche basée seulement sur le texte OCR et l'expansion de la requête ($\lambda = 1$) répond à des interrogations de type "logo du constructeur VOLVO". La précision moyenne pour ce type de requêtes est de 75%. Les documents pertinents se trouvent plutôt dans le haut rappel où la précision reste faible. La présence de zones graphiques implique une fréquence moins élevée des mots contrairement aux images à dominance textuelle et influe donc sur le classement des images considérées comme pertinentes.

La recherche basée sur les zones graphiques seulement ($\lambda = 0$) traite des requêtes portant sur des images exemples. La précision est faible dans le haut rappel à cause de la différence des tailles avec les zones à rechercher, des graphiques mal reconnus lors de la segmentation et des classes importantes de la classification automatique. En effet, les classes qui contiennent un nombre important d'objets présentent un faible facteur discriminant, ce qui retourne une liste des images pertinentes où les logos, photographies et graphes sont présents. Toutefois, la précision moyenne, toutes requêtes confondues, atteint 75%.

Enfin, la combinaison texte - non texte permet de remédier à certains problèmes sus-cités. Le texte reconnu par OCR rend pertinentes des images dont les zones graphiques sont mal reconnues. Le pouvoir discriminant se trouve renforcé par l'indexation multi-niveau qui nous informe sur les graphiques présents dans la base d'images. La combinaison de ces résultats avec la liste des images considérées comme pertinentes textuellement améliore la performance de la recherche. La précision moyenne, toutes requêtes confondues, atteint 85% et l'amélioration par rapport aux deux autres approches est très remarquée dans la zone de haut rappel (la précision commence à 35% pour atteindre les 70%), ce qui est très encourageant. Des techniques comme la mise à jour de l'indexation non textuel par un retour de la pertinence ou par la constitution d'une base des faits (collection des jugements des utilisateurs) amélioreraient grandement la précision pour ce type de recherche.

7.5 Conclusion

La localisation et la reconnaissance des différentes régions informationnelles de l'image de document est une étape primordiale pour une recherche d'information adéquate. Les premiers résultats obtenus sur la base d'apprentissage sont très prometteurs. Par ailleurs, l'estimation des paramètres de la segmentation et des classifieurs a été effectué sur la base des réalisations pratiques mais d'autres techniques d'estimations sont envisageables. Une décomposition classique des objets en étudiant l'apport des caractéristiques et la dimentionalité doit être un axe de recherche fort.

Pour les zones textuelles, la reconnaissance par OCR n'est pas exempt d'erreurs. Ainsi, l'extension de la requête concentre en tête de liste les documents pertinents, plus qu'elle n'agit sur la précision au détriment du rappel comme cela est souvent affirmé en recherche d'information. D'un point de vue pratique, cette extension de requêtes par les erreurs-grams permet d'améliorer légèrement les performances globales de notre système de recherche. L'extension par erreur-grams permet donc de faire émerger des images qui n'auraient pas nécessairement été trouvées par une extension manuelle. Nous avons discuté de l'utilisation de l'OCR et de la qualité de la reconnaissance qui est fonction de la nature des dégradations subies par les images. Une centaine de pages Web et 700 images dégradées ont été examinées afin de mesurer l'impact de la qualité de l'image et de l'extension des requêtes sur la performance de la reconnaissance du texte et de la recherche d'information. Les expériences menés ont montré une amélioration de la recherche par rapport aux méthodes basées sur la recherche exacte ou l'appariement par chevauchement de parties de mots de la requête. Beaucoup de perspectives restent ouvertes à l'issue de ces expériences. Tout d'abord, notre mise-en-oeuvre de l'extension de requête pourrait se renforcer par un dictionnaire sémantique. Il serait en effet intéressant de choisir des mots reliés à ceux de la requête par l'utilisation de Wordnet. WordNet est une base de données lexique, où les sens des mots sont structurés sur des relations lexico-sémantiques.

Enfin, on peut également inclure ces ressources sémantiques non plus seulement à l'interrogation, mais dès la phase d'indexation. Cela nécessite alors de définir une représentation plus complexe que la simple représentation vectorielle usuellement utilisée et reste un problème ouvert.

CHAPITRE 8

CONCLUSION GÉNÉRALE ET PERSPECTIVES

Nous avons présenté dans ce travail de recherche quatre approches pour la recherche d'information dans les images de documents à partir de données textuelles et graphiques. Le pari est d'assister l'utilisateur des images de documents scannés et des archives en le soulageant des travaux routiniers et fastidieux qui sont la localisation, la reconnaissance et la recherche d'information.

La première approche présentée dans cette thèse traite de la localisation et de l'identification de l'information contenue dans l'image. Le problème de la localisation des régions informatives a tout d'abord été abordé par une méthode basée sur le concept multi-échelle pour identifier les régions à partir des formes obtenues lors de la segmentation. Elle est formulée comme un problème de détection de formes et de textures résolu par un algorithme de lissage de bruit à différents niveaux. Nous avons utilisé dans un premier lieu une méthode ascendante pour la détection d'objets basée sur la Gaussienne pour lisser le bruit et créer un espace multi-échelle et le Laplacien pour localiser les surfaces informationnelles de l'image. La réutilisation de toute l'information pertinente à des échelles successives est une approche hybride pour suivre l'évolution des régions informationnelles. Cette méthodologie permet la détection de la forme et de la position des régions à partir d'une hiérarchie indicée. La méthode proposée est validée sur une base de 512 images provenant de différents médias qui ont été segmentées par l'opérateur Laplacien de la Gaussienne. Les résultats montrent que la segmentation par cette méthode fournit des informations qui aident à l'analyse du contenu.

Par contre, certains blocs de texte restent collés aux régions graphiques. C'est pour cela que nous avons migré vers l'opérateur SKCS (Separable Kernel with Compact Support) dans le but d'isoler les régions graphiques pour des fins de reconnaissance. C'est un opé-

rateur à noyau avec support compact qui offre plusieurs degrés de liberté au lieu du seul paramètre de déviation σ pour l'opérateur LoG. Nous sommes arrivés pour certains paramètres de cet opérateur à réduire les sur-segmentations par un processus de fusion de blocs et à porter le taux de rappel des zones graphiques à des proportions dépassant les 90% pour le rappel et environ 60% pour la précision. Notre approche fonctionne bien dans le cas général mais ne prend pas en compte tous les cas particuliers et surtout les graphes avec des courbes dont les zones d'informations sont peu étendues et difficilement localisables. Les observations sur les résultats obtenus montrent que ces mesures de la localisation restent subjectives et la segmentation reste sensible à beaucoup de paramètres pour améliorer la détection et l'interprétation des régions segmentées. Ces problèmes restent un axe de recherche prometteur et offre beaucoup d'opportunités. Par ailleurs, il n'est pas raisonnable de chercher à traiter tous les cas. Nous en concluons qu'une bonne approche de localisation doit être capable d'analyser et de synthétiser le comportement à adopter face à des nouvelles situations.

La deuxième approche concerne la recherche d'information reliée à l'information textuelle. Le traitement de chaînes de caratères dans un corpus textuel est un axe de recherche prometteur et très fertile pour la communauté de l'analyse du document. Les textes reconnus par OCR ne sont pas exempts d'erreurs surtout pour les images de mauvaise qualité. Nous avons discuté de l'utilisation de l'OCR et de la qualité de la reconnaissance qui est fonction de la nature des dégradations subies par les images. Un millier d'images de documents pour l'apprentissage, une centaine de pages Web et 700 images dégradées pour tester la robustesse ont été examinés afin de mesurer l'impact de la qualité de l'image sur la reconnaissance du contenu textuel par un OCR.

La reconnaissance du texte par OCR est de 93.8% pour les images de la base d'apprentissage. L'utilisation d'une petite distance d'édition a permis de corriger 5185 mots, d'augmenter le taux de reconnaissance à environ 95% et de construire 2822 erreur-grams et règles de production à partir de la base d'apprentissage. Nous avons utilisé trois diffé-

rentes collections pour tester la robustesse de notre approche. Le taux de reconnaissance sur la première base de test, composée de 200 pages web est d'environ 72.26%, à cause de la qualité de l'impression et de la scanérisation, et a atteint 37.59% lorsque les images ont été dégradées par des photocopies successives ou par l'ajout de bruit ou de flou. La reconnaissance atteint même des valeurs au dessous de 10% lorsque les images sont de très mauvaise qualité.

Notre contribution réside dans le repérage de ces erreurs et l'utilisation de l'expansion de la requête pour augmenter le rappel et améliorer la performance de la recherche d'information dans un contexte de modèle vectoriel. Nous avons montré que les n-grames et les probabilités correspondantes influencent grandement les résultats et le classement des images pertinentes à une requête donnée. Nous avons comparé les résultats de notre approche à des méthodes comme le modèle vectoriel sans expansion ou le recouvrement 3-grames.

Les expériences menées ont montré que notre approche réalise une recherche plus efficace comparée aux autres méthodes grâce à l'utilisation de critères statistiques et d'un classement des mots ajoutés selon l'importance des erreurs-grams utilisées. Malgré la capacité de l'approche 3-grames d'extraire des parties de mots, nous avons apprécié la performance de notre approche qui surpasse 3-grames et SMART de 4% pour le bas-rappel et l'écart augmente jusqu'à 13.08% pour le haut-rappel. La performance de notre méthode sur les images dégradées réside dans la sélection, lors de l'expansion de la requête, de grames qui sont susceptibles d'être erronés. Il est intéressant de voir comment la qualité d'image affecte la recherche et que toutes les approches ont beaucoup de mal à améliorer la performance lorsqu'il s'agit d'images très dégradées.

La troisième approche élaborée dans cette thèse est le traitement relié aux objets non textuels. Nous avons tenté d'utiliser d'autres aspects aussi significatifs que les formes et les caractéristiques des objets, notamment la classification pour aider à différencier les

différents types de contenus. Il s'agit de prédire, à partir des caractéristiques extraites, les logos, les illustrations et les graphes. La nouvelle méthode est basée sur les spécifications de deux classifieurs et appliquée à une base d'apprentissage de 425 images contenant 1114 zones graphiques et à une base de test de 237 images contenant 561 zones graphiques.

Le premier classifieur automatique, K-moyennes, est non supervisé et repose sur une répartition grossière des objets dans différentes classes. Il est intéressant de noter que le nombre de classes ne résout pas le problème de l'homogénéité et de l'étendu des objets dans les classes. En effet, trois classes regroupent 95% des objets et malgré l'augmentation du nombre de classes, la répartition des objets demeure presque identique.

Le deuxième classifieur automatique, MKL, est aussi non supervisé et traite la projection des vecteurs caractéristiques sur un espace multi-dimensionnel. Une analyse en composantes principales basée sur l'algorithme de Karhunen-Loeve permet de réduire la dimensionnalité avant de regrouper les objets proches en sous-espaces homogènes. Nous notons par ailleurs que l'application de ce classifieur avec un nombre variable de classes ne résout pas le problème de l'homogénéité (2 classes regroupent 85% des objets) ni la différentiacion entre les logos, les illustrations et les graphes.

La combinaison de ces deux classifieurs réside dans la réutilisation des classes importantes de K-moyennes pour réduire la dimensionnalité et constituer des sous groupes minimisant l'éparpillement spatial à l'aide de MKL. Les résultats obtenus ont montré une nette amélioration de la précision (de 75% pour K-moyennes à 90% pour la combinaison) tout en maintenant le rappel autour de 76%. La répartition en sous classes réduit la dominance d'une classe sur les autres et améliore le pouvoir discriminant. En effet, 52% des photographies sont dans une sous-classe, 24% des logos dans une autre sous-classe et enfin 14% des graphes dans une troisième sous-classe. La classification sur la base des tests justifie le pouvoir discriminatoire (environ 58% des logos sont regroupés dans une sous-classe, 26% des photographies dans deux autres sous-classes et 25% des graphes dans une quatrième

sous-classe). Ces chiffres sont motivants et des voies de recherche dans ce domaine reste à explorer pour d'éventuelles améliorations.

La quatrième approche met en avant des travaux intéressants en indexation qui suggèrent la représentation hiérarchique des catégories d'objets résultat de l'approche CAH appliquée lors de la classification. L'architecture basée sur le modèle d'indexation multi-niveau permet d'organiser les objets de la segmentation tout en réduisant le recouvrement entre les différents types de graphiques. La sélection d'un noeud, à partir des objets de la requête, permet de retourner rapidement les images susceptibles d'être pertinentes. Cette arborescence a montré qu'il existe un lien entre l'indexation textuelle et les caractéristiques visuelles des objets de l'image. Ce lien nous permet d'augmenter la puissance d'interrogation et d'améliorer la performance de la recherche par l'extraction d'informations graphiques à partir des vecteurs de la structure d'indexation et l'utilisation des vecteurs des mots déduits de la reconnaissance par OCR.

Finalement, une cinquantaine de requêtes portant à la fois sur le texte et les zones graphiques des images, utilisées lors de la classification, ont montré que l'apport de l'information non textuelle à la recherche d'images n'est pas négligeable. La précision moyenne est de 75% pour l'interrogation basée sur le texte et l'expansion de la requête et elle est d'environ 50% pour les requêtes portant sur le graphique seulement. La combinaison du texte et du non texte a amélioré les réponses aux requêtes pour atteindre des valeurs dépassant les 85%. Les résultats restent globalement satisfaisants, bien que des améliorations restent possibles, notamment dans les zones de haut rappel où des techniques restent à mettre en oeuvre pour augmenter la précision.

Pour assurer la continuité de ce travail de recherche, nous proposons les améliorations potentielles suivantes :

- la taille de la base d'images, sur laquelle nous avons appliqué notre analyse de régions obtenues, reste insuffisante et ne contient pas tous les types d'informations

à considérer. Notre méthode s'applique à d'autres types informationnelles pourvu qu'une base d'images du type en question soit disponible

- la validation des méthodes de segmentation proposées en utilisant des mesures d'indices géométriques (inclinaison, étirement etc.)
- la modification de l'algorithme "distance d'édition" pour le rendre itératif et utiliser différentes valeurs pour les coûts des opérations d'édition
- la combinaison de plusieurs OCRs pour améliorer nos erreurs-grams
- l'utilisation d'un dictionnaire sémantique pour ajouter les mots reliés aux termes de la requête lors de l'expansion
- l'amélioration de la discrimination des logos, des illustrations et des graphiques pour améliorer la recherche d'information portant sur les zones graphiques
- l'étude des différents paramètres de la combinaison textuelle et graphique pour d'éventuelles optimisations
- la prise en considération de l'évaluation des réponses du système par l'utilisateur
- l'élaboration d'interfaces graphiques pour la représentation de l'information textuelle et visuelle de la requête et des réponses.

BIBLIOGRAPHIE

Amlani, M. et Kasturi, R. (1988). A query processor for information extraction from images of paper-based maps. *Proceedings of the RIAO 88*, pages 991–1000.

Babaud, J., Withkin, A., Baudin, M., et Duda, R. (1986). Uniqueness of the gaussian kernel for scale-space filtring. *IEEE Trans. PAMI*, 8(1) : 26–33.

Baird, H. (1987). The skew of printed documents. *Proceedings of the SPIE's 40 th Annual Conference and Symposium on Hybrid Imaging Systems, New York, USA*, pages 21–24.

Baird, H. (1999). Document image quality : Making fine discriminations. *Proceedings IAPR, Int. Conf. on Document Analysis and Recognition*, number 11, pages 1209–1223, Bangalore, India.

Beitzel, S., Jensen, E., et Grossman, D. (August 2002). Retrieving ocr text : A survey of current approaches. *Proceedings of the SIGIR 2002 Workshop on Information Retrieval and OCR*, Tampere, Finland.

Belaid, A. (1994). Reconnaissance des formes : méthodes et applications. Le traitement électronique de documents. Paris : ADBS, pages 49–92.

Bellot, P. et El-Beze, M. (1999). A clustering method for information retrieval. Technical report IR-0199. Laboratoire d'informatique d'Avignon, France.

Ben Braiek, E., Cheriet, M., et Doré, V. (2005). SKCS, a separable kernel family with compact support to improve visual segmentation of handwritten data. *Electronic Letters on Computer Vision and Image Analysis*, 5(1) : 14–29.

Bocchieri, E. et Wilpon, J. (1993). Discriminative feature selection for speech recognition. *Computer Speech and Language*, pages 229–246.

Bunke, H. et Csirik, J. (1975). Parametric string edit-distance and its application to pattern recognition. *IEEE Transactions Systems, Man and Cybernetics*, 25(1) : 202–206.

Cappelli, R., Maio, D., et Maltoni, D. (1999). Similarity search using multi-space kl. Dans *Proceedings of the First International Workshop Similarity Search Database and Expert Systems Applications*, pages 155–160.

Cappelli, R., Maio, D., et Maltoni, D. (2001). Multispace kl for pattern representation

and classification. *IEEE Trans. PAMI*, 23(9) : 977–996.

Carson, C., Belongie, S., Greenspan, H., et Malik, J. (1999). Blobworld : Image segmentation using expectation-maximization and its application to image querying. *Third International Conference on Visual Information Systems.*

Chalmond, B. et Stéphane, C. (1999). Nonlinear modeling of scattered multivariate data and its application to shape change. *IEEE Trans. PAMI*, 21(5) : 422–432.

Chang, E. (2003). Csba : Content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Transactions on Circuits Systems Video Technol.*, 13(1) : 26–38.

Chang, F. (August 2001). Retrieving information from document images : Problems and solutions. *Int. Journal on Document Analysis and Recognition IJDAR*, 4(1) : 46–55.

Cheriet, M. (1999). Extraction of handwritten data from noisy gray-level images using a multi-scale approach. *IJPRAI*, 13(5) : 665–685.

Cutting, D., Karger, J., Pedersen, J., et Tukey (1992). Scatter/gather : a cluster-based approach to browsing large document collections. *Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329.

Doermann, D. (1998). The indexing and retrieval of document images : A survey. Computer Vision and Image Understandind, pages 287–298.

Fataicha, Y., Cheriet, M., Nie, J., et Suen, C. (2002). Content analysis in document images : a scale space approach. Dans *16th IEEE International Conference on Pattern Recognition*, volume 3, pages 335–338, Quebec, Canada.

Fataicha, Y., Cheriet, M., Nie, J., et Suen, C. (2003). Information retrieval based on OCR errors in scanned documents. *IEEE Conference on Computer Vision and Pattern Recognition Workshops CVPR'03*, volume 3, Madison, USA.

Fataicha, Y., Cheriet, M., Nie, J., et Suen, C. (2005). Retrieving poorly degraded ocr documents. *International Journal on Document Analysis and Recognition (IJDAR).*

Fataicha, Y., Nie, J., Cheriet, M., et Suen, C. (2001). Détection multi-échelle d'objets dans des images de documents composites. Dans *International Conference on Image and Signal Processing "ICISP"*, volume 1, pages 538–546, Agadir, Morocco.

Halkidi, M., Batistakis, Y., et Vazigiannis, M. (2001). On clustering validation techniques. Journal of Intelligent Information Syst., 17(2-3) : 107–145.

Harding, S., Croft, W., et Weir, C. (1997). Probabilistic retrieval of OCR degraded text using n-grams. *Research and Advanced Technology for Digital Libraries.*, number 1324, pages 345–359, First European Conference ECDL, Pisa, Italy.

Hartigan, J. et Wong, M. (1979). A k-means clustering algorithm. 28 :100–108.
Hull, J. (1996). Performance evaluation for document analysis. *IJIST*, 7(4) : 357–362.

Hull, J. (1996). Performance evaluation for document analysis. *IJIST*, 7(4) : 357–362.

Ingold, R. et Armangil, D. (1991). A top-down document analysis method for logical structure recognition. Dans *Proceedings of the Fourth Int. Conf. On Document Analysis and Recognition*, pages 41–49, St. Malo, France.

Ingold, R., Hitz, O., et Robadey, L. (2000). Segmentation de documents à structure complexe. *Proceedings of the Colloque Internat. Fracophone sur l'Écrit et le Document. CIFED'2000*, Lyon, France.

Jain, A., Murty, M., et Flynn, P. (1999). Data clustering : a review. ACM Comput. Survey, 31(3) : 264–323.

Jain, A. et Yu, B. (1998). Document representation and its application to page decomposition. *IEEE Transactions on Pattern Analysis And Machine Intelligence*, pages 294–308.

Jain, A. et Zhong, Y. (1996). Page segmentation using texture analysis. Pattern Recognition, 29(5) : 743–770.

Jing, F., Minging, L., Zhang, H., et Zhang, B. (2005). A unified framework for image retrieval using keyword and visual features. *IEEE Transactions on Image Processing*, 14(7) :979–989.

Kanungo, T., Baird, H., et Haralick, R. (2002). Special issue on performance evaluation: Theory, practice, and impact. *Int. Journal on Document Analysis and Recognition IJDAR*, 4(3) :139–139.

Kerpedjiev, S. (1997). Automatic extraction of information structures for document. Dans *Proceedings of the Fourth Int. Conf. On Document Analysis and Recognition*, pages 32–40, ULM- Germany.

Kida, H., Iwaki, O., et Kawada, K. (1986). Document recognition system for office automation. *Proceedings of the Eighth International Conference on Pattern Recognition, Paris, France*, pages 446–448.

Koenderink, J. (1984). The structure of images. Bio. Cybernetics, 53 : 363–370.

Krishnamoorthy, M., Nagy, G., Seth, S., et Viswanathan, M. (1993). Syntactic segmentation and labelling of digitized pages from technical journals. *IEEE Computer Vision, Graphics and image processing*, 47 : 327–352.

Kyong-Ho, L., Yoon-Chul, C., et Sung-Bae, C. (2000). Geometric structure analysis of document images : A knowledge-based approach. *IEEE Trans. PAMI*, 22(11) : 1224–1240.

Lindeberg, T. (1994). Scale-space theory in computer vision. *Kluwer Academic Publishers, Boston.*

Ma,W. et Manjunath, B. (1997). Netra : A toolbox for navigating large image databases. *Proc. IEEE Int'l Conf. Image Processing.*

Makinen, V., Navarro, G., et Ukkonen, E. (2003). Algorithms for transposition invariant stringmatching. Dans *STACS 2003 Proceedings : 20th Annual Symposium on Theoretical Aspects of Computer Science, Berlin, Germany*, pages 191–202, Lecture Notes in Computer Science, Spring-Verlag, Heidelberg.

Makinen, V., Baeza-Yates, R., et Riberro-Neto, B. (1999). Modern information retrieval. *Addison-Wesley, Paperback*, 513 pages.

Mokhtarian, F., Abbasi, S., et Kittler, J. (1996a). Efficient and robust retrieval by shape through curvature scale space. *Proceedings of the first International Workshop on Image Databases and Multimedia Search.*

Mokhtarian, F., Abbasi, S., et Kittler, J. (1996b). Robust and efficient shape indexing through curvature scale space. Dans *British Machine Vision Conference.*

Muller, S. et Rigoll, G. (1999). Improved stochastic modeling of shapes for contentbased image retrieval. *IEEE Workshop on Content-based Access of Image and Video Libraries(CBAIVL'99)*, pages 23–27, Fort Collins, CO.

Nagy, G. et Seth, S. (1984). Hierarchical representation of optical scanned documents. *Proceedings of the 7th IEEE Joint Conference on Pattern Recognition, Montréal, Canada*, pages 347–349.

Nastar, C., Boujemaa, N., Mitschke, M., et Meilhac, C. (1998a). Surfimage : un système flexible d'indexation et de recherche d'images. Dans *Journées CNET, CORESA*, Lannion, France.

Nastar, C., Mitschke, M., Meilhac, C., et Boujemaa, N. (1998b). Surfimage : A flexible

content-based image retrieval system. Dans *Proceedings of the ACM International Multimedia Conference, Bristol, England*, pages 339–344.

Niblack, W. (1993). The qbic project : Querying images by content color texture and shape. Dans *Proc. SPIE Conf. Storage and Retrieval for Image and Video Databases*, pages 173–197.

Ohta, O., Takasu, A., et Adachi, J. (1998). Probabilistic retrieval methods for text missrecognized ocr characters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11) : 1224–1240.

Pentland, A., Picard, R., et Sclaroff, S. (1994). Photobook : Content-based manipulation of image databases, storage and retrieval for image and video databases. *Proc. SPIE.*

Phillips, I. (Aug. 1993). User's reference manual for the uw english/technical document image database. *UW-I English-Technical Document Image Database, University of Washington.*

Remaki, L. et Cheriet, M. (2000). Kcs-new kernel family with compact support in scale space : Formulation and impact. *IEEE Transactions on Images Processing*, 9(6) : 970–981.

Rijsbergen, C., Harper, D., et Porter, M. (1981). The selection of good search terms. *Information Processing and Management*, 17 : 77–91.

Risjsbergen, C. (1979). Information retrieval. *Editor Butterworths*, London.

Roccio J.J. (1971). Relevance feedback in information retrieval. Englewood Cliffo, NJ : Prentice Hall (ISBN 0-13-814525-3) : 313–323.

Salton, G. (1971). The smart retrieval system- experiments in automatic document processing. *Prentice Hall Inc., Englewood Cliffs NJ.*

Salton, G. et McGill, M. (1983). Introduction to modern information retrieval. *McGraw-Hill New York*, volume 1.

Saporta, G. (1990). Probabilités, analyse des données et statistique. *Édition Technip, Paris, France*, page 241.

Sclaroff, S. et Pentlab, A. (1995). Modal matching for correspondence and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(6).

Seong-Whan, L. et R., D.-S. (2001). Parameter-free geometric document layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1240–1256.

Smeaton, A. (1998). Retrieval images of scanned text documents. *Proceedings of the Optical Engineering Society of Ireland and Irish Machine Vision and Image Processing Joint Conference, Vernon Ed.*, pages 271–286, U.K.

Smith, J. et Chang, S. (1996). Visualseek : A fully automated content-based image query system. *Proc. ACM Int'l Conf. Multimedia*, pages 87–98.

Soffer, A. et Samet, H. (1997). Negative shape features for image databases consisting of geographic symbols. Dans *Third International Workshop on Visual Form, Capri, Italy.*

Soffer, A. et Samet, H. (1998). Using negative shape features for logo similarity matching. Dans *proceedings of the 14th International Conference on Pattern Recognition*, volume 1, Brisbane, Australia.

Souza, A., Cheriet, M., Naoi, S., et Suen, C. (2003). Automatic filter selection using image quality assessment. Dans *Proceedings of the 7th International Conference on Document Analysis and Recognition ICDAR'03*, pages 508–512, Edinburgh, Scotland.

Spink, A. et Saracevic, T. (1997). Interactive information retrieval : Sources and effectiveness of search terms during mediated online searching. *Journal of the American Society for Information Science*, 48(8) : 741–761.

Srihari, R. (1995). Automatic indexing and content-based retrieval of captioned photographs. IEEE Comput., pages 49–56.

Strohmaier, C., C. Ringlstetter, C., Schulz, K., etMihov, S. (2003). Lexical postcorrection of OCR-results : The web as a dynamic secondary. Dans *Proceedings of the 7th International Conference on Document Analysis and Recognition ICDAR'03*, pages 1133–1137, Edinburgh, Scotland.

Suen, C. (1979). N-gram statistics for natural language understanding and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1 : 164–171.

Taghva, K., Borsack, J., et Condit, A. (1996a). Effects of ocr errors on ranking and feedback using the vector space model. *Information Processing and Management*, 32(3) : 317–327.

Taghva, K., Borsack, J., et Condit, A. (1996b). Evaluation of model-based retrieval effectiveness with ocr text. *ACM Transactions on Information Systems*, 14(1) : 64–93.

Taghva, K., Borsack, J., et Condit, S. (2002). Hairetes : A search engine for OCR documents. Dans *Proceedings of the 5th. IAPR International Workshop on Document Analysis Systems*, pages 412–422, Princeton, NY, USA, August 2002.

Taghva, K. et Stofsky, E. (2001). Ocrspell : an interactive spelling correction system for ocr errors in text. *Int. Journal on Document Analysis and Recognition IJDAR*, 3(3) : 125–137.

Tang, Y., Yan, C., Cheriet, M., et Suen, C. (1997). Automatic analysis and understanding of documents. *Handbook of Pattern Recognition and Computer Vision*, pages 625–654.

Ukkonen, E. (1983). On approximate string matching. Dans *Proceedings International Conference on Foundations of Computer Theory, Spring-Verlag, LNCS*, number 158, pages 487–495.

Venters, C. et Cooper, M. (1999). A review of content-based image retrieval systems. *Manchester Visualization Center, University of Manchester.*

Wang, D. et Srihari, S. (1989a). Classification of newspaper image blocks using texture analysis. Dans *Computer Vision, Graphics and image processing*, volume 47, pages 327–352.

Wang, D. et Srihari, S. N. (1989b). Classification of newspaper image blocks using texture analysis. *Computer Vision, Graphics, and Image Processing*, 47 : 327–352.

Winter, A. et Nastar, C. (1999). Differential feature distribution maps for image segmentation and region queries in image databases. *CBAIVL Workshop at CVPR, Fort Collins, Colorado.*

Zen, H. et Osawa, S. (1985). Extraction of thr fair document from mixed mode manuscript. Dans *Proceeding of the CVPR, San Francisco, USA*, pages 544–549.