HAL
open science

Représentations visuelles de concepts textuels pour la recherche et l'annotation interactives d'images

Nhu Van Nguyen

To cite this version:

Nhu Van Nguyen. Représentations visuelles de concepts textuels pour la recherche et l'annotation interactives d'images. Ordinateur et société [cs.CY]. Université de La Rochelle, 2011. Français. NNT : 2011LAROS338 . tel-00730707

HAL Id: tel-00730707

https://theses.hal.science/tel-00730707v1

Submitted on 10 Sep 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE DE DOCTORAT

présentée devant

L'Université de La Rochelle
École Doctorale S2I

en vue de l'obtention du grade et du titre de

Docteur de l'Université de La Rochelle

soutenue publiquement le 9 septembre 2011 par

Nhu-Van Nguyen

Représentations visuelles de concepts textuels pour la recherche et l'annotation interactives d'images

Travail effectué conjointement au MSI-IFI (IRD, UMI 209 UMMISCO), Hanoi, Viêtnam et au L3i, Université de La Rochelle et au LORIA, Université de Nancy 2

sous la co-direction de MM. Jean-Marc Ogier, Alain Boucher et Salvatore Tabbone

Composition du jury

Florence Sèdes	Professeur des Universités, Université Paul Sabatier	*examinateur*
Sylvie Philipp-Foliguet	Professeur des Universités, Université de Cergy/-Pontoise	*rapporteur*
Philippe Mulhem	Chargé de Recherche, Centre National de la Recherche Scientifique	*rapporteur*
Jean-Marc Ogier	Professeur des Universités, Université de La Rochelle	*directeur de thèse*
Alain Boucher	Professeur AUF, IFI-AUF, Hanoi, Viêtnam	*co-encadrant scientifique*
Salvatore Tabbone	Professeur des Universités, Université de Nancy 2	*co-encadrant scientifique*

Ce travail de thèse a été effectué au sein des laboratoires suivants :

- MSI-IFI (IRD, UMI 209 UMMISCO), Hanoi, Viêtnam
- L3i (EA 2118), Université de La Rochelle, France
- LORIA, Université de Nancy 2, France

Dedicace

Résumé

En recherche d'images aujourd'hui, nous manipulons souvent de grands volumes d'images, qui peuvent varier ou même arriver en continu. Dans une base d'images, on se retrouve ainsi avec certaines images anciennes et d'autres nouvelles, les premières déjà indexées et possiblement annotées et les secondes en attente d'indexation ou d'annotation. Comme la base n'est pas annotée uniformément, cela rend l'accès difficile par le biais de requêtes textuelles. Nous présentons dans ce travail différentes techniques pour interagir, naviguer et rechercher dans ce type de bases d'images. Premièrement, un modèle d'interaction à court terme est utilisé pour améliorer la précision du système. Deuxièmement, en se basant sur un modèle d'interaction à long terme, nous proposons d'associer mots textuels et caractéristiques visuelles pour la recherche d'images par le texte, par le contenu visuel, ou mixte texte/visuel. Ce modèle de recherche d'images permet de raffiner itérativement l'annotation et la connaissance des images.

Nous identifions quatre contributions dans ce travail. La première contribution est un système de recherche multimodale d'images qui intègre différentes sources de données, comme le contenu de l'image et le texte. Ce système permet l'interrogation par l'image, l'interrogation par mot-clé ou encore l'utilisation de requêtes hybrides. La deuxième contribution est une nouvelle technique pour le retour de pertinence combinant deux techniques classiques utilisées largement dans la recherche d'information : le mouvement du point de requête et l'extension de requêtes. En profitant des images non pertinentes et des avantages de ces deux techniques classiques, notre méthode donne de très bons résultats pour une recherche interactive d'images efficace. La troisième contribution est un modèle nommé "Sacs de KVR" (*Keyword Visual Representation*) créant des liens entre des concepts sémantiques et des représentations visuelles, en appui sur le modèle de Sac de Mots [Sivic 2008]. Grâce à une stratégie d'apprentissage incrémental, ce modèle fournit l'association entre concepts sémantiques et caractéristiques visuelles, ce qui contribue à améliorer la précision de l'annotation sur l'image et la performance de recherche. La quatrième contribution est un mécanisme de construction incrémentale des connaissances à partir de zéro. Nous ne séparons pas les phases d'annotation et de recherche, et l'utilisateur peut ainsi faire des requêtes dès la mise en route du système, tout en laissant le système apprendre au fur et à mesure de son utilisation.

Les contributions ci-dessus sont complétées par une interface permettant la vi-

sualisation et l'interrogation mixte textuelle/visuelle. Même si pour l'instant deux types d'informations seulement sont utilisées, soit le texte et le contenu visuel, la généricité du modèle proposé permet son extension vers d'autres types d'informations externes à l'image, comme la localisation (GPS) et le temps.

Mots-clés : recherche d'images multimodale, annotation interactive d'images, retour de pertinence, représentation de concepts, apprentissage par renforcement

Abstract

As regard image retrieval today, we often manipulate large volumes of images, which may vary or even update continuously. In an image database, we end up with both old and new images, the first possibly already indexed and annotated and the latter waiting for indexing or annotation. Since the database is not annotated consistently, it is difficult to use text queries. We present in this work different techniques to interact, navigate and search in this type of image databases. First, a model for short term interaction is used to improve the accuracy of the system. Second, based on a model of long term interaction, we propose to combine semantic concepts and visual features to search for images by text, visual content or a mix between text and visual content. This model of image retrieval can iteratively refine the annotation of images.

We identify four contributions in this work. The first contribution is a system for multimodal retrieval of images which includes different kinds of data, like visual content and text. This system can be queried by images, by keywords or by hybrid text/visual queries. The second contribution is a novel technique of relevance feedback combining 2 classic techniques : query point movement and query expansion. This technique profits for non-pertinent feedback and combines the advantages of both classic techniques and improve performance for interactive image retrieval. The third contribution is a model based on visual representations of keywords (KVR : Keyword Visual Representation) that create links between text and visual content, based on long term interaction. With the strategy of incremental learning, this model provides an association between semantic concepts and visual features that help improve the accuracy of image annotation and image retrieval. Moreover, the visual representation of textual concept gives users the ability to query the system by text queries or mixed queries text / images, even if the image database is only partially annotated. The fourth contribution, under the assumption that knowledge is not available early in most image retrieval systems, is a mechanism for incremental construction of knowledge from scratch. We do not separate phases of retrieval and annotation, and the user can make queries from the start of the system, while allowing the system to learn incrementally when it is used.

The contributions above are completed by an interface for viewing and querying mixing textual and visual content. Although at present only two types of information are used, the text and visual content, the genericity of the proposed

model allows its extension to other types of external information, such as location (GPS) and time.

Keywords : multimodal image retrieval, interactive image annotation, relevance feedack, concept representation, reinforcement learning

Remerciements

Tout d'abord, j'exprime ma gratitude aux membres de mon jury de thèse : le professeur Florence Sedes, qui a présidé mon jury de thèse, ainsi que le professeur Sylvie Philipp-Foliguet et Philippe Mulhem, rapporteurs, pour la relecture qu'ils ont faite de mon manuscrit.

Je tiens à exprimer ma profonde gratitude à mon directeur de thèse, le professeur Jean-Marc Ogier pour sa supervision enthousiaste lors de mon travail. J'attribue la réussite de mon doctorat à son orientation intellectuelle, ses encouragements continus et d'effort et sans lui cette thèse, n'aurait pas été achevée. Je ne pouvais pas souhaiter un meilleur superviseur ou plus amical.

J'ai également une grande dette de gratitude à mon superviseur, le professeur Alain Boucher pour son soutien important et sa convivialité tout au long de ce travail. Son expertise, la compréhension et la patience dont il a fait preuve ont considérablement enrichi mon expérience de la recherche. Si il n'avait pas été présent par son soutien et sa motivation, ma thèse n'aurait pas été possible.

Je tiens à remercier le professeur Salvatore Tabbone, mon superviseur, pour m'avoir fait profiter de ses connaissances par ses commentaires détaillés et constructifs, pour m'avoir donné l'opportunité de travailler à l'équipe QGAR de LORIA, et pour m'avoir constamment soutenu pendant mon temps de travail à Nancy.

Mes remerciements vont à Thomas, un collègue et un ami cher Thomas et sa famille m'ont beaucoup aidé depuis le premier jour j'ai été à La Rochelle. J'aimerais adresser mes remerciements à mes collègues et amis au L3i et à MSI pour des discussions stimulantes et pour tout le plaisir que nous avons eu au cours des quatre dernières années.

Je suis tellement reconnaissante à mes parents, ma femme et ma fille bien-aimée pour me soutenir et m'encourager m'avoir soutenu et encouragé à poursuivre ce travail. Cette thèse est particulièrement dédiée à mon père, un chercheur très compétent et passionné de probabilités et statistiques.

Enfin, je tiens à souligner l'appui financier, technique et académique de l'Institut Francophonie pour l'Informatique, de l'Université de La Rochelle et de leurs personnels, en particulier dans l'attribution d'une bourse d'études IFI-AUF, qui a fourni le soutien financier nécessaire à cette recherche.

Table des matières

1 Introduction **1**
1.1 Contexte 1
1.2 Problèmes et Objectifs 3
1.3 Contributions 5
1.4 Structure de la dissertation 6

2 État de l'art **7**
2.1 Introduction 7
2.2 Indexation 8
2.2.1 Information internes 9
2.2.1.1 Signatures globales d'images 9
2.2.1.2 Signatures locales d'images 11
2.2.2 Informations externes 12
2.2.2.1 Signatures textuelles 13
2.2.2.2 Métadonnées de l'image 13
2.3 Recherche d'images 14
2.3.1 La formation de requêtes 14
2.3.2 Modèle de recherche par le texte 15
2.3.2.1 Modèle vectoriel 15
2.3.2.2 Modèle probabiliste 16
2.3.3 Modèle de recherche par le contenu 17
2.3.3.1 Modèle vectoriel 18
2.3.3.2 Modèles probabilistes 19
2.3.3.3 Autres modèles 19
2.3.4 Avantages et inconvénients des modalités par le texte et par le contenu 20
2.3.4.1 Modèle par le texte 20
2.3.4.2 Modèle par le contenu 21
2.4 Recherche multimodale 22
2.4.1 Fusion précoce 22
2.4.1.1 Intégration dans un modèle probabiliste 23
2.4.1.2 Indexation Sémantique Latente (*LSI - Latent Semantic Indexing*) 23

2.4.1.3 Analyse des corrélations canoniques (*CCA - Canonical Correlation Analysis*) 24
2.4.2 Fusion tardive 25
2.4.2.1 Raffinement des classements 25
2.4.2.2 Combinaison des classements 25
2.4.3 Approche basée sur les transformations 25
2.5 Annotation d'images 26
2.6 Systèmes intégrés 27
2.6.1 SCENIQUE 27
2.6.2 MAMI 28
2.6.3 ALIPR 28
2.6.4 SnapToTell 28
2.7 Systèmes commerciaux 29
2.8 Discussion 31
2.9 Système de recherche d'images proposé 32
2.10 Conclusion 36

3 Interaction **37**
3.1 Introduction 38
3.2 État de l'art 39
3.2.1 La spécification des requêtes 39
3.2.2 Exploration 40
3.2.2.1 Structures statiques hiérarchiques 42
3.2.2.2 Réseaux statiques 42
3.2.3 Retour de pertinence à court terme 42
3.2.3.1 Modification de la requête 43
3.2.3.2 Optimisation de distance métrique 46
3.2.3.3 Apprentissage de classifieurs 48
3.2.3.4 Retour de pertinence pour la recherche multimodale 48
3.2.4 Conclusion 50
3.3 Retour de pertinence à court terme dans notre système 52
3.3.1 Retour de pertinence basé sur les groupes pour la recherche par le contenu 52
3.3.2 Sélection de la méthode de regroupement 59
3.3.2.1 K-moyennes adaptatif 59
3.3.2.2 Agglomération compétitive 60
3.3.2.3 Sélection de la méthode de regroupement 61
3.4 Visualisation 61
3.4.1 État de l'art 61
3.4.2 Discussion sur la visualisation 64
3.4.3 Notre visualisation pour la recherche mixte texte / images . 66

3.4.3.1 Les exigences de la visualisation pour la recherche mixte texte/images 66
3.4.3.2 L'interface de visualisation 68
3.5 Interaction pour la recherche mixte texte/image 69
3.5.1 Notre technique de retour de pertinence pour la recherche mixte texte/images 69
3.6 Evaluation 71
3.6.1 Protocole d'expérimentation 71
3.6.1.1 Base d'experimentation 72
3.6.1.2 Discussion des protocoles utilisés dans d'autres systèmes : cas de MARS et QCluster 72
3.6.1.3 Protocole pour notre expérimentation 73
3.6.2 Résultats et discussion 74
3.7 Conclusion 77

4 Modèle Sacs de KVR pour l'apprentissage de connaissances 79
4.1 État de l'art de l'association du texte et du contenu 80
4.1.1 Association du texte et du contenu 80
4.1.1.1 Modèle de cooccurrences 80
4.1.1.2 Modèles de traduction 82
4.1.1.3 Modèle de pertinence cross-média 82
4.1.1.4 Modèle LSA (Latent Semantic Analysis) 84
4.1.1.5 Modèle de transformation 84
4.1.2 Verrous et difficultés de l'existant de la littérature 86
4.1.3 Positionnement et argumentation pour notre modèle 88
4.2 La représentation Sac de KVR - (Représentation visuelle de mots textuels) 89
4.2.1 Le modèle Sacs de Mots (BoW) 89
4.2.2 La construction du modèle Sac de mots 92
4.2.3 La représentation Sac de KVR 93
4.3 Apprentissage des Sacs de KVR 96
4.3.1 Apprentissage incrémental des Sacs de KVR 98
4.3.1.1 Sélection de la méthode de regroupement 101
4.3.1.2 Opérateurs de KVR 102
4.3.1.3 Algorithmes proposés 104
4.4 Utilisation de Sacs de KVR 107
4.4.1 Propagation d'annotations 107
4.4.2 La recherche d'images par la requête textuelle 109
4.4.3 Comparaison avec les autres modèles 110
4.5 Conclusion 111

5 Expérimentation **113**
5.1 Les difficultés de l'évaluation du système 114
5.1.1 Orientations retenues pour l'expérimentation et l'évaluation 115
5.2 Présentation de la base de données pour l'expérimentation 115
5.3 Protocole d'expérimentation 118
5.3.1 Validations croisées pour l'expérimentation 118
5.3.2 Scénarios 120
5.3.3 Méthode d'évaluation pseudo interactive 121
5.3.3.1 Simulation de connaissances des utilisateurs/experts 122
5.3.3.2 Simulation de la spécification de la requête 122
5.3.3.3 Simulation du retour de pertinence 123
5.4 Résultats 124
5.4.1 L'évolution des connaissances, l'évaluation du caractère "dynamique" 124
5.4.2 La quantité de connaissances 126
5.4.2.1 Le nombre de concepts appris 126
5.4.2.2 Le nombre d'annotations 128
5.4.2.3 Conclusion sur la quantité de connaissances 129
5.4.3 La qualité de connaissances d'annotations 130
5.4.3.1 La partie initiale 130
5.4.3.2 La partie nouvelle 132
5.4.3.3 La propagation d'annotations dans la base Corel 30K 134
5.4.4 La qualité de connaissances de niveau "image" 136
5.4.5 Evaluation sur le retour de pertinence utilisé dans l'apprentissage de connaissances 138
5.4.6 Evaluation de la méthode de regroupement 139
5.4.7 Utilisation de Sacs de KVR : Evaluation du caractère "statique" 140
5.4.7.1 Annotation d'images 141
5.4.7.2 La recherche d'images 142
5.5 Conclusion 145

6 Conclusion et Perspectives **149**
6.1 Conclusion 149
6.1.1 Résumé des contributions 149
6.1.2 Problèmes restants 150
6.2 Perspectives 151
6.2.1 A court terme 151
6.2.2 A long terme 153
6.3 Conclusion finale 158

Travaux de l'auteur **159**

Bibliographie **161**

Table des figures

1.1 Vue d'ensemble du projet IDEA. 3

2.1 Exemple d'architecture de système de recherche d'images. 7
2.2 Les multiples couches de PLSA pour combiner les caractéristiques visuelles et textuelles en un vecteur de caractéristiques unique. . . . 24
2.3 Architecture client/serveur du système SnapToTell. 29
2.4 Le moteur de recherche d'images Bing Images. 30
2.5 Le moteur de recherche d'images Google Images. 30
2.6 Le moteur de recherche d'images eBay Images. 31
2.7 Vue d'ensemble du système de recherche d'images proposé. Au début de la vie du système (étape 1), la base de données est sans connaissances et ne contient que des images (couche A de la base). Des images sont envoyées au système en temps réel (flèche rouge à droite entrant dans la base). Ensuite viennent le retour de pertinence (étapes 2 et 3), l'apprentissage de KVR (étape 4) et la propagation d'annotations (étape 5). Les connaissances du système sont créées et évoluent dans le temps (annotation manuelle, ajout de KVR, annotation propagée). Les flèches rouges entrantes vers la base signifie que des connaissances/images sont (r)envoyées au système, la flèche rouge sortante de la base signifie que des connaissances/images sont utilisées pour la recherche d'images ou pour l'apprentissage de connaissances. 34
2.8 Vue d'ensemble du système de recherche d'images proposé lorsque des connaissances sont disponibles. Les flèches rouges entrantes dans la base signifient que des connaissances/images sont renvoyées au système et les flèches rouges sortantes de la base signifient que des connaissances/images sont utilisées pour la recherche d'images ou pour l'apprentissage de connaissances. 35

3.1 Système de recherche d'images avec le retour de pertinence à court terme. 37
3.2 Définition de l'unimodalité d'un groupe d'images. 44
3.3 Le mouvement de requête. 44

3.4 Extension de requêtes, (a) une requête à un seul point est remplacée par (b) une requête à points multiples. 46

3.5 Mouvement du point de requête et extension de la requête. Dans le mouvement du point de requête, le point idéal de requête comprend des exemples pertinents en raison de l'unimodalité des exemples pertinents. Dans l'extension de requêtes, les points idéaux de la requête convergent lentement lorsque les exemples non pertinents ne sont pas utilisés et ils peuvent entraîner le résultat dans un piège de maximum local. 51

3.6 Système de recherche d'images : le retour de pertinence à court terme. 52

3.7 Combinaison du mouvement de requête et de l'extension de requêtes. Les points idéaux de requête sont atteint plus efficacement et plus rapidement. Les exemples non pertinents sont éliminés des groupes locaux. 53

3.8 Retour de pertinence basé sur les groupes pour la recherche par le contenu (CBIR). Les signes "+" et "-" sont les exemples pertinents/non-pertinents. La couleur rouge montre les exemples non annotés pendant l'interaction, les autres couleurs montrent les exemples annotés. 55

3.9 Visualisation composée : Google Swirl. 64

3.10 Visualisation multimodale d'un ensemble d'images [Camargo 2010]. 65

3.11 La représentation en coordonnées polaires pour des requêtes visuelles et textuelles. Le vecteur V est la distance visuelle entre la requête et l'image. L'angle α est la distance entre le mot textuel et l'image. Les mots textuels sont les annotations manuelles fournies par les experts. Les images en bleu/rouge sont pertinentes/non pertinentes. Dans l'exemple ci-dessus, les images proches de l'axe des X partagent avec la requête le contenu visuel et les mots textuels. Plus on s'éloigne de l'axe des X, plus les images divergent par leurs mots textuels de la requête, la similarité du contenu visuel dépendant quant à elle de la distance, sur l'interface, par rapport à la requête. 67

3.12 La visualisation avec l'interaction de l'utilisateur pour le retour de pertinence pour la recherche mixte texte/image. A gauche, la requête mixte se compose d'un exemple d'images et de 3 concepts ("sunset", "horizon" et "tree"). Les images résultantes sont visualisées dans les 3 quadrants qui correspondent aux 3 concepts. Pendant l'interaction, l'utilisateur fournit des connaissances/informations qui sont des exemples pertinents (rectangles bleus) et des exemples non pertinents (en rouge). Le système reformule la requête et fait une nouvelle recherche pour tenter d'obtenir de meilleurs résultats. . . . 70

3.13 QE signifie la technique de l'extension de requêtes (*Query expansion*), QPM signifie la technique du mouvement de requête (*Query point movement*), CR et CNR sont les méthodes Regroupement-répétition (*Clustering-Repeat*) et Regroupement-non-répétition (*Clustering-Non-Repeat*) que nous avons décrites précédemment. Cette figure illustre les précisions moyennes pour les 100 premières images récupérées de 4 techniques de retour de pertinence avec 10 exemples retournés pour chaque itération. Les deux techniques CNR et CR montrent une très bonne performance comparées aux méthodes de modification de la requête. 75

3.14 QE signifie la technique de l'extension de requêtes (*Query expansion*), QPM signifie la technique du mouvement de requête (*Query point movement*), CR et CNR sont les méthodes Regroupement-répétition (*Clustering-Repeat*) et Regroupement-non-répétition (*Clustering-Non-Repeat*) que nous avons décrites précédemment. Cette figure illustre les précisions moyennes pour les 100 premières images extraites de 4 techniques de pertinence avec 20 exemples de retour de pertinence pour 1 itération. La méthode CNR donne le meilleur résultat. 76

3.15 Les précisions moyennes pour les 50 premières images récupérées des 4 techniques de retour de pertinence avec 10 exemples de retour pour chaque itération. 77

3.16 Les précisions moyennes pour les 50 premières images récupérées des 4 techniques de retour de pertinence avec 20 exemples de retour pour chaque itération. 77

4.1 L'architecture de notre système : l'apprentissage de connaissances (en bleu). 79

4.2 Le modèle de cooccurences de Mori et al. [Mori 1999]. 81

4.3 Extension du modèle de traduction : modèle hiérarchique de Barnard et al. [Barnard 2003]. 83

4.4 La recherche d'images multilingues [Lin 2007]. 85

4.5 Traduire une requête image en requête textuelle [Chang 2008]. . . . 86

4.6 Notre système de recherche d'images avec en bleu l'utilisation de notre modèle Sacs de KVR pour l'association texte/image pour la recherche d'images multimodales et pour l'annotation d'images. . . 89

4.7 Le modèle BoW dans le traitement du langage naturel. Chaque document est représenté par un sac de mots. 90

4.8 Les images sont traitées comme des documents textuels, et les caractéristiques extraites des images sont considérées comme des "mots" [Fei-Fei 2007] 90

4.9 Comme un dictionnaire d'un langage ou un verbe peut se présenter par des mots différents lorsqu'il est conjugué, les caractéristiques de régions semblables d'une image sont représentées par un même mot visuel. 91

4.10 3 étapes : (1) détection et description des mots (Extraction de caractéristiques), (2) formation d'un dictionnaire de mots visuels, (3) représentation des images [Fei-Fei 2007]. 92

4.11 Exemple d'une représentation BoW du mot textuel "tortue". 94

4.12 La représentation Sac de KVR. "Ciel" pourrait être l'un des trois types : "Clair", "Nuageux" et "Coucher de soleil". Le mot textuel est alors représenté par un sac contenant les trois KVR correspondants. 94

4.13 Le modèle Sac de KVR pour l'association texte/image (notre contribution en rouge) basé sur la représentation existante Sac de Mots (en bleu). 95

4.14 Les régions pertinentes sont sélectionnées au cours du processus de recherche. Des mots visuels extraits des régions sont utilisés pour représenter le mot-clé. 97

4.15 Un scénario pour l'apprentissage des Sacs de KVR des mots textuels. 98

4.16 Un scénario pour l'apprentissage des Sacs de KVR des mots textuels. 99

4.17 Scénario pour l'apprentissage des Sacs de KVR de mots textuels : Fusion de deux KVR, c'est-à-dire la fusion de 2 groupes d'images correspondant à 2 KVR. 100

4.18 Scénario pour l'apprentissage des Sacs de KVR des mots textuels : Ajout d'un nouveau KVR. 100

4.19 Apprentissage incrémental des KVR fondé sur le retour de pertinence basé sur des groupes. 107

4.20 La propagation d'annotation d'images par le modèle Sacs de KVR (étape 5). 108

4.21 La recherche d'images par le contenu dans notre système. 109

5.1 Système de recherche d'images 113

5.2 Une vue générale de la base Corel 30K. 116

5.3 Quelques catégories de la base Corel 30K avec des mots clés. 116

5.4 Agents pour simuler les interactions. La section en rouge est la simulation des interactions des utilisateurs. La section en vert est le système que l'on veut évaluer. La section en noir représente le processus d"évaluation du système. La base d'images est dynamique avec des nouvelles images arrivant dans le temps. 121

5.5 Nombre de concepts appris en fonction du temps (itérations). Dans un premier temps, avec le point à $t = 330$, on voit qu'on apprend 200 concepts. Nous observons donc un taux rapide jusqu'à environ la moitié des concepts (200) identifié par la flèche noire en pointillé. Dans un second temps, nous observons un taux beaucoup plus lent identifié par la flèche bleue en pointillé, jusqu'à environ 400 concepts appris à la fin. 126
5.6 La distribution des concepts dans la base Corel 30K. Certains concepts sont associés avec beaucoup d'images (concepts grands - à gauche), tandis que d'autres sont associés avec seulement quelques images (concepts petits - à droite). 127
5.7 Nombre d'annotations propagées dans la partie initiale (en haut à droite), dans la partie nouvelle (en haut à gauche), dans la base totale (en bas à droite) et le nombre d'annotations manuelles (en bas à gauche). 129
5.8 L'évolution de la précision pour la propagation d'annotation sur la partie initiale. La précision d'annotation augmente rapidement dans un premier temps (flèche noire) où les Sacs de KVR des concepts grands sont appris (la connaissance augmente rapidement). Dans un second temps (flèche rouge), les Sacs de KVR des concepts grands sont améliorés (la connaissance est améliorée) et les autres concepts petits sont appris (la connaissance augmente moins rapidement que dans le premier temps). Alors la précision d'annotation augmente plus lentement. 131
5.9 L'évolution du rappel de la propagation d'annotations sur la partie initiale. 132
5.10 L'évolution de la précision de la propagation d'annotations sur la partie nouvelle. 133
5.11 Evolution du rappel de la propagation d'annotations sur la partie nouvelle. 134
5.12 L'évolution de la précision de la propagation d'annotations (annotation automatique) sur la base Corel 30K. Dans un premier temps (jusqu'au temps $t = 280$) l'expérimentation 2 donne la meilleure performance, les autres expérimentations ayant presque la même performance. Dans un second temps, la quantité de connaissances commence à influencer la performance de la propagation. 135
5.13 Evolution du rappel de la propagation d'annotations sur la base Corel 30K. 135
5.14 Evolution de la précision de la recherche d'images par des Sacs de KVR sur la partie initiale. En général, la précision de recherche est meilleure dans le temps ou en d'autres termes, les Sacs de KVR s'améliorent dans le temps. 137

5.15 Evolution de la précision de la recherche d'images par des Sacs de KVR sur la partie nouvelle. 137
5.16 Evolution de la précision de l'annotation d'images pour les méthodes CR et CNR. 138
5.17 Evolution de la précision de la recherche d'images pour les méthodes CR et CNR. 139
5.18 L'évolution de la précision de l'annotation d'images selon la methode de regroupement : K-moyennes adaptif et Agglomération compétitive.140
5.19 L'évolution de la précision de la recherche d'images selon la methode de regroupement : K-moyennes adaptif et Agglomération compétitive.140
5.20 La recherche d'images par le contenu dans notre système. 143
5.21 Précision/rappel de la recherche d'images par l'exemple (en noir), de la recherche d'images par requête textuelle pour tous les concepts (en rouge) et pour les concepts avec plus de 10 images associées dans la base de test (en bleue). 144
5.22 Résultat de la recherche d'images par requête textuelle : "people+street".145
5.23 Résultat de la recherche d'images par le contenu (pour trouver des gens dans la rue). 146
5.24 Résultat de la recherche d'images par requête textuelle : "bear+snow".147
5.25 Résultat de la recherche d'images par le contenu (pour trouver des ours sur la neige). 148

6.1 Notre proposition basée sur la mise en page radiale utilisée dans Google Swirl. 152
6.2 Structuration des données par des SR-tree. 154
6.3 Groupes de situations d'urgence dans la ville. 155
6.4 Comparatif entre méthode traditionnelle / nouvelle méthode à base d'agents pour la recherche d'images par le contenu. 156
6.5 Nature et puissance des forces selon la similarité et la proximité des agents. 156
6.6 Visualisation basée sur des agents. 157

Liste des tableaux

4.1 Comparaison des modèles d'association du texte et du contenu. . . 111

5.1 Nombre d'images dans les bases d'images pour les 4 expérimentations.130
5.2 Tableau comparatif des méthodes d'annotation d'images. 142

Chapitre 1

Introduction

1.1 Contexte

De nos jours, l'avènement largement déployé du numérique contribue à une augmentation considérable d'informations stockées dans des bases de données, avec un objectif croissant de partage et d'accès rapide à des informations faiblement structurées (images, musique, ...). Les progrès des techniques de stockage de données et des technologies d'acquisition d'images ont permis la création de grandes bases de données d'images qui sont utilisées dans des domaines aussi divers que le commerce, le secteur médical, culturel, militaire, le monde du divertissement, la gestion des catastrophes naturelles... Tous ces secteurs d'activités partagent la nécessité d'élaborer des systèmes d'informations pour gérer efficacement ces bases d'images et convergent tous vers un objectif commun : la recherche d'images (Ces systèmes soulèvent des questions allant d'une fonction de similarité d'images à un moteur d'annotation ou de classification d'images).

Parmi les systèmes existant, on distingue deux grandes catégories de méthodes utilisées pour rechercher des images : d'une part les méthodes basées sur l'information textuelle et d'autre part celles basées sur l'information visuelle. La première catégorie de méthodes est basée sur des métadonnées textuelles de chaque image pour procéder à la recherche par mots-clés. La seconde est basée sur le contenu de chaque image pour rechercher des images similaires à celle choisie par l'utilisateur. La recherche par mots-clés fournit généralement de meilleurs résultats que celle par le contenu en termes de temps de réponse et de précision. Toutefois, ce type d'approche nécessite une annotation a priori de la base de données d'images, tâche très laborieuse, chronophage et souvent subjective. La recherche par le contenu est souvent nécessaire lorsque des annotations textuelles sont inexistantes ou incomplètes. En outre, la recherche par le contenu peut potentiellement améliorer

la précision même si des annotations textuelles pré-existent, notamment grâce à l'apport informationnel du contenu des images.

Dans le contexte général des applications pour la recherche d'images, nous nous intéressons aux systèmes de recherche d'images pour l'aide à la décision dans le cas de catastrophes naturelles. Nous verrons un peu plus tard dans la description du projet qui a conduit à cette thèse que ce type d'application n'utilise pas seulement le contenu des images et le texte éventuellement associé, mais s'appuie également sur des informations externes comme la localisation et le temps pour rechercher des images.

Le contexte de cette thèse est le projet IDEA (*Images of natural Disasters from robot Exploration in urban Area* - financé par le programme STIC-Asie du MAE/CNRS/INRIA) sur la gestion d'images provenant d'une zone urbaine après une catastrophe naturelle (tremblement de terre) visant l'aide à la coordination des secours. Ce projet se situe dans un cadre plus vaste nommé AROUND (*Autonomous Robots for Observation of Urban Networks after a Disaster*) visant l'ajout de technologies robotiques et logicielles pour aider la collecte d'informations et la gestion des secours en zone urbaine. Le but ultime d'AROUND est la conception et l'implémentation d'un système temps réel d'aide à la décision, intégrant une combinaison à plusieurs échelles de SIG, de systèmes d'aide à la décision et de capteurs robotisés sur le terrain. Dans ce cadre, de multiples caméras (fixes ou montés sur robots) seront utilisées afin de renseigner sur la situation en temps réel aux différents points de la ville (figure 1.1). Ce système a donc pour caractéristique principale une évolution constante de la base, notamment du fait de l'arrivée continue de nouvelles images captées en temps réel à différents endroits dans une ville.

Toutes ces sources fournissent une quantité considérable d'images et d'informations difficiles à gérer par des interventions humaines du fait des grands volumes considérés, en particulier dans une situation de crise et de stress. Des outils d'aide à la décision et de gestion d'images deviennent donc nécessaires afin d'aider les décideurs locaux à prendre les bonnes décisions en fonction des situations, sans être dépassés par une quantité ingérable d'informations. Des méthodes pour organiser et indexer de telles bases de données d'images, ainsi que des outils de recherche interactive efficace sont donc fondamentaux. Nous nous concentrons donc sur l'élaboration d'un système complexe doté d'une capacité d'organisation et d'indexation d'une base d'images par différents types d'information. Nous envisageons également que ce système soit en mesure de rechercher des images de façon interactive et d'apprendre interactivement de la connaissance.

Cependant, nous tenons à préciser que, dans le cadre de ce document, nous n'allons pas traiter les aspects applicatifs du projet IDEA, mais uniquement des

problématiques scientifiques que ce projet a engendrées. Entre autres, bien qu'un travail ait été fait pour constituer une base d'images propre à ce domaine d'application, celle-ci n'est pas encore assez bien structurée pour être exploitable pour valider nos travaux de recherche, en particulier par le manque de vérité terrain provenant d'experts et par le manque de complétude de certains aspects de la base. Nous nous replions donc sur des bases d'images d'images existantes, principalement la base d'images Corel. Ceci apporte aussi l'avantage de rester dans un cadre plus générique et donc de montrer l'utilité de notre travail pour d'autres applications. Dans les perspectives (chapitre 6), nous revenons sur une extension de notre travail qui a été faite et qui montre d'autres problématiques applicatives provenant du projet IDEA.

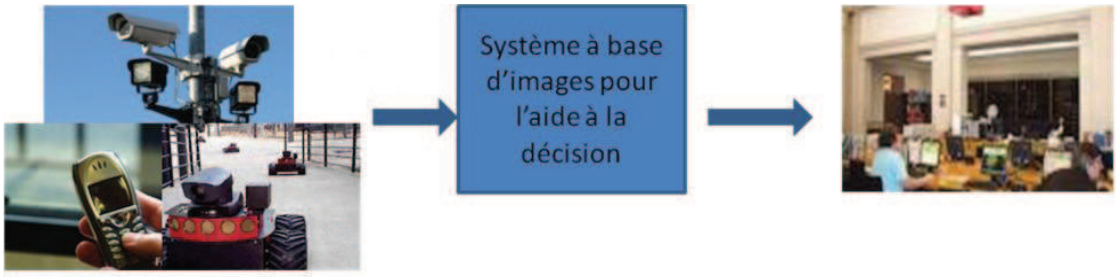

FIGURE 1.1 – Vue d'ensemble du projet IDEA.

1.2 Problèmes et Objectifs

Comme nous venons de l'évoquer, le système que nous souhaitons réaliser est un système complexe car la base de données d'images évolue en temps réel. Cette hypothèse est par ailleurs vraie pour de nombreuses applications de surveillance ou de contrôle. Dans la plupart de ces applications, le volume des images n'est pas fixe, et la base de données contient généralement des images anciennes et des nouvelles images, des images déjà traitées et indexées et des images en attente d'être transformées ou indexées. Les informations sur les images sont également en constante évolution, en fonction d'évènements extérieurs nouveaux pouvant intervenir à tout moment. Les images contiennent non seulement de l'information visuelle, mais aussi des informations géographiques (GPS par exemple), ou encore le cachet de l'heure d'arrivée au sein du système et des informations complémentaires liées à des d'experts. Nous résumons ci-dessous les hypothèses sur lesquelles nous nous appuyons pour l'élaboration de notre système, ainsi que son contexte d'utilisation :

– Travail sur une grande base d'images sans connaissance préalable.

- Le volume d'images n'est pas fixe, et le traitement des nouvelles images doit se faire en temps réel.
- La connaissance du système s'appuie sur les annotations des images (images représentées par le texte) mais également sur le contenu, et nous traduirons l'un vers l'autre dans ce travail (texte représenté par l'image).
- L'apprentissage est interactif et peut se faire par renforcement ou de manière incrémentale. L'interaction entre les utilisateurs, les experts du domaine et le système pour améliorer la connaissance globale de système doit se faire de façon simple en cliquant des images pertinentes/non pertinentes pendant la recherche/l'exploration.
- Très peu de données d'entraînement (renforcement/approche incrémentale). Le nombre d'images cliquées pour chaque interaction doit être très faible (20 maximum).
- L'annotation d'images doit se faire en temps réel.
- Le nombre d'annotations pour chaque image varie de 1 à 5. Ce point est discuté dans la partie consacrée aux perspectives.

Considérant l'ensemble de ces hypothèses, à un instant t, il existe trois types d'images dans notre système :

1. Des images sans annotation.
2. Des images avec informations supplémentaires fournies par des experts (annotation manuelle).
3. Des images avec informations ajoutées automatiquement (que nous appelerons annotations "propagées").

La recherche de solution pour la mise en oeuvre de système aussi complexe passe par l'utilisation de méthodes interactives d'aide à la décision basées sur des annotations semi-automatiques d'images, mais également par l'utilisation de techniques de recherche combinée d'images basée sur le contenu et sur des informations externes (telles que le texte, la localisation, le temps), et enfin sur l'adoption de techniques exploratoires de recherche dans une base d'images. La décision sur une situation n'a pas vocation à être automatisée (interprétation automatique des images), mais le but est de fournir un ensemble d'informations sur les situations observées et d'aider un opérateur humain à se retrouver rapidement dans une énorme masse d'informations en fonction de ses besoins (requêtes) immédiats.

D'un point de vue scientifique (hors applicatif), les problématiques étudiées sont multiples : la recherche d'images par le contenu dans des bases d'images faiblement indexées ou avec de multiples indexations par image ; l'apprentissage interactif de connaissances (entre autres pour l'indexation) par des techniques adaptatives de type retour de pertinence, ou de type renforcement, visant un plus long terme ;

l'organisation interne de données multidimensionnelles et provenant de plusieurs sources (internes ou externes à l'image) dans ce contexte ; la visualisation et l'exploration de la base d'images en se basant sur des informations internes et externes.

Nous nous intéressons donc dans ce travail à un système de recherche multimodale d'images (la recherche multimodale correspond à la recherche combinée texte/contenu/informations externes). Considérant l'évolution de la base d'images et de la base de connaissance liées à notre application de gestion de crise, nous avons identifié différents scénarios possibles pour la recherche :

- Scénario 1 : Une image sans annotation est utilisée comme requête, visant à trouver des images similaires à partir d'une base de données.
- Scénario 2 : Un ensemble de mots est utilisé comme requête.
- Scénario 3 : Une image et des mots sont utilisés comme requête.

Le problème des deux derniers scénarios se pose en particulier lorsque tout ou partie de la base d'images n'est pas annotée, ce qui rend cette partie inaccessible par le biais des requêtes textuelles. Dans ce travail, nous étudions donc comment utiliser l'association entre mots textuels et caractéristiques visuelles pour la recherche multimodale et pour l'annotation d'images dans ce cas.

1.3 Contributions

Nous avons présenté les hypothèses du système et ainsi que son contexte d'utilisation. Nos contributions scientifiques et applicatives dans l'élaboration de ce système portent sur les quatre points suivants :

1. Un système de recherche multimodale d'images qui intègre différentes sources de données, comme le contenu de l'image et le texte. Ce système permet l'interrogation par l'image, l'interrogation par mot-clé ou encore l'utilisation de requêtes hybrides.
2. Une nouvelle technique pour le retour de pertinence combinant deux techniques classiques utilisées largement dans la recherche d'information : le mouvement de requête et l'extension de requêtes. En profitant des images non pertinentes et des avantages de ces deux techniques classiques, notre méthode donne de très bons résultats pour une recherche interactive d'images efficace.
3. Un modèle nommé "Sacs de KVR" (*Keyword Visual Representation*) créant des liens entre des concepts sémantiques et des représentations visuelles, en appui sur le modèle de Sac de Mots [Sivic 2008]. Grâce à une stratégie d'apprentissage incrémental, ce modèle fournit l'association entre concepts sémantiques et caractéristiques visuelles, ce qui contribue à améliorer la précision

de l'annotation sur l'image et la performance de recherche.

4. Un mécanisme de construction incrémentale des connaissances à partir de zéro. Nous ne séparons pas les phases d'annotation et de recherche, et l'utilisateur peut ainsi faire des requêtes dès la mise en route du système, tout en laissant le système apprendre au fur et à mesure de son utilisation.

1.4 Structure de la dissertation

En plus de ce chapitre introductif, ce manuscrit est composé de cinq autres chapitres.

Le chapitre 2 présente l'état de l'art de la recherche d'images. Dans ce chapitre, des caractéristiques visuelles et textuelles, des fonctions de similarité, des modèles de la recherche d'images, ainsi que des systèmes de recherche d'images sont présentés. Nous introduisons ensuite notre système de recherche d'images en proposant un aperçu de l'architecture de ce système.

Dans le chapitre 3, nous présentons les interactions possibles dans les systèmes de recherche d'images. Nous introduisons la méthode d'interaction utilisée dans notre système pour la recherche d'images par l'exemple et la recherche d'images par utilisation de requête mixte image/texte.

L'association du texte et de l'image est présentée dans le chapitre 4. Après l'introduction de l'état de l'art des modèles de l'association du texte et de l'image, nous proposons notre modèle Sacs de KVR. Ensuite, l'utilisation du modèle Sacs de KVR pour l'annotation d'images et la recherche d'images est présentée.

Le chapitre 5 est consacré à l'expérimentation de notre système et à son évaluation. Des évaluations sur des jeux de données significatifs sont proposés, mettant en évidence l'intérêt de la démarche proposée.

Enfin, nous concluons notre manuscrit et proposons des perspectives à ce travail dans le dernier chapitre 6.

Chapitre 2

État de l'art

2.1 Introduction

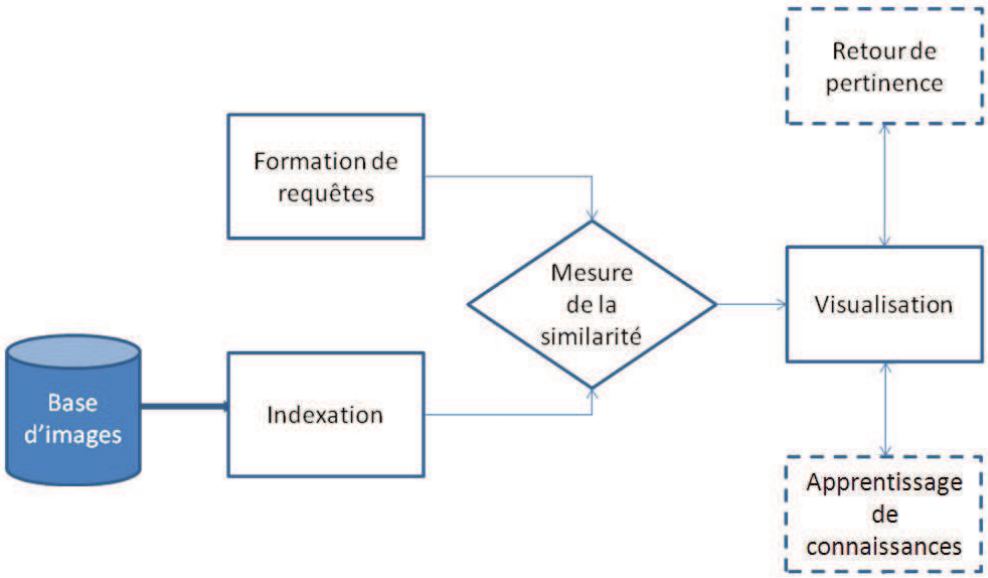

FIGURE 2.1 – Exemple d'architecture de système de recherche d'images.

Au cours de ces dernières années, le volume des données multimédias a rapidement augmenté grâce aux nouvelles technologies, notamment avec les nombreux outils de production d'images (ou de vidéos) qui se sont banalisés, et se sont intégrés dans les réseaux sociaux et les dispositifs de partage d'images. Il existe beaucoup de source d'images : images diffusées sur l'internet, images personnelles, images d'art, images médicales, images de l'espace. . .Ces sources d'images demandent des applications pour organiser, rechercher, explorer et profiter de ces images pour répondre à des besoins divers et variés. Dans ce contexte général et avec cette optique

de développement d'applications, les techniques de traitement d'images et de vision par ordinateur ont été largement étudiées. Ces techniques se sont récemment concentrées autour d'une problématique principale : la recherche d'images.

Dans la figure 2.1, nous présentons de façon générale les composants principaux de la recherche d'images, c'est-à-dire : l'indexation, la formation de requêtes, le modèle de la recherche, la visualisation, l'interaction et l'apprentissage.

L'indexation d'images est l'opération qui consiste à extraire une signature numérique ou textuelle d'une image, signature qui décrit son contenu sémantique d'une manière précise et concise, afin de permettre une recherche efficace dans une base de données. Selon la terminologie généralement utilisée, l'indexation intègre également la structuration des données dans l'espace dans lequel les images sont projetées, afin d'améliorer la vitesse d'accès à l'information pertinente (clustering, structures d'index, ...). La formation de requêtes est l'opération qui permet de représenter les intérêts de l'utilisateur. Le modèle de la recherche consiste en des processus de recherche et de mesure de similarités qui permettent de comparer des signatures d'images. La visualisation permet à l'utilisateur de voir les images sur une interface (très souvent sur un écran). L'interaction consiste en des techniques permettant d'intégrer l'utilisateur dans le processus de recherche d'images. L'apprentissage peut être mis en œuvre pour améliorer la performance de la recherche d'images.

Dans ce chapitre, l'indexation d'images en appui sur des informations internes/externes (terminologie que nous préciserons) est d'abord présentée. Puis la formation de requêtes est présentée, suivie des modèles de recherche d'images selon deux catégories : la recherche unimodale et la recherche multimodale. Notre choix d'indexation et du modèle de recherche est aussi discuté. Des systèmes spécialisés sur la base d'un choix relatif à notre intérêt sont présentés ensuite. Nous résumons après cela quelques systèmes commerciaux de recherche d'images. Enfin, nous proposons notre système de recherche d'images plutôt orienté pour une application de catastrophes naturelles. Dans cette partie, la visualisation, l'interaction et l'apprentissage ne sont pas immédiatement approfondis, car nous les abordons en détails dans les chapitres 3 et 4, ces problématiques constituant nos contributions principales.

2.2 Indexation

Les différents types d'informations qui sont classiquement associés aux images sont les suivants :

– Les informations internes qui comprennent des données basées sur le contenu

de l'image : couleur, texture, forme...
- Les informations externes, parfois indépendantes du contenu purement pixellaire et qui sont souvent représentées par des métadonnées, ou des données qui ne sont pas directement reliées au contenu de l'image, mais qui s'y rapportent, comme par exemple : la description de l'image, le format de l'image, le nom de l'auteur, la date et l'heure d'acquisition ou encore la localisation. Ces informations externes peuvent être utilisées indépendamment ou avec les informations internes.

Dans notre travail, nous proposons un système de recherche d'images qui permet d'utiliser plusieurs types d'informations internes/externes dont le texte (annotations), le contenu, le temps et la localisation. En fait, dans le cadre de cette thèse, nous nous concentrons sur un système de recherche d'images par le texte et le contenu. Les autres informations sont discutées dans les perspectives.

2.2.1 Information internes

Les images sont indexées par des caractéristiques visuelles qui peuvent être extraites à l'échelle globale, à l'échelle locale ou à l'échelle semi-locale. Les caractéristiques globales décrivent l'image dans son intégralité. Les caractéristiques locales sont extraites à partir des régions ou des objets d'une image. Pour obtenir les caractéristiques locales de l'image, une image est souvent divisée en différentes parties. La méthode la plus simple pour la division est la grille régulière [Vogel 2006] [Li 2005]. L'image est segmentée par des lignes horizontales et verticales. Une autre méthode très populaire est le détecteur de points d'intérêt [Li 2005] [Sivic 2008]. En outre, des méthodes de segmentation sont aussi souvent utilisées [Vidal-Naquet 2003] [Barnard 2003]. Dans les parties suivantes, nous présentons les signatures globales et locales les plus connues dans la recherche d'images par le contenu.

2.2.1.1 Signatures globales d'images

Il existe de nombreuses de signatures globales d'images, mais nous nous limiterons à celles issues de la couleur, la texture, la forme et/ou la disposition spatiale qui représentent les caractéristiques visuelles les plus courantes.

Couleur : La caractéristique de couleur est largement utilisée dans la recherche d'images pour des images couleurs de scènes naturelles. Trois types principaux de signatures de couleurs existent dans la littérature : l'histogramme des couleurs, les moments de couleurs et le corrélogramme de couleurs. Un histogramme des couleurs représente la répartition du nombre de pixels pour chaque partition quantifiée dans

chaque canal de couleur. L'espace couleur est partitionné et pour chaque partition les pixels de même couleur sont comptés, ce qui entraîne une représentation de la fréquence relative des couleurs présentes dans l'image. Les moments de couleur (équation 1) : la moyenne (E), la variance (σ) et l'asymétrie (s) sont premièrement utilisées dans [Flickner 1995] pour caractériser la distribution des couleurs.

$$E_i = \frac{1}{N}\sum_{j=1}^{N} p_{ij}, s_i = \left(\frac{1}{N}\sum_{j=1}^{N}(p_{ij} - E_i)^3\right)^{\frac{1}{3}}, \sigma_i = \left(\frac{1}{N}\sum_{j=1}^{N}(p_{ij} - E_i)^2\right)^{\frac{1}{2}} \quad (1)$$

Un corrélogramme des couleurs [Huang 1997] est un histogramme à trois dimensions, dont la première et la seconde sont les distributions des paires de pixels et la troisième dimension est leur distance spatiale. Il est utilisé pour caractériser non seulement les distributions de couleur des pixels, mais aussi la corrélation spatiale des paires de couleurs.

Texture : Les signatures de textures sont très nombreuses dans la littérature et il est difficile d'en dresser une liste exhaustive. Aussi nous nous limiterons à certaines d'entre elles fréquemment utilisées, c'est-à-dire les matrices de cooccurrences de Haralick, les caractéristiques de texture de Tamura et les filtres de transformation en ondelettes de Gabor. Les matrices de cooccurrences de Haralick représentent la dépendance spatiale des niveaux de gris de texture [Haralick 1973] en considérant les valeurs des pixels à différentes orientations et distances. Différentes statistiques sont alors calculées (14 au total) à partir de la matrice. Ces statistiques comprennent entre autres le moment différentiel inverse et l'entropie qui sont utilisées en tant que caractéristiques de Haralick. Les caractéristiques de texture de Tamura comprennent la grossièreté, le contraste, la directionnalité, la linéarité, la régularité et la rugosité [Tamura 1978]. Les filtres de transformation en ondelettes [Mallat 1989] fournissent une approche multi-résolution pour l'analyse et la classification de textures.

Forme : En général, les représentations de la forme peuvent être divisées en deux catégories : basées sur les contours ou basées sur les régions. Des caractéristiques basées sur les contours n'utilisent que les contours extérieurs d'une forme. Des caractéristiques très utilisées sont celles basées sur les contours sur lesquels sont calculés les descripteurs de Fourier. L'idée principale d'un descripteur de Fourier est d'utiliser la transformée de Fourier des contours en tant que caractéristiques de la forme. Les caractéristiques basées sur les régions utilisent l'intérieur de la région de la forme. Les caractéristiques basées sur la région sont présentées par une collection de primitives comme le centroïde, la superficie, l'excentricité, la circularité et les moments statistiques. Parfois, pour répondre à certaines contraintes de multi-orientations ou de multi-échelles, d'autres descripteurs de forme, plus élaborés sont utilisés, comme les invariants de Zernike, ou encore les invariants de

Fourier Mellin.

Disposition spatiale : La disposition spatiale [Philipp-Foliguet 2009], [Pham 2010] encode la position absolue ou relative des objets ou régions segmentées en des relations topologiques et/ou directionnelles. Les relations topologiques décrivent les relations entre les frontières des objets, telles que "près de", "à l'intérieur de" ou "à côté de". Les relations directionnelles montrent les positions relatives des objets les uns par rapport aux autres, comme par exemple "devant", "à gauche" ou "en haut de"... Ces relations spatiales sont souvent décrites avec un graphe relationnel attribué (ARG). Les nœuds du graphe représentent des objets, et un arc entre deux nœuds représente une certaine relation entre eux, souvent de nature topologique.

Dans les systèmes de recherche d'images, les caractéristiques visuelles sont généralement calculées hors ligne et stockées pour chaque image. Ces caractéristiques sont choisies en se basant sur la base d'images et sur l'objectif de chaque système. Pour une base d'images couleurs, la caractéristique de couleur est fréquemment utilisée. Pour des bases de patterns, d'images satellites, d'images médicales ou d'images de scènes naturelles comme les nuages, la caractéristique de texture est souvent utilisée. La caractéristique de forme est adaptée pour représenter et décrire des images d'objets déjà segmentés (par exemple la base Columbia), des dessins ou des images de marques.

2.2.1.2 Signatures locales d'images

Pour l'extraction de caractéristiques locales, une image peut être divisée en petits blocs et les caractéristiques sont alors calculées individuellement pour chaque bloc. Les signatures sont associées à des caractéristiques différentes de l'image (calculées sur les spécificités locales : régions, frontières ou points d'intérêt) et sont appelées descripteurs locaux. Les signatures locales sont très robustes aux occultations et aux transformations.

Tout d'abord, pour extraire les signatures locales, il est souvent nécessaire de détecter des régions locales (régions, points d'intérêt). [Lindeberg 1998] a développé un détecteur de "blobs" invariant aux changements d'échelles, où un blob est défini par un maximum du Laplacien normalisé dans l'espace d'échelles. [Matas 2002] a introduit l'extraction de régions extrêmes maximalement stables (*Maximally Stable Extremal Regions*) avec un algorithme de segmentation de lignes de partage des eaux. [Lowe 2004] rapproche le Laplacien avec le filtre Différence de gaussiennes (DoG) et détecte également les extrémums locaux dans l'espace d'échelles. [Tuytelaars 2004] construit deux types de détecteurs de régions invariant aux transformations affines, l'un basé sur une combinaison de points d'intérêt et les bords

et l'autre basé sur l'intensité de l'image. [Kadir 2004] mesure l'entropie d'histogrammes d'intensités des pixels calculés pour les régions elliptiques pour trouver des maximas locaux dans l'espace des transformations affines. Une comparaison de l'état de l'art de détecteurs affines de régions peut être trouvée dans [Mikolajczyk 2005a]. Les points d'intérêt (POI), des points de l'image qui peuvent être localisés de façon unique, sont les caractéristiques les plus populaires, en raison de leur robustesse.

Quand un point ou une région est détecté, une petite fenêtre autour du point est utilisée pour calculer le descripteur local. Des descripteurs différents sont comparés dans [Mikolajczyk 2005b] : contexte de forme, filtres orientables, PCA-SIFT, invariants différentiels, spin images, SIFT, filtres complexes, moments invariants. Selon l'auteur, le classement des descripteurs est majoritairement indépendant du détecteur de régions d'intérêt et le descripteur SIFT donne la meilleure performance.

Le descripteur SIFT [Lowe 2004] est un histogramme 3D des localisations du gradient et des orientations, la localisation étant quantifiée dans une grille 4x4 et l'angle du gradient étant quantifié en 8 orientations. Le descripteur qui en résulte est en 128 dimensions.

Il existe d'autres descripteurs de caractéristiques locales qui ont été proposés récemment, soit le descripteur flou géométrique [Berg 2005], le descripteur de l'auto-similitude [Shechtman 2007] ou encore le descripteur SURF [Bay 2006].

Afin de calculer la similarité entre des images, un système basé sur une approche à base de descripteur local doit comparer des ensembles de descripteurs. Il s'agit d'une opération complexe. La plupart des systèmes adoptent un critère de "vote", qui est simple et qui permet d'éviter la comparaison (de coût élevé) entre l'image requête et chaque image dans la base de données. Plus récemment, un nouveau modèle pour la représentation des descripteurs locaux a été proposé basé sur les "sacs de mots". Ce modèle Sacs de Mots a l'avantage de permettre l'utilisation des techniques de recherche de texte pour représenter et comparer des descripteurs locaux. Ce modèle de sacs de mots sert de base à notre travail et nous y revenons en détails dans le chapitre 4.

2.2.2 Informations externes

Des informations externes indépendantes du contenu sont des données qui ne concernent pas directement le contenu de l'image, mais qui s'y rapportent, comme par exemple : le format de l'image, le nom de l'auteur, le temps, la localisation et

la description. Ces informations externes peuvent être utilisées indépendamment ou avec les informations internes. Dans notre travail, nous proposons un système de recherche d'images qui permet d'utiliser comme types d'informations principalement le contenu et le texte (annotations), mais qui peut être étendu facilement aux informations de temps et de localisation. Quelques pistes d'extensions possibles sont présentées dans les perspectives de notre travail.

2.2.2.1 Signatures textuelles

Les images peuvent être indexées par des descriptions textuelles. Ces informations textuelles sont issues d'opérations humaines ou produites par des machines, ces deux cas étant définis comme annotation manuelle ou annotation automatique d'images. Aujourd'hui la plupart des systèmes commerciaux de recherche d'images sont basés sur des informations textuelles d'images.

Tandis que les caractéristiques visuelles de bas niveau sont directement liées à des aspects perceptifs du contenu de l'image, les informations textuelles/conceptuelles de haut niveau des images sont normalement représentées à l'aide de descripteurs offrant un cadre simplifié et efficace pour la recherche d'images à partir de requêtes textuelles. Dans certaines techniques de recherche basée sur le contenu, les descripteurs de texte sont également utilisés pour modéliser les aspects perceptifs. Toutefois, l'insuffisance de la description textuelle est problématique. En effet, la plupart des bases de données d'images ne sont pas indexées par le texte. Pour effectuer la recherche d'images par le texte, l'indexation d'images utilisant des mots, des phrases ou des textes complets doit être réalisée en premier lieu. La tâche d'affectation des textes aux images est appelée annotation d'images. Il y a un manque significatif d'annotations des images parce que l'annotation d'images est un processus fastidieux pour l'homme. Par conséquent, de nombreux travaux sur l'annotation automatique / semi-automatique des images sont proposés aujourd'hui. L'état de l'art de l'annotation d'images est présenté en détails dans le chapitre 4 (section 4.1).

2.2.2.2 Métadonnées de l'image

Les métadonnées de l'image peuvent souvent être catégorisées en différentes parties distinctes : celles relatives à la technique, au contenu et à la description et les autres métadonnées. Les métadonnées techniques concernent par exemple des informations sur le système d'acquisition d'images ; ces informations sont soit disponibles soit intégrées dans le fichier image ou à la disposition d'applications externes. Les métadonnées de contenu décrivent le contenu de l'image, comme par exemple les couleurs, les textures, les formes ou encore d'autres caractéristiques de

l'image originale, telles que la disposition des objets mis en forme par exemple. Les métadonnées de description comprennent la date, l'heure, la localisation, les personnes, les événements ou les sentiments subjectifs. D'autres métadonnées peuvent se référer à la localisation réelle de l'image, à des droits de propriété intellectuelle, et même à des questions d'intérêt pour le stockage et la distribution de l'image. Notons qu'il existe un certain nombre de normes de métadonnées sur l'image actuellement acceptées par différents organismes de normalisation, comme par exemple : VRA Core 3.0, MOA2, METS, MPEG-7...

2.3 Recherche d'images

Dans cette partie, nous présentons en détails des modèles pour la recherche d'images en utilisant les informations présentées dans la section précédente.

2.3.1 La formation de requêtes

Pour la recherche d'images, les utilisateurs doivent fournir des requêtes. Les modalités de requêtes et les traitements associés peuvent être catégorisés comme suit [Datta 2008] :

1. Mots-clés : L'utilisateur pose une simple requête sous la forme de mots textuels. C'est actuellement le moyen le plus populaire pour la recherche d'images, comme par exemple, celui utilisé par Google et Yahoo !, deux moteurs commerciaux de recherche d'images. Ce type de système demande une indexation par signatures textuelles de la base d'images. La recherche d'images est réalisée en utilisant le modèle de la recherche par le texte.
2. Texte libre : L'utilisateur fournit une phrase complexe, une question ou une histoire sur ce qu'il désire à partir du système. Comme dans le cas précédent, il s'agit ici de recherche par le texte, et il faut donc une indexation par signatures textuelles sur la base d'images et un modèle de recherche par le texte.
3. Images : L'utilisateur souhaite rechercher une image semblable à une image requête. L'utilisation d'une image comme exemple est peut-être le moyen le plus représentatif de l'interrogation d'un système CBIR (*Content-Based Image Retrieval*) en l'absence de données fiables de type métadonnées. Avec l'indexation par signatures visuelles, la recherche d'images est réalisée par le modèle de la recherche par le contenu.
4. Esquisse : Des dessins à la main ou des images générées par ordinateur ou des graphiques peuvent être présentés en tant que requête. Comme la requête

par images, ce type de requête demande l'indexation visuelle et le modèle de recherche par le contenu.

5. Composé : Ce sont des méthodes qui combinent l'utilisation d'une ou plusieurs modalités de requête. Elles couvrent également l'interrogation interactive telle que dans les systèmes de retour de pertinence. Pour utiliser la requête composée, il faut généralement procéder à l'indexation sur la base d'une exploitation combinée des signatures textuelles et des signatures visuelles. De plus, les deux modèles de recherche par le texte et par le contenu doivent être combinés, ou sinon les 2 signatures visuelles et textuelles doivent être combinées, par concaténation ou encore par sélection.

2.3.2 Modèle de recherche par le texte

2.3.2.1 Modèle vectoriel

Le modèle d'espace vectoriel a été proposé par Salton et al. en 1975 [Salton 1975]. Chaque document dans une base de données D, ainsi que la requête Q (Q est aussi bien un document ou un ensemble de documents), est représenté comme un vecteur pondéré de termes :

$$d_i = (w_{i,1}, w_{i,2} \dots w_{i,t}) \tag{2}$$

Pour calculer la similarité entre la requête Q et le document D, on utilise le produit scalaire entre les vecteurs correspondants :

$$S(d_i, q) = \sum_j w_{i,j} \times w_{q,j} S(d_i, q) = \sum_j w_{i,j} \times w_{q,j} \tag{3}$$

Différentes méthodes d'estimation de $w_{i,j}$ et $w_{q,j}$ ont permis la création de plusieurs fonctions de classement pour le modèle d'espace vectoriel. Parmi celles-ci, le modèle de pondération classique TF-IDF et l'état de l'art de la pondération et de la normalisation par pivot qui sont décrits ci-dessous.

Le modèle vectoriel classique proposé dans [Salton 1975] calcule le poids d'un terme dans un document comme un produit de la fréquence du terme (TF) et de la fréquence inverse de document (IDF).

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$idf_i = \log \frac{N}{|\{d : t_i \in d\}|}$$

où $n_{i,j}$ est le nombre d'occurrences du terme considéré (t_i) dans le document d_j et le dénominateur est la somme du nombre d'occurrences de tous les termes du

document d_j. N est le nombre total de documents dans le corpus, $|\{d : t_i \in d\}|$ est le nombre de documents où le terme t_i apparaît (qui a $n_{i,j} = 0$).

Nous obtenons la fonction de classement pour cette méthode :

$$S(d_i, q) = \frac{\sum_j D_i \times D_q}{|D_i|\,|D_q|} S(d_i, q) = \frac{\sum_j D_i \times D_q}{|D_i|\,|D_q|} \tag{4}$$

avec $D_k = tf_k.idf_k$ et $|D_i|\,|D_q|$ le dénominateur pour la normalisation.

Récemment, cette fonction de classement a été utilisée pour la recherche d'images [Sivic 2008], [Yang 2007], [Tirilly 2008], [Karthik 2006].

La deuxième méthode est proposée dans [Singhal 2001], [Fang 2005]. C'est une méthode pour la normalisation de la longueur du document :

$$S(d_j, q) = \sum_{t=1}^{n} \frac{1 + \log\left(1 + \log\left(tf_{t,D}\right)\right)}{(1-s) + s\frac{dl}{avdl}} tf_{t,Q} \log \frac{N+1}{n_i} \tag{5}$$

avec dl étant la longueur du document, $avdl$ étant la longueur moyenne des documents, N étant le nombre de documents du corpus, n_i étant le nombre de documents qui contiennent le terme, et s étant une constante, généralement 0,20.

2.3.2.2 Modèle probabiliste

Dans le modèle probabiliste, on estime la probabilité pour qu'un document d_j soit pertinent pour une requête q spécifique [Jones 2000], [Fuhr 1992], notée $P\left(R|q, d_j\right)$:

$$P\left(R|q, d_j\right) = \frac{P\left(d_j relevant - to - q\right)}{P\left(\left(d_j non - relevant - to - q\right)\right)} \tag{6}$$

L'ensemble des termes utilisés dans un document d_j peut être représenté comme un vecteur binaire $x = (x_1, x_2 ... x_n)$ avec $x_i = 1$, si le terme i est présent dans d_j et $x_i = 0$ sinon. Ensuite, les documents sont classés par ordre décroissant en fonction de l'expression suivante :

$$S(d_j, q) = \sum_{i=1}^{n} \log \frac{P\left(x_i|R\right)\left(1 - P\left(x_i|\overline{R}\right)\right)}{\left(1 - P\left(x_i|R\right)\right) P\left(x_i|\overline{R}\right)} \tag{7}$$

où R est l'ensemble des exemples pertinents (exemples positifs) et $\overline{R}$ est l'ensemble des exemples non pertinents (exemples négatifs). $P\left(x|R\right)$ et $P\left(x|\overline{R}\right)$ sont les probabilités d'un élément pertinent ou non pertinent conditionnellement à la représentation vectorielle x.

La fonction de classement de l'état de l'art pour le modèle probabiliste est la

fonction Okapi BM25 par Sparck Jones et al. [Jones 2000] :

$$S(d_j, q) = \sum_{t \in Q} W_t \frac{(k_1 + 1) tf_{t,D}}{k_1((1-b) + \frac{b.dl}{avdl}) + tf_{t,D}} \frac{(k_3 + 1) tf_{t,Q}}{k_3 + tf_{t,Q}} \tag{8}$$

avec k_1 (entre 1,0 et 2,0), b (souvent égal à 0,75) et k_3 souvent fixé à 7 ou 1000 (infini dans la pratique) étant des constantes. L'ensemble des paramètres ($k_1 = 1,2$, $b = 0,75$, $k_3 = 1000$) est utilisé souvent dans les expérimentations. W_t est le poids de Robertson / SPARCK Jones du terme t dans la requête tel que :

$$W_t = \log \frac{(r_t + 0.5)(N - R - n_t + r_t + 0.5)}{(n_t - r_t + 0.5)(R - r_t + 0.5)} \tag{9}$$

où N est la taille du corpus, R est la taille de l'ensemble pertinent, n_t est le nombre de documents dans N contenant le terme t, r_t est le nombre de documents dans R contenant le terme t. Ce poids est utilisé pour le retour de pertinence. La recherche initiale utilise la formule de réduction du poids de Robertson / SPARCK Jones (avec $R = 0$, $r_t = 0$) :

$$W_t = \log \frac{N - n_t + 0.5}{n_t + 0.5} \tag{10}$$

2.3.3 Modèle de recherche par le contenu

Dans l'approche CBIR, la recherche d'images est basée sur les caractéristiques extraites automatiquement du contenu visuel. L'approche CBIR s'appuie sur une requête par l'image où l'image requête est donnée par l'utilisateur et le système mesure la similarité entre les caractéristiques de l'image requête et celles des images dans la base de données. Un système CBIR peut utiliser différents ensembles de caractéristiques, nécessitant de faire le choix d'une mesure de similarité appropriée pour chaque ensemble de caractéristiques de l'image. La combinaison optimale des similarités calculées à partir d'un ensemble de caractéristiques différentes de l'image est devenue un enjeu important. Le modèle de recherche CBIR est dépendant de la fonction de similarité et des signatures visuelles utilisées qui peuvent être des vecteurs de caractéristiques, des signatures basées sur les régions ou des caractéristiques locales résumées. Le modèle d'espace vectoriel et le modèle probabiliste de la recherche par le texte sont adaptés pour l'approche CBIR. D'autres techniques spéciales ont été développées pour rechercher des images en utilisant différentes représentations des caractéristiques. Les mesures de similarité pour la recherche d'images par le contenu sont résumées dans [Liu 2008].

2.3.3.1 Modèle vectoriel

La plupart des systèmes CBIR sont construits sur un modèle d'espace vectoriel. Les corpus d'images et de requêtes sont représentés comme des vecteurs de caractéristiques dans un espace vectoriel à n dimensions. La fonction de classement est dépendante des caractéristiques utilisées. Par exemple, la similarité des caractéristiques de texture est souvent mesurée en utilisant la distance de Minkowski ou la distance de Mahalanobis (qui considère la corrélation des caractéristiques différentes). L'intersection d'histogrammes est également populaire pour mesurer la similarité des couleurs en se basant sur les histogrammes couleurs.

Considérons deux images, indexées par leurs vecteurs respectifs $I = (I_1, \ldots, I_n)$ et $J = (J_1, \ldots, J_n)$. Le calcul de similarité entre 2 images revient à calculer la similarité entre I et J. Les métriques de Minkowski (ou normes Lp) sont les distances géométriques les plus courantes. Leur forme générale est la suivante :

$$L_p = \sqrt[p]{\sum_{i=1}^{n} (I_i - J_i)^p)}$$

La distance de Mahalanobis prend en compte la corrélation entre vecteurs et la distribution des données au lieu des distances dans le cas quadratique. La matrice C est la matrice de covariance de la distribution de I et J : $d = \sqrt{(\vec{I} - \vec{J})^T C^{-1} (\vec{I} - \vec{J})}$.

Considérons d'abord une signature d'image sous la forme d'un ensemble de vecteurs de caractéristiques $((z_1, p_1), (z_2, p_2), \ldots, (z_n, p_n))$ où les z_i sont les vecteurs de caractéristiques et les p_i sont les poids correspondants qui leur sont assignés. Supposons deux signatures $I_m = ((z_1^m, p_1^m), (z_2^m, p_2^m), ..., (z_n^m, p_n^m))$, $m = 1, 2$. Une approche naturelle à la définition d'une mesure de similarité est de mesurer la similarité entre z_i^1 et z_i^2, puis de combiner les distances entre ces vecteurs en une distance entre des ensembles de vecteurs.

Une approche pour la mesure [Wang 2000] est d'attribuer un poids à chaque paire z_i^1 et z_i^2 avec $1 \leq i \leq n$ et $1 \leq j \leq n$, tel que le poids $s_{i,j}$ indique l'importance d'associer z_i^1 et z_i^2. Les poids sont soumis à des contraintes, les plus communes étant $\sum_i S_{i,j} = p_j^2$ et $\sum_j S_{i,j} = p_i^1$. Une fois que les poids sont déterminés, la distance entre I_1 et I_2 est agrégée à partir des distances entre les paires de vecteurs différents. Plus précisément :

$$D(I_1, I_2) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} s_{i,j} d\left(z_i^{(1)}, z_j^{(2)}\right)$$

la distance d(x, y) pouvant être définie de diverses manières. D'autres méthodes utilisent la distance de Hausdorff, où chaque z_i^1 est adaptée à son plus proche vecteur dans I_2, c'est-à-dire z_i^2, et la distance entre I_1 et I_2 est le maximum parmi tous les $d\left(z_i^1, z_i^2\right)$.

2.3.3.2 Modèles probabilistes

Les modèles probabilistes ont été adaptés à l'approche CBIR [Vasconcelos 2004], [Zhang 2005], [Westerveld 2003]. Chaque image est considérée comme un modèle probabiliste génératif qui définit une distribution de probabilité sur les caractéristiques visuelles. Pour la recherche d'images, une image est traitée comme une observation de l'un des modèles et les modèles sont classés par l'ordre décroissant de probabilité en générant cette observation. Cela suppose qu'il existe un sous-ensemble d'images R pertinentes que l'utilisateur veut retrouver parmi celles disponibles, les autres images $\bar{R}$ étant considérées comme non pertinentes. Si $P(R|\vec{Q})$ est la probabilité que l'image I soit pertinente pour la requête Q et si $P(\bar{R}|\vec{Q})$ est la probabilité que l'image I ne soit pas pertinente pour la requête Q, alors la similarité entre l'image I et la requête Q est exprimée par : $sim(I, Q) = \frac{P(R|\vec{Q})}{P(\bar{R}|\vec{Q})}$.

L'idée de [Vasconcelos 2000] est d'intégrer la sélection de variables, la représentation fonctionnelle et des mesures de similarité dans une formulation bayésienne, avec l'objectif de minimiser la probabilité d'erreur. Un problème connu de cette approche est la complexité des calculs impliqués dans l'estimation des mesures de similarité probabiliste. La complexité est réduite dans [Vasconcelos 2004] en utilisant la quantification vectorielle pour modéliser des distributions de probabilité des caractéristiques d'images.

2.3.3.3 Autres modèles

Pour les caractéristiques qui ne sont pas représentables sous la forme d'un vecteur de caractéristiques, telles que les descripteurs de la disposition spatiale, la mesure de similarité est plus complexe. Les techniques mises en oeuvre sont souvent liées à l'objet à reconnaître et font appel à des méthodes issues de la vision par ordinateur, en particulier aux méthodes structurelles. Veltkamp et Hagedoom [Veltkamp 2001] donnent un aperçu des techniques d'appariement de formes, telles que l'appariement basé sur la transformation des fonctions d'angle, l'appariement de modèles déformables et l'appariement de graphes. Par exemple, dans [Philipp-Foliguet 2009] la mesure de similarité est basée sur l'appariement de graphes. Les auteurs proposent d'utiliser une recherche d'images par noyaux sur graphes de régions pour effectuer l'appariement de graphes en tenant compte à la fois de la similitude entre les régions et des relations spatiales entre elles. Une autre approche

est l'utilisation de la quantification vectorielle (VQ) sur des blocs de l'image pour générer un vocabulaire pour la représentation et la recherche, en s'inspirant de la compression de données et des stratégies basées sur le texte [Sivic 2008]. Cette approche s'appelle le modèle "Sac de Mots" qui peut être utilisé avec le modèle vectoriel ou le modèle probabiliste (voir le chapitre 4 pour plus de détails).

2.3.4 Avantages et inconvénients des modalités par le texte et par le contenu

2.3.4.1 Modèle par le texte

La stratégie la plus largement acceptée et mise en œuvre pour la recherche d'images est basée sur le texte. La recherche d'images par le texte a été utilisée avec succès dans de grosses bases d'images, telles que la recherche des images sur le web. Les avantages sont que les images peuvent être relativement facilement représentables en utilisant un ensemble de mots, et les utilisateurs représentent leurs informations naturellement en utilisant des *mots textuels*[1]. Plus précisément, en utilisant la modalité textuelle, nous pouvons trouver des images avec des concepts sémantiques, trouver des images appartenant à certains événements (par exemple montrer toutes les images du tournoi de Wimbledon 2010), trouver des images prises à un endroit particulier...Beaucoup d'images sont créées avec des légendes, des descriptions, des tags...et sont donc ouvertes à la recherche textuelle.

La recherche d'images par texte a été utilisée en ignorant complètement le contenu visuel. Cette stratégie de recherche d'images est basée sur l'hypothèse que la description textuelle est suffisante pour répondre aux requêtes de l'utilisateur. Toutefois, les annotations ne sont pas suffisantes pour décrire le contenu visuel d'une image, comme la couleur, la texture, la forme ou la composition spatiale d'une image. [Sclaroff 1999] indique que certaines images ne peuvent pas être annotées, car il est difficile de décrire leur contenu visuel avec des mots, et un système d'extraction de texte peut facilement confondre deux images juste en regardant leurs descriptions. En outre, représenter les images par un ensemble de termes conduit inévitablement à une certaine perte de sémantique dans les images. La polysémie et la synonymie représentent effectivement deux difficultés importantes, pour l'utilisation de textes pour la recherche. De plus, la recherche d'images par le texte présuppose une annotation préalable des images, opération très chronophage et extrêmement subjective (voir le chapitre 4 pour plus d'informations sur l'annotation d'images).

1. Nous employons dans ce travail l'expression "mot textuel" pour désigner des mots-clés afin de bien faire la distinction avec les "mots visuels". Nous opposons souvent ces deux concepts, mot textuel et mot visuel, tout au long de notre travail.

2.3.4.2 Modèle par le contenu

La recherche d'images basée sur le contenu visuel a été un sujet actif de recherche au cours des dernières années. Avec l'hypothèse *"Une image vaut mille mots"*, la recherche d'images par le contenu est considérée comme un modèle puissant. La recherche d'images par le contenu permet aux utilisateurs de rechercher des images en fonction du contenu image visuel, comme la couleur, la texture et la forme, qui sont indépendants du domaine, et non limité à un domaine particulier. En outre dans ce contexte, nous pouvons rechercher des images sans aucune annotation. Par exemple, il est possible de trouver des images contenant une personne en particulier sur la base d'une détection appropriée des visages et d'une reconnaissance appropriée des visages pour trouver des images similaires. On peut également trouver des images avec certaines couleurs ou rechercher les images en double ou redondantes. Mais un des arguments principaux qui milite pour l'utilisation de l'indexation par le contenu est son objectivité, en comparaison avec un système d'annotation manuelle.

Toutefois, le dicton *"Une image vaut mille mots"* illustre bien l'idée que les images sont complexes à décrire. Le modèle de la recherche par le contenu a donc des limites. Il n'a pas été utilisé en application commerciale aussi largement que la recherche par le texte. Il a été rapidement montré par les chercheurs que les caractéristiques visuelles ne sont pas suffisantes pour la recherche d'images en raison du fossé sémantique, à savoir le manque de liens entre l'information sémantique et les caractéristiques visuelles de bas niveau, calculées sur les images [Smeulders 2000]. Les images résultats basées sur les similarités de contenu visuel ne possèdent pas nécessairement les valeurs sémantiques qui intéressent l'utilisateur. Par exemple, deux images avec des distributions de couleurs et des organisations spatiales très similaires, comme le jaune au centre, le bleu en haut et le vert en bas, peuvent signifier deux choses complètement différentes d'un point de vue humain. La première peut illustrer le soleil dans un ciel bleu avec quelques herbes au sol, tandis que la seconde montre une orange sur un fond bleu avec un ensemble de légumes. Même si elles partagent de nombreuses propriétés visuelles du point de vue numérique, la sémantique de ces images est extrêmement différente. En outre, le syndrome de la "page vide" existe aussi dans la recherche d'images par le contenu, c'est-à-dire que dans de nombreux cas, les utilisateurs n'ont pas d'images à disposition pour fournir un exemple de requête. Dans certains systèmes, une interface d'interrogation complexe doit être conçue pour permettre à l'utilisateur de dessiner une esquisse. Mais une interface d'interrogation complexe peut par ailleurs ne pas être parfaitement satisfaisante pour l'utilisateur.

2.4 Recherche multimodale

La recherche multimodale se réfère à la combinaison des différentes sources d'information comme un moyen d'améliorer les résultats de la recherche d'images.

Dans notre travail, nous proposons un système de recherche multimodale d'images capable d'intégrer plusieurs types d'informations pour la recherche d'images. Pour l'instant, nous nous concentrons sur le texte et le contenu visuel, les autres informations étant discutées en perspectives de notre travail.

En général, le texte et le contenu visuel sont souvent considérés comme des informations très complémentaires. Les caractéristiques visuelles peuvent être utilisées pour trouver des images similaires. L'information textuelle peut être utilisée pour trouver "sémantiquement" des images annotées. En se basant sur le fait que chacune de ces approches de recherche d'images (textuelle et visuelle) a des avantages et des inconvénients propres, comme nous l'avons présenté dans la partie 2.3.4, beaucoup de travaux essaient de les combiner afin d'améliorer la qualité de la recherche [Belkhatir 2005], [Lin 2007], [Lau 2007], [Tollari 2008], [Chang 2008]. La recherche multimodale peut combiner les forces des deux modalités et surmonter leurs limitations respectives. L'utilisation des images peut aider à lutter contre l'ambiguïté en recherche de texte. L'utilisation de texte peut aider à réduire le fossé sémantique en recherche d'images. Un des avantages de la recherche multimodale des images est que l'utilisateur peut exprimer sa requête soit par des mots *textuels* soit par des images exemples.

Il y a eu plusieurs approches d'intégration multimodale rapportées dans la littérature qui varient selon leur contexte et selon l'application. En règle générale, on peut classer ces approches en (a) Fusion précoce (*Early Fusion*) (b) Fusion tardive (*Late Fusion*) ou (c) Approche basée sur la transformation.

2.4.1 Fusion précoce

La fusion précoce fait référence aux méthodes qui intègrent les modalités avant de faire la recherche, par exemple en fusionnant les caractéristiques liées à l'indexation de manière préalable et en utilisant un algorithme de recherche qui fonctionne sur la nouvelle représentation. Cette approche combine toutes les caractéristiques et considère ainsi implicitement la mise en relation de caractéristiques différentes. L'approche la plus simple est de normaliser et de concaténer des vecteurs de représentation de caractéristiques de chaque modalité. Différentes méthodes de fusion précoce sont décrites ci-dessous.

2.4.1.1 Intégration dans un modèle probabiliste

Le texte et les images sont des données qui peuvent être facilement intégrées dans un modèle probabiliste. Les modèles probabilistes combinent non seulement les résultats des modalités texte-image en même temps, mais ils peuvent aussi associer le texte avec l'image. Les modèles probabilistes proposés dans la littérature ont de nombreuses variantes. Ils peuvent différer sur l'hypothèse de probabilité a priori, et sur les méthodes pour estimer les distributions de probabilités à la fois par le texte et les images. Certains modèles utilisent des caractéristiques de l'image segmentées en régions, alors que certains modèles utilisent des caractéristiques de l'image sans la segmentation d'images. Dans ce qui suit, nous résumons quelques modèles probabilistes multimodaux représentatifs en recherche d'images.

Les modèles probabilistes estiment la probabilité conjointe du texte et des images $P(T, I)$. La plupart des modèles supposent que les modalités par le texte et par l'image sont dépendantes. La probabilité conditionnelle des images connaissant le texte T, notée $P(I|T)$, et la probabilité conditionnelle du texte connaissant les images, notée $P(T|I)$, peuvent être obtenues à partir de la probabilité conjointe. Ainsi, les modèles peuvent être utilisés pour la recherche d'informations en modalités croisées (*cross-modal information retrieval*), pour la recherche avec une requête composée et pour l'annotation automatique d'images.

Par exemple, Monay et Gatica-Perez [Monay 2003] ont appliqué l'analyse sémantique probabiliste latente (PLSA) [Hofmann 1998] à l'annotation automatique d'images. Le texte et la caractéristique visuelle sont considérés comme des termes. La technique de la fusion précoce est utilisée pour représenter l'image par les termes. On suppose ici que chaque terme peut provenir d'un certain nombre de sujets latents et chaque image peut contenir plusieurs sujets.

Plus de détails sur l'association texte et image sont présentés dans le chapitre 4.

2.4.1.2 Indexation Sémantique Latente (*LSI - Latent Semantic Indexing*)

La technique LSI est utilisée pour combiner les caractéristiques visuelles et textuelles en un vecteur de caractéristiques unique [Westerveld 2000], [Theobald 2002], [Zhao 2002], [Agrawal 2006], [Pham 2007]. L'approche LSI est utilisée pour déterminer des groupes de mots-clés co-occurrents qu'une requête qui utilise un mot-clé donné peut alors utiliser pour récupérer des images ne contenant peut-être pas a priori ce mot-clé, mais contenant d'autres mots-clés à partir du même cluster de concept.

Dans [Zhao 2002], l'indexation sémantique latente (LSI) utilise conjointement des caractéristiques textuelles et visuelles pour extraire la structure sous-jacente sémantique des documents Web. L'amélioration de la performance de la recherche est attribuée à la synergie des deux modalités. Dans [Pham 2007], une image est considérée comme un document avec des données texte et des motifs visuels (qui peuvent être des régions ou des points d'intérêt). Les deux représentations sont projetées ensemble dans un espace latent dans lequel la recherche d'images similaires est effectuée. Dans [Lienhart 2009], les auteurs étendent cette approche qu'ils nomment *Probabilistic Latent Semantic Analysis* (PLSA) [Hörster 2008] à plusieurs couches (figure 2.2).

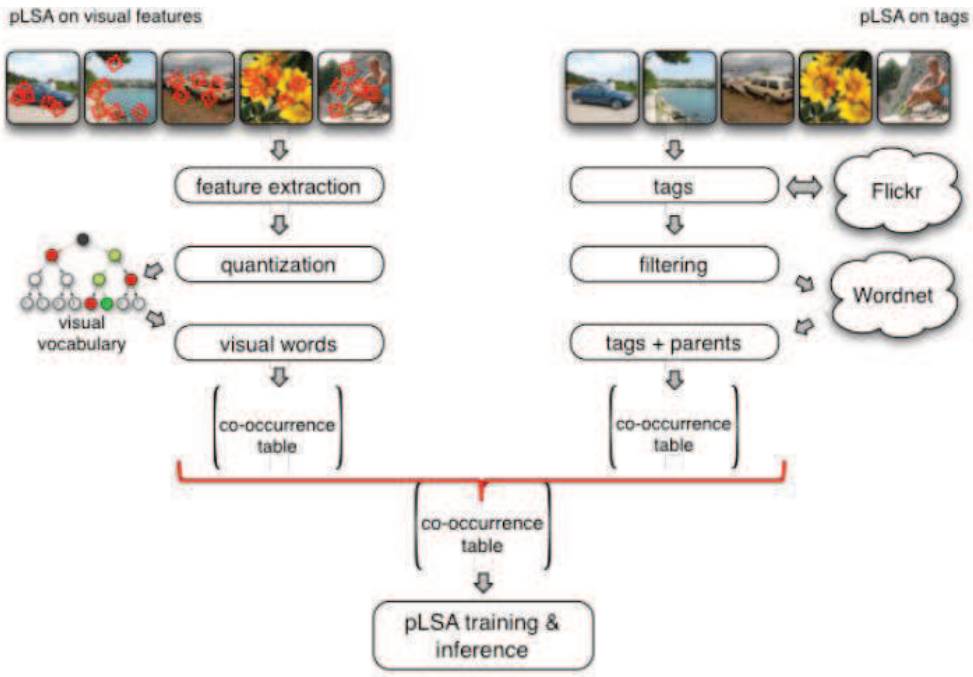

FIGURE 2.2 – Les multiples couches de PLSA pour combiner les caractéristiques visuelles et textuelles en un vecteur de caractéristiques unique.

2.4.1.3 Analyse des corrélations canoniques (*CCA - Canonical Correlation Analysis*)

L'approche *Canonical Correlation Analysis* (CCA) a également été proposée pour trouver des relations entre les motifs visuels et des textes descriptifs. Par exemple, Vinokourov et al. [Vinokourov 2003] ont appliqué un noyau CCA à une collection d'images du Web pour identifier les liens entre le visuel et les représentations textuelles afin de résoudre les requêtes en modalités croisées. Plus récemment, le problème de la fusion précoce a été reformulé comme un sous-espace d'apprentissage qui offre à la fois de la réduction de dimensionnalité et de la fusion des

caractéristiques [Fu 2008].

2.4.2 Fusion tardive

La fusion tardive se réfère aux méthodes qui préservent chaque modalité de données séparément [Lau 2007], [Escalante 2008], [Villena-Román 2008]. Pour une requête, deux algorithmes de recherche sont exécutés, puis les résultats sont combinés juste avant d'être livrés à l'utilisateur.

2.4.2.1 Raffinement des classements

Le raffinement des classements, ou en d'autres termes la fusion séquentielle des deux modalités, consiste à utiliser les informations textuelles ou bien visuelles pour faire la première recherche, et ensuite utiliser les caractéristiques de l'autre type d'informations pour exclure les images non pertinentes.

2.4.2.2 Combinaison des classements

Il s'agit d'une fusion naïve des deux techniques : c'est l'approche la plus simple, qui consiste à utiliser les deux techniques séparément (en parallèle) et à fusionner les résultats de recherche des deux [Lau 2007], [Escalante 2008], [Villena-Román 2008]. Il existe plusieurs algorithmes pour combiner les classements qui sont bien connus dans la communauté de recherche d'informations, comme la combinaison linéaire de classement, l'addition de tous les scores de similarité pour chaque document et les algorithmes de vote. Ces algorithmes ont été évalués en recherche d'images, en utilisant un moteur de recherche d'images par le texte et un moteur de recherche d'images par le contenu [Villena-Román 2008]. En outre, Lau et al. [Lau 2007] ont montré que des combinaisons linéaires des classements de texte et de contenu peuvent conduire à de meilleurs résultats que chaque système individuellement :

$$score\left(q_k, di\right) = \alpha score_V\left(q_k, di\right) + \left(1 - \alpha\right) score_T\left(q_k, di\right)$$

2.4.3 Approche basée sur les transformations

Cette approche est basée sur l'extraction des relations entre les images et le texte afin de les utiliser pour transformer les informations textuelles en informations visuelles et vice versa. Les requêtes peuvent être textuelles, visuelles ou les deux. Nous pouvons n'utiliser que la requête textuelle pour la recherche multimodale

(texte / contenu). Partant des requêtes à la fois textuelles et visuelles, les auteurs de [Chang 2006] transforment les requêtes visuelles en textuelles et obtiennent de nouvelles requêtes textuelles. Ils appliquent des techniques textuelles pour traiter les requêtes textuelles initiales et les nouvelles requêtes textuelles construites à partir des requêtes visuelles. Enfin, ils fusionnent les résultats obtenus. [Lin 2007] s'est basé sur l'extraction des relations entre les images et le texte afin de les utiliser pour transformer les informations textuelles en informations visuelles et vice versa. Ce modèle est utilisé pour la recherche d'images et aussi pour l'annotation d'images.

2.5 Annotation d'images

La plupart des bases de données d'images ne sont pas indexées par mots-clés. Pour effectuer la recherche d'images par mots-clés, l'indexation d'images à base de mots-clés est en général une étape préalable. La tâche d'affectation des mots-clés aux images est appelé "annotation" d'image. Cette tâche peut être effectuée par des utilisateurs/experts dans un environnement interactif leur permettant de saisir ces données, parfois suivant un thésaurus, parfois de façon totalement libre. Cette phase d'annotation étant particulièrement fastidieuse, de nombreuses bases d'images souffrent d'un manque significatif d'annotations. Par conséquent, on trouve dans la littérature de nombreux ouvrages sur l'annotation automatique ou semi-automatique d'images.

Les méthodes d'annotation automatique basées sur le contenu d'images peuvent être scindées en 2 catégories : les approches à base d'annotation utilisant les caractéristiques des régions [Barnard 2003], [Jeon 2003], [Wang 2004], [Monay 2003] et celles utilisant les caractéristiques globales [Chang 2003], [Feng 2004], [Barrat 2009]. L'annotation basée sur les caractéristiques des régions se fait en 3 étapes : la segmentation d'images, l'extraction de caractéristiques des régions et leur regroupement, et la génération du modèle d'annotation. On peut utiliser les algorithmes de segmentation ou tout simplement la partition de l'image en un ensemble fixe de régions (dans certains cas). Les régions sont ensuite souvent décrites en utilisant la couleur, la texture, les formes..., et elles sont regroupées en différents ensembles de régions similaires, appelées blobs. La dernière étape consiste à cartographier les régions et appliquer le modèle d'annotation. L'annotation globale quant à elle contient 2 étapes : l'extraction de caractéristiques d'images et la génération du modèle d'annotation. Un système d'annotation automatique peut être mis en œuvre en combinant une méthode d'extraction de caractéristiques et une technique d'apprentissage, souvent utilisée pour déterminer la probabilité du mot-clé attribué à l'image sur la base de ces caractéristiques. Par example, dans [Barrat 2009], l'auteur a proposé une modèle "Gaussian-Mixtures and Bernoulli mixture" pour la

propogation (extension) d'annotations dans des images faiblement annotées (où le nombre d'annotations est inférieur au maximum défini dans la vérité terrain) en se basant sur une distribution de probabilité conjointe sur un dictionnaire de mots-clés et les caracteristiques visuelles extraites de collection d'images (en niveau de gris et couleur).

L'annotation semi-automatique est souvent proposée pour améliorer la faible précision de l'annotation automatique [Yang 2005], [Wenyin 2001]. Elle contient souvent 2 parties. Tout d'abord, les images sont annotées par une méthode d'annotation automatique, puis les annotations sont mises à jour en utilisant les techniques de retour de pertinence lors du processus de recherche. Lors du processus de recherche d'images par le contenu et / ou par mots-clés, l'utilisateur fournit les informations concernant les images qu'il juge pertinentes / non pertinentes. L'annotation des images est alors mise à jour en utilisant cette information. Cependant, cette stratégie d'annotation repose grandement sur les performances de recherche d'images par le contenu (CBIR), ce qui conduit parfois à des annotations erronées et des images non considérées en raison de l'imperfection du processus CBIR.

L'état de l'art de l'annotation d'images est abordé dans la section 4.1 où la problématique plus générale de l'association du texte et du contenu est présentée.

2.6 Systèmes intégrés

Dans cette section, nous présentons des systèmes intégrés qui utilisent différentes sources d'informations pour la recherche d'images. Les systèmes présentés sont relativement récents et proviennent d'équipes de recherche dans le domaine de la vision par ordinateur.

2.6.1 SCENIQUE

SCENIQUE (*Semantic and ContENt-based Image QUErying*) [Bartolini 2010] est un système de recherche d'images multimodales (mixte image/texte) qui fournit à l'utilisateur deux composantes principales : (1) un annotateur image, qui est capable de prédire les mots-clés pour les nouvelles images, et (2) une recherche intégrée d'images qui permet à l'utilisateur la recherche d'images en utilisant des requêtes visuelles et des requêtes textuelles. La recherche d'images dans SCENIQUE peut prendre trois formes de base : requête visuelle, requête textuelle et requête mixte texte/image. SCENIQUE est basé sur un framework "multiple-structure" qui se compose d'un ensemble d'objets avec schémas qui spécifient la classification d'objets selon plusieurs critères distincts. Les mots-clés sont organisés comme des

dimensions qui prennent la forme d'arbres. Lorsque l'interrogation par une requête mixte texte/image est réalisée, le système retourne des images dans l'intersection de la recherche par le contenu et de la recherche par le texte, suivi par les résultats de la requête textuelle et enfin des résultats de la requête visuelle. Les auteurs ont démontré que la précision obtenue par une requête mixte est meilleure que par interrogation à base de requête textuelle unique ou à base d'une seule requête visuelle.

2.6.2 MAMI

MAMI (*Multimodal Automatic Mobile Indexing*) [Anguera 2008] est un prototype de téléphone portable qui permet aux utilisateurs d'annoter et de rechercher des photos numériques sur leur smart-phones via une commande vocale. MAMI est implémenté comme une application mobile qui fonctionne en temps réel sur le téléphone. Les utilisateurs peuvent ajouter des annotations audios (parole) au moment de la saisie de photos ou bien à une date ultérieure. Des métadonnées supplémentaires sont également stockées avec les photos, comme la localisation, l'identification de l'utilisateur, la date ou encore l'heure de capture des images. Les utilisateurs peuvent rechercher des photos dans leur dossier personnel par le biais de la parole.

2.6.3 ALIPR

Le système ALIPR [Li 2008] présenté depuis 2006 est un service en temps réel d'annotation et de recherche d'images. Selon les auteurs, ce travail est le premier à atteindre des performances temps réel avec un niveau de précision utile dans certaines applications réelles. Il est également la première tentative d'évaluation de la performance à grande échelle d'un système d'annotation d'images. Ce système a été soumis à une évaluation rigoureuse, incluant des tests approfondis en utilisant des images issues du Web et totalement indépendantes des images d'apprentissage. La performance du système a également été évaluée en fonction de l'entrée de milliers d'utilisateurs en ligne.

2.6.4 SnapToTell

SnapToTell [Chevallet 2005], [Joo-Hwee Lim 2004], [Chevallet 2007] est un système qui fournit de l'information multimédia aux touristes en se basant sur les images prises par des téléphones portables et sur l'information de localisation.

En utilisant le service de SnapToTell, les touristes peuvent avoir des informations concernant un endroit via des photos de cet endroit.

L'idée de ce système est qu'il tente d'abord d'utiliser l'information de localisation géographique du téléphone portable de l'utilisateur pour réduire le nombre d'images qui doivent être examinées. Ensuite, il utilise les informations de contenu pour trouver l'image la plus semblable dans la base de données images réduite pour trouver le site qui intéresse l'utilisateur. La base de données de ce système collecte les images de tous les sites célèbres d'un pays (ici Singapour), chaque site correspondant à plusieurs images prises à plusieurs distances différentes et de plusieurs points de vue différents (figure 2.3).

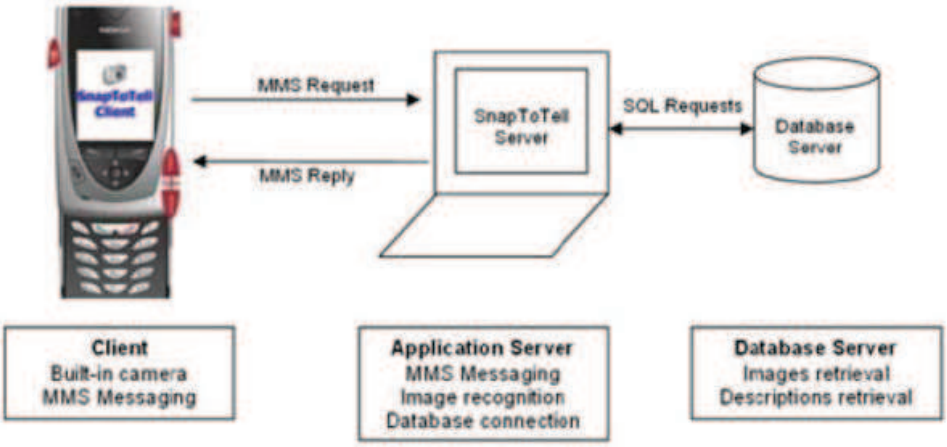

FIGURE 2.3 – Architecture client/serveur du système SnapToTell.

2.7 Systèmes commerciaux

Les trois systèmes commerciaux publics les plus connus en recherche d'images sont Bing Images, Google Images et eBay Images.

Bing Images est un moteur de recherche d'images développé par la société Microsoft. Il a été rendu public le 1er juin 2009. Bing Images peut rechercher des images par le contenu et le texte. La reconnaissance de visages est aussi implémentée (figure 2.4).

Google Images est un moteur de recherche d'images développé par la société Google. Comme Bing Images, il peut servir à la recherche d'images par le texte, par le contenu et la reconnaissance de visages (figure 2.5).

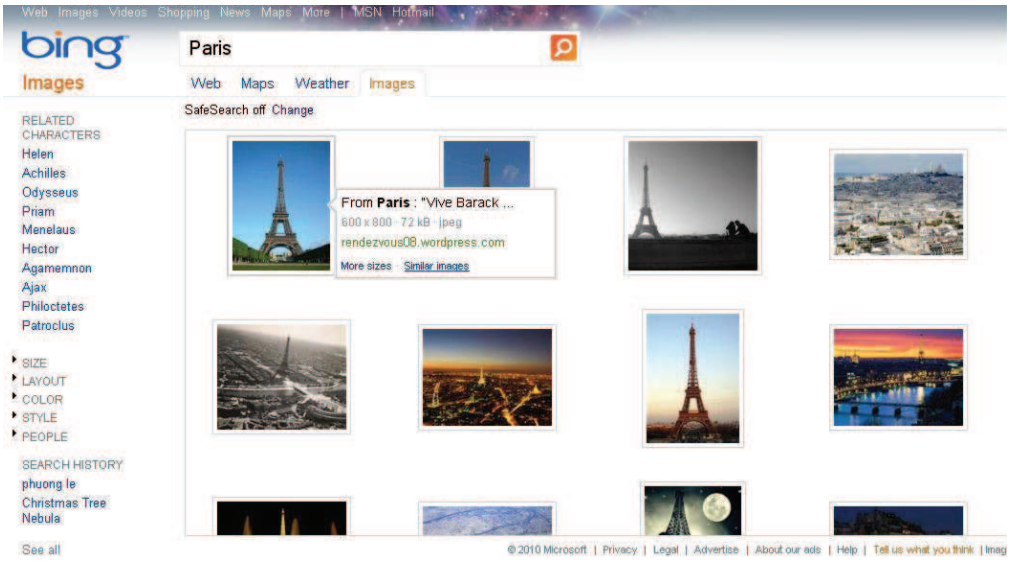

FIGURE 2.4 – Le moteur de recherche d'images Bing Images.

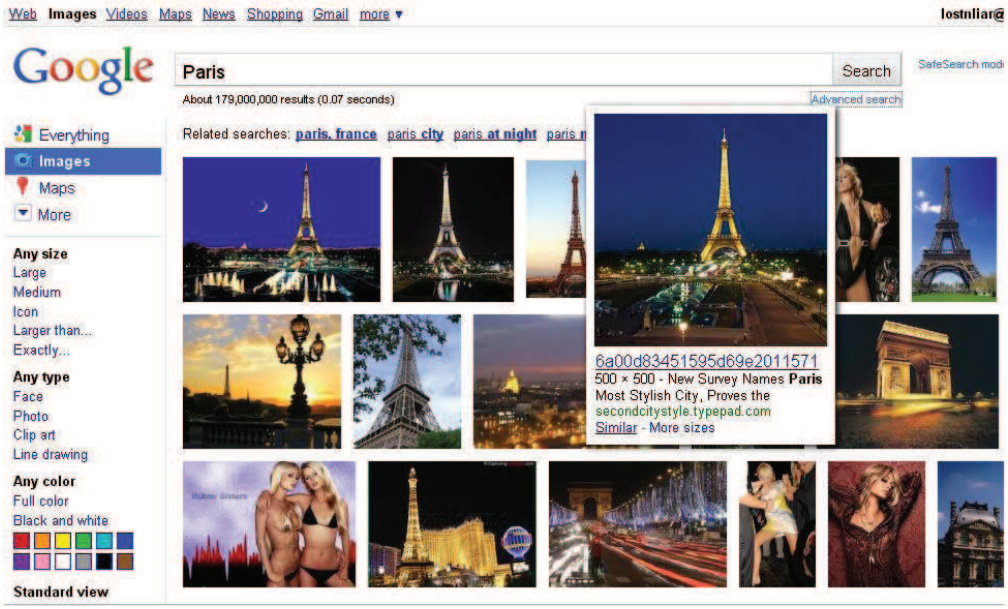

FIGURE 2.5 – Le moteur de recherche d'images Google Images.

eBay Images est un moteur de la recherche d'images qui est utilisé particulièrement pour la mode. L'utilisateur peut trouver des articles à acheter (figure 2.6).

Il existe d'autres systèmes de recherche d'images comme : Gazopa Images, Imense Images, Incogna Images, Pixsta, Shopachu, TinEye...Nous disposons de très peu d'informations concernant les mécanismes internes de ces systèmes.

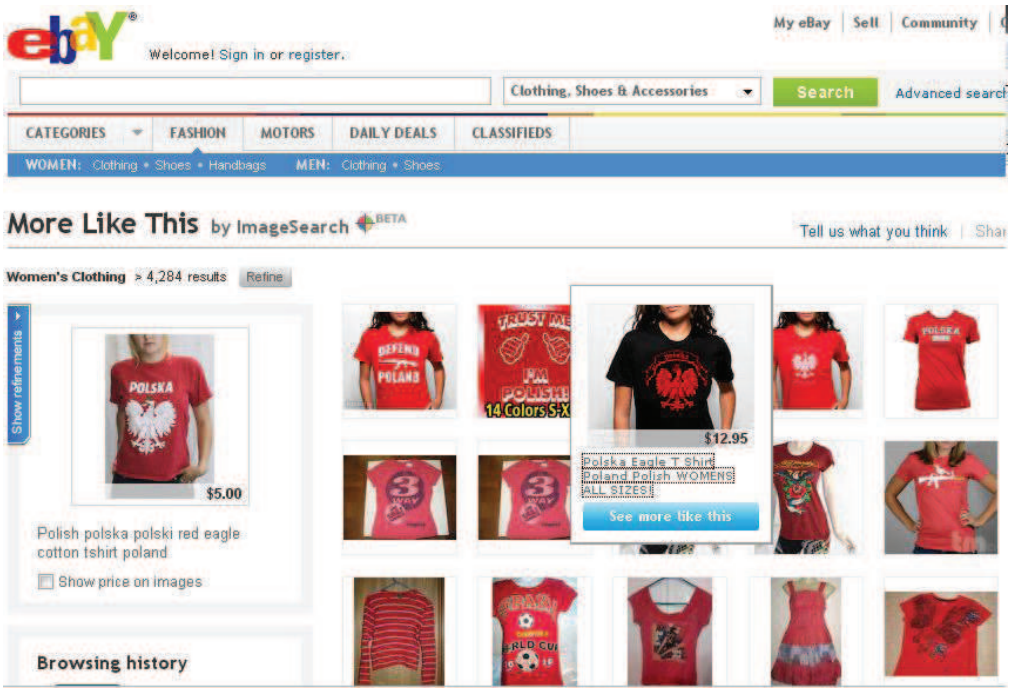

FIGURE 2.6 – Le moteur de recherche d'images eBay Images.

2.8 Discussion

L'indexation est une phase très importante dans un système de recherche d'images. Une bonne indexation donne une bonne performance pour le système et vice versa. Les inconvénients et les avantages de l'indexation et de la recherche par le texte et par le contenu sont donnés dans la section 2.3.4. Bien que la recherche d'images par le contenu ait des avantages clairement exprimés, une indexation par le contenu visuel n'est généralement pas suffisante et les systèmes de recherche d'images par le contenu ont souvent une performance limitée, du fait du faible pouvoir descriptif des primitives visuelles. Cela est principalement dû à deux facteurs : l'impossibilité d'exprimer pleinement toute l'intention de l'utilisateur en utilisant une requête visuelle simple pour la recherche d'une part, et la différence entre l'interprétation de l'utilisateur et la description bas niveau du contenu d'une image, souvent appelé fossé sémantique, d'autre part. La recherche par le contenu ne procurant pas d'excellents résultats, la recherche par mots *textuels* est utilisée par la quasi-totalité des systèmes commerciaux de recherche, comme Yahoo ! ou Google. L'ensemble de ces considérations conduisent au développement de la recherche d'images combinée par le contenu et le texte (mixte contenu et texte) étudiée intensivement au cours des dernières années. De façon complémentaire, d'autres types d'informations, comme la localisation et le temps, sont également utilisées pour l'indexation multimodale et la recherche multimodale d'images.

Dans les systèmes de recherche multimodale d'images actuels, des informations

externes des images (ou des connaissances dans notre point de vue) sont nécessaires a priori. Dans la plupart des cas, les connaissances ne sont pas complètes (par exemple la base d'images est partiellement annotée) et il faut un apprentissage hors ligne au début pour que le système soit utilisable. De plus, la base d'images est souvent fixée, ce qui ne correspond pas à une caractéristique "réaliste" d'un point de vue opérationnel, car les volumes d'images varient très souvent dans les applications concrètes. Dans la base d'images, on se retrouve ainsi avec certaines images anciennes et d'autres nouvelles, les premières déjà indexées et possiblement annotées et les secondes en attente d'indexation ou d'annotation. Comme la base n'est pas annotée uniformément, cela rend l'accès difficile par le biais de requêtes textuelles.

Dans le cadre de cette thèse, notre contribution est un système de recherche multimodale d'images qui ne demande pas de connaissances au début de la vie du système. Ce système permet d'apprendre et de raffiner itérativement l'annotation et la connaissance des images. Nous ne séparons pas les phases d'apprentissage et de recherche, et l'utilisateur peut faire des requêtes dès la mise en route du système, tout en laissant le système apprendre au fur et à mesure de son utilisation. Ce système est utilisable avec une base d'images variée où il existe toujours des parties qui ne sont pas annotées. Nous présentons notre système proposé dans la section suivante.

2.9 Système de recherche d'images proposé

Notre système s'inspire d'une application dans le domaine des catastrophes naturelles au sein de laquelle un système CBIR vise à fournir des informations stratégiques pour les experts en secours à partir d'images de la catastrophe. Nous faisons l'hypothèse que la base de données images évolue au cours du temps, avec l'arrivée dynamique de nouvelles images prises par les différents systèmes d'acquisition. L'information captée sur les lieux du désastre (par des caméras de surveillance, par des téléphones portables, par imagerie aérienne ou autres) entre en temps réel dans le système qui doit fournir des aides à la décision pour gérer les situations post-catastrophe. Cette hypothèse, qui a servi de point de départ à notre réflexion, est également vraie pour de nombreuses applications de surveillance ou de détection. Dans la plupart de ces applications, le volume d'images n'est pas fixé et les images peuvent être identifiées selon leur heure d'arrivée. Dans la base de données, il y a toujours des images anciennes et des images nouvelles, des images déjà traitées et indexées et des images en attente d'être transformées ou entièrement indexées. Les informations sur les images sont également en constante évolution. Les images contiennent non seulement de l'information visuelle, mais aussi des informations géographiques (GPS par exemple), le cachet sur l'heure d'arrivée dans le système

et quelques informations complémentaires éventuellement saisies par des experts. Partant de ces hypothèses, à un instant t, il existe globalement trois types d'images dans notre système :

1. des images sans annotations
2. des images avec informations complémentaires émanant des experts (annotations manuelles)
3. des images avec des informations ajoutées automatiquement (annotations propagées)

Dans le contexte de notre projet, nous considérons que les annotations manuelles saisies par les experts sont plus fiables (valeur de confiance plus élevée) que celles propagées automatiquement, si nous faisons une distinction claire entre ces deux types d'annotations. Cependant, les annotations propagées automatiquement peuvent en outre être manuellement validées par des experts (à travers leurs interactions avec le système) et voir ainsi leur statut changer pour devenir ainsi des annotations manuelles.

Le système proposé comporte deux composantes principales, la première correspondant à la couche visuelle (figure 2.7, étapes 1-2-3, en rouge) qui comprend la recherche d'images et le retour de pertinence, et la seconde correspondant à la couche d'apprentissage (figure 2.7, étapes 4-5, en bleu) comprenant l'apprentissage des représentations visuelles des mots-clés (KVR - ce concept est expliqué plus loin) et la propagation d'annotations.

Au début de l'exécution du système (figure 2.7, étapes 1 et 2), nous faisons l'hypothèse qu'il n'y a que des images sans aucune connaissance (figure 2.7, couche A dans la base), que des experts peuvent annoter manuellement en profitant du retour de pertinence via l'interface du système (figure 2.7, couche B dans la base). La formation de requêtes et l'interaction dans le système (incluant les mécanismes de retour de pertinence) sont présentées en détails dans le chapitre 3.

Le système apprend ensuite les associations entre les mots-clés et les représentations visuelles, appelées représentations visuelles de mot-clé (KVR - *Keyword Visual Representation*), grâce à ces annotations (figure 2.7, étape 4). Nous disposons alors de connaissance au niveau "image", correspondant aux KVR dans la base de données (figure 2.7, couche C dans la base). Lorsqu'il y a encore des images sans annotations, notre système propose un mécanisme de propagation d'annotations, qui est appliqué pour ces images en utilisant les KVR (figure 2.7, étape 5). Il y a dès lors 2 types d'images annotées dans le système : les images avec annotations manuelles (figure 2.7, couche B dans la base) et les images avec annotations automatiques (figure 2.7, couche D dans la base). Outre les images annotées, conformément à notre hypothèse de fonctionnement concernant l'arrivée

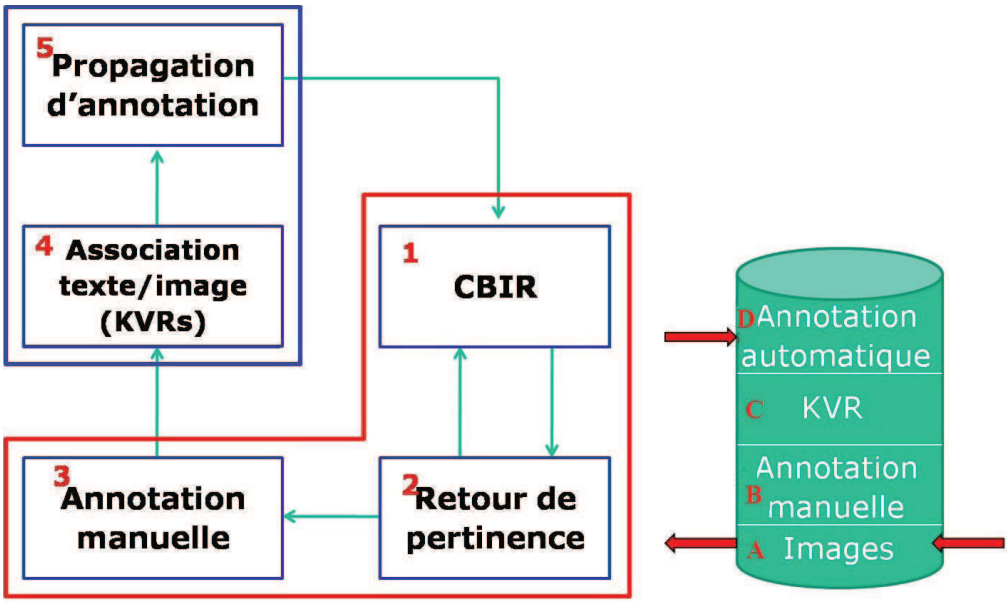

FIGURE 2.7 – Vue d'ensemble du système de recherche d'images proposé. Au début de la vie du système (étape 1), la base de données est sans connaissances et ne contient que des images (couche A de la base). Des images sont envoyées au système en temps réel (flèche rouge à droite entrant dans la base). Ensuite viennent le retour de pertinence (étapes 2 et 3), l'apprentissage de KVR (étape 4) et la propagation d'annotations (étape 5). Les connaissances du système sont créées et évoluent dans le temps (annotation manuelle, ajout de KVR, annotation propagée). Les flèches rouges entrantes vers la base signifie que des connaissances/images sont (r)envoyées au système, la flèche rouge sortante de la base signifie que des connaissances/images sont utilisées pour la recherche d'images ou pour l'apprentissage de connaissances.

en continu de nouvelles images, il peut encore y avoir de nouvelles images arrivées sans annotations pendant les phase précédentes. L'apprentissage de l'association du texte et du contenu et l'annotation d'images sont présentées dans le chapitre 4.

Considérant l'ensemble des étapes précédentes, il existe à ce stade les connaissances suivantes au sein du système : celles issues de l'annotation manuelle (figure 2.7, couche B), les KVR (couche numéro 4) et celles issues de l'annotation propagée (figure 2.7, couche D). Lorsque l'information textuelle est disponible pour certaines images, le système peut dès lors effectuer des recherches basées respectivement sur le contenu et les mots-clés (figure 2.8, étape 1). Il existe dès lors 3 types de requête pour les utilisateurs :

1. Requête par des mots-clés
2. Requête par des images
3. Requête combinant des images exemples avec des mots clés.

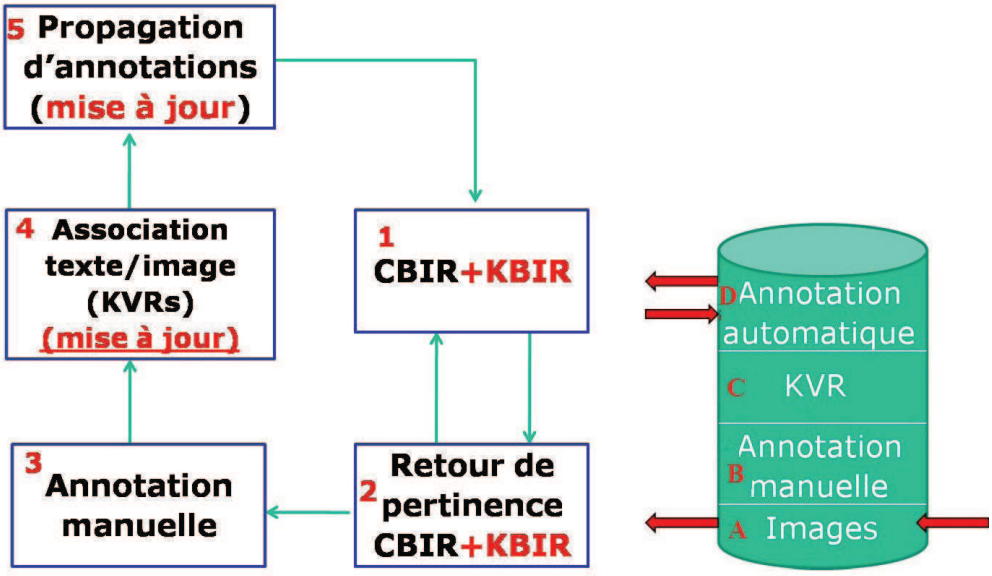

FIGURE 2.8 – Vue d'ensemble du système de recherche d'images proposé lorsque des connaissances sont disponibles. Les flèches rouges entrantes dans la base signifient que des connaissances/images sont renvoyées au système et les flèches rouges sortantes de la base signifient que des connaissances/images sont utilisées pour la recherche d'images ou pour l'apprentissage de connaissances.

La recherche multimodale d'images Les résultats de la recherche sont la combinaison de la recherche visuelle à partir d'exemples d'images et la recherche textuelle par mots-clés. La similitude visuelle V_d entre une requête utilisant l'image Q et une image I_i est le produit scalaire entre les vecteurs correspondants :

$$Vd(\bar{I}_i, \bar{Q}) = \sum_j w_{i,j} \times w_{q,j} \tag{8}$$

La similitude textuelle S_d entre une requête utilisant l'image Q et une image I_i est calculée comme suit :

$$Sd(I_i, q) = \frac{|K_{i,q}|}{|K_q|} \tag{9}$$

où $|K_{i,q}|$ est le nombre de mots-clés communs entre l'image Q et l'image I_i et $|K_q|$ est le nombre de mots-clés de l'image Q.

La similarité finale $S_{final} = r * S_d + (1 - r) * V_d$ où r est la pondération des deux similitudes. Dans notre système, différentes valeurs de r sont utilisées pour les 3 types d'images, comme suit :

1. Images avec annotations manuelles : $r = 0,8$
2. Images avec annotations propagées : $r = 0,3$
3. Images sans annotations : $r = 0,0$

Ce choix repose sur notre hypothèse que les annotations manuelles sont plus fiables que les annotations automatiquement propagées.

2.10 Conclusion

Dans ce chapitre, nous avons présenté un état de l'art concernant l'indexation d'images par des signatures visuelles, textuelles et aussi par des informations externes comme la localisation et le temps. La recherche unimodale ou multimodale d'images qui utilisent un ou de multiples types d'informations pour l'indexation est aussi présentée dans ce chapitre.

Le système que nous proposons dans ce travail offre la capacité d'indexer des images sur la base d'informations différentes. Nous avons combiné le contenu visuel et le texte pour l'indexation et la recherche d'images. Une généralisation vers d'autres types d'informations comme la localisation et le temps est montré en perspectives de ce travail. Pour l'indexation par le contenu, nous utilisons la signature locale SIFT de [Lowe 2004] et le modèle Sacs de Mots de l'état de l'art [Sivic 2008] est utilisé pour la recherche. Nous présentons en détails ce modèle dans le chapitre 4. Pour ce dispositif, nous nous appuierons sur la signature SIFT, du fait de sa performance reconnue dans de nombreux ouvrages de la littérature [Mikolajczyk 2005b]. Pour l'indexation par le texte, nous proposons une indexation / annotation semi-automatique d'images basée sur l'association entre texte et image. Cette association est apprise incrémentalement (et interactivement) via des interactions utilisateur (par des techniques adaptatives de type retour de pertinence, visant un plus long terme - présentées dans le chapitre 3) en utilisant un nouveau modèle de Sacs de KVR (*Keyword Visual Representation*) qui est présenté dans le chapitre 4. Pour la recherche mixte par le contenu et le texte, nous utilisons dans un premier temps une approche basée sur les transformations (section 2.4.3) pour créer des associations entre texte et images, et ensuite une fusion tardive qui est utilisée pour combiner des classements (section 2.4.2.2).

Dans ce chapitre, le lecteur aura pu constater qu'il manque de l'état de l'art sur l'interaction dans les systèmes de recherche d'images. Cet aspect est dans le chapitre 3. Nous y présentons aussi notre contribution sur l'interaction en recherche d'images.

Chapitre 3

Interaction

Dans le chapitre précédent, nous avons évoqué l'indexation et la recherche unimodale/multimodale d'images ainsi que notre proposition de système de recherche d'images. Dans ce chapitre, nous présentons d'abord l'état de l'art de l'interaction dans les systèmes de recherche d'images : l'exploration, le retour de pertinence à court terme et le retour de pertinence à long terme. Ensuite, nous présentons en détails notre contribution relative à la technique d'interaction de notre système. Cette contribution correspond à la couche visuelle (figure 3.1, étapes 1, 2 et 3, en rouge), en particulier le retour de pertinence à court terme.

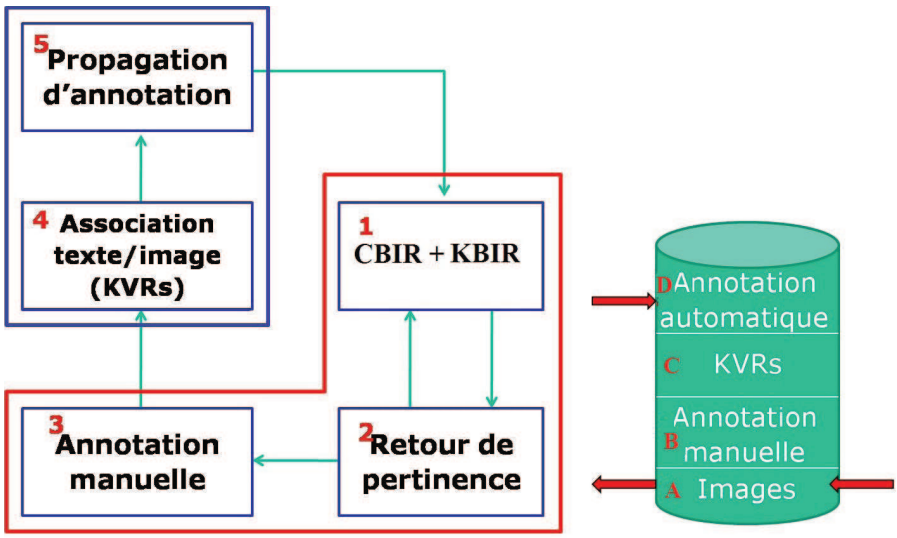

FIGURE 3.1 – Système de recherche d'images avec le retour de pertinence à court terme.

3.1 Introduction

L'objectif principal des systèmes de recherche d'images étant de fournir des outils efficaces de navigation et de recherche pour un utilisateur, il est donc fondamental que la conception des systèmes soit centrée sur l'humain/l'utilisateur. Nous considérons que la compréhension des intentions d'un utilisateur joue un rôle clé dans un système de recherche d'images, ainsi que l'identification de la raison de son interaction. Il peut aussi agir comme un guide pour la conception du système. Pour définir les modes possibles d'interaction de façon pertinente, il nous faut analyser le comportement et les besoins de l'utilisateur [Lew 2006], [Jaimes 2006]. Les différents types d'utilisateurs sont souvent décrits de la façon suivante [Datta 2008] :

- Navigateur : Cet utilisateur n'a pas d'informations bien définies en premier lieu et souhaite simplement explorer la collection d'images. Il a besoin d'une méthode rapide et cohérente pour explorer la base.
- Surfeur : Cet utilisateur a un léger objectif de prospection au début. Au fil de son exploration de la base d'images des images, il va trouver ce qu'il veut de temps en temps.
- Chercheur : Il s'agit d'un utilisateur qui est très clair sur ce qu'il cherche dans le système. La session d'un chercheur est généralement courte, avec des recherches cohérentes conduisant à un résultat final.

Le Navigateur parcourt des images sans objectif précis en tête. C'est une session avec des recherches indépendantes qui traverse plusieurs catégories diversifiées. Son interaction avec le système est seulement l'exploration dans la base d'images. Le Surfeur veut rechercher des images dans une catégorie quelconque mais sans savoir précisément exprimer sa recherche et par conséquent sans savoir à quoi ça ressemble. Le Chercheur est en mesure d'exprimer la formulation d'une requête, et donne au système la tâche de trouver des images pertinentes par rapport à sa requête. Dans ce contexte, son interaction est la spécification des requêtes. Le retour de pertinence est utilisé par le Surfeur et quelquefois par le Chercheur pour raffiner le résultat dans une session d'interrogation ou pour améliorer la précision/performance du système. L'interaction d'un utilisateur dans un retour de pertinence est de juger la pertinence/non pertinence des images.

En résumé, nous considérons que l'interaction dans les systèmes CBIR (*Content-Based Image Retrieval*) se compose des éléments suivants : la spécification des requêtes, l'exploration dans une masse d'images - la façon de présenter les images et de guider l'utilisateur d'une manière significative, et le retour de pertinence - l'outil pour interagir avec le système afin d'affiner les requêtes ou de faire progresser le système.

Dans notre travail nous faisons un lien entre le retour de pertinence, l'annotation d'images et l'apprentissage interactif. Nous définissons le retour de pertinence comme l'apprentissage interactif à court terme et l'annotation d'images comme l'apprentissage interactif à long terme.

L'apprentissage interactif à court terme est une technique interactive utilisée pour améliorer la précision des systèmes de recherche d'informations pour un utilisateur spécifique. Le mot « court terme » correspond à l'adaptation du processus de la recherche pour un utilisateur spécifique et une requête spécifique. Pendant l'interaction, le système peut mémoriser la requête et ses résultats, mais une fois qu'il a terminé, le système nettoie sa mémoire et l'utilisateur suivant recommence à partir de zéro. Dans cette thèse, nous référons à l'apprentissage interactif à court terme ou au retour de pertinence à court terme comme étant la même chose. Nous présentons le retour de pertinence à court terme dans la partie suivante.

L'apprentissage interactif à long terme ou l'annotation interactive d'images est une technique interactive utilisée pour perfectionner la connaissance pour un groupe d'utilisateurs ou pour tous les utilisateurs. Par exemple, l'utilisateur peut trouver quelque chose d'intéressant dans une image et il souhaite que le système s'en rappelle et que cela puisse être ré-utilisé par lui-même une prochaine fois ou par les utilisateurs suivants. Dans cette thèse, nous référons à l'apprentissage interactif à long terme comme étant le retour de pertinence à long terme. Le retour de pertinence à long terme n'est pas presenté dans ce chapitre mais dans le chapitre suivant, soit le chapitre 4.

3.2 État de l'art

Dans cet état de l'art pour l'interaction dans les systèmes CBIR, nous présentons d'abord la spécification des requêtes. Ensuite, nous présentons les techniques d'exploration. Enfin nous évoquons le retour de pertinence à court terme.

3.2.1 La spécification des requêtes

La notion de spécification de requêtes fait référence aux dispositifs offerts à l'utilisateur pour l'expression de ses requêtes. Plusieurs mécanismes d'interrogation ont été créés pour aider les utilisateurs à définir leurs besoins d'informations [Aslandogan 1999], [Torres 2006], [Datta 2008]. Nous présentons ici différentes stratégies possibles de recherche qui peuvent être employées dans les systèmes CBIR :

- Interrogation par exemples d'images : L'utilisateur fournit des images exemples au système. Les images renvoyées peuvent être jugées positives seulement ou parfois à la fois positives et négatives. Pour ces dispositifs, le système de recherche des images pertinentes se base généralement sur des caractéristiques visuelles.
- Interrogation par mots-clés : L'utilisateur fournit des mots-clés. Le système recherche des images pertinentes en se basant sur des signatures textuelles (ou annotations) ou le système transforme les mots-clés en des représentations visuelles et recherche des images pertinentes en se basant sur des caractéristiques visuelles. Ce type d'interrogation demande une transformation entre images et texte. On peut donc ainsi trouver différents types de transformation : texte à images ou images en texte.
- Interrogation par combinaison de caractéristiques : L'utilisateur sélectionne différentes caractéristiques. Par exemple : "50% rouge et 50% bleu". Ce type d'interrogation est similaire à l'interrogation par l'exemple si nous considérons une image comme un ensemble de caractéristiques visuelles.
- Interrogation par région(s) localisée(s) : L'utilisateur fournit des régions d'images en dessinant ou en cliquant sur une ou plusieurs images. Ce type d'interrogation demande plus d'interaction à l'utilisateur, mais lui permet normalement de trouver rapidement les images qu'il recherche.
- Interrogation par esquisse : Il s'agit d'une requête dessinée à la main ou issue d'une image générée par ordinateur. Ce type d'interrogation est ici beaucoup plus chronophage pour l'utilisateur parce qu'il doit dessiner son intention en images.
- Interrogation combinée : La combinaison de différentes techniques présentées ci-dessus.

En résumé, la spécification de requêtes permet à l'utilisateur de formuler son intérêt par un modèle de données qui joue le rôle d'une requête. Ce type d'interaction trouve différentes représentations puisqu'il peut s'agir d'entrer de simple mots-clés, mais également de fournir des exemples d'images ou encore, plus compliqué, de dessiner un graphique ou de combiner différentes méthodes.

3.2.2 Exploration

La plupart des systèmes de recherche d'images (CBIR) ont traditionnellement été étudiées dans un cadre qui assume la disponibilité de la requête. Néanmoins, l'hypothèse que les utilisateurs sont toujours en mesure de formuler une requête appropriée peut parfois être discutable. Les problématiques d'exploration et de navigation dans une base d'images restent donc importantes, même s'il y a peu de soutien pour la recherche exploratoire de grandes collections d'images. Pour

défendre la nécessité de l'exploration, [Heesch 2008] développe l'argumentaire suivant :

- requête mentale : L'interrogation par exemples d'images est insuffisante lorsque les images requêtes ne sont pas à portée de main. En effet, les utilisateurs peuvent avoir besoin d'accéder d'abord à une collection pour identifier des images de requête. L'exploration peut donc dans certains cas être accomplie avec seulement une représentation mentale de la requête sous forme d'un guide, et la requête peut-être aussi simple ou aussi complexe que nous le souhaitons.
- requête vague : Dans certains cas, les utilisateurs n'ont pas d'informations concrètes et précises pour l'expression de leur requête. L'information peut d'abord être très vague et se développer au fur et à mesure de l'exploration avec la base d'images. Obtenir un aperçu d'une collection et être capable de naviguer rapidement entre les différents types d'images devient alors crucial. Selon la façon dont la collection est structurée, la navigation peut apporter un soutien beaucoup plus important pour la recherche non orientée et aider les utilisateurs à développer leurs besoins en informations.
- possibilité d'exploiter les capacités cognitives des utilisateurs : Les humains sont beaucoup plus à même de reconnaître si quelque chose est pertinent par rapport à leur intention, plutôt que pour décrire ce qui peut être considéré comme pertinent.

Des travaux comme [Heesch 2008] présentent des modèles qui permettent la navigation/l'exploration de manière interactive de grandes collections d'images. Les modèles de navigation pour la recherche d'images ont tendance à organiser la collection dans une structure qui peut être parcourue de manière interactive et efficace. Considérant ces différentes problématiques, l'une des plus grandes difficultés d'une approche de navigation est d'identifier les structures qui sont propices à la recherche efficace, dans le sens qu'elles favorisent une navigation/exploration rapide, et fournissent une organisation significative pour choisir un chemin de navigation et pour permettre aux utilisateurs de se positionner dans une zone d'intérêt [Cox 1992]. Intuitivement, nous voudrions que les objets soient à proximité les uns les autres et facilement accessibles s'ils sont similaires. Dans le contexte général de la littérature sur ces sujets, nous pouvons distinguer des modèles de navigation/exploration selon 2 classes : (i) les structures statiques hiérarchiques, (ii) les réseaux statiques. Nous présentons de façon très synthétique ces deux modèles parce que ceux-ci ne touchent pas beaucoup notre travail.

3.2.2.1 Structures statiques hiérarchiques

Les structures hiérarchiques ont été étudiées depuis de nombreuses années afin de proposer des alternatives à la recherche par plus proche voisin, du fait de sa complexité algorithmique [Chen 2000]. L'idée générale ici est de trouver les voisins les plus proches en parcourant (depuis la racine) un arbre de centroïdes de clusters hiérarchiquement organisé. À chaque étape, une comparaison est effectuée entre la représentation de la requête et la représentation stockée au niveau du nœud donné. Pour utiliser les structures hiérarchiques pour la navigation, c'est l'utilisateur qui doit comparer l'image requête avec les centroïdes de clusters à un niveau particulier de la hiérarchie et décider quel chemin suivre pour continuer.

Les structures hiérarchiques sont construites par des techniques de regroupement (clustering). Certaines méthodes de regroupement simples sont habituellement utilisées pour construire des hiérarchies : regroupement par divisions, k-moyennes de haut en bas, méthodes ascendantes.

3.2.2.2 Réseaux statiques

Les réseaux statiques sont de la même forme que les structures associatives dont les concepts sont reliés entre eux sur la base de relations sémantiques. Dans ces structures décentralisées, la nature hiérarchique de nombreuses données est implicitement intégrée dans les poids associés à des paires de nœuds connectés. En règle générale, les réseaux sont construits sur la base de similitudes entre les objets. Les réseaux les plus populaires sont : les réseaux des plus proches voisins, les graphes seuillés [Jin 2003a], les réseaux de recherche de chemins [Schvaneveldt 1990] et les réseaux NN^k [Heesch 2006].

3.2.3 Retour de pertinence à court terme

Le retour de pertinence à court terme (RF - *Relevance Feedback*) est une technique interactive généralement utilisée pour améliorer la précision des systèmes de recherche d'information, notamment dans les systèmes CBIR [Ortega Binderberger 2004], [Kim 2005], [Ishikawa 1998]. Le mot "court terme" correspond à l'adaptation du processus de recherche pour un utilisateur spécifique et une requête spécifique, qui ne reste pas toujours valable pour d'autres utilisateurs ou pour d'autres requêtes. Sur la base d'une requête, l'utilisateur interagit avec les résultats de manière à les modifier en demandant au système de changer le poids des paramètres ou de modifier la requête elle-même pour adapter le résultat selon son intention. Pendant ce temps d'interaction, le système peut mémoriser ses résultats, mais une fois

qu'il a terminé, le système nettoie sa mémoire et l'utilisateur suivant recommence à partir de zéro. Il existe d'autres types de retour de pertinence ou le système peut faire du "profilage" de l'utilisateur et peut mémoriser, pour chaque utilisateur, les interactions qu'il a eues avec le système, pour les ré-utiliser ultérieurement. Ces types de retour de pertinence à "long terme" sont présents dans les systèmes de recommandation ou les systèmes d'annotation d'images, dont nous parlerons dans le chapitre 4

Du fait des difficultés pour exprimer pleinement les intentions de l'utilisateur en utilisant une requête simple et également à cause du problème du fossé sémantique, de nombreux travaux ont mis l'accent sur la technique de retour de pertinence à court terme. L'idée fondamentale derrière cette notion de retour de pertinence est de montrer à l'utilisateur une liste d'images candidates, de lui demander de décider si chaque image est pertinente ou non, et de modifier l'espace des paramètres, l'espace sémantique, l'espace de représentation ou encore l'espace de classification pour tenir compte des exemples pertinents et non pertinents. Diverses techniques de retour de pertinence ont été proposées dans la littérature : optimisation de distances métriques [Ishikawa 1998], modification de la requête [Ortega Binderberger 2004], apprentissage de classificateur [Tao 2006]. L'optimisation de distances métriques améliore la fonction de distance qui est utilisée pour calculer les similarités entre la requête et la base de données images, la modification de la requête recherche le "point idéal" de la requête et l'apprentissage de classificateur utilise les images pertinentes / non pertinentes pour construire un classificateur probabiliste.

3.2.3.1 Modification de la requête

Diverses techniques de retour de pertinence ont été proposées pour améliorer la performance de la recherche. Parmi elles, la modification de la représentation de la requête est la méthode la plus populaire ; elle est largement utilisée dans la recherche d'images et la recherche de texte. Deux techniques de modification de requête sont disponibles dans la littérature : l'extension de requêtes et le mouvement du point de requête (en termes courts : mouvement de requête). Pour présenter ces deux techniques de modification de la requête, nous devons au préalable définir quelques concepts liés à l'unimodalité d'un groupe d'images :

Définition La notion d'unimodalité d'un groupe d'images signifie que toutes les images de ce groupe sont similaires et qu'elles forment un groupe distinct des autres images dans la base dans l'espace des caractéristiques. Par exemple, le groupe d'images à gauche dans la figure 3.2 est unimodal, tandis que le groupe d'images à droite est non unimodal.

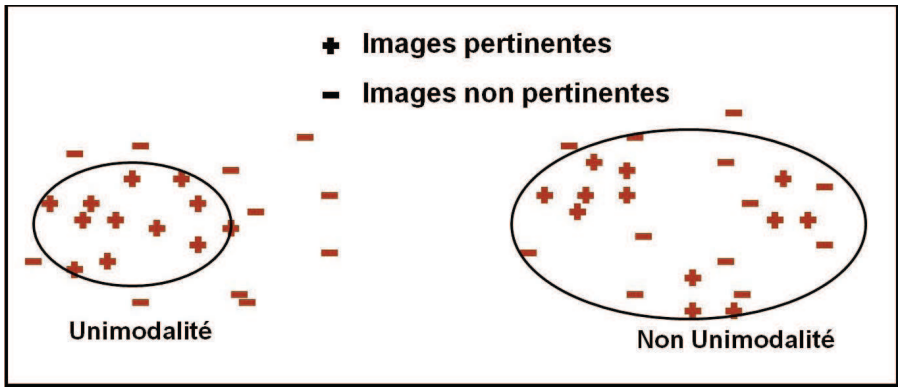

FIGURE 3.2 – Définition de l'unimodalité d'un groupe d'images.

Mouvement de requête Le mouvement de requête [Ortega Binderberger 2004], [Ishikawa 1998] est utilisé pour raffiner la requête par l'intermédiaire des images pertinentes et non pertinentes, à partir de l'hypothèse d'unimodalité des images pertinentes [Wu 2004]. Le mouvement de requête atteint le point idéal de requête en le déplaçant vers des images pertinentes et en l'éloignant de celles qui ne le sont pas.

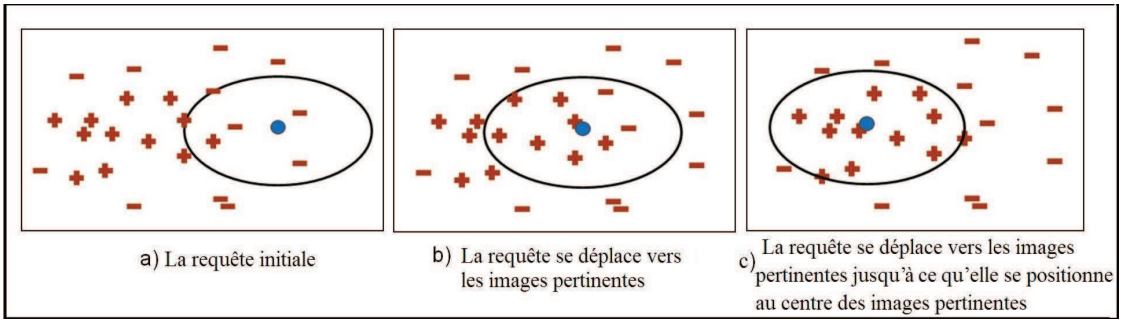

FIGURE 3.3 – Le mouvement de requête.

Pour le mouvement du point de requête en recherche d'images en utilisant la requête par l'exemple [Ortega Binderberger 2004], [Ishikawa 1998], une requête est représentée par un seul point dans un espace de caractéristiques et le processus de raffinement tente de reformuler le vecteur de caractéristiques de la requête de sorte qu'elle se rapproche de l'espace contenant les images pertinentes (voir figure 3.3). Sur la base d'une hypothèse d'unimodalité des images pertinentes, la requête optimale maximise la similitude des images pertinentes et minimise celle des images non pertinentes [Kim 2005]. La technique de Rocchio [Rocchio 1971] est souvent utilisée pour calculer la requête optimale :

$$\vec{q}_{i+1} = \alpha\vec{q}_i + \frac{\beta}{|D_r|}\sum_{\vec{d}\in D_r}\vec{d} - \frac{\gamma}{|D_n|}\sum_{\vec{d}\in D_n}\vec{d} \tag{1}$$

où $\vec{q}_i$ est la requête à la i^{eme} interaction du retour de pertinence, D_r est l'ensemble pertinent, D_n est l'ensemble non pertinent, α, β et γ sont les poids relatifs de q, D_r et D_n. En pratique, l'ensemble des paramètres $\alpha = \beta = \gamma = 1$ est largement utilisée pour la recherche d'images.

Extension de requêtes Différemment du mouvement de point de requête, l'extension de requêtes [Ortega Binderberger 2004], [Kim 2005] modifie la requête par l'ajout de nouveaux points sélectifs pertinents à la représentation de la requête. Une requête à un seul point est remplacée par une requête à points multiples (voir figure 3.4). Plutôt que de supposer une distribution unimodale comme dans le cas du mouvement de point de requête, l'extension de requêtes suppose de nombreuses petites distributions unimodales pour construire plusieurs groupes locaux à partir d'images pertinentes. Les représentants des groupes locaux sont utilisés pour effectuer des requêtes à points multiples. Le regroupement des images pertinentes est répété à chaque interaction de pertinence. L'interrogation par de multiples points est étudiée dans [Jin 2003b], [Natsev 2003], [Westerveld 2004], [Tahaghoghi 2002], [Natsev 2005] qui se sont concentrés sur la fonction de similarité et la fusion de plusieurs requêtes en un seul point. L'évaluation expérimentale dans [Ortega Binderberger 2004] montre que l'extension de requêtes surpasse le mouvement de point de requête en efficacité de recherche. L'extension de requêtes est sans doute l'une des approches les plus efficaces de retour de pertinence [Tahaghoghi 2002].

Récemment, de nouvelles approches sont apparues visant à améliorer la technique de modification de la requête. Le système QCluster [Kim 2005] notamment utilise une nouvelle classification d'adaptation et une méthode de fusion de clusters pour trouver de multiples régions. L'étape de regroupement ne se répète pas comme dans l'extension de requêtes. QCluster classe les exemples pertinents dans des groupes déjà existants ou forme un nouveau groupe. Le nombre de groupes est limité à un nombre fixe en utilisant une méthode de fusion de clusters. Cependant, cette approche complexe présente des difficultés pour faire un usage efficace des exemples non pertinents. Dans [Liu 2009], les auteurs proposent une technique de mouvement rapide de requête pour adapter la recherche objective dans le domaine CBIR. Toutes les méthodes ci-dessus ont toutefois des inconvénients identifiés, comme par exemple le risque de tomber dans des pièges de maxima locaux et une convergence lente.

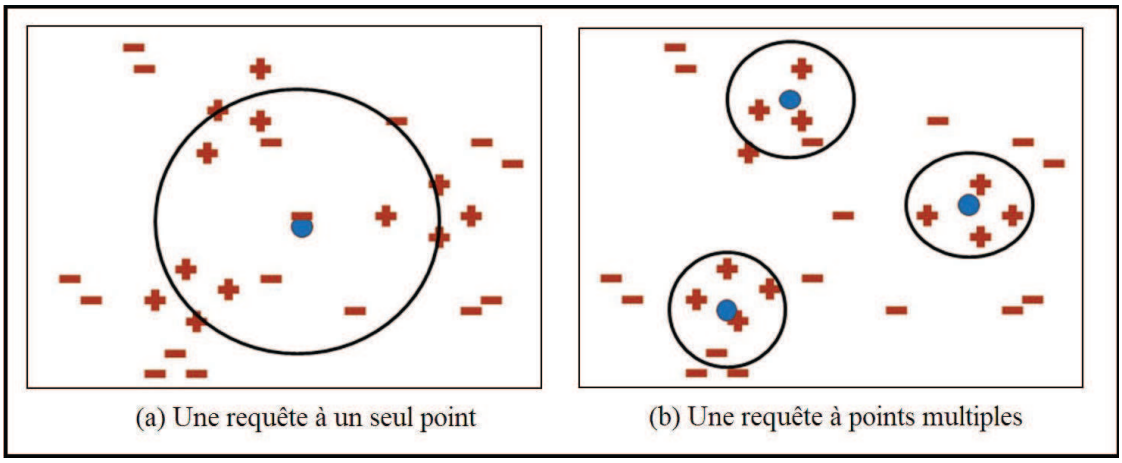

FIGURE 3.4 – Extension de requêtes, (a) une requête à un seul point est remplacée par (b) une requête à points multiples.

L'interrogation par points multiples L'extension de requêtes statue nécessairement sur l'interrogation par points multiples. Une requête est représentée par un point dans l'espace de caractéristiques, donc l'interrogation par de multiples requêtes est représentée par des points multiples dans l'espace des caractéristiques. L'interrogation par des points multiples est étudiée dans [Westerveld 2004], [Tahaghoghi 2002], [Natsev 2005] qui se sont concentrés sur la fonction de similarité et de fusion de plusieurs requêtes en un seul point. La similitude des images pour chaque requête en un seul point est déterminée de manière indépendante. Le résultat d'une requête en un seul point est une liste ordonnée. Les listes de toutes les requêtes doivent être combinées pour déterminer le classement final des requêtes multiples à un seul point. Une fonction de combinaison est donc nécessaire pour réduire les valeurs de similarité multiples en une seule valeur. Lorsque cette réduction a été effectuée pour toutes les images de la collection, une liste d'images est présentée dans l'ordre décroissant de similarité. Toutes les fonctions intégrées dans ce processus s'appuient généralement sur trois types d'opération : minimum, maximum et somme. Dans notre expérience, la fonction minimum se trouve être la meilleure fonction de combinaison en terme de robustesse. Ceci est également confirmé par Tahaghoghi et al. [Tahaghoghi 2002].

3.2.3.2 Optimisation de distance métrique

Un des problèmes relatifs à la notion de similarité se rapporte à la question de savoir comment mettre en balance les différentes caractéristiques. Il faut déterminer quelle caractéristique est importante et quelle caractéristique ne l'est pas à partir du retour de pertinence.

La similarité entre les représentations de 2 images est calculée par combinaison des similarités des vecteurs de chaque caractéristique individuelle. En recherche d'images, les mesures de distance couramment utilisées sont des instances de la métrique générale de Minkowski :

$$D(x,y) = \left[\sum_i |x_i - y_i|^\alpha\right]^{\frac{1}{\alpha}}, \alpha > 0 \tag{2}$$

avec $\alpha = 1$ on a la métrique L_1, $\alpha = 2$ la distance euclidienne. On peut obtenir la distance pondérée de Minkowski en ajoutant un poids à chaque composante. Ci-dessous est décrite la distance pondérée euclidienne (la somme des w_i est égale à 1).

$$D(x,y) = \left[\sum_i w_i \left(x_i - y_i\right)^2\right]^{\frac{1}{2}} \tag{3}$$

La combinaison la plus connue des similarités de chaque caractéristique est la somme pondérée linéaire :

$$D = \sum_i p_i d_i \tag{4}$$

ou p_i est le poids pour la similarité d_i de la caractéristique i et la somme des p_i est égale a 1.

La technique de retour de pertinence qui optimise la fonction de distance consiste à fournir des poids dans les deux métriques ci-dessus [Ishikawa 1998], [Ortega Binderberger 2004], [He 2004], [Urban 2004], [Aggarwal 2002]. Les deux types de poids sont le poids des composantes (p_i) et le poids des caractéristiques (w_i) respectivement. L'interprétation de cette technique est de donner plus d'importance à certaines composantes de la caractéristique (w_i) ou à certaines caractéristiques de la représentation d'images (p_i). Par exemple, pour un utilisateur donné, le pourcentage de pixels verts dans une image peut être plus important que le pourcentage de pixels rouge dans l'image ($w_{vert} > w_{rouge}$), ou la caractéristique de couleur peut être plus importante que celle de texture pour une image ($p_{couleur} > p_{texture}$).

3.2.3.3 Apprentissage de classifieurs

Récemment, des techniques de classification ont été introduites pour le retour de pertinence [Tao 2006], [Jing 2004], [Chen 2005], [Tieu 2004] parmi lesquelles les classificateurs de type SVM (*Support Vector Machine*) sont considérés comme les classificateurs les plus prometteurs. Dans ce type de retour de pertinence, la recherche d'images consiste à classifier des images en pertinentes et non pertinentes. Pour ce faire, on exige la connaissance de la densité de probabilité de toutes les images : $p(x|P)$ et $p(x|N)$, la probabilité pour qu'une image soit pertinente/non pertinente. Une fois que les densités sont estimées on peut rechercher des images pertinentes. Après le retour de pertinence de l'utilisateur, l'ensemble du processus de la recherche réitère sa phase d'apprentissage pour une nouvelle phase, en considérant les images pertinentes et non pertinentes précédemment traitées comme des exemples à apprendre par un classificateur probabiliste binaire. La technique du SVM associe les signatures d'images à un espace de caractéristiques de dimensions supérieures en utilisant une transformation non linéaire associée à un noyau, puis effectue implicitement la discrimination linéaire entre images pertinentes et images non pertinentes dans cet espace. Les SVM ont un certain nombre d'avantages qui les rendent particulièrement appropriés pour le retour de pertinence par rapport aux autres classificateurs. Plus particulièrement, les SVM évitent les hypothèses trop restrictives sur la répartition (l'unimodalité) des données. Néanmoins, les SVM demandent un grand nombre d'exemples d'apprentissage, ce point étant l'inconvénient le plus important.

3.2.3.4 Retour de pertinence pour la recherche multimodale

Le retour de pertinence a été initialement utilisé pour la recherche unimodale. Récemment, quelques auteurs ont étudiés le retour de pertinence pour la recherche multimodale [Jing 2004], [Ferecatu 2008], [MM 2010].

Dans [Jing 2004], les auteurs proposent un schéma efficace pour le retour de pertinence dans la recherche multimodale d'images. Ils utilisent 2 modèles : le modèle statistique par mots-clés et le modèle SVM en ligne. Pour chaque mot-clé, le modèle statistique est appris hors ligne en se basant sur les caractéristiques visuelles d'un petit ensemble d'images étiquetées manuellement et il est utilisé pour propager les mots-clés à d'autres images non marquées. Les résultats initiaux de la recherche sont donnés par le tri décroissant des images selon la probabilité estimée d'un mot-clé requête K estimé en utilisant le modèle statistique par mots-clés. Lorsque l'utilisateur marque quelques images comme des exemples pertinents/non pertinents, un modèle SVM est appris en ligne en considérant les caractéristiques visuelles des images marquées comme l'ensemble d'apprentissage, pour étendre l'es-

pace de recherche. Autrement dit, la nouvelle similarité d'une image I (probabilité que I soit étiquetée avec le mot clé requête K) est la fusion de 2 probabilités : une probabilité estimée par le modèle statistique par mots-clés et une probabilité estimée par le modèle SVM en ligne.

Dans [Ferecatu 2008], la technique de retour de pertinence SVM est utilisée pour la recherche multimodale d'images. Les auteurs créent un descripteur de concept nouveau basé sur la sélection de l'ensemble des concepts clés de l'image. Pour trouver de bons candidats pour ces concepts clés, les auteurs s'appuient sur l'ontologie générale WordNet [1] pour définir des relations sémantiques entre les concepts. Ce vecteur de caractéristiques conceptuelles peut être employé directement par un système CBIR existant comme tous les autres vecteurs de caractéristiques visuelles. En utilisant ce système hybride visuel / textuel pour la représentation d'images, nous pouvons utiliser n'importe quelle technique de retour de pertinence pour améliorer la performance de la recherche.

Dans [MM 2010], une nouvelle méthode d'extension de requêtes multimodales est présentée pour un cadre de recherche d'images médicales en appui sur une intégration visuelle et sur des mots-clés. Dans ce travail, les images sont représentées par un ensemble de mots visuels et de mots textuels. La similarité entre une requête multimodale, (contenant une partie de l'image I_q et une partie du texte D_q) et une image j, (contenant également deux parties liées à l'image I_j et le texte D_j), est définie comme :

$$Sim(q, j) = w_I Sim_I(I_q, I_j) + w_D Sim_D(D_q, Dj) \tag{5}$$

Pour une requête q donnée, l'ensemble S_l des premières images récupérées (ou des images pertinentes de l'utilisateur) s'appelle l'ensemble des images locales. En outre, l'ensemble $C_l \subseteq C$ de tous les mots visuels distincts $c_i \in C_l$ et l'ensemble $T_l \subseteq T$ de tous les mots textuels distincts $t_i \in T_l$ dans l'ensemble des images locales S_l sont appelés respectivement le vocabulaire local des mots visuels et le vocabulaire local des mots textuels. Une matrice de corrélation locale est construite en se basant sur la co-occurrence des mots visuels à l'intérieur des images et des mots textuels des annotations associées. Puis, en se basant sur cette matrice de corrélation locale, les éléments du vecteur de la requête sont ajoutés et re-pondérés.

1. WordNet : http ://wordnet.princeton.edu/

3.2.4 Conclusion

Nous avons présenté différentes interactions possibles dans les systèmes de recherche d'images : la spécification de requêtes, l'exploration et le retour de pertinence. Dans notre système, nous n'utilisons que des exemples d'images et des mots-clés pour la spécification de requêtes. L'exploration de notre système n'est pas abordée encore dans ce chapitre. Celle-ci sera traitée dans le chapitre 6. En effet, dans ce chapitre, nous nous concentrons sur le retour de pertinence à court terme.

Parmi des techniques de retour de pertinence à court terme, la modification de requête basée sur la recherche de texte est considéré comme l'approche majoritaire du retour de pertinence dans les systèmes de recherche d'images. Ce type d'approche est encore très efficace, comparé à toutes les autres méthodes dans les deux domaines : la recherche de texte et la recherche d'images. Dans le contexte général des processus de recherche d'images et du développement de méthodes de bouclage de pertinence, un des problèmes reconnus est le petit nombre d'exemples disponibles. Un utilisateur peut étiqueter seulement jusqu'à 30 images lorsque la plupart des méthodes d'apprentissage en requièrent beaucoup plus. Si nous comparons l'algorithme simple de modification de requêtes de Rocchio avec des algorithmes d'apprentissage (d'optimisation de métrique ou de classificateur), tels que les réseaux de neurones par exemple, il est clair que la popularité de la modification de requêtes est liée au fait qu'elle nécessite très peu d'exemples en apprentissage. Cependant, il faut toutefois noter des problèmes de modification de requêtes, concernant ces deux méthodes de mouvement de requête et d'extension de requêtes.

Le désavantage principal du mouvement de requête est la contrainte d'unimodalité (voir définition 3.2.3.1) sur les exemples pertinents ; par ailleurs l'inconvénient reconnu de l'extension de requêtes est sa difficulté à utiliser efficacement des images non pertinentes. Dans le mouvement de requête, le point de requête tente de se déplacer vers des exemples pertinents et de s'éloigner de ceux qui sont non pertinents. Dans le cas où les images pertinentes sous forme visuelle sont membres de sous-ensembles distincts (c'est-à-dire suivant une distribution non unimodale), se pose alors le problème du besoin de couvrir les clusters multiples avec une seule requête. Dans ces cas, le point idéal de requête comprend donc des exemples non pertinents. La figure 3.5a montre l'ellipse représentant la gamme à égale distance d'une nouvelle requête. Nous pouvons voir quelques exemples non pertinents inclus dans les ellipses.

L'extension de requêtes et sa meilleure version améliorée QCluster [Kim 2005] utilisent seulement des exemples pertinents pour former des requêtes à points multiples. La technique d'extension de requêtes n'utilise pas les exemples non pertinents car nous ne pouvons regrouper des exemples pertinents et non pertinents,

qui donnent de faux groupes. Notre analyse sur le sujet laisse penser que sans les exemples non pertinents, la convergence vers les points de requête idéaux peut potentiellement être très lente, et par ailleurs le risque de tomber dans un minimum local est non négligeable. En effet, un point faussement idéal de requête peut être atteint lorsque le groupe local est proche de quelques exemples pertinents, situés à proximité de nombreux exemples non pertinents (voir figure 3.5b). Nous pouvons constater sur cette figure que des exemples non pertinents peuvent être inclus dans les groupes locaux, car ces derniers sont construits en se fondant uniquement sur des exemples pertinents sans se soucier de la présence ou non de ceux qui ne le sont pas. En général, les techniques de retour de pertinence utilisent souvent des exemples pertinents, la gestion des exemples non pertinents reste un facteur de progression important, représentant ainsi une question scientifique très ouverte [Wang 2008].

Dans la section suivante, nous présentons notre approche de retour de pertinence qui tente précisément d'apporter des réponses aux questions préalablement identifiées, en se basant sur les exemples non pertinents et sur la combinaison du mouvement de requête et de l'extension de requêtes. Cette technique de retour de pertinence est utilisée dans la composante 2 de l'architecture de notre système pour l'interrogation par exemples d'images et l'interrogation combinée par exemples d'images et mots-clés. Nous présentons ainsi la visualisation car la visualisation et le retour de pertinence sont étroitement liés dans les mécanismes de recherche d'images (voir partie 3.4).

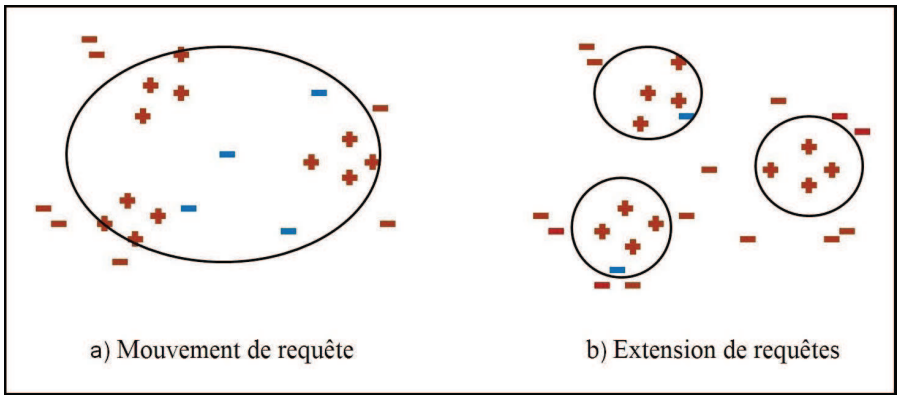

FIGURE 3.5 – Mouvement du point de requête et extension de la requête. Dans le mouvement du point de requête, le point idéal de requête comprend des exemples pertinents en raison de l'unimodalité des exemples pertinents. Dans l'extension de requêtes, les points idéaux de la requête convergent lentement lorsque les exemples non pertinents ne sont pas utilisés et ils peuvent entraîner le résultat dans un piège de maximum local.

3.3 Retour de pertinence à court terme dans notre système

Nous présentons dans ce chapitre notre contribution concernant le retour de pertinence à court terme. Cette contribution (correspondant à l'étape 2, en rouge, figure 3.6) permet d'améliorer la performance du système par une recherche interactive, et ainsi fournir un composant basique pour le retour de pertinence à long terme, ou en d'autres termes l'apprentissage de connaissances, que nous verrons dans le chapitre suivant (chapitre 4).

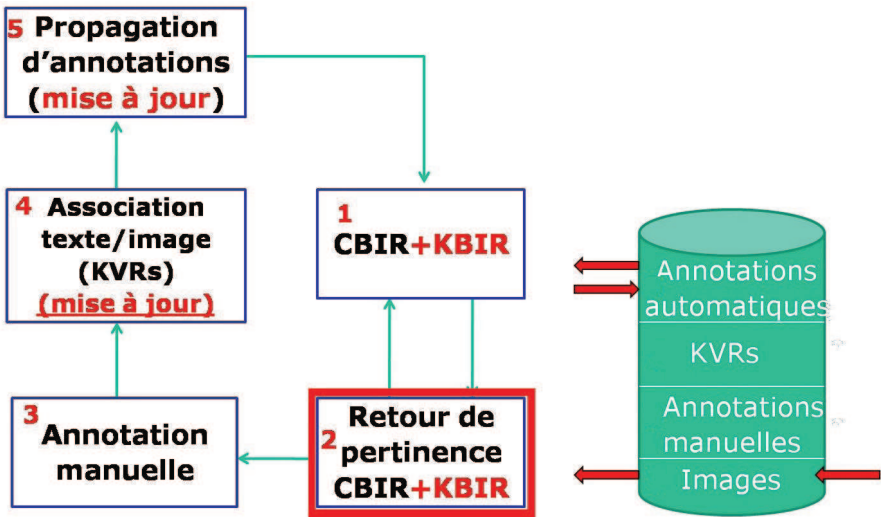

FIGURE 3.6 – Système de recherche d'images : le retour de pertinence à court terme.

3.3.1 Retour de pertinence basé sur les groupes pour la recherche par le contenu

Dans cette partie, une combinaison de mouvement de requête et d'extension de requêtes est proposée pour surmonter les problèmes liés à l'extension de requêtes et au mouvement de requête pris isolément. L'inconvénient principal du mouvement de requête est la contrainte de l'unimodalité sur des exemples pertinents qui n'est pas vraie en réalité. Nous résolvons ce problème en utilisant une technique de regroupement pour construire de multiples ensembles locaux qui assurent l'unimodalité à partir des exemples pertinents. L'inconvénient principal de l'extension de requêtes est l'incapacité à faire un usage efficace des exemples non pertinents. Dans notre approche, nous tentons de profiter des exemples non pertinents en utilisant la technique de mouvement de requête sur de multiples ensembles locaux. Selon

nous, cette combinaison est la meilleure parmi toutes les combinaisons possibles parce qu'elle peut assurer l'unimodalité des exemples (figure 3.7b) et profiter des exemples non pertinents (figure 3.7c) pour atteindre efficacement la requête idéale.

Il est difficile d'éliminer des exemples non pertinents dans le résultat par le mouvement de requête à cause de la contrainte de l'unimodalité (figure 3.5a) ou par l'extension de requêtes à cause de l'utilisation uniquement d'exemples pertinents (figure 3.5b). Pour atteindre la requête idéale nous acceptons le fait que l'ensemble des exemples pertinents n'assure pas l'unimodalité et nous tentons de construire des groupes locaux avec les exemples pertinents qui assureront l'unimodalité (figure 3.7b). Dans ces groupes locaux, il est possible d'obtenir des exemples non pertinents. Pour éliminer ces exemples non pertinents, il nous faut identifier des exemples non pertinents qui sont présents autour de chaque groupe local (dans l'espace des caractéristiques). Nous pouvons le faire par la classification des exemples non pertinents. En déplaçant ces groupes locaux vers les exemples pertinents et en éloignant les exemples non pertinents par la technique du mouvement de requête, notre approche peut éliminer alors les exemples non pertinents (figure 3.7c). Les groupes locaux finaux sans aucun exemple non pertinent représentent la requête idéale à points multiples. Notre nouvelle méthode qui combine le mouvement de requête et l'extension de requêtes tout en intégrant les exemples non pertinents est décrite ci-dessous.

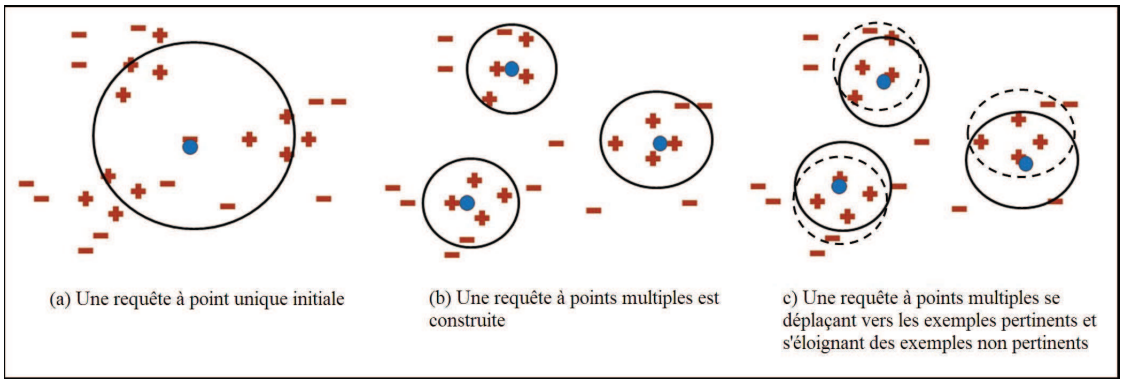

FIGURE 3.7 – Combinaison du mouvement de requête et de l'extension de requêtes. Les points idéaux de requête sont atteint plus efficacement et plus rapidement. Les exemples non pertinents sont éliminés des groupes locaux.

Le but de cette technique est d'atteindre la requête idéale par des interactions avec l'utilisateur. La première interaction du retour de pertinence est illustrée dans la figure 3.8. Les exemples pertinents dans le résultat d'une requête n'assurent pas l'unimodalité (étapes 1 et 2, figure 3.8). La requête composée d'un seul point est remplacée par une requête à points multiples en utilisant la technique d'extension de requêtes. D'abord, les exemples pertinents sont regroupés en c groupes :

$C_1, C_2, \ldots, C_c$ pour assurer l'unimodalité (étape 3, figure 3.8). Le paramètre c est sélectionné automatiquement par un algorithme de regroupement limité à une valeur maximale. Les deux algorithmes de regroupement utilisés dans notre système sont présentés dans la section 3.3.2. Les exemples non pertinents sont ensuite classés dans ces c groupes (étape 4, figure 3.8) pour identifier des exemples non pertinents présents dans chaque groupe local. Les exemples pertinents et non pertinents dans chaque groupe sont alors utilisés pour construire la requête à points multiples par l'équation 7 ci-dessous (étape 5, figure 3.8). Le classificateur des k plus proches voisins ($k - NN$) est utilisé dans l'étape 4 pour la classification des exemples non pertinents en raison de son efficacité et sa simplicité, le paramètre k de ce classificateur étant sélectionné comme suit :

$$k_i = \min(|C_j|, j = 1 : c) \tag{6}$$

Soit $I_1, I_2, \ldots, I_n$: n exemples non pertinents et $R_1, R_2, \ldots, R_m$: m exemples pertinents du groupe local C_i, le point de requête $\overrightarrow{q_i}$ de ce groupe est calculé en utilisant la technique de mouvement de requête [Rocchio-2] :

$$\overrightarrow{q_i} = \frac{\sum_{j=1}^{m} \overrightarrow{R_j}}{m} - \frac{\sum_{j=1}^{n} \overrightarrow{I_j}}{n} \tag{7}$$

Ces c points de requête forment la requête à points multiples finale.

Comme nous l'avons présenté ci-dessus, dans la première interaction, la requête initiale (d'un seul point) est remplacée par une requête à points multiples en construisant des groupes locaux (l'étape de regroupement). Pour les interactions suivantes, il existe deux possibilités pour améliorer la requête à points multiples. La première possibilité est de déplacer ces points de requête vers des points idéaux en se basant sur les nouveaux exemples pertinents/non pertinents provenant des interactions suivantes. Cette méthode assume qu'on peut arriver aux points idéaux de la requête à partir de ces premiers points de la requête. Comme on ne reconstruit pas les groupes locaux, l'étape de regroupement est effectuée une seule fois au début (la première interaction), la requête des interactions suivantes est construite en se basant sur les points multiples de la requête de la première itération. D'autre part, la deuxième possibilité ne se base pas sur les points multiples de la requête (l'étape de regroupement de la première itération), mais consiste à en regrouper à nouveau les exemples pertinents. Cette méthode tente d'ajouter les points pertinents de la requête en supprimant les points non pertinents de cette même requête en se basant sur tous les exemples pertinents/non pertinents venant de toutes les interactions. Le regroupement et la classification sont répétés à chaque interaction pour cette méthode.

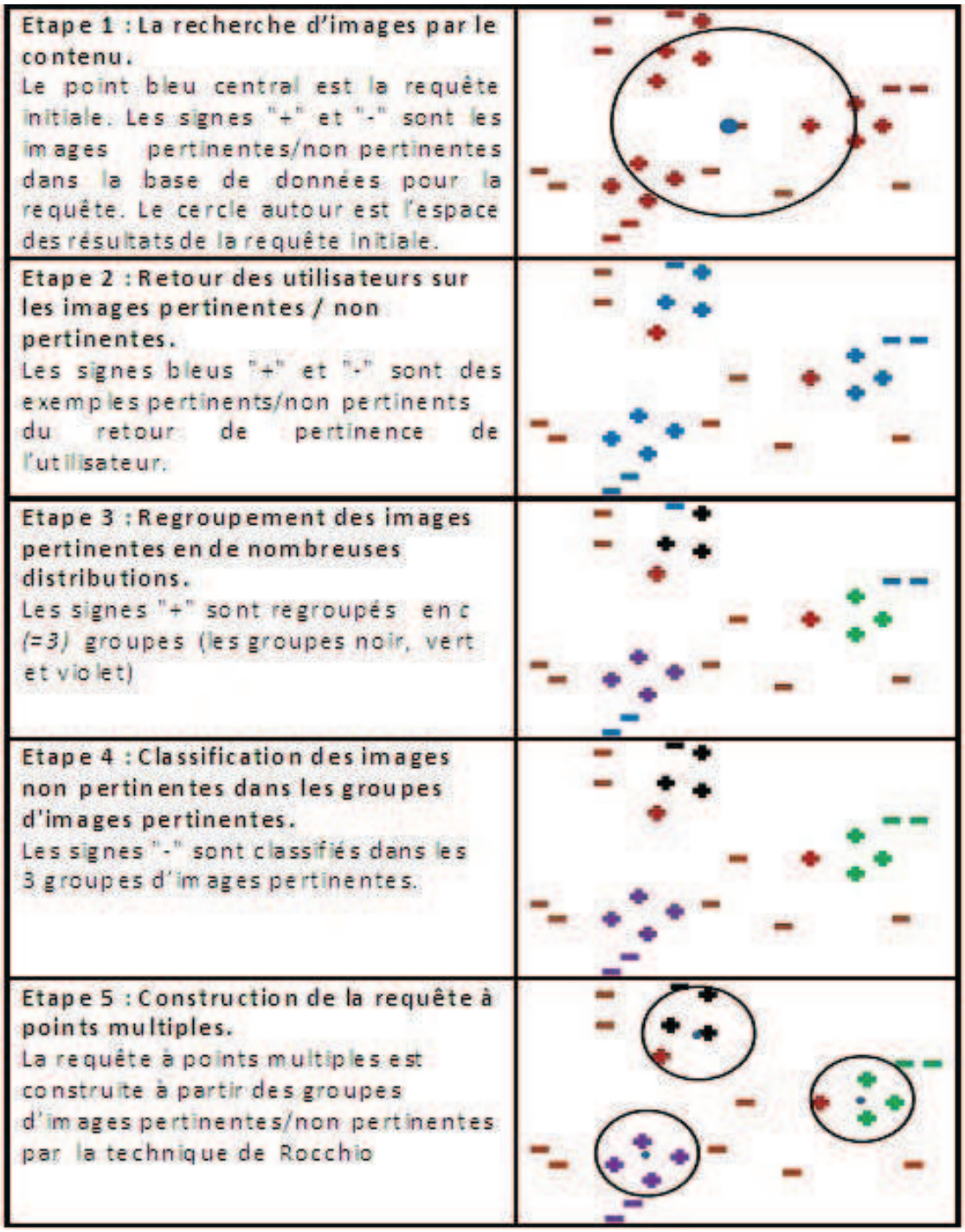

FIGURE 3.8 – Retour de pertinence basé sur les groupes pour la recherche par le contenu (CBIR). Les signes "+" et "-" sont les exemples pertinents/non-pertinents. La couleur rouge montre les exemples non annotés pendant l'interaction, les autres couleurs montrent les exemples annotés.

Nous pouvons observer que la première méthode est plus influencée par le mouvement de requête que par l'extension de requêtes, du fait cette méthode tente de

déplacer la requête à points multiples vers la requête idéale. En revanche, la seconde méthode est plus influencée par l'extension de requêtes que par le mouvement de requête du fait qu'elle tente de créer les points idéaux de la requête en se basant sur le regroupement. Nous appellerons les deux méthodes respectivement : regroupement-répétition (CR - *Clustering-Repeat*) et regroupement-non-répétition (CNR - *Clustering-Non-Repeat*). Les deux algorithmes de ces méthodes sont décrits ci-dessous.

Regroupement-répétition (CR) Dans cette approche, l'étape de regroupement des exemples pertinents, l'étape de classification des exemples non pertinents et l'étape de la construction de la requête à points multiples sont répétées à chaque itération du retour de pertinence. Le système réalise ainsi le même processus pour toutes les itérations. La requête de l'itération précédente n'influence pas directement la requête nouvelle. Des exemples provenant de l'itération précédente sont également intégrés dans l'itération courante. Implicitement, les points pertinents sont ajoutés et les points non pertinents sont abandonnés lorsque nous passons d'une itération à la suivante.

Algorithme 1 : Regroupement-répétition (*Clustering-Repeat*)
Entrée : une requête q
Sortie : une liste d'images
Début
Étape 1 : Rechercher des images similaires à la requête courante
Si c'est la première itération (requête à point unique) :
(1) calculer les distances entre les images et la requête q
Sinon (requête à points multiples) :
(1) calculer les distances entre les images et chaque point de requête
(2) calculer la distance finale en combinant toutes les distances calculées
Étape 2 : Interaction avec l'utilisateur
Afficher les $N(= 100)$ premières images pour l'utilisateur
Si l'image I_i est marquée comme pertinente, ajouter I_i à l'ensemble pertinent
Si l'image I_i est marquée comme non pertinente, ajouter I_i à l'ensemble non pertinent
Étape 3 : Regroupement
Regrouper les images de l'ensemble pertinent en c clusters
Étape 4 : Classification
Classifier les images de l'ensemble non pertinent en clusters
Étape 5 : Mouvement de requête

Construire la requête à points multiples en utilisant l'équation (7)
Aller à l'étape 1 pour l'itération suivante

Fin

Regroupement-non-répétition Selon cette démarche, les requêtes précédentes affectent directement les nouvelles requêtes. L'étape de regroupement des exemples pertinents est effectuée une seule fois au début (première itération). Ensuite, durant les itérations suivantes, au lieu d'effectuer un nouveau regroupement comme c'est le cas dans la méthode CR, tous les exemples pertinents / non pertinents sont classés dans les groupes émanant de la précédente requête, afin de tirer profit de la requête précédente. Les nouvelles requêtes sont raffinées à partir des exemples pertinents / non pertinents des groupes en utilisant la technique du mouvement de requête :

$$\overrightarrow{q_i} = \overrightarrow{q}_{i,previous} + \frac{\sum_{j=1}^{m} \overrightarrow{R_j}}{m} - \frac{\sum_{j=1}^{n} \overrightarrow{I_j}}{n} \tag{8}$$

où $\overrightarrow{q}_{i,previous}$ est la requête précédente.

Algorithme 2 : Regroupement-non-répétition (*Clustering-Non-Repeat*)
Entrée : une requête q
Sortie : une liste d'images
Début
Étape 1 : Rechercher des images similaires à la requête courante
Si c'est la première itération (requête à point unique) :
(1) calculer les distances entre les images et la requête q
Sinon (requête à points multiples) :
(1) calculer les distances entre les images et chaque point de requête
(2) calculer la distance finale en combinant toutes les distances calculées
Étape 2 : Interaction avec l'utilisateur
Afficher les $N(= 100)$ première images pour l'utilisateur
Si l'image I_i est marquée comme pertinente, ajouter I_i à l'ensemble pertinent
Si l'image I_i est marquée comme non pertinente, ajouter I_i à l'ensemble non pertinent
Étape 3 : Regroupement
Si c'est la première itération
(1) Regrouper les images de l'ensemble pertinent en c clusters
Étape 4 : Classification

Classifier les images des ensembles non pertinent et pertinent en c clusters
Étape 5 : Mouvement de requête
Construire la requête à points multiples en utilisant l'équation (8)
Supprimer les ensembles pertinent et non pertinent
Aller à l'étape 1 pour l'itération suivante
Fin

Dans les deux algorithmes précédents, nous voyons que les différences portent sur les étapes 3, 4 et 5. Dans le cas du Regroupement-non-répétition, l'étape 3 est effectuée une seule fois (à la première itération) tandis qu'elle se répète à toutes les itérations pour le Regroupement-répétition. Pour l'étape 4, seul l'ensemble non pertinent est divisé en clusters pour l'algorithme Regroupement-répétition, tandis que tous les deux ensembles (pertinent et non pertinent) sont divisés en clusters pour l'algorithme Regroupement-non-répétition. Pour l'étape 3 de l'algorithme Regroupement-répétition, l'ensemble pertinent est utilisé pour reconstruire les groupes locaux (l'étape 3 se répète). Enfin la formule utilisée pour construire la requêtes à points multiples est différente pour les deux algorithmes. Comme il s'agit de méthodes dépendantes des retours de l'utilisateur, qui constitutent les entrées des deux algorithmes, la convergence globale des résultats dépend de la qualité des retours donnés par celui-ci. Bien que l'algorithme CNR soit plus rapide d'exécution que l'algorithme CR, cela ne fait pas de différence du point de vue de l'utilisateur, car le nombre de points à calculer pour le clustering est en général assez faible (10 à 20 retours de l'utilisateur).

Discussion Dans cette partie, nous avons présenté notre approche avec ses deux variantes pour le retour de pertinence à court terme. Notre approche combine les deux méthodes de modification de requête : le mouvement de requête et l'extension de requêtes, afin de profiter des exemples pertinents et de traiter le problème de l'unimodalité en tentant d'éliminer tous les exemples non pertinents dans le résultat. Les deux variantes de notre approche sont la méthode CNR et la méthode CR qui visent à atteindre les points idéaux de requête que nous passons d'une itération/interaction à l'autre. La méthode CNR (*Clustering-Non-Repeat*) vise à déplacer les points de requête vers les points idéaux, la méthode CR (*Clustering-Repeat*) vise à remplacer les points non pertinents par des points pertinents. La première méthode dépend de la construction des points initiaux. Par exemple, si les exemples pertinents dans la base peuvent être présentés dans n groupes distincts mais que les exemples pertinents utilisés pour construire les points initiaux appartiennent à $c << n$ groupes distincts, ceci peut produire la perte d'une partie du résultat. La deuxième méthode dépend plus de la performance de la méthode de regroupement utilisée que la première méthode parce que dans cette méthode le regroupement (clustering) est répété pour toutes les itérations.

3.3.2 Sélection de la méthode de regroupement

Dans notre méthode de retour de pertinence, une étape importante est celle du regroupement (ou clustering). La méthode de regroupement est utilisée pour regrouper les images pertinentes/non pertinentes en des groupes distincts. Dans notre système, le nombre de groupes est inconnu. Nous nous intéressons donc aux méthodes de regroupement qui sont capables de déterminer le nombre de groupes de façon autonome. Nous utilisons 2 méthodes de regroupement pour notre expérimentation : la méthode des K-moyennes adaptatif présentée par [Kothari 1999] et la méthode d'Agglomération compétitive présentée par [Frigui 1997]. Ces deux méthodes sont choisies en raison de leur capacité à déterminer automatiquement le nombre de groupes, et elles sont représentatives de deux types connus de méthodes de regroupement dans la littérature : les méthodes hiérarchiques et les méthodes de partitionnement.

3.3.2.1 K-moyennes adaptatif

L'algorithme le plus connu pour le regroupement est la méthode des k-moyennes. Pour p modèles $\{x^\mu : \mu = 1, 2, \ldots, p\}, x^\mu \in \mathbb{R}^n$ la methode des k-moyennes obtient la position des k centres de cluster y^ν en minimisant la fonction de coût donnée par :

$$J = \sum_{\mu=1}^{p} \sum_{\nu=1}^{k} I(y^\nu|x^\mu)||x^\mu - y^\nu||^2, \tag{9}$$

$||.||$ désigne une métrique de distance, $I(y^\nu|x^\mu)$ est une fonction indicatrice qui vaut 1 si $\mu = argmin||x^\mu - y||^2$ et 0 autrement.

Dans la méthode des k-moyennes adaptatif [Kothari 1999], la fonction de coût proposée est :

$$J = \sum_{\mu=1}^{p} \sum_{\nu=1}^{k} I(y^\nu|x^\mu)||x^\mu - y^\nu||^2 + \sum_{\mu=1}^{p} \sum_{\nu=1}^{k} \tilde{\lambda}_\nu \tilde{I}(y^\nu|x^\mu)||y^\nu - y^\omega||^2 \tag{10}$$

$\tilde{I}(y^\nu|x^\mu)$ est une fonction indicatrice qui vaut 1 si $y^\nu \in N_{y^\omega}, \omega = argmin||x^\mu - y||^2$, et N_{y^ω} sont les voisinages du centre du cluster y^ω.

Il y a 2 termes dans cette fonction de coût : le premier est similaire à celui de la méthode des k-moyennes, le deuxième est un terme supplémentaire. Le terme supplémentaire tente de répartir les centres de clusters de manière à minimiser la distance des sommes des carrés d'un centre de cluster aux centres des clusters

voisins.

Des valeurs plus petites du voisinage encouragent la formation de plusieurs centres de clusters distincts, alors que les grandes valeurs du voisinage encouragent la formation de moins de centres de clusters distincts. La méthode des k-moyennes adaptatif identifie le voisinage comme un paramètre d'échelle et permet d'obtenir le nombre de centres de clusters à différentes valeurs du paramètre d'échelle. Le nombre de centres de clusters dans les données est alors obtenu en se basant sur la stabilité des clusters en faisant varier le paramètre d'échelle.

3.3.2.2 Agglomération compétitive

La plupart des méthodes de regroupement ont pour inconvénient qu'il faut connaître le nombre de groupes C, ou bien alors de tenter de déterminer ce nombre en répétant le processus de regroupement pour plusieurs valeurs de C et en sélectionnant une partition selon un critère de validité particulier. L'algorithme d'agglomération compétitive [Frigui 1997] propose de regrouper les données automatiquement suivant un nombre optimal de groupes.

L'agglomération compétitive minimise une fonction objective qui intègre les avantages des techniques de regroupement hiérarchiques et partitionnelles. L'algorithme d'agglomération compétitive produit une séquence de partitions avec une diminution du nombre de groupes. L'agglomération compétitive commence par le partitionnement des données sur un nombre spécifié de groupes, et permet d'obtenir finalement le nombre "optimal" de groupes. Pendant les phases de regroupement, les groupes adjacents jouent les uns contre les autres pour s'approprier les points de données, et les groupes qui perdent peu à peu dans la compétition s'épuisent et disparaissent, jusqu'à ce que seuls les groupes à grande cardinalité survivent. L'algorithme peut intégrer différentes mesures de distance dans la fonction objective pour trouver un nombre de groupes de formes diverses.

Greater competitive minimizes an objective function that integrates the advantages of hierarchical clustering techniques and partitional. The competitive agglomeration algorithm produces a sequence of partitions with a decrease in the number of groups. Greater competitive begins with data partitioning on a specified number of groups, and finally provides the number of "best" of groups. During the consolidation phase, the adjacent groups playing against each other to capture the data points, and groups that are gradually losing in the competition run out and disappear, until only groups large cardinality survive. The algorithm can incorporate different distance measures in the objective function to find a number of groups in various forms.

3.3.2.3 Sélection de la méthode de regroupement

Lors de nos expériences, différentes méthodes de regroupement ont été étudiées pour calculer les groupes locaux. En tirant partie des avantages de deux types de regroupement (hiérarchique et partitionnel), la méthode d'agglomération compétitive [Frigui 1997] semble produire la meilleure performance, au regard de nos nombreux tests. Un autre avantage de cette méthode de regroupement est la sélection automatique du nombre de groupes. Nos expériences ont montré que le choix de la méthode de regroupement et la méthode de classification n'influencent pas beaucoup le résultat, car le nombre total d'échantillons (pertinents / non pertinents) est très faible. Rappelons en effet qu'ici, l'utilisateur ne marque que quelques exemples comme pertinents ou non pertinents lors de la phase de retour de pertinence.

3.4 Visualisation

Dans le domaine de la recherche d'images, la problématique de la visualisation représente un des challenges très importants puisqu'elle conditionne la perception du système par l'utilisateur. Il s'agit en effet de visualiser les résultats (images) sur une interface (2D par exemple) pour les utilisateurs après une phase de recherche. La visualisation prend d'avantage d'importance dans le cas de recherche d'images avec interaction (par exemple avec retour de pertinence) car elle influence directement la recherche des images similaires via des interactions avec les utilisateurs. Les deux sujets de la visualisation et de l'interaction dans la recherche d'images sont donc étroitement liés. Dans cette section nous présentons quelques idées sur la visualisation avec retour de pertinence, notamment pour la recherche mixte texte/image. Nos propositions ne sont que préliminaires et demandent à être améliorées. Cependant, nos investigations nous ont permis d'engager de premières démarches pour le retour de pertinence en recherche mixte texte/image, ce qui représente une tâche difficile dans le domaine.

3.4.1 État de l'art

La visualisation des grandes collections d'images est une problématique difficile car elle doit permettre d'utiliser efficacement un espace d'affichage limité et donner à l'utilisateur une vue d'ensemble, tout en ayant la possibilité d'accéder à certains détails. La plupart des dispositifs d'affichage sont à deux dimensions - le papier et l'écran - expliquant ainsi le fait que les recherches de la littérature se concentrent principalement sur la visualisation des images sur des interfaces 2D. La façon la plus populaire de présenter des images est une matrice rectangulaire de vignettes.

L'arrangement spatial est généralement basé sur le rang dans le résultat de la recherche, de manière fortement inspirée des moteurs traditionnels de recherche d'informations.

Ce principe est résumé par Shneiderman, rappelant la nécessité d'offrir en premier lieu une vue d'ensemble, accompagnée d'un zoom et d'un filtre, et d'éventuels détails sur demande [Shneiderman 1996]. Les méthodes de visualisation sont basées sur le principe de la corrélation des différentes informations : l'évolution des données dans le temps [Havre 2002], la similitude des contenus sémantiques [Janecek 2005], [Carey 2003], [Yee 2003] ou hiérarchiques [Andrews 2007]. D'un point de vue général, les méthodes de visualisation sont difficiles à comparer et leur efficacité dépend à la fois de la tâche à accomplir et de la perception des utilisateurs. Dans cette partie, nous présentons différents types de visualisation avec leurs possibilités d'interaction.

Pertinence C'est probablement la façon la plus populaire de présenter les résultats de recherche. Les résultats sont classés en se basant sur la pertinence par rapport à la requête. Cette démarche consiste à présenter les images sur une matrice rectangulaire de vignettes. Cela permet la navigation et laisse aux utilisateurs la possibilité d'analyser tout simplement la grille d'images comme s'ils lisaient un texte. [Combs 1999] et [Bederson 2001] essaient d'améliorer cette structure visuelle en étudiant les propriétés de zoom pour améliorer la navigation sur l'image. Rodden et al. [Rodden 2001] déterminent s'il est profitable aux utilisateurs d'avoir des ensembles de vignettes organisées en fonction de leur similitude, c'est-à-dire en regroupant ensemble les images qui se ressemblent. Ils décrivent des expériences qui consistent à examiner si des dispositions basées sur la similitude des images candidates peuvent aider à la sélection de photos. L'observation est que les deux arrangements basés sur les caractéristiques visuelles et sur les dispositions de concepts ont leurs propres avantages et inconvénients. Google Images et Yahoo ! Images utilisent ce type de visualisation pour leur moteurs de recherche d'images.

Interaction : L'interaction pour ce type d'approche est limitée à la sélection du degré de pertinence et à la possibilité de zoom.

Regroupement D'autres techniques de visualisation essaient de considérer non seulement la similitude entre la requête et chaque image récupérée, mais aussi entre toutes les images récupérées elles-mêmes [Santini 2001], [Stan 2003]. Le regroupement des résultats de la recherche, en plus d'être une forme intuitive et souhaitable de présentation, a également été utilisé pour améliorer les performances de la recherche dans Chen et al. [Chen 2005]. Ces initiatives présentent pour inconvénient le fait que visuellement des images similaires qui sont placées à côté les unes des

autres peuvent parfois donner le sentiment de fusion, du fait de leur chevauchement, ce qui les rend moins attirantes que si elles étaient séparées [Rodden 2001].

Interaction : Ce type de visualisation est capable de sélectionner des groupes d'images pertinentes/non pertinentes.

Hiérarchique Stan et al. [Stan 2003] décrivent un système d'exploration pour une banque d'images, qui traite d'un outil pour la visualisation de la base de données à différents niveaux de détails, reposant sur une technique de mise à l'échelle multi-dimensionnelle. Ce groupe de techniques de visualisation est proposé pour une hiérarchie de clusters pour la perception des images similaires. Des images peuvent parfois se chevaucher, tout comme dans la base d'images El Niño [Santini 2001]. Dans ce contexte, Tian et al. [Tian 2001] proposent d'utiliser l'ACP (Analyse en Composantes Principales) et se penchent sur une stratégie d'optimisation pour ajuster la position et la taille des images afin de minimiser les chevauchements (maximiser la visibilité), tout en maintenant la fidélité par rapport à la position d'origine qui reste un indicateur de similitude.

Interaction : Ce type de visualisation est capable de sélectionner des groupes d'images pertinentes/non pertinentes à différents niveaux.

Temps/Localisation Les images sont affichées dans un ordre chronologique plutôt que par leur pertinence. Le système de Google Picasa [1] pour les collections personnelles fournit une option pour visualiser selon un axe temporel. Pour les images personnelles, des arrangements combinant les résultats de la recherche basée sur le contenu visuel, de l'horodatage et de l'utilisation efficace de l'espace de l'écran ajoutent de nouvelles dimensions à la navigation [Huynh 2005]. Lai et al. [Lai 2010] proposent un système qui met en évidence l'intérêt du couplage contenu/localisation dans les mécanismes de recherche d'images.

Interaction : L'interaction pour ce type est limitée à la sélection du degré de pertinence (du contenu, du temps ou de la localisation).

Composée Ce type d'approche consiste à combiner ensemble deux ou plusieurs des types de visualisation présentés ci-dessus, en particulier pour des systèmes spécialisés. La classification hiérarchique et la visualisation de groupes conceptuels sont des exemples de visualisation composée (ex. Google Swirl, figure 3.9). Torres et al. présentent dans [Torres 2003] deux techniques de visualisation basées sur des spirales et des anneaux concentriques afin d'explorer les résultats des requêtes. Ces structures visuelles sont centrées sur le maintien de la focalisation de l'utilisateur

1. Google Picasa : http ://picasa.google.com/

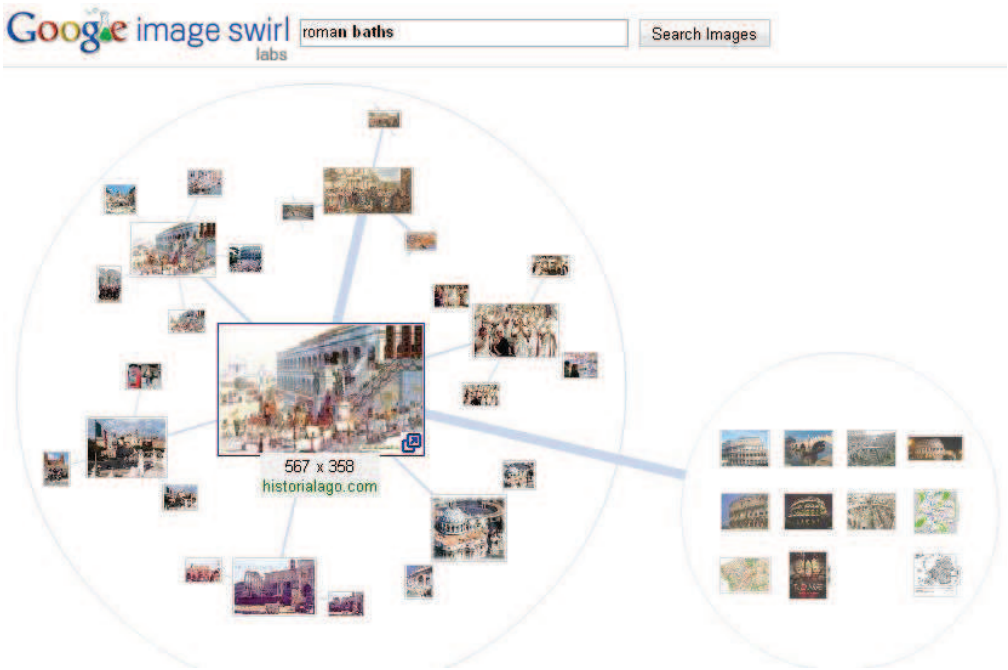

FIGURE 3.9 – Visualisation composée : Google Swirl.

sur l'image requête (au milieu de l'interface) et sur les images trouvées les plus similaires (autour de l'image requête). En évitant les chevauchements d'images couramment rencontrés dans les systèmes CBIR, ces stratégies améliorent la présentation de la grille 2D.

Janecek [Janecek 2005] a proposé un système basé sur la technique de visualisation *fish-eye* [Furnas 1986] pour la recherche et la navigation dans une collection d'images qui ont été étiquetées manuellement avec des métadonnées textuelles par des professionnels. Les similitudes entre les images correspondent à la distance des étiquettes par rapport à la hiérarchie WordNet [1]. Les images renvoyées sont disposées selon un modèle de forces (*spring layout*) qui déterminent leur poids visuels à l'égard de leur pertinence à la requête.

3.4.2 Discussion sur la visualisation

La plupart des types de visualisation sont conçus pour la recherche monomodale d'images. Pour la recherche multimodale, les techniques de visualisation d'images résultats ne montrent souvent que la similitude globale entre les images et la requête. Cependant, elles ne permettent pas d'identifier les relations entre les différentes modalités de la requête par rapport aux images et de projeter ces relations dans un même plan 2D. Cette problématique engendre une perte d'in-

1. WordNet : http ://wordnet.princeton.edu/

formation sur la structure des images résultats en terme de visualisation pour l'utilisateur. On note actuellement un manque global de recherche approfondie sur la visualisation pour la recherche multimodale d'images, même si on commence à trouver quelques travaux comme ceux de Camargo et al. [Camargo 2010] qui utilisent à la fois le texte et l'information visuelle pour visualiser les images. Camargo et al. utilisent une factorisation en matrices non négatives (*Non-negative Matrix Factorization*) pour construire un espace latent multimodal dans lequel les caractéristiques visuelles et les termes du texte sont représentés ensemble. En utilisant la représentation latente proposée, une visualisation de la collection d'images est construite, dans laquelle les images et le texte peuvent être projetés simultanément. Ce type d'approche permet de mieux comprendre la distribution des images dans la collection. La figure 3.9 montre la visualisation multimodale de [Camargo 2010]. On peut noter sur cette figure qu'il est particulièrement remarquable que certains termes similaires du texte, d'un point de vue sémantique, soient représentés en étroite relation dans l'espace latent, et donc ainsi dans la visualisation. Malheureusement toutefois, la visualisation de Camargo et al. présente pour limite le fait qu'elle ne soit pas conçue pour intégrer de l'interaction avec l'utilisateur.

FIGURE 3.10 – Visualisation multimodale d'un ensemble d'images [Camargo 2010].

Dans la section suivante, nous proposons une nouvelle démarche de visualisation multimodale qui présente pour intérêt d'intégrer de l'interaction pour la recherche multimodale d'images avec le retour de pertinence.

3.4.3 Notre visualisation pour la recherche mixte texte / images

Les méthodes multimodales sont devenues très populaires en recherche d'images et de vidéos au cours des dernières années [Snoek 2005], [Lienhart 2009], [Chandrika 2010]. La visualisation pour la recherche multimodale reste un domaine ouvert. Dans tous les travaux sur la recherche multimodale d'images présentés dans la partie précédente, la visualisation reste indépendante du formalisme de la requête. Les images sont généralement représentées dans une interface traditionnelle, très souvent une matrice rectangulaire de vignettes. Les images sont placées près de "leurs similaires" par rapport à la requête, la similarité étant la combinaison des similarités unimodales (ex. texte et image). Dans le cadre de nos travaux, un de nos objectifs est de représenter les résultats sur une interface qui permette de visualiser les relations entre les images et la requête. Par exemple, nous souhaiterions pouvoir représenter : le temps, la localisation, le texte et le contenu visuel.

Dans cette partie, nous présentons une interface qui permet de visualiser les relations entre les images et la requête composée de texte et d'images. Le temps et la localisation ne sont pas traités pour l'instant. La technique de retour de pertinence pour la recherche mixte texte/image est présenté dans la prochaine section.

3.4.3.1 Les exigences de la visualisation pour la recherche mixte texte/images

Une interface pour la recherche d'images par le contenu ou par le texte ne visualise souvent que la similarité visuelle ou la similarité textuelle entre les images résultantes et la requête. Une interface pour la recherche mixte texte/image peut être construite en suivant les mêmes principes, avec la similarité entre des images et la requête, sans considérer les composantes dans la requête mixte texte/image. Ce type d'approche peut entraîner une perte d'information pour des utilisateurs qui sont intéressés par un couplage de ces deux types de données, dans le cadre d'une requête mixte. Comme évoqué à plusieurs reprises dans ce manuscrit, notre objectif est ici de construire une interface 2D capable de visualiser les relations visuelles et conceptuelles, et par ailleurs offrant des services d'interactions pour la recherche d'images par requête mixte texte/image.

Partant de l'hypothèse que la requête mixte contient une sous-requête textuelle (mots textuels) et une sous-requête visuelle (mots visuels), les coordonnées cartésiennes 2-dimensions permettent de visualiser les relations entre les images et la requête. Par exemple, dans ce système de coordonnées, la relation visuelle peut

être projetée sur l'axe X, et la relation conceptuelle sur l'axe Y. Cependant, cette relation conceptuelle n'illustre pas les relations entre des images et chaque mot / concept de la requête, mais représente la relation entre des images et tous les mots pris simultanément. Ce type d'approche pose problème notamment lorsque la pertinence d'une image par rapport à la requête ne porte que sur quelques mots, et non sur la liste exhaustive. Dans le cadre de nos travaux, nous proposons une interface de visualisation qui peut permettre de voir non seulement la relation visuelle et la relation conceptuelle mais aussi les relations entre des images et chaque mot / concept en utilisant les coordonnées polaires (figure 3.11). Selon cette représentation, la relation entre des images résultantes et la requête textuelle (mots textuels) d'une part et les relations entre des images résultantes et la requête visuelle d'autre part sont présentées dans une interface 2D qui est décrite ci-dessous.

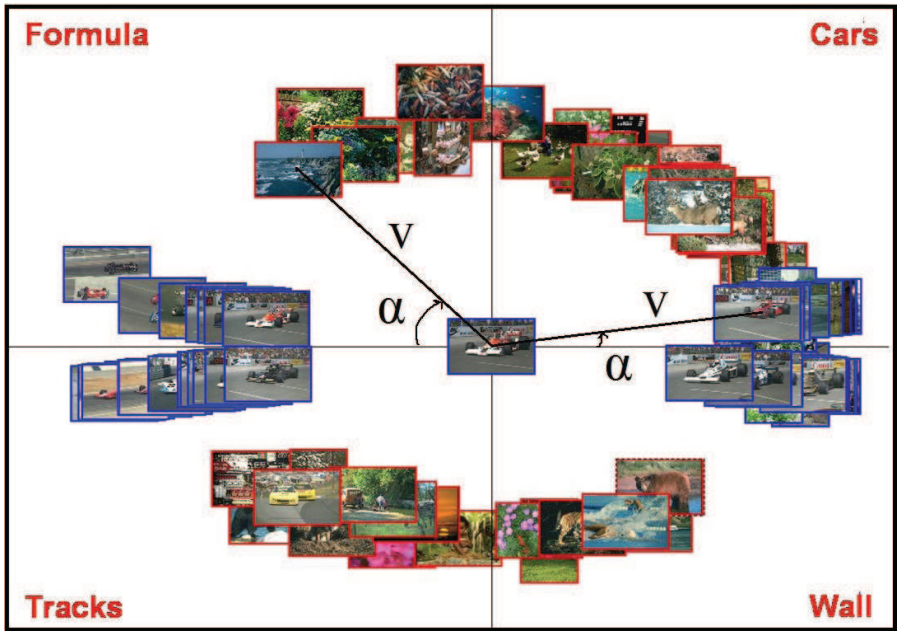

FIGURE 3.11 – La représentation en coordonnées polaires pour des requêtes visuelles et textuelles. Le vecteur V est la distance visuelle entre la requête et l'image. L'angle α est la distance entre le mot textuel et l'image. Les mots textuels sont les annotations manuelles fournies par les experts. Les images en bleu/rouge sont pertinentes/non pertinentes. Dans l'exemple ci-dessus, les images proches de l'axe des X partagent avec la requête le contenu visuel et les mots textuels. Plus on s'éloigne de l'axe des X, plus les images divergent par leurs mots textuels de la requête, la similarité du contenu visuel dépendant quant à elle de la distance, sur l'interface, par rapport à la requête.

3.4.3.2 L'interface de visualisation

Pour visualiser les relations entre des images et chaque mot/concept, nous visons à distinguer les images liées à chaque mot. Ces images sont alors affichées dans différents endroits de l'interface. Tout d'abord, une recherche multimodale d'images par une requête mixte texte/image est effectuée. Chaque image résultat a une valeur de similarité "mixte" intégrant la mixité de la requête, une valeur de similarité pour la sous-requête textuelle et une valeur de similarité pour la sous-requête visuelle. Les similarités conceptuelles entre les images et chaque mot sont également caculées. Les N (typiquement 100) premières images du résultat en terme de la similarité mixte sont présentées dans cette représentation. Nous proposons l'interface pour une requête mixte qui inclut au maximum 4 mots-clés, cette limite étant discutée dans les perspectives (chapitre 6). Le principe que nous avons retenu est de diviser l'interface 2D en 4 quadrants, qui correspondent aux 4 mots-clés de la sous-requête textuelle. Les images pertinentes pour chaque mot sont affichées dans chacun des 4 quadrants de l'interface en se basant sur les similarités conceptuelles.

La figure 3.11 montre le résultat de la recherche par une requête mixte texte/image qui comprend une image exemple avec ses 4 mots textuels. Le cosinus de l'angle α ($\cos(\alpha)$) de chaque image avec l'axe horizontal est la similarité textuelle entre l'image et la sous-requête textuelle. D'autre part, la distance euclidienne (rayon polaire) entre une image et l'origine (0,0) (le vecteur V) représente la similarité visuelle entre l'image et la sous-requête visuelle (orientée vers les mots visuels). L'intérêt de cette visualisation est d'illustrer en détails les relations visuelles et conceptuelles entres les images et la requête.

Outre la possibilité de visualiser la pertinence par rapport à des mots textuels et des images, cette représentation en coordonnées polaires présente pour autre avantage d'offrir des services d'interactions utilisant les groupes sur lesquelles est fondée cette représentation, sujet qui est abordé dans la section suivante.

Pour l'instant, notre visualisation est limitée en représentant jusqu'à 4 groupes images qui correspondent aux 4 mots textuels. En terme de perspectives pour ces travaux, nous souhaitons réutiliser la technique de la visualisation de graphes conceptuels (ex. Google Swirl [1]) et deux techniques de visualisation basées sur les spirales et les anneaux concentriques présentés dans [Torres 2006] pour surmonter cette limitation.

1. Google Swirl : http ://image-swirl.googlelabs.com/

3.5 Interaction pour la recherche mixte texte/image

La visualisation présentée ci-dessus peut offrir des services d'interaction pour la recherche d'images par la requête mixte texte/image. Dans la section 3.3.2, nous avons présenté notre technique de retour de pertinence pour la recherche par le contenu d'images. En profitant de la visualisation en coordonnées polaires, nous pouvons appliquer cette technique de retour de pertinence pour la recherche mixte d'images. Nous présentons ci-dessous la technique de retour de pertinence pour la recherche mixte que nous avons retenue. Nous verrons que celle-ci s'appuie sur les mêmes principes que la méthode CR présentée dans la section 3.3.2. La différence majeure avec celle présentée dans ce chapitre est liée au fait que l'étape de regroupement est effectuée par l'utilisateur, et non par des algorithmes.

3.5.1 Notre technique de retour de pertinence pour la recherche mixte texte/images

Comme évoqué dans la partie 3.2.3.4, le retour de pertinence pour la recherche multimodale (ou particulièrement, la recherche mixte texte/image) est un challenge récent. Dans cette partie, nous proposons une autre approche en utilisant notre technique de retour de pertinence pour la recherche d'images par le contenu (section 3.3).

Le retour de pertinence pour la recherche mixte est basé principalement sur le jugement de pertinence de l'utilisateur sur des images renvoyées par le système. Dans notre système, la requête mixte se compose de 2 sous-requêtes : la sous-requête visuelle qui consiste en une image exemple et la sous-requête textuelle qui consiste en des mots-clés. Le jugement de la pertinence/non pertinence des images ne se limite pas à la sous-requête visuelle, mais inclut également la sous-requête par mots-clés. Les utilisateurs peuvent notifier les images qui sont pertinentes ou non pertinentes par rapport à l'image exemple de la requête, et également celles qui sont pertinentes / non pertinentes pour les mot-clés (figure 3.1, étapes 2 et 3). Cela signifie que les images pertinentes / non pertinentes sont regroupées dans des ensembles différents correspondant aux mots-clés fournis.

Notre technique de retour de pertinence pour la recherche par le contenu de la section 3.3.2 intègre deux phases : 1) la construction des groupes locaux pertinents qui peuvent assurer l'unimodalité et 2) la construction des points de requête en se basant sur des exemples pertinents et non pertinents. Cette technique de retour de pertinence peut être appliquée sur la requête mixte en se basant sur les ensembles

d'images pertinentes et non pertinentes regroupées ci-dessus. Nous résumons notre retour de pertinence pour la recherche mixte texte/image ci-dessous.

Retour de pertinence pour la recherche mixte texte/images : D'abord, venant du retour de pertinence de l'utilisateur et de notre interface mixte, nous avons (jusqu'à) 4 ensembles d'images pertinentes et non pertinentes qui correspondent aux (jusqu'à) 4 mots textuels de la sous-requête textuelle. Les unimodalités des ensembles d'images pertinentes sont assurés parce que : 1) les images dans un ensemble sont pertinentes pour un mot (sémantiquement similaire) et 2) les images sont visuellement similaires par rapport à la sous-requête visuelle. Ensuite, la requête à points multiples est créée en se basant sur la technique de Rocchio. Ce processus est répété pour toutes les itérations jusqu'au contentement de l'utilisateur.

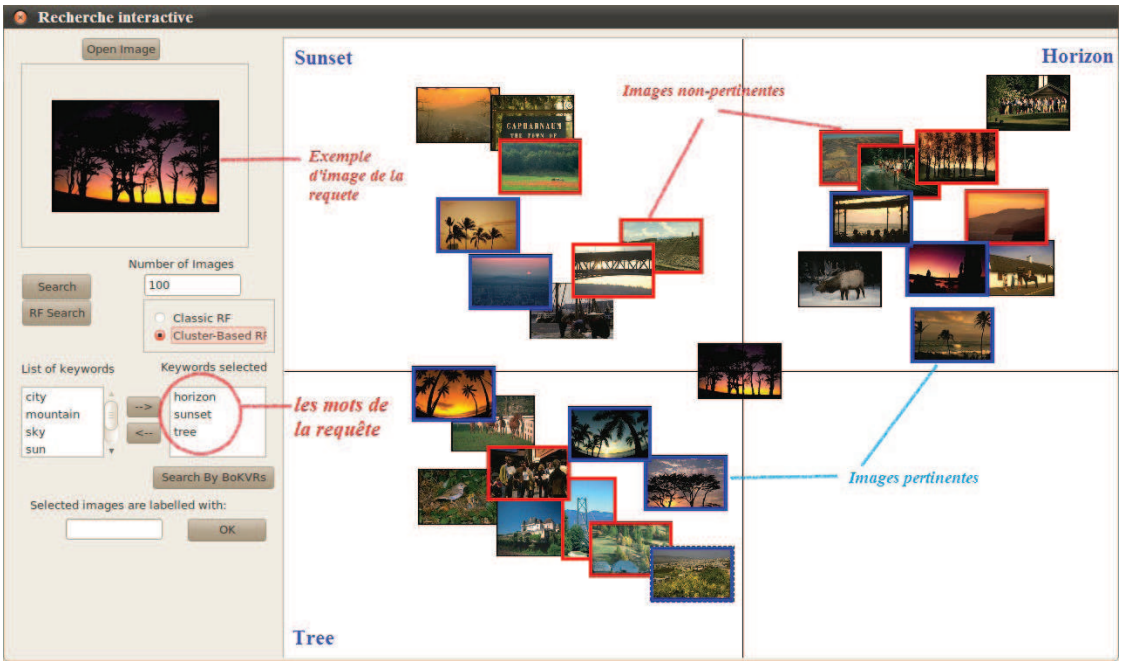

FIGURE 3.12 – La visualisation avec l'interaction de l'utilisateur pour le retour de pertinence pour la recherche mixte texte/image. A gauche, la requête mixte se compose d'un exemple d'images et de 3 concepts ("sunset", "horizon" et "tree"). Les images résultantes sont visualisées dans les 3 quadrants qui correspondent aux 3 concepts. Pendant l'interaction, l'utilisateur fournit des connaissances/informations qui sont des exemples pertinents (rectangles bleus) et des exemples non pertinents (en rouge). Le système reformule la requête et fait une nouvelle recherche pour tenter d'obtenir de meilleurs résultats.

Dans le cas de notre système, les utilisateurs doivent fournir plus d'informations que dans le cas classique, pour lequel ils fournissent uniquement des images perti-

nentes et non pertinentes par rapport à l'image exemple E. Cependant, même si l'effort cognitif de l'utilisateur est plus important (réflexion pus approfondie sur ses propres objectifs), le nombre effectif d'interactions n'est pas augmenté (en nombre de clics souris par exemple). Les ensembles d'images pertinentes / non pertinentes sont utilisés pour modifier les requêtes en utilisant la technique de Rocchio rappelée ci-dessous :

$$Q_{k,new} = \alpha Q_{k,old} + \frac{\beta}{|D_r|} \sum_{i \in D_r} \bar{I}_i - \frac{\gamma}{|D_n|} \sum_{i \in D_n} \bar{I}_i \tag{10}$$

où D_r et D_n sont respectivement l'ensemble des images pertinentes et l'ensemble des images non pertinentes pour le mot-clé k et l'image exemple E. α=β= γ=1. À la première itération du retour de pertinence, $Q_{k,old}$ est la requête initiale. $\bar{I}_i$ est le vecteur de caractéristiques de l'image i.

3.6 Evaluation

Nous avons présenté dans ce chapitre notre contribution sur le retour de pertinence pour la recherche d'images par le contenu avec deux variantes. Ces méthodes sont basées sur une combinaison des deux techniques populaires : mouvement de requête et extension de requêtes (section 3.3). Le principe de notre approche est d'éviter les problèmes liés au mouvement de requête et à l'extension de requêtes pour améliorer le résultat de la recherche d'images. Nous avons également proposé une visualisation pour la recherche mixte texte/contenu offrant des possibilités de visualiser les relations visuelles / conceptuelles des images et la requête (section 3.4). De plus, les relations conceptuelles sont visualisées en détails sur chaque mot-clé. En profitant de cette visualisation, nous pouvons appliquer notre technique de retour de pertinence pour la recherche mixte texte/image (section 3.5). Cette approche donne un bon outil pour améliorer la performance de la recherche d'images par la requête mixte texte/image. Dans cette section, nous présentons nos expérimentations pour évaluer nos méthodes pour le retour de pertinence.

3.6.1 Protocole d'expérimentation

Dans notre expérimentation, l'interface de visualisation n'est pas évaluée. Plusieurs raisons expliquent le fait que notre visualisation n'a pas pu être évaluée de manière suffisament pertinente : tout d'abord, il existe une très grande dépendance entre les intérêts des utilisateurs et leur environnement applicatif de recherche. La recherche d'images de portraits peut par exemple être très différente d'une re-

cherche d'images de paysages, infuençant considérablement la perception des interfaces de visualisation. D'autre part, le caractère extrêmement chronophage de l'expérimentation (vérité terrain, nombre d'utilisateurs significatifs, ...) est également un paramètre important pour justifier notre absence d'évaluation, considérant les moyens dont nous disposons pour notre système et nos échéances temporelles. Nous discuterons toutefois de notre l'interface de visualisation dans la partie perspectives du chapitre 6. Considérant ce contexte, pour notre expérimentation, les interactions utilisateurs sont simulées par des connaissances externes correspondant aux annotations manuelles de la base Corel 30K. La méthode de simulation des interactions est discutée beaucoup plus en détails dans le chapitre 5.

Trois méthodes de retour de pertinence sont évaluées en utilisant la recherche d'images par le contenu dans cette expérimentation : le mouvement de requête, l'extension de requêtes et notre nouvelle méthode combinant les deux techniques avec ses deux variantes que sont le regroupement-répétition (CR - *Clustering-Repeat*) et le regroupement-non-répétition (CNR - *Clustering-Non-Repeat*).

3.6.1.1 Base d'experimentation

Comme mentionné préalablement, nous appuyons notre expérience sur un sous-ensemble de la base Corel, plus précisément la base de données Corel 30K [Carneiro 2007]. Cette base de données contient 30000 images divisées en différentes catégories par des experts du domaine et il y a 100 images dans chaque classe. La taille de chaque image est de 384×256 pixels.

Pour l'expérimentation, nous nous appuyons en fait sur une simulation des interactions humaines, en utilisant les données déjà présentes dans la base Corel, jouant un rôle un peu équivalent à celui d'un humain. Une technique de retour de pseudo-pertinence est utilisée pour simuler automatiquement les interactions humaines de retour de pertinence. Notre démarche s'appuie donc sur l'utilisation des données autre que images de la base Corel, pour laquelle différentes possibilités existent pour spécifier une vérité terrain pour la validation.

3.6.1.2 Discussion des protocoles utilisés dans d'autres systèmes : cas de MARS et QCluster

Plusieurs travaux ont utilisé la base de données Corel, mais la majeure partie d'entre eux utilisent seulement quelques catégories (10, 20 ou 50) sans fournir aucune motivation pour cela [Huiskes 2008]. En outre, le nombre de requêtes effectuées est faible (100 pour Mars, QCLUSTER...). Ces choix peuvent grandement influencer l'évaluation des résultats.

Dans le système MARS [Ortega Binderberger 2004], les images pertinentes d'une requête image sont sélectionnées comme suit. Ils sélectionnent une requête image Q au hasard parmi l'ensemble des données et recherchent les 50 premières images résultats. Cet ensemble de 50 images est dénommé l'ensemble $relevant(Q)$. Ils construisent ensuite la requête en se déplacant autour de Q (cette requête est proche de Q dans l'espace des caractéristiques). On considère alors Q comme la requête idéale (soit celle qui représente le plus fidèlement possible les intentions de l'utilisateur). Des requêtes sont choisies autours de Q en espérant que celles-ci atteindront la requête idéale Q (par retour de pertinence). Ensuite, ils recherchent les 100 premières images résultats, qui deviennent l'ensemble $retrieved(Q)$. Dans MARS, la précision et le rappel sont calculés en utilisant les ensembles $relevant(Q)$ et $retrieved(Q)$ par les formules classiques rappelées ci-dessous :

$$precison = \frac{relevant(Q) \bigcap retrieved(Q)}{retrieved(Q)}$$

$$rappel = \frac{relevant(Q) \bigcap retrieved(Q)}{relevant(Q)}$$

Dans le système MARS [Ortega Binderberger 2004], l'ensemble pertinent est sélectionné en assurant l'unimodalité car toutes les images sont visuellement similaires à une image requête. Les auteurs présument que toutes les images pertinentes forment un ensemble unimodal, hypothèse pas tout à fait réaliste, engendrant implicitement une limite de l'approche. De plus, ces travaux appuient toutes les mesures sur des moyennes sur une base de 100 requêtes, ce nombre étant très faible par rapport au nombre d'images dans la base.

Dans un autre exemple (le système QCluster [Kim 2005]), la vérité terrain est relativement simple, car ils utilisent l'information de catégorie de haut niveau dans la base Corel comme vérité terrain pour simuler le retour de pertinence. Les images de la même catégorie sont considérées comme les plus pertinentes et les images des catégories apparentées (comme les fleurs et les plantes) sont considérées comme pertinentes. Cette hypothèse crée une condition facile pour le retour de pertinence, parce que le nombre d'images pertinentes est plus élevé par rapport autres approches (ex. MARS [Ortega Binderberger 2004]), expliquant la bonne qualité (subjective) des résultats du système QCluster.

3.6.1.3 Protocole pour notre expérimentation

Pour notre expérience, nous considérons comme vérité terrain la classe des images, ce qui peut produire une grande diversité de classes, mais qui nous semble représentative de la vie réelle. Nous mesurons la qualité de la recherche avec les

critères classiques de rappel/précision en récupérant les 100 premières réponses (nous posons l'hypothèse que l'utilisateur ne peut voir que les 100 premiers résultats sur l'interface de l'écran). Les mesures présentées sont des moyennes calculées sur un grand nombre de requêtes (5000), qui sont effectuées parmi les 30000 images de la base de données Corel pour notre expérimentation.

Une des caractéristiques du retour de pertinence est la question du nombre d'échantillons. Le nombre d'exemples d'apprentissage est normalement faible. Dans nos expériences, nous nous appuyons sur l'hypothèse qu'un maximum de 20 images peut être sélectionné par l'utilisateur. En fait ces images sont les P premiers exemples pertinents et les N premiers exemples non pertinents dans les 100 premières réponses, où $P + N \leq 20$. Ces exemples sont automatiquement retournées par le système en utilisant la vérité terrain. Nous proposons 2 stratégies pour le nombre d'exemples :

1. 10 exemples pertinents, 10 exemples non pertinents dans le cas du mouvement de requête, CR et CNR. Et 20 exemples pertinents dans le cas de l'extension de requêtes. L'extension de requêtes n'utilise pas d'exemples non pertinents parce que cette méthode tente de regrouper les exemples pertinents pour former la requête à points multiples.
2. 5 exemples pertinents, 5 exemples non pertinents dans le cas du mouvement de requête, CR et CNR. Et 10 exemples pertinents dans le cas de l'extension de requêtes.

3.6.2 Résultats et discussion

Dans cette section, les 4 techniques de retour de pertinence sont comparés suivant le protocole décrit ci-dessus. Comme evoqué ci-dessus, nous calculons les critères classiques de rappel/précision en récupérant les 100 premières réponses (nous posons l'hypothèse que l'utilisateur ne peut voir que les 100 premiers résultats sur l'interface de l'écran et parce que la précision des images récupérées qui sont présentées aux utilisateurs est la plus importante). Comme le nombre d'images de chaque classe est 100 (donc le nombre d'exemples pertinents est égal au nombre d'exemples récupérés), le rappel pour les 100 premières images récupérées est égal à la précision :

$rappel = \frac{relevant \bigcap retrieved}{relevant}$ et $precision = \frac{relevant \bigcap retrieved}{retrieved}$

Dans le cas d'expériences à base de 10 exemples d'images (figure 3.13), nos techniques de retour de pertinence basées sur les groupes sont meilleures que celles de l'extension de requêtes et du mouvement de requête. La méthode CNR est légèrement meilleure que la méthode CR. Après deux itérations du retour de pertinence,

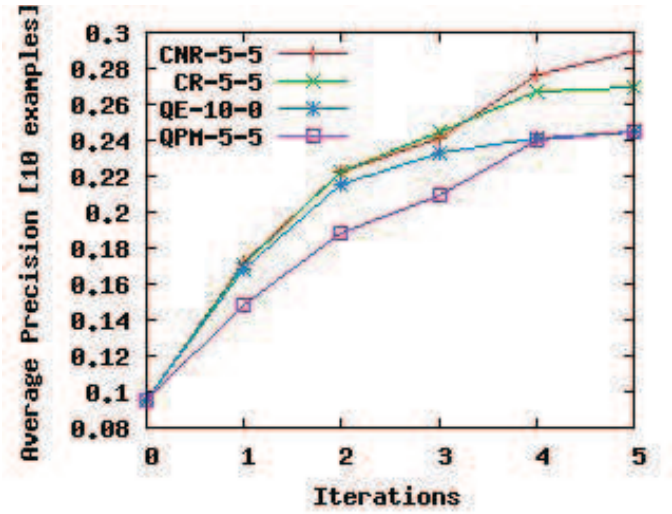

FIGURE 3.13 – QE signifie la technique de l'extension de requêtes (*Query expansion*), QPM signifie la technique du mouvement de requête (*Query point movement*), CR et CNR sont les méthodes Regroupement-répétition (*Clustering-Repeat*) et Regroupement-non-répétition (*Clustering-Non-Repeat*) que nous avons décrites précédemment. Cette figure illustre les précisions moyennes pour les 100 premières images récupérées de 4 techniques de retour de pertinence avec 10 exemples retournés pour chaque itération. Les deux techniques CNR et CR montrent une très bonne performance comparées aux méthodes de modification de la requête.

le mouvement de requête a la pire performance, les trois autres méthodes ayant des performances équivalentes. Lors des itérations suivantes, les deux méthodes CR et CNR deviennent meilleures que les techniques classiques. La précision moyenne des techniques classiques est environ de 0,244 après 5 itérations tandis que la méthode CNR a une précision moyenne de 0,288 et la méthode CR a une précision moyenne de 0,279. L'amélioration en terme de précision de notre méthode par rapport aux méthodes classiques est donc de 18% selon ces résultats.

Dans le cas d'expériences avec 20 images de retour (figure 3.14), la méthode CNR surpasse largement toutes les autres méthodes. Les deux techniques basées sur les groupes ont de meilleures précisions pour les premières itérations, mais la précision de la technique CR n'est pas meilleure que le mouvement de requête pour les itérations suivantes. Dans ce cas, l'extension de requêtes donne la pire performance, le mouvement de requête et la méthode CR ont la même performance, avec des précisions moyennes d'environ 0,305. La méthode CNR a la meilleure précision moyenne avec 0,39. L'amélioration en terme de précision de la méthode CNR par rapport aux méthodes classiques est donc de 28% dans cette expérimentation.

Les figures 3.15 et 3.16 montrent la précision des 4 méthodes pour les 50 premières images récupérées. La méthode CNR et la méthode CR donnent toujours

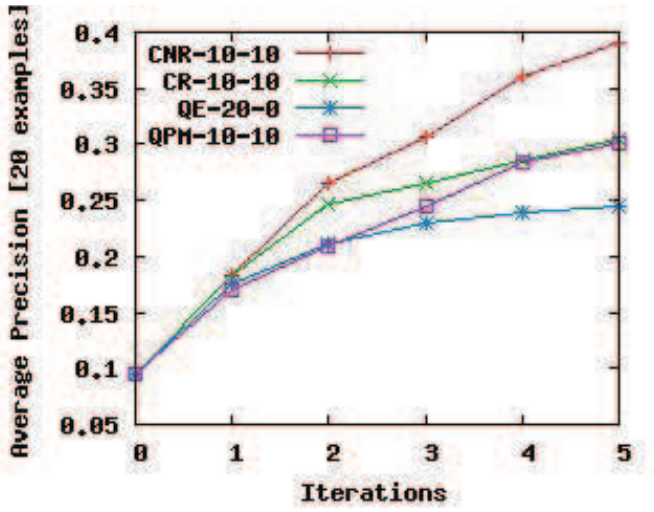

FIGURE 3.14 – QE signifie la technique de l'extension de requêtes (*Query expansion*), QPM signifie la technique du mouvement de requête (*Query point movement*), CR et CNR sont les méthodes Regroupement-répétition (*Clustering-Repeat*) et Regroupement-non-répétition (*Clustering-Non-Repeat*) que nous avons décrites précédemment. Cette figure illustre les précisions moyennes pour les 100 premières images extraites de 4 techniques de pertinence avec 20 exemples de retour de pertinence pour 1 itération. La méthode CNR donne le meilleur résultat.

de meilleurs résultats. Dans la plupart des cas, la méthode CNR donne le meilleur résultat.

Nos méthodes (les deux variantes) donnent de meilleurs résultats par rapport aux techniques de modification de requête utilisées dans MARS [Ortega Binderberger 2004] (mouvement de requête). Notre méthode procure également une nette amélioration en terme de précision moyenne par rapport à QCluster [Kim 2005] (expansion de requête). Notre méthode montre des améliorations de 18% et 28% (avec respectivement 10 et 20 exemples de retour de pertinence dans les premières 100 images extraites) comparée aux techniques classiques. Qcluster a une amélioration de 20% par rapport aux techniques classiques, mais pour cette approche, le nombre d'exemples est le nombre maximum d'images pertinentes dans les 100 premières images de résultat. Ce nombre est plus grand que le nombre d'exemples proposé dans notre système (et nos 2 méthodes - 20 maximum). En réalité, la démarche proposée par QCluster nous semble irréaliste du point de vue des usages, car il est inacceptable de demander trop d'interactions à l'utilisateur. Un système qui demande 20 interactions à l'utilisateur nous semble plus réaliste par rapport à celui qui en exige 100. En outre, Qcluster et MARS sont évalués sur seulement 100 requêtes et leurs vérités terrain sont choisies uniquement pour leurs propres méthodes. Notre méthode est évaluée sur un nombre de 5000 requêtes (en calculant la moynenne) qui assure donc bien plus de généricité que QCluster et MARS.

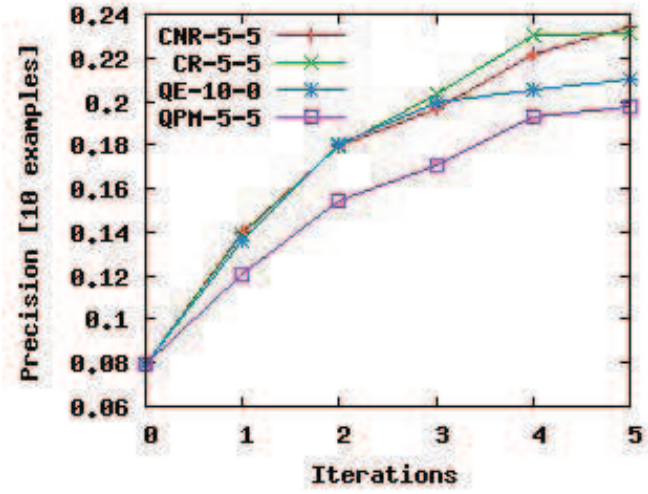

FIGURE 3.15 – Les précisions moyennes pour les 50 premières images récupérées des 4 techniques de retour de pertinence avec 10 exemples de retour pour chaque itération.

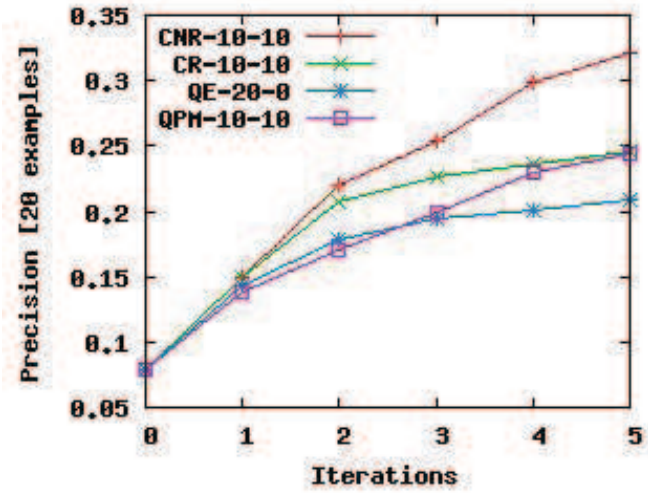

FIGURE 3.16 – Les précisions moyennes pour les 50 premières images récupérées des 4 techniques de retour de pertinence avec 20 exemples de retour pour chaque itération.

3.7 Conclusion

Dans ce chapitre, nous avons présenté différents types d'interaction dans les systèmes de recherche d'images : l'exploration, le retour de pertinence à court terme et le retour de pertinence à long terme. Nous avons proposé une nouvelle technique de retour de pertinence à court terme appelé retour de pertinence basé sur les groupes. Notre méthode est inspirée de deux techniques existantes de retour de pertinence : le mouvement de requête et l'extension de requêtes. En profitant

des images non pertinentes et des avantages de ces deux techniques classiques, notre méthode donne un meilleur résultat. Nous avons discuté du fait que notre méthode de retour de pertinence est bien utilisable dans le cas de la recherche mixte texte/image.

Le retour de pertinence basée sur les groupes est proposé avec deux approches différentes : le regroupement-répétition (CR - *Clustering-Repeat*) et le regroupement-non-répétition (CNR - *Clustering-Non-Repeat*). Ces deux approches combinent les deux méthodes de modification de requête que sont le mouvement de requête et l'extension de requêtes, et profitent des exemples non pertinents. Dans tous les cas, la méthode du regroupement-non-répétition (CNR) donne le meilleur résultat. La méthode du regroupement-répétition (CR) donne de bons résultats lorsque le nombre d'exemples de retour est faible. Notre méthode ne nécessite pas de calcul complexe, mais offre de très nettes améliorations en terme de précision par rapport aux techniques traditionnelles.

Dans le chapitre suivant, nous présentons un modèle pour l'apprentissage des connaissances basé sur notre retour de pertinence basé sur les groupes. Cet apprentissage de connaissances est considéré comme un retour de pertinence à long terme, où les connaissances sont apprises et mémorisées dans le système pour une utilisation à long terme. Dans ces chapitres, la technique de l'exploration n'est pas présentée. Nous allons l'aborder dans les perspectives de ce document.

Chapitre 4

Modèle Sacs de KVR pour l'apprentissage de connaissances

Dans ce chapitre, nous présentons notre modèle *Sac de KVR* (Keyword Visual Representation) pour l'apprentissage de connaissances. Un KVR est une représentation visuelle possible pour un mot. Cet apprentissage est basé sur l'association du contenu visuel des images avec des concepts textuels. Dans la première section, l'état de l'art de l'association texte/image est présenté. Ensuite, nous présentons en détails de notre modèle Sac de KVR et des algorithmes associés. Enfin, des utilisations du modèle pour l'annotation d'images et la recherche d'images sont présentées avec quelques discussions sur les résultats obtenus. Notre contribution dans ce chapitre correspond à la couche d'apprentissage (en bleu dans la figure 4.1 montrant l'architecture du système).

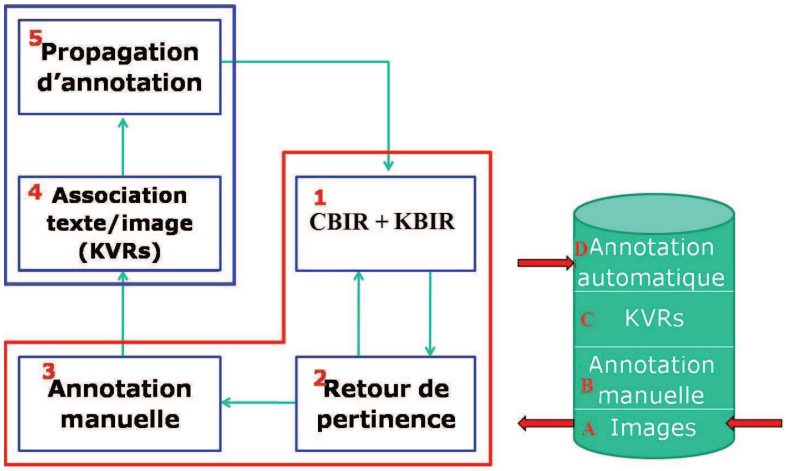

FIGURE 4.1 – L'architecture de notre système : l'apprentissage de connaissances (en bleu).

4.1 État de l'art de l'association du texte et du contenu

Nous avons parlé dans la section 2.5 des problématiques générales liées aux méthodes d'annotation d'images. L'annotation d'images est généralement basée sur un modèle d'annotation qui apprend les relations entre l'image ou les régions de l'image et des concepts sémantiques. Apprendre les relations entre les régions de l'image et les concepts sémantiques (mots) est un problème intéressant lié à la fouille de données multimodales. Cette problématique est notamment intéressante parce qu'elle peut être utilisée pour l'annotation d'images mais également pour la recherche d'images multimodale. Dans cette partie, nous présentons des modèles pour l'association entre les concepts sémantiques et les caractéristiques visuelles qui sont étudiés depuis longtemps pour la fouille de données multimodales.

4.1.1 Association du texte et du contenu

Dans cette section, nous présentons d'abord des modèles de l'état de l'art pour l'association du texte et du contenu. Ensuite nous parlons des difficultés et des problématiques résiduelles. Notre proposition d'un modèle d'association du texte et du contenu est discutée en fin de cette partie.

4.1.1.1 Modèle de cooccurrences

Le modèle de cooccurrences proposé par Mori, Takahashi and Oka [Mori 1999] représente une première approche pour l'association entre texte et contenu. Tout d'abord, les images de la base d'apprentissage sont divisées en régions qui héritent de tous les mots textuels des images originales dont elles dépendent. Des descripteurs visuels sont alors extraits à partir de chaque région. Tous les descripteurs sont regroupés en un certain nombre de groupes, dont chacun est représenté par son centre de gravité. Généralement c groupes sont créés par quantification vectorielle, et les probabilités (probabilité conditionnelle) $P(w|c_j), (j = 1, 2, \ldots, W)$ pour chaque mot w et chaque c sont estimées à partir de statistiques liées à leur fréquence d'apparition :

$$P(w_i|c_j) = \frac{P(c_j|w_i)P(w_i)}{\sum_{k=1}^{W} P(c_j|wk)P(w_k)}$$

$$= \frac{(m_{ji}/n_i)(n_i/N)}{\sum_{k=1}^{W}(m_{jk}/n_k)(n_k/N)}$$

$$= \frac{m_{ji}}{\sum_{k=1}^{W} m_{jk}} = \frac{m_{ji}}{M_j} \tag{1}$$

où $m_{i,j}$ est le nombre total de mot w_i dans le groupe c_j, M_j désigne le nombre total de mots dans le groupe c_j, n_i le nombre total des mot w_i dans toutes les images, et N est le nombre total de mots pour toutes les images $N = \sum_{i=1}^{W} n_i$.

Ce modèle est utilisé pour l'annotation d'images. Étant donné une image I non annotée, il consiste à diviser l'image en régions, à en extraire des caractéristiques comme évoqué précédemment, et à trouver le plus proche cluster pour chaque région. Lors de cette étape, la probabilité de chaque mot relatif à chacune des régions peut être mesurée. Les mots présentant les probabilités moyennes les plus fortes sur toutes les régions d'une image de test sont choisis comme les prédictions. Mathématiquement, cela peut être formalisé comme suit :

$$p(w|I) = \frac{1}{|I|} \sum_{r \in I} (p(w|c_r) \tag{2}$$

où $p(w|I)$ est la probabilité moyenne de w étant donné une image I, c_r est le plus proche cluster d'une région r de l'image I, et $|I|$ est le nombre de régions de l'image I.

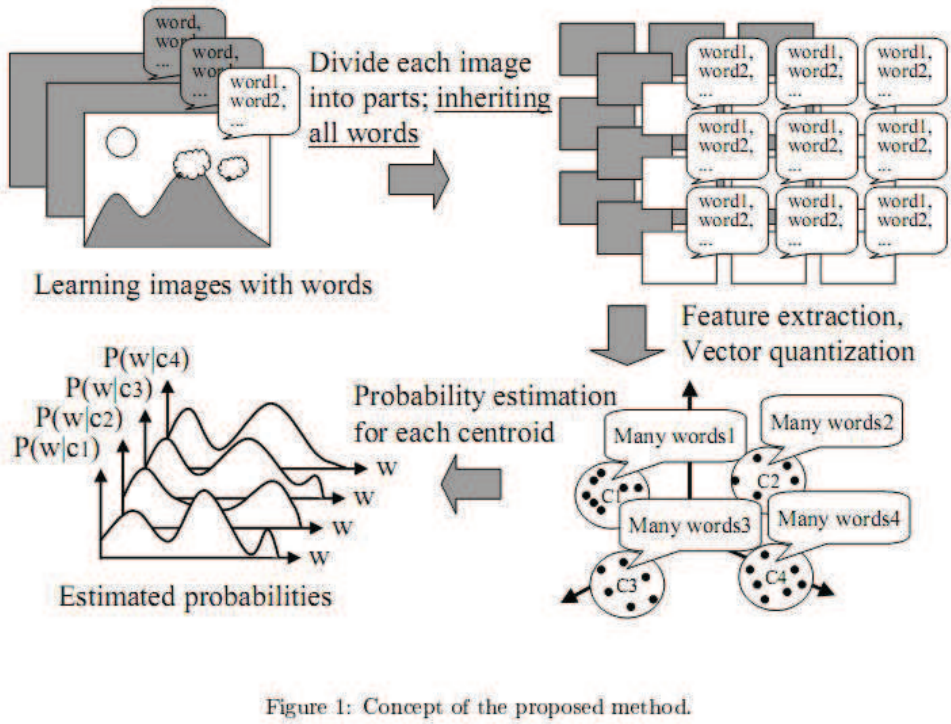

Figure 1: Concept of the proposed method.

FIGURE 4.2 – Le modèle de cooccurences de Mori et al. [Mori 1999].

4.1.1.2 Modèles de traduction

Duygule et al. [Duygulu 2002] ont proposé un modèle de traduction pour représenter la relation entre texte et contenu. Selon leur point de vue, la caractéristique visuelle et le texte sont deux langages (ex. régions de l'image = français et concepts = anglais) que l'on peut traduire de l'un à l'autre.

Ils ont d'abord utilisé un algorithme de segmentation pour segmenter des images en régions. Ensuite, la quantification de caractéristiques est appliquée aux descripteurs qui sont extraits de toutes les régions, pour construire un vocabulaire visuel appelé "blobs". Un blob est en fait un représentant d'un groupe de régions d'images visuellement similaires. Enfin, un modèle de traduction automatique qui a été initialement proposé pour la traduction linguistique est adapté pour construire un "lexique", une table de traduction contenant les estimations de probabilité de la traduction entre les régions de l'image et les mots textuels. Une image est alors annotée en choisissant le mot le plus probable pour chacune de ses régions.

Barnard et al. [Barnard 2003] ont étendu le modèle de traduction de Duygule [Duygulu 2002] au modèle hiérarchique. Ce modèle combine le modèle "aspect" [Hofmann 1998] avec un modèle de "regroupement doux" (*soft clustering model*). Comme illustré à la figure 4.3, les images et les textes sont générés par des nœuds disposés dans une structure arborescente. Les nœuds génèrent des régions d'images à l'aide d'une distribution gaussienne, et les mots en utilisant une distribution multinomiale. Chaque groupe est associé à un chemin d'accès à partir d'une feuille à la racine. Les nœuds près de la racine sont partagés en de nombreux groupes, et les nœuds plus près des feuilles sont partagés en quelques groupes.

4.1.1.3 Modèle de pertinence cross-média

Jeon et al. [Jeon 2003] ont proposé des améliorations aux résultats de Duygulu [Duygulu 2002] en introduisant un modèle de génération de langage, dénommé le modèle de pertinence cross-média (*CMRM-Cross Media Relevance Model*). Tout d'abord, ils utilisent le même processus que Duygulu [Duygulu 2002] pour calculer la représentation d'images (représentation par blobs). La différence par rapport aux travaux de Duygulu et al. réside dans le fait qu'ils font l'hypothèse qu'il existe une correspondance d'un à un entre les régions et les mots, alors que Jeon et al. [Jeon 2003] supposent qu'un ensemble de blobs est lié uniquement à un ensemble de mots. Ainsi, au lieu de chercher une table de traduction probabiliste, CMRM rapproche tout simplement la probabilité d'observer un ensemble de blobs et de mots dans une image donnée. Pour une image non annotée I, on suppose qu'il existe une distribution de probabilité notée $P(I)$ de tous les blobs et des mots possibles qui pourrait apparaître dans l'image I. Si la représentation des blobs de l'image

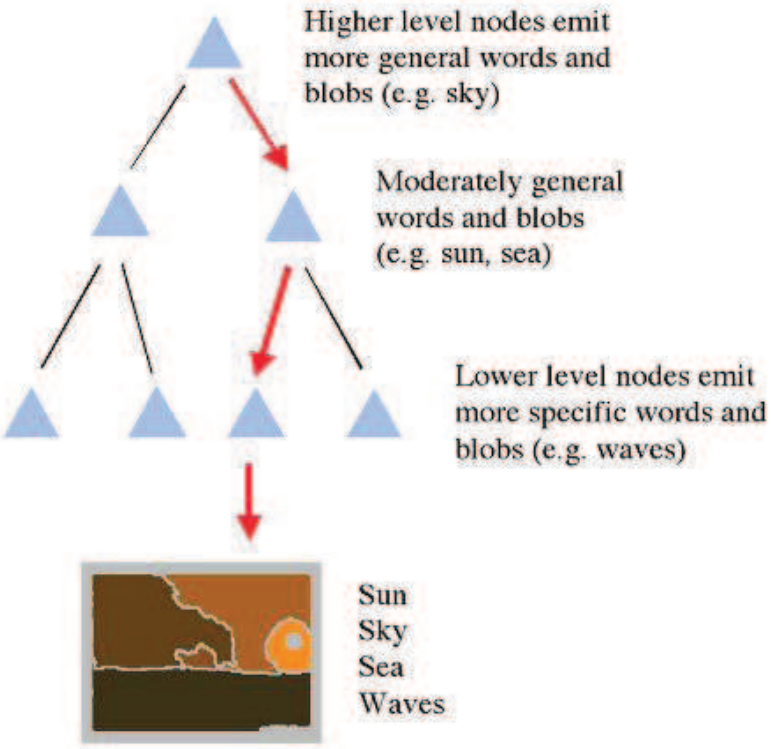

FIGURE 4.3 – Extension du modèle de traduction : modèle hiérarchique de Barnard et al. [Barnard 2003].

$I = (b_1, \ldots, b_m)$, et m est le nombre de blobs dans I, la probabilité d'observer un mot w est estimée de la façon suivante :

$$P(w|I)\ P(w|b_1, \ldots, b_m) \tag{3}$$

Le calcul de $P(w|b_1, \ldots, b_m)$ est équivalent à calculer la probabilité conjointe $P(w, b_1, \ldots, b_m)$, qui est estimée comme l'espérance sur l'ensemble d'apprentissage. Dans l'hypothèse où les mots et les blobs sont produits indépendamment d'une image donnée d'apprentissage J, $P(w, b_1, \ldots, b_m)$ est calculée comme suit :

$$P(w, b_1, \ldots, b_m) = \sum_{J \in T} P(J) P(w|J) \prod_{i=1}^{m} P(b_i|J) \tag{4}$$

où T est l'ensemble d'apprentissage. Les probabilités $P(J)$ sont conservées uniformes sur toutes les images d'apprentissage, tandis que $P(w|J)$ et $P(b_i|J)$ sont estimés par maximum de la probabilité.

Lavrenko et al. [Lavrenko 2004] ont fait valoir que le processus de quantification

des caractéristiques de l'image en continu (CRM - *Continuous Relevance Model*), pour éviter la perte d'informations liées à la production du dictionnaire dans le modèle CMRM [Duygulu 2002]. En utilisant des caractéristiques continues de densités de probabilités pour estimer la probabilité d'observer une région donnée dans une image, ils ont montré que la performance du modèle sur le même ensemble de données est beaucoup plus efficace que les modèles proposés par Duygulu et al. [Duygulu 2002] et par [Jeon 2003].

4.1.1.4 Modèle LSA (Latent Semantic Analysis)

Quelques travaux ont tenté d'utiliser la technique LSA pour combiner les caractéristiques visuelles et textuelles parmi lesquels ceux de Monay et Gatica-Perez [Monay 2003] qui ont appliqué l'analyse sémantique probabiliste latente (*PLSA-Probabilistic Latent Semantic Analysis*) [Hofmann 1998] pour l'annotation automatique d'images. Selon cette approche, le texte et les caractéristiques visuelles sont considérés comme des termes. Une technique de fusion précoce (*early fusion*, voir la partie 2.4.1) est utilisée pour représenter l'image par les termes. Elle repose sur l'hypothèse que chaque terme peut provenir d'un certain nombre de sujets latents et que chaque image peut contenir plusieurs sujets.

4.1.1.5 Modèle de transformation

Dans [Lin 2007], la requête textuelle est transformée automatiquement en des représentations visuelles. Tout d'abord, les relations entre texte et images sont extraites d'un ensemble d'images annotées avec des descriptions textuelles. Un dictionnaire trans-média qui est semblable à un dictionnaire bilingue est mis en place dans la base d'apprentissage.

Les auteurs considèrent les blobs [Carson 2002] comme une représentation visuelle d'images. Blobworld [Carson 2002] est utilisé pour segmenter une image en régions. Blobworld regroupe en régions les pixels d'une image qui sont cohérents par rapport à des caractéristiques de bas niveau telles que la couleur ou la texture, et qui correspondent à des objets ou à des parties d'objets. Pour chaque région, un ensemble de caractéristiques basé sur la couleur, la texture, la forme, la position et la taille est extrait. Les régions de toutes les images sont regroupées par une méthode de regroupement. Chaque groupe se voit affecter un numéro unique, c'est-à-dire un jeton blob (un code du blob, comme un *codeword* dans un vocabulaire des groupes de régions) et chaque image est représentée par des jetons blobs. En tenant compte de la description textuelle et des jetons blobs d'images, les auteurs font l'apprentissage de la corrélation entre les informations textuelles et visuelles. Un modèle d'information mutuelle (MI) est adopté pour mesurer la corrélation

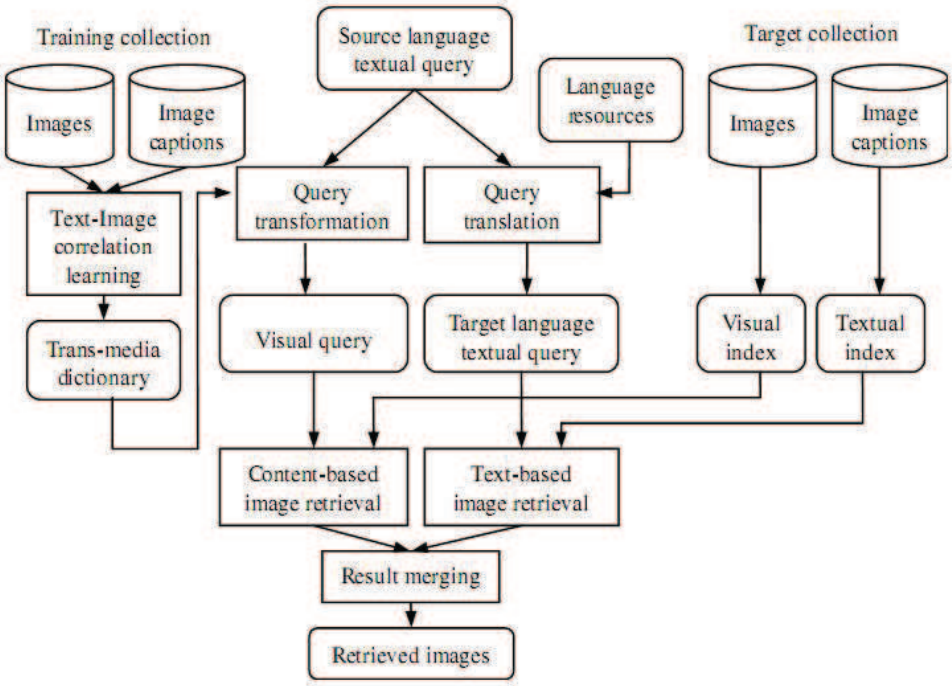

FIGURE 4.4 – La recherche d'images multilingues [Lin 2007].

entre un blob d'image et un mot. Soit x un mot et y un blob d'image, le modèle d'information mutuelle MI de x et y est défini comme suit :

$$MI(x,y) = p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{6}$$

où $p(x)$ est la probabilité d'occurrence du mot x dans les descriptions de texte, $p(y)$ est la probabilité d'occurrence du blob y dans les blobs d'images et $p(x,y)$ est la probabilité que x et y coexistent dans une description de l'image.

Les MI entre les mots et les blobs sont alors calculés et on peut générer des blobs liés à un mot w_i. Les blobs dont les valeurs MI liées à w_i dépassent un certain seuil sont associés à w_i. Les blobs produits peuvent être considérés comme la représentation visuelle de w_i. De cette façon, un dictionnaire trans-média "mots-blobs" est mis en place.

Les auteurs dans [Chang 2008] proposent également de faire l'inverse, c'est-à-dire de traduire une requête image en requête textuelle. Partant des requêtes à la fois textuelles et visuelles, ils transforment les requêtes visuelles en textuelles, et obtiennent de nouvelles requêtes textuelles. Ensuite, ils appliquent des techniques textuelles pour traiter les requêtes textuelles initiales et les nouvelles requêtes textuelles construites à partir des requêtes visuelles (voir figure 4.5). Enfin,

ils fusionnent les résultats obtenus.

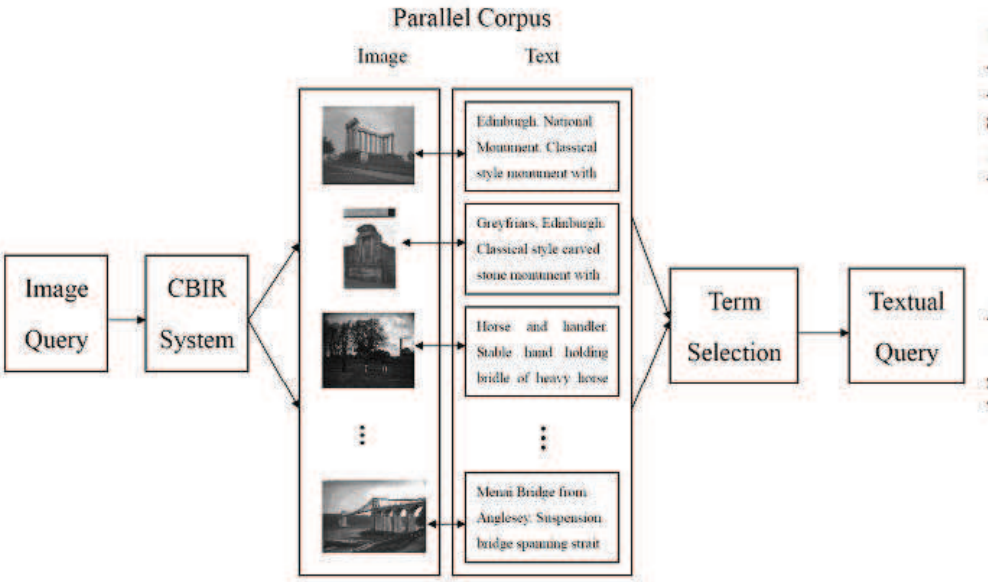

FIGURE 4.5 – Traduire une requête image en requête textuelle [Chang 2008].

4.1.2 Verrous et difficultés de l'existant de la littérature

La segmentation d'images Le but de la segmentation d'images est de diviser l'image en régions saillantes, comme par exemple les régions correspondantes aux surfaces individuelles, aux objets ou à des parties d'objets naturels. C'est une tâche très difficile dans le domaine du traitement d'images. Il n'y a pas de solution générale au problème de la segmentation d'images. Il faut souvent combiner avec des connaissances du domaine afin de résoudre efficacement un problème de segmentation d'images pour un domaine donné.

La plupart des modèles ci-dessus utilisent la segmentation d'images (modèle de cooccurrences, modèle de traduction, modèle de pertinence cross-média, modèle de transformation). La performance dépend donc de la segmentation d'images qui est une opération extrêmement délicate en général.

Indisponibilité de la transformation de texte en images La transformation de texte en images est très utile dans le cas de recherche d'images non annotées par une requête textuelle. Nous pouvons rechercher une base d'images non annotée ou annotée partiellement en utilisant une requête textuelle si nous la transformons en une requête visuelle. La plupart des méthodes d'association texte/contenu ont pour objectif de transformer des caractéristiques visuelles de bas niveau en des mots

textuels pour l'annotation d'images. Seul le modèle de transformation [Lin 2007] propose la transformation de texte en images. Ce modèle présente une représentation visuelle des mots textuels en utilisant l'information mutuelle entre texte et blobs, offrant la possibilité d'annoter des images et de rechercher des images par le texte. Toutefois, ce modèle a pour inconvénient la problématique de la segmentation des images.

Toutes les méthodes statistiques ci-dessus (modèle de cooccurrences, modèle de traduction, modèle de pertinence cross-média...) utilisent la probabilité conjointe d'une image et d'un ensemble de mots, la probabilité des mots dans une image donnée, ou la probabilité des mots étant donnée une région d'image spécifique. Ces probabilités sont utilisées pour annoter des images par des mots, ou en d'autres termes, pour transformer des images en texte.

Par rapport aux méthodes statistiques, les méthodes basées sur les espaces vectoriels (modèle LSA) donnent une vision plus claire de la façon dont les images, les zones d'images et les mots sont liés les uns aux autres, en fonction de leur position dans l'espace des caractéristiques. Toutefois, ces méthodes ne donnent pas les relations entre texte et contenu mais seulement des représentations d'images combinées par le texte et le contenu visuel. Ces méthodes sont utilisées souvent pour l'annotation d'images. Quelques méthodes proposent des approches de recherche d'images en réduisant la représentation d'images à des mots textuels et en appliquant des approches de recherche de texte standard.

Toutes ces méthodes statistiques et méthodes basées sur les espaces vectoriels sont utilisées pour transformer des images en texte. Néanmoins, ces méthodes ne peuvent pas être utilisées pour transformer du texte en images.

Contrainte de la disponibilité de la connaissance Quels que soient les modèles d'association texte/image, l'existence de connaissances a priori, souvent représentées sous forme d'annotations, est absolument indispensable pour les phases d'apprentissage. Cette phase d'annotation est extrêmement chronophage pour l'utilisateur, et particulièrement complexe pour les applications spécialisées. Ces phases d'apprentissage se font essentiellement hors ligne et le problème est particulièrement délicat, notamment lorsque les connaissances sont amenées à évoluer, par exemple pour intégrer de nouvelles connaissances. La performance de ces modèles n'est pas facile à améliorer. La problématiques du développement d'approches permettant d'enrichir en connaissances le système est donc particulièrement cruciale, et ces approches doivent pouvoir s'effectuer sans demander de calculs hors ligne.

4.1.3 Positionnement et argumentation pour notre modèle

Dans le cadre de ce travail, nous nous donnons pour objectif d'apporter des réponses aux différentes problématiques évoquées en conclusion du paragraphe précédent. Tout d'abord, afin d'éviter une dépendance à la qualité de cette phase de traitement toujours délicate, nous nous sommes placés dans un contexte sans segmentation. Ensuite, afin de supporter la recherche d'images par des requêtes textuelles indépendamment de toute annotation manuelle, nous proposons d'ajouter une stratégie à base de transformation de texte en images à celle, plus classique, de transformation d'images en texte. Enfin, nous nous plaçons dans un système avec apprentissage incrémental des connaissances, qui ne demande pas de connaissances particulières au début de la vie du système. Cette contrainte nous semble en effet à la fois essentielle, mais également réaliste, car la majeure partie des applications spécialisées n'ont pas de connaissances en début de vie.

Nous proposons donc un modèle d'association texte/image pour la recherche d'images et pour l'annotation d'images (apprentissage de connaissances) dénommé "modèle Sac de KVR" (*Keyword Visual Representation*), (voir figure 4.6, rectangle en bleu) où un KVR est une représentation visuelle possible pour un mot. Pour éviter la segmentation d'images, nous utilisons le modèle connu de Sac de Mots (*BoW - Bag Of Words model*) pour représenter les images. Dans ce modèle, l'image n'est pas représentée par des régions mais par des points d'intérêts (détaillé dans la partie suivante). Une autre raison de l'utilisation du modèle BoW est l'efficacité de ce modèle, confirmée par les tendances actuelles de la recherche sur ce modèle [Yang 2007], [Tirilly 2008], [Sivic 2008], [Fei-Fei 2007].

Comme le modèle de transformation, notre modèle tente de représenter les mots textuels par un contenu visuel. Le modèle de transformation (et aussi les autres) utilise des régions d'images. Par contre, en appui sur le modèle Sacs de mots de l'état de l'art (*Bag of Words* - [Sivic 2008]), notre modèle représente les mots textuels par des caractéristiques visuelles (des points d'intérêts) via le modèle BoW, évitant ainsi les phases de segmentation.

Selon notre modèle, profitant des interactions des utilisateurs/experts pendant la recherche / l'exploration d'images, les représentations des mots textuels sont apprises par une méthode d'apprentissage incrémental via le retour de pertinence à long terme sans aucune connaissance au début. Contrairement aux autres modèles où les connaissances sont disponibles a priori, dans notre système, la connaissance vient des interactions des utilisateurs. Par conséquent, la connaissance de notre système est améliorée dans le temps grâce aux interactions, sans demander d'apprentissage hors ligne.

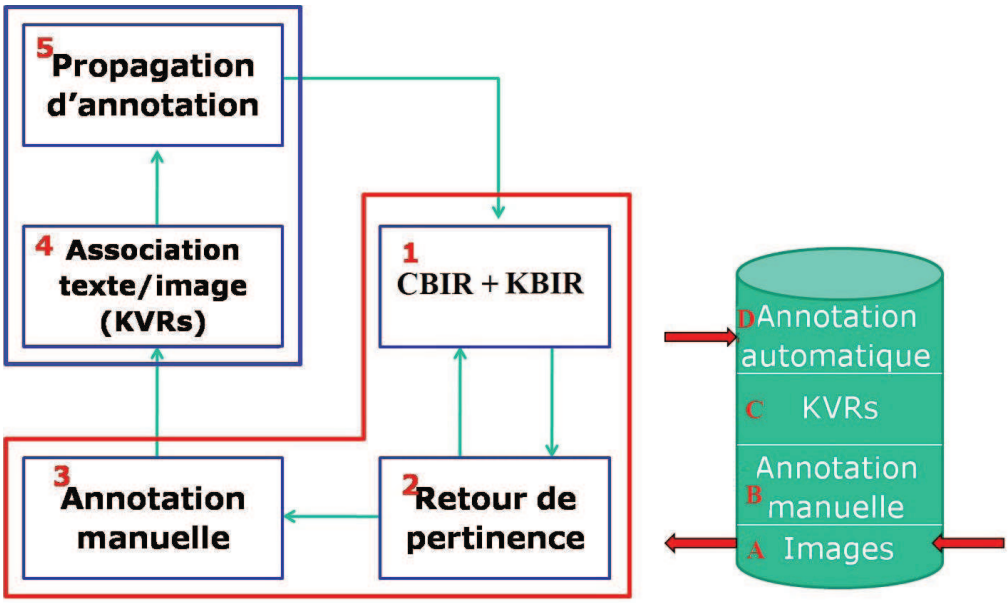

FIGURE 4.6 – Notre système de recherche d'images avec en bleu l'utilisation de notre modèle Sacs de KVR pour l'association texte/image pour la recherche d'images multimodales et pour l'annotation d'images.

4.2 La représentation Sac de KVR - (Représentation visuelle de mots textuels)

Dans cette partie, nous présentons une nouvelle représentation visuelle pour des concepts textuels. Notre représentation Sac de KVR évite le premier problème de l'état de l'art mentionné précédemment (la segmentation d'images) en s'inspirant de l'efficacité et de la simplicité du modèle Sac de Mots de l'état de l'art. Dans cette section, nous parlons d'abord du modèle existant de Sacs de Mots. Ensuite, nous évoquons la construction de ce modèle dans notre système. Enfin, nous présentons notre nouvelle représentation Sac de KVR.

4.2.1 Le modèle Sacs de Mots (BoW)

Le modèle BoW dans le traitement du langage naturel est une méthode populaire pour représenter les documents. Chaque document est représenté par un sac de mots caractérisant la fréquence d'apparition des mots d'un dictionnaire (voir figure 4.7). Tous les documents de la base de données sont résumés sur la base de mots d'un dictionnaire. Pour le processus de recherche de documents, un texte est récupéré par le calcul de la similitude entre son vecteur de fréquences de mots et ceux des autres documents.

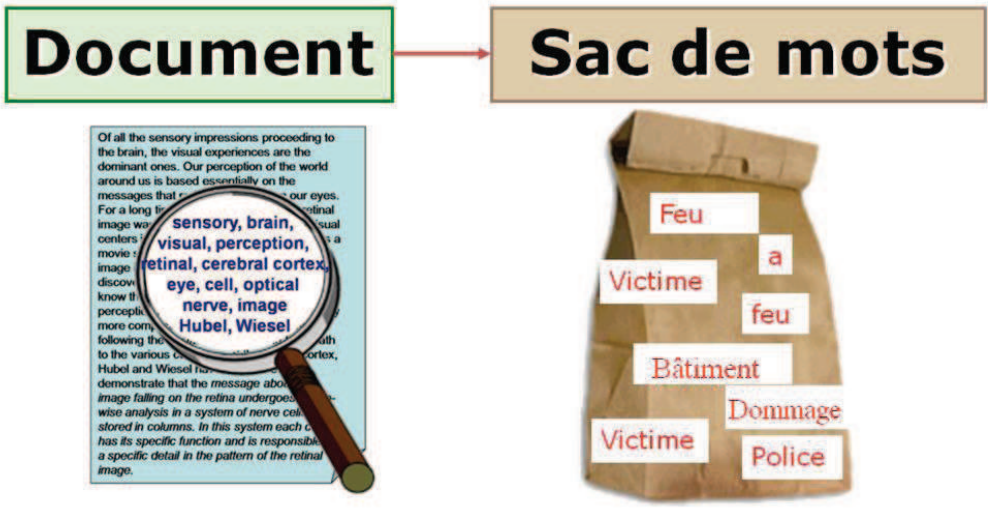

FIGURE 4.7 – Le modèle BoW dans le traitement du langage naturel. Chaque document est représenté par un sac de mots.

Les chercheurs en vision par ordinateur ont utilisé la même idée pour la représentation des images [Sivic 2008]. Les images sont traitées comme des documents, et les caractéristiques extraites des images sont considérées comme des "mots". Les figures 4.8 et 4.9 présentent l'idée de l'analogie des mots visuels avec les mots textuels.

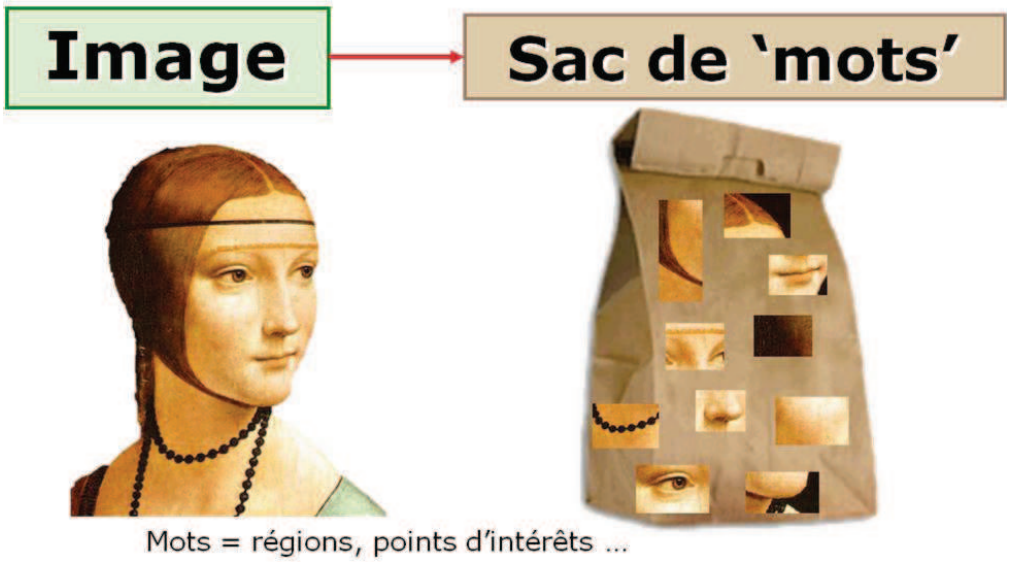

FIGURE 4.8 – Les images sont traitées comme des documents textuels, et les caractéristiques extraites des images sont considérées comme des "mots" [Fei-Fei 2007]

Comme le dictionnaire d'un langage, les caractéristiques de régions semblables d'une image sont représentées par un même mot visuel.

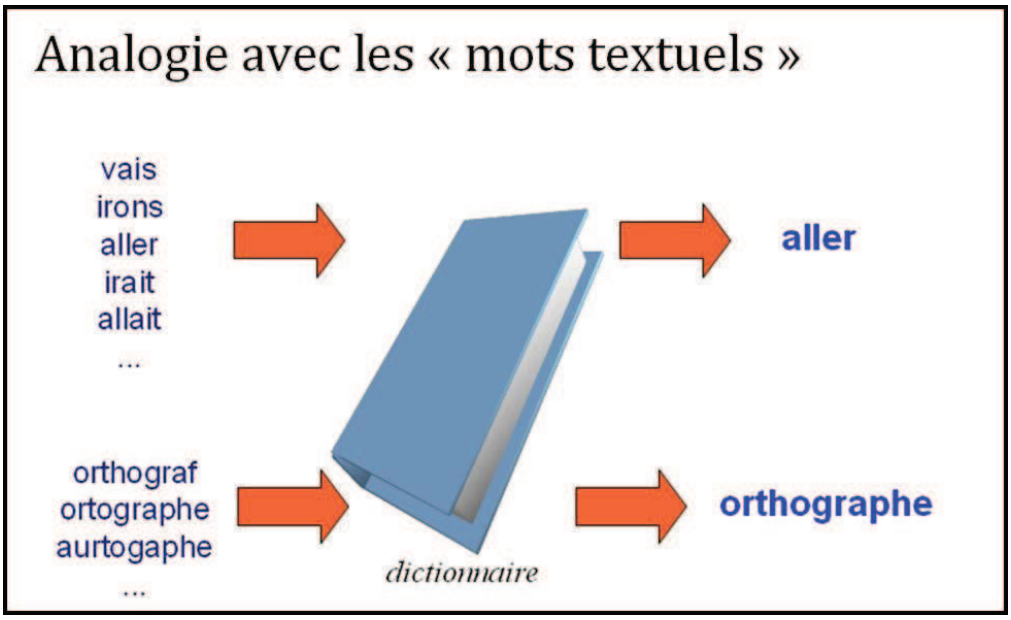

FIGURE 4.9 – Comme un dictionnaire d'un langage ou un verbe peut se présenter par des mots différents lorsqu'il est conjugué, les caractéristiques de régions semblables d'une image sont représentées par un même mot visuel.

Le modèle BoW est généralement assuré par 3 étapes : détection et description des mots (figure 4.10, étape 1), formation du dictionnaire (figure 4.10, étape 2) et enfin représentation de l'image (figure 4.10, étape 3). Une des problématiques importantes de ce modèle est la détection de caractéristiques extraites des régions locales, qui représentent des candidats pour les "mots". La méthode la plus simple pour cette fonctionnalité de détection est une grille régulière [Li 2005], [Vogel 2006]. L'image est segmentée par des lignes horizontales et verticales et chaque élément résultant de cette matrice est considéré comme un mot potentiel. Une autre méthode très populaire est le détecteur de points d'intérêt [Li 2005], [Sivic 2008]. L'échantillonnage aléatoire et les méthodes de segmentation pour la fonction de détection sont utilisées dans [Vidal-Naquet 2003] et [Barnard 2003]. La méthode la plus populaire pour la description du mot est celle qui s'appuie sur le descripteur SIFT [Lowe 2004]. Le descripteur SIFT décrit chaque région comme un vecteur de dimension 128. La deuxième étape consiste à convertir les vecteurs de descriptions dans un dictionnaire de mots (*codewords*). Chaque code de mot résulte d'une méthode de regroupement, comme par exemple la méthode des k-moyennes, et représente ainsi plusieurs régions similaires. La dernière étape consiste à représenter l'image comme un histogramme de mots visuels.

Récemment, le modèle Sac de mots (BoW) est devenu une référence pour les systèmes CBIR. Le modèle BoW est celui que nous avons retenu pour notre système, en raison de son potentiel pour améliorer les performances de la recherche d'images en utilisant la richesse des connaissances et les pratiques déjà existantes

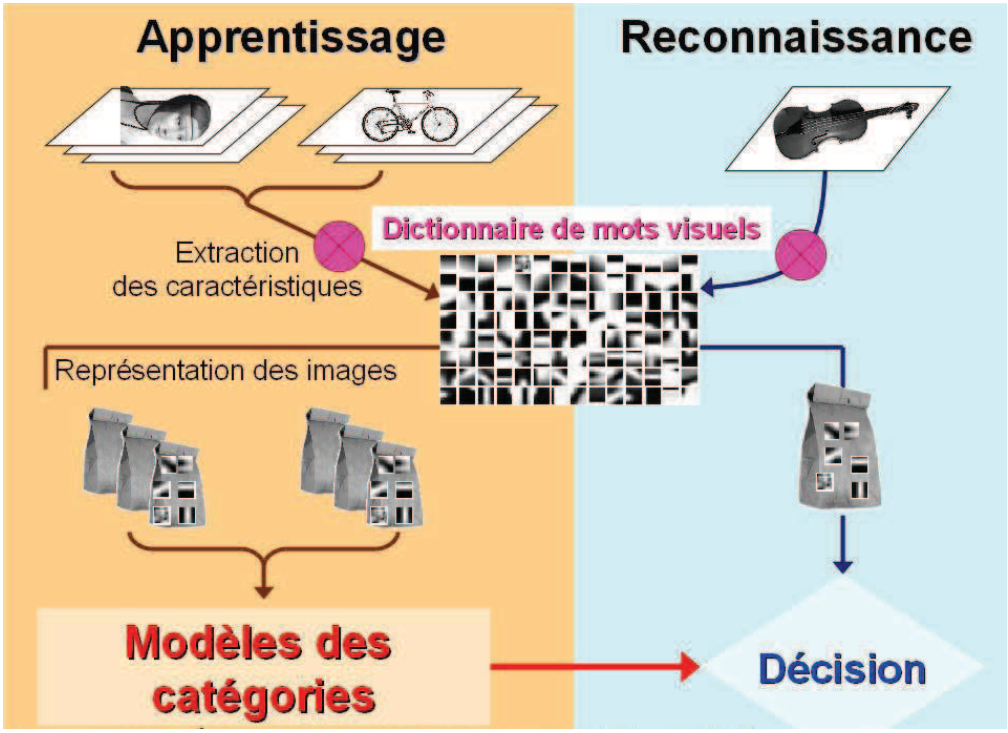

FIGURE 4.10 – 3 étapes : (1) détection et description des mots (Extraction de caractéristiques), (2) formation d'un dictionnaire de mots visuels, (3) représentation des images [Fei-Fei 2007].

en recherche de texte. La construction du modèle Sac de mots dans notre système est présentée dans la partie suivante.

4.2.2 La construction du modèle Sac de mots

Dans notre modèle, nous estimons un mot visuel à partir d'un point d'intérêt [Li 2005], [Sivic 2008]. La description d'un point d'intérêt est obtenue en utilisant le descripteur SIFT [Lowe 2004]. La méthode de regroupement des k-moyennes est alors utilisée pour la construction du dictionnaire de mots visuels. Une représentation BoW d'une image j est un vecteur pondéré de mots visuels :

$$\bar{I}_j = (w_{j,1}, w_{j,2} \dots w_{j,t}) \tag{7}$$

où w_j est calculé comme le poids $TF * IDF$ souvent utilisé en recherche d'information et en fouille de texte [Sivic 2008] et la longueur du vecteur est la taille du dictionnaire de mots visuels.

Soit l'image j et le mot i, alors la fréquence TF du mot visuel dans l'image

est :

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

où $n_{i,j}$ est le nombre d'occurrences du mot visuel i dans l'image j. Le dénominateur est le nombre d'occurrences de tous les mots visuels dans l'image j.

La fréquence inverse de document IDF du mot visuel i est :

$$IDF_i = \log \frac{|I|}{|\{I_j : i \in I_j\}|}$$

où $|I|$ est le nombre total d'images dans le corpus, $|\{I_j : i \in I_j\}|$ est le nombre d'images où le mot visuel i apparaît (c'est-à-dire $n_{i,j} \neq 0$)

4.2.3 La représentation Sac de KVR

Notre nouvelle représentation Sac de KVR est basée sur la représentation existante Sac de Mots (ou BoW - *Bag Of Words*). Une représentation visuelle d'un mot textuel (ou d'un KVR) est une représentation BoW d'une région correspondant à un mot textuel. Par exemple, dans la figure 4.11, une représentation du mot "tortue" est un sac des mots visuels qui sont présents dans la région de la tortue. Cependant, en réalité, un mot textuel peut correspondre à plusieurs régions de l'image. Par exemple, dans la figure 4.12, le mot "Ciel" peut être interprété par trois types différents de ciel : ciel clair (bleu), ciel avec nuages (blanc), et coucher de soleil (ciel rouge). Dans notre modèle, le Sac de KVR est créé avec l'hypothèse selon laquelle un mot textuel correspond à une ou plusieurs régions dans les images. Un mot textuel est alors représenté par un ensemble de régions ou un sac de KVR dans lequel un KVR correspond à une région (ou un sac de mots visuels dans notre système). La figure 4.12 est un exemple de sac de KVR du mot "Sky" (ciel).

La similarité visuelle entre 2 mots textuels ou un mot textuel avec une image est la similarité visuelle entre 2 Sacs de KVR. Pour comparer la similarité visuelle entre 2 Sacs de KVR nous définissons une fonction de similarité de la manière suivante. Considérons 2 sacs de KVR S_1, S_2 de 2 mots textuels M_1, M_2 :

$S_1 = (KVR_1^1, KVR_2^1, .., KVR_{k1}^1)$, $S2 = KVR_1^2, KVR_2^2, .., KVR_{k2}^2)$

où $k1$, $k2$ sont les nombres de KVR dans S_1, S_2.

La similarité visuelle entre S_1, S_2 est définie comme :

$$Sim_visuelle(S_1, S_2) = max(Sim_visuelle(KVR_i^1, KVR_j^2)) \quad (8)$$

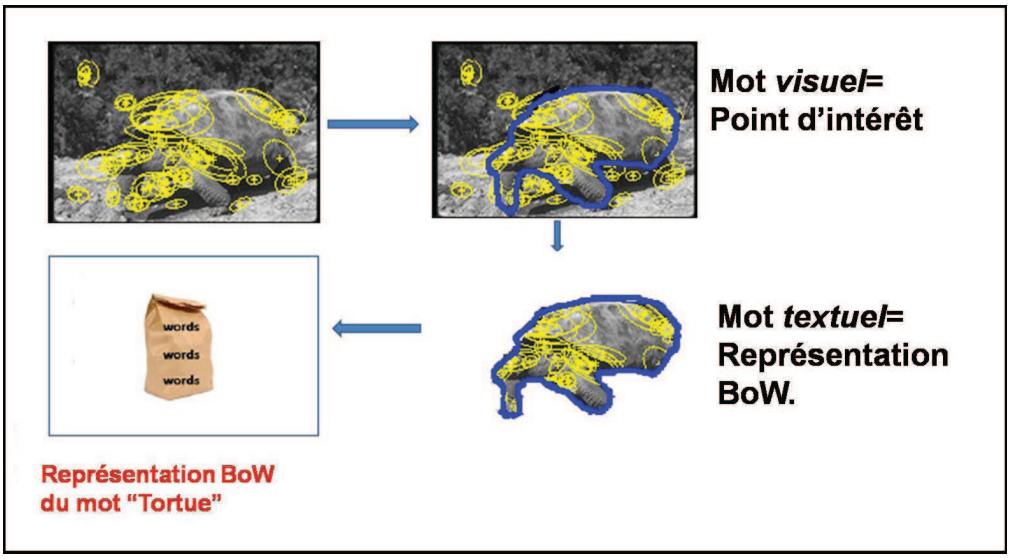

FIGURE 4.11 – Exemple d'une représentation BoW du mot textuel "tortue".

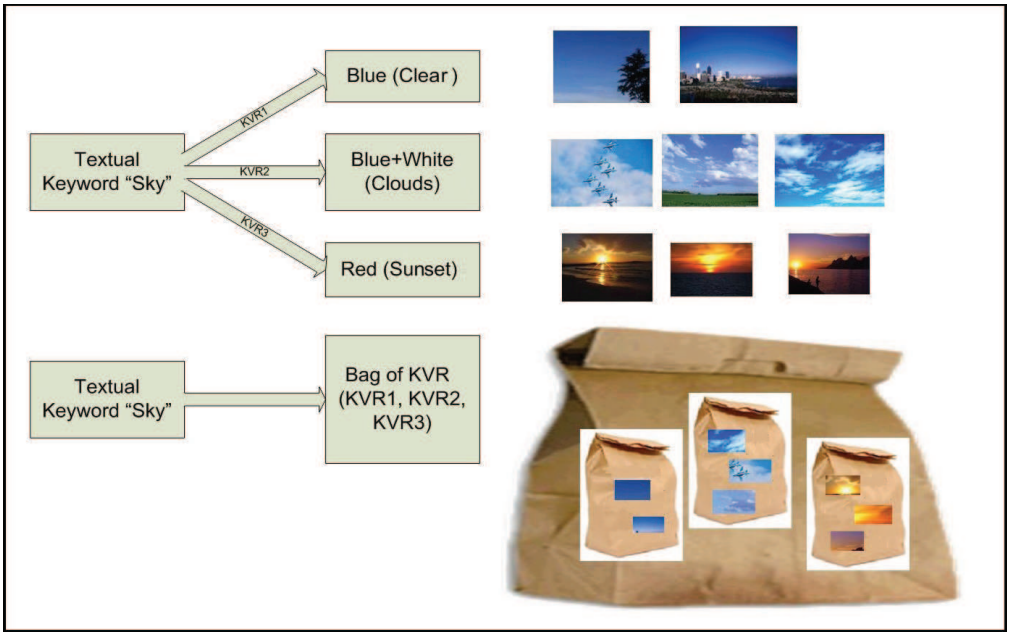

FIGURE 4.12 – La représentation Sac de KVR. "Ciel" pourrait être l'un des trois types : "Clair", "Nuageux" et "Coucher de soleil". Le mot textuel est alors représenté par un sac contenant les trois KVR correspondants.

avec $i = 1 : k_1, j = 1 : k_2$

La similarité visuelle de 2 KVR ou en d'autres termes, 2 BoW, est présentée dans le modèle BoW ci-dessus.

Dans la figure 4.13, la représentation Sac de KVR est présentée avec ses différentes caractéristiques. On peut résumer comme suit la représentation Sac de KVR :

1. Chaque image est représentée par un Sac de mots visuels
2. Chaque groupe de régions similaires des images dans une catégorie est représenté par un Sac de mots visuels qui s'appelle KVR (*Keyword Visual Representation*)
3. Avec l'hypothèse selon laquelle un mot textuel correspond à une ou plusieurs régions dans les images, chaque mot textuel est donc représenté par 1 ou plusieurs KVR (1 Sac de KVR)

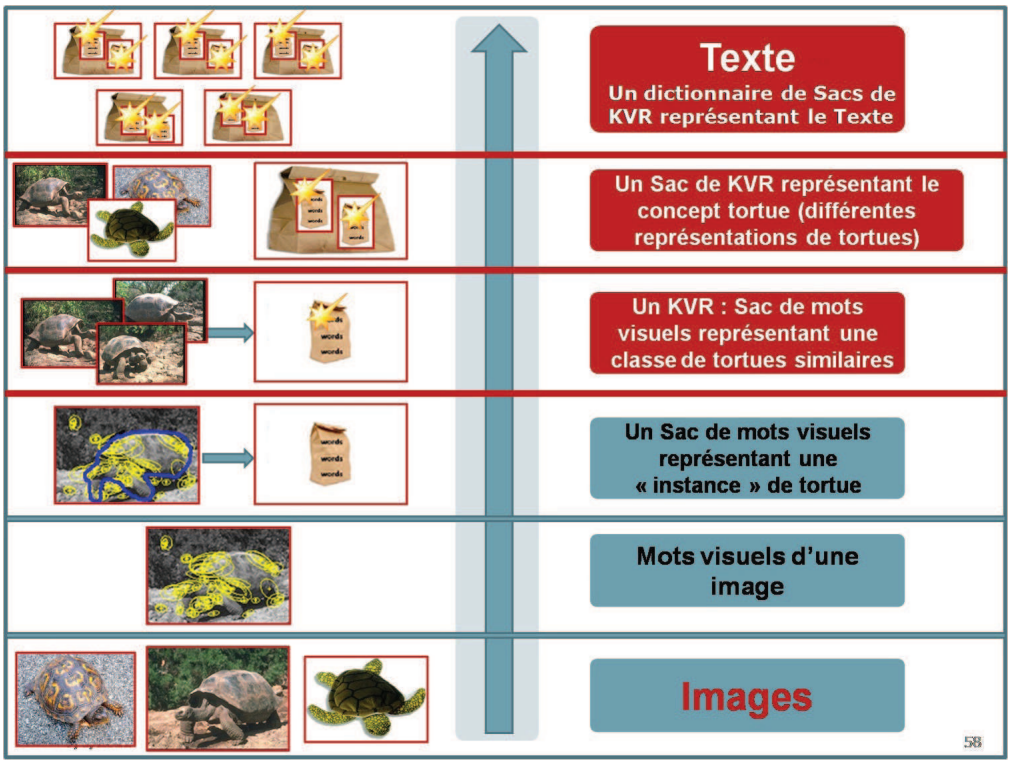

FIGURE 4.13 – Le modèle Sac de KVR pour l'association texte/image (notre contribution en rouge) basé sur la représentation existante Sac de Mots (en bleu).

La représentation Sac de KVR est utilisée pour l'annotation d'images ou la recherche multimodale d'images. Elle est particulièrement efficace en cas d'insuffisance d'annotations textuelles dans la base d'images parce que les requêtes textuelles peuvent être représentées par des caractéristiques visuelles (la transformation de texte en images qui est le deuxième problème existant dans l'état de l'art). Il s'agit d'un modèle du type de transformation comme le modèle de [Lin 2007] pour l'association du texte et du contenu. Alors que le modèle de [Lin 2007] utilise l'information mutuelle et la segmentation d'images pour transformer une requête textuelle en une requête visuelle, notre modèle profite de l'efficacité et de la simplicité du modèle BoW pour représenter la requête textuelle par des caractéristiques visuelles, qui dans ce cas correspond à la représentation BoW. Notre modèle peut

ainsi profiter de l'efficacité du modèle BoW et peut éviter le problème de la segmentation d'images. De plus, il peut être utilisé pour l'annotation d'images dont nous allons parler dans la section 4.5.1 (la transformation d'images en texte). Pour construire des représentations Sacs de KVR des concepts textuels, nous proposons d'utiliser une méthode d'apprentissage par renforcement/incrémental que nous présentons dans la section suivante.

4.3 Apprentissage des Sacs de KVR

Avant d'évoquer l'apprentissage des représentations Sacs de KVR, nous résumons ici notre système de recherche d'images. Rappelons que nous partons de l'hypothèse que la base d'images du système est de dimension importante et qu'elle évolue au cours du temps, sans connaissances a priori au début de la vie du système. Dans le chapitre 3, nous avons présenté notre méthode pour le retour de pertinence qui permet la recherche efficace interactive dans notre système. Nous consacrons cette section à l'apprentissage de connaissances à partir des interactions de l'utilisateur. Trois types de connaissances peuvent être distingués dans notre système :

1. Des connaissances liées à l'annotation manuelle : des images se voient affecter des concepts/mots textuels par les utilisateurs/experts pendant l'interaction.
2. Des connaissances liées à l'annotation propagée : d'autres images dans la base se voient affecter dynamiquement des concepts textuels en utilisant des Sacs de KVR des concepts (voir la partie 4.4).
3. Des connaissances de niveau "image" : Les représentations visuelles de concepts ou en d'autres termes, les Sacs de KVR de concepts.

Bien évidemment, les connaissances du système apprises pendant les interactions sont directement liées à l'apprentissage des représentations de Sacs de KVR qui est présenté dans la partie suivante.

Après avoir présenté la représentation Sac de KVR pour les mots (concepts) dans la partie précédente, nous présentons ici l'apprentissage des Sacs de KVR des concepts. Nous avons retenu la méthode d'apprentissage par renforcement incrémental pour résoudre le problème identifié dans la littérature , et relatif à la contrainte de la disponibilité de la connaissance. En fait, l'apprentissage par renforcement est une méthode qui n'utilise pas nécessairement de connaissances pré-existantes mais uniquement des connaissances issues d'interactions avec les utilisateurs. Par opposition aux autres modèles d'association texte/image qui demandent des connaissances a priori, notre modèle tente d'apprendre des Sacs de

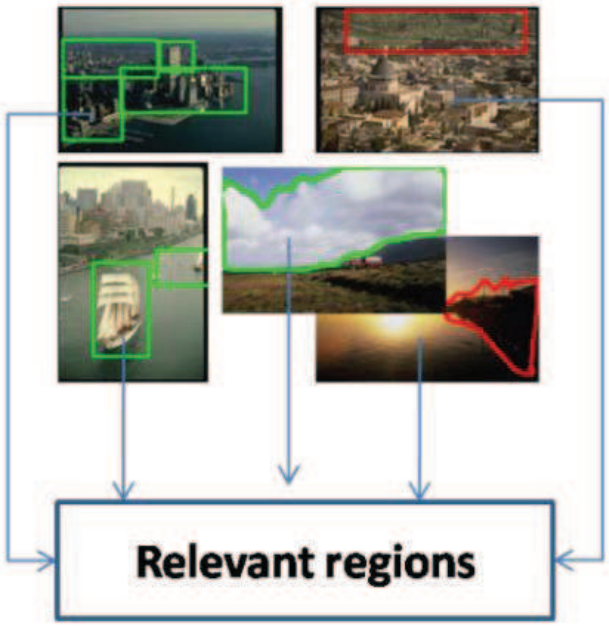

FIGURE 4.14 – Les régions pertinentes sont sélectionnées au cours du processus de recherche. Des mots visuels extraits des régions sont utilisés pour représenter le mot-clé.

KVR de concepts sans demander de connaissances au début de la vie du système. La connaissance dans notre modèle vient des interactions des utilisateurs/experts pendant l'exploration d'images et la recherche d'images.

L'apprentissage des représentations de concepts est difficile du fait de l'absence de connaissances entre les concepts et les régions des images. Le problème devient plus aisé si on offre la possibilité de sélectionner les régions et les mots correspondants, car cela offre la possibilité de construire la représentation visuelle des mots en utilisant des mots visuels dans les régions correspondantes. Une première approche que l'on peut imaginer pour cela est d'utiliser une technique d'interaction au sein de laquelle l'utilisateur sélectionne les régions concernées avec des formes plus ou moins complexes (figure 4.14). Pendant l'exploration / la recherche dans la base d'images, l'utilisateur marque les régions par des informations d'intérêt. En phase de retour de pertinence, les représentations des mots-clés sont ainsi mises à jour en construisant les caractéristiques visuelles de ces régions.

Bien que cette première approche offre une bonne précision, cela nécessite un travail important pour l'utilisateur. Considérant que cette méthode constitue une tâche complexe pour les utilisateurs à qui il est difficile de faire accepter de lourdes charges dans les systèmes CBIR, nous proposons une seconde approche basée sur le retour de pertinence basé sur les groupes d'images pour l'apprentissage des Sacs de KVR des mots textuels dans la partie suivante. Par rapport à la première approche

de sélection des régions, cette seconde approche demande seulement à l'utilisateur de cliquer sur des images, rendant ainsi sa tâche moins complexe et beaucoup plus agréable.

4.3.1 Apprentissage incrémental des Sacs de KVR

Dans cette partie, nous procédons à des apprentissages de Sacs de KVR des mots textuels en se basant sur le retour de pertinence décrit dans le chapitre 3. L'apprentissage de Sacs de KVR est illustré dans la figure 4.15, incluant une boucle d'apprentissage incrémental et une boucle d'interaction. Il existe une boucle d'interaction entre les utilisateurs/experts et la machine pour chaque phase de l'apprentissage. Apres une boucle d'interaction, les Sacs de KVR des mots textuels ainsi que les annotations textuelles sont réactualisés de manière incrémentale. Partant du principe que la recherche d'images est la tâche principale demandée à l'utilisateur dans notre système, nous identifions tout d'abord différents scénarios possibles d'interaction des utilisateurs pour la tâche de recherche d'images.

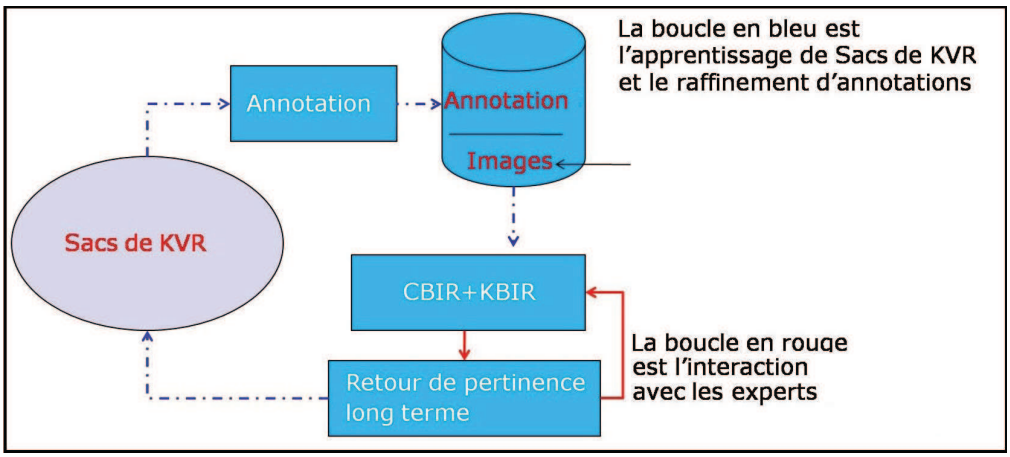

FIGURE 4.15 – Un scénario pour l'apprentissage des Sacs de KVR des mots textuels.

Scénario d'interaction 1 Au début, nous faisons l'hypothèse que la base de données ne contient pas d'annotations. Les utilisateurs ne font que rechercher des images par le contenu. En explorant les images, ils peuvent trouver des informations intéressantes, et souhaitent étiqueter des images avec des mots textuels réutilisables par eux-mêmes ou par d'autres utilisateurs. Le système va utiliser ces informations pour créer ou mettre à jour des Sacs de KVR de mots textuels.

La figure 4.16 montre un exemple du scénario 1 d'apprentissage du Sac de KVR pour le concept "Ciel". Tout d'abord, en explorant des images l'utilisateur trouve quelques images et peut marquer celles-ci du concept "Ciel". Dans la figure 4.16,

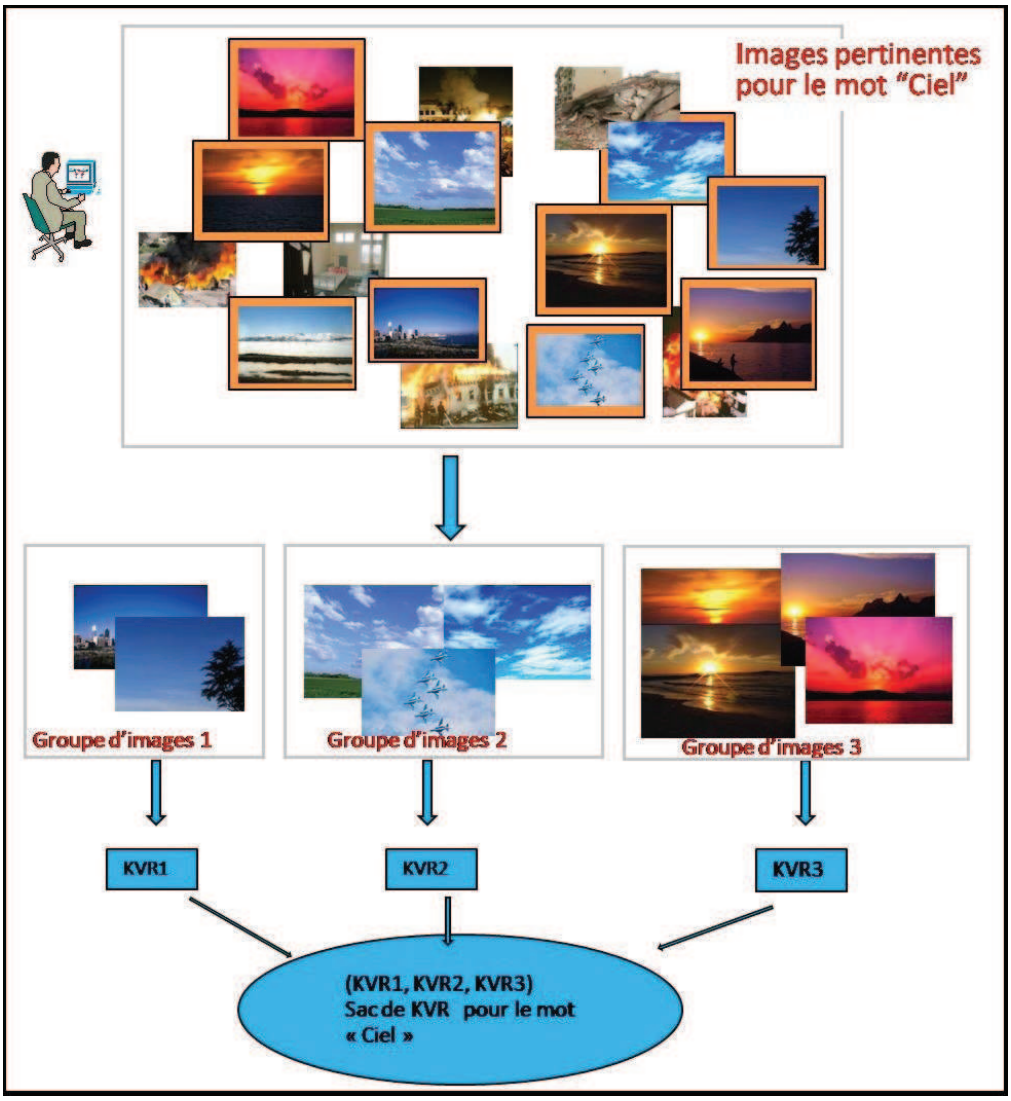

FIGURE 4.16 – Un scénario pour l'apprentissage des Sacs de KVR des mots textuels.

on peut constater qu'il est possible de trouver des images de différents types mais restant cohérentes avec le concept "Ciel" : des ciels de coucher de soleil, des ciels bleus, des ciels bleus avec nuages blancs. Pour identifier manuellement ces groupes distincts d'images, l'utilisateur doit procéder à un gros travail qu'il est difficile de faire accepter dans les systèmes CBIR. Nous proposons donc d'utiliser une méthode de regroupement (clustering) pour grouper automatiquement ces images en différents paquets représentant des types du concept "Ciel". Cette démarche facilite la tâche de l'utilisateur, grâce à cette stratégie de clustering. Enfin, le Sac de KVR du concept "Ciel" est construit par la technique de Rocchio [Rocchio 1971]. Le choix d'une méthode de regroupement est discuté dans la partie 4.4.1.

Scénario d'interaction 2 Pour ce scénario, nous supposons que des annotations manuelles/automatiques et des Sacs de KVR de mots-clés sont disponibles dans la base de données. L'utilisateur peut procéder à la recherche d'images par le contenu et le texte. Il peut alors réactualiser les connaissances liées aux images en utilisant le retour de pertinence basé sur les groupes (chapitre 3). Basé sur l'interaction utilisateur permettant d'identifier les images pertinentes et les images non pertinentes par rapport à la requête, des Sacs de KVR de mots textuels sont mis à jour et l'annotation automatique d'images est mise à jour également.

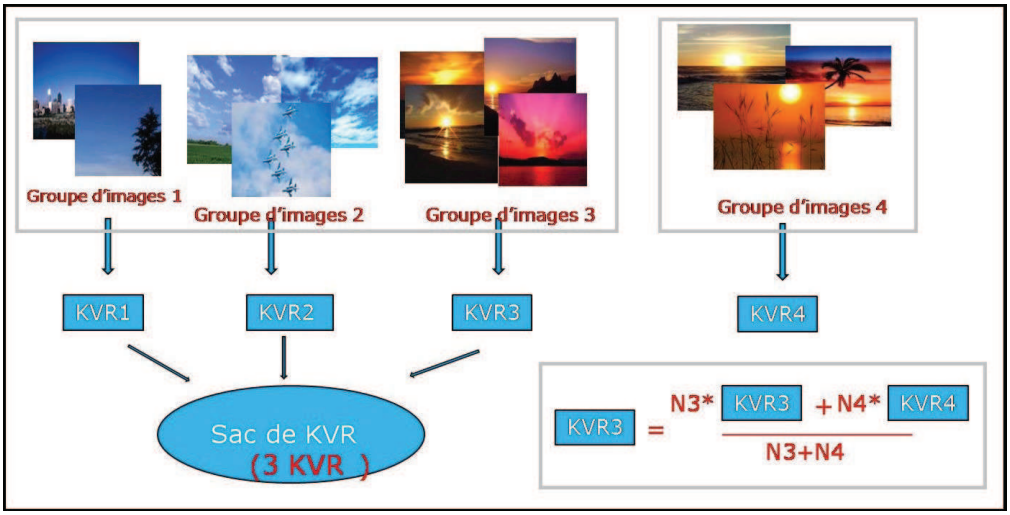

FIGURE 4.17 – Scénario pour l'apprentissage des Sacs de KVR de mots textuels : Fusion de deux KVR, c'est-à-dire la fusion de 2 groupes d'images correspondant à 2 KVR.

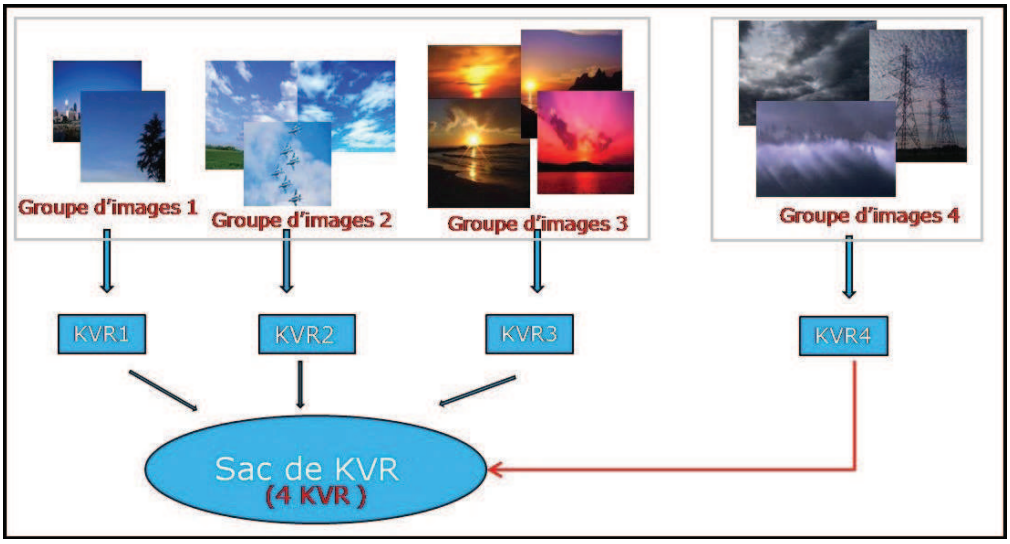

FIGURE 4.18 – Scénario pour l'apprentissage des Sacs de KVR des mots textuels : Ajout d'un nouveau KVR.

Les figures 4.17 et 4.18 montrent des exemples du scénario 2 pour l'apprentissage du Sac de KVR pour le concept "Ciel". Faisant l'hypothèse que le concept "Ciel" a déjà un Sac de KVR et qu'on vise à mettre à jour ce Sac de KVR, en utilisant la requête textuelle ou la requête multimodale texte et images, l'utilisateur peut modifier ou valider des images de résultats de recherche. Grâce aux images pertinentes / non pertinentes sélectionnées par l'utilisateur lors du retour de pertinence, de nouveaux KVR du concept "Ciel" sont éventuellement créés. Dans le cas de la figure 4.17, le nouveau KVR du concept "Ciel" est similaire avec un KVR dans le Sac de KVR. Une fusion des deux KVR est réalisée, et le Sac de KVR est mis à jour. Cette fusion est définie comme un opérateur de KVR qui est présenté dans la partie 4.4.2 avec les autres opérateurs. Dans le cas de la figure 4.18, le nouveau KVR est différent des KVR existants dans le Sac de KVR, ce KVR est donc ajouté dans le Sac de KVR.

Dans les parties suivantes, nous discutons d'abord notre choix pour la méthode de regroupement. Ensuite les opérateurs de KVR dans notre modèle Sacs de KVR sont présentés. Les algorithmes d'apprentissage pour les deux scénarios sont présentés à la fin.

4.3.1.1 Sélection de la méthode de regroupement

Nous avons déjà discuté une première fois de choix de méthodes de regroupement dans le chapitre 3 dans le cadre de notre stratégie de retour de pertinence. Ici, dans notre méthode pour l'apprentissage des Sacs de KVR, une étape importante est encore le regroupement. La méthode de regroupement est utilisée pour regrouper des images pertinentes / non pertinentes en des groupes distincts. Ces groupes distincts sont utilisés pour construire des représentations visuelles KVR pour des concepts textuels distincts. La qualité du résultat du regroupement conditionne la qualité de la représentation. Dans notre système, comme évoqué précédemment, le nombre de groupes est inconnu. Nous nous sommes donc intéressés aux méthodes de regroupement capables de déterminer le nombre de groupes de façon autonome. Pour ce faire, nous avons utilisé 2 méthodes de regroupement dans notre expérimentation : la méthode des k-moyennes adaptatif présentée par [Kothari 1999] et la méthode d'agglomération compétitive (*Competitive Agglomeration*) présentée par [Frigui 1997]. Ces deux méthodes ont été choisies en raison de leur capacité à déterminer automatiquement le nombre de groupes. Elles sont présentées en retenant les méthodes de regroupement les plus populaires : les méthodes hiérarchiques et les méthodes de partition. Voir la section 3.3.2 pour plus de détails et d'état de l'art sur les méthodes de regroupement.

4.3.1.2 Opérateurs de KVR

Le modèle Sacs de KVR est optimal si des concepts textuels sont bien représentés par des Sacs de KVR. Les méthodes de regroupement sont utilisées dans l'apprentissage des Sacs de KVR des concepts. Néanmoins, il reste possible de rencontrer des éléments de bruit dans le calcul des Sacs de KVRs si le regroupement ne donne pas des groupes d'images qui assurent l'unimodalité (voir partie 3.2.3 pour la définition de ce concept). Pour bien identifier des Sacs de KVR de concepts textuels et éliminer le plus possible de bruit, nous définissons les 4 opérateurs suivants pour les KVR : ADD, EQUAL, MERGE et SPLIT. Ces opérateurs sont utilisés dans l'apprentissage des Sacs de KVR des concepts.

ADD L'opérateur ADD est utilisé pour ajouter un KVR dans un Sac de KVR avec comme condition qu'il soit différent de tous les KVR déjà existants dans le Sac de KVR. Cette condition est verifiée par l'opérateur EQUAL qui suit.

EQUAL L'opérateur EQUAL est utilisé pour déterminer si deux KVR sont considérés comme suffisamment proches pour être combinés. Nous proposons d'utiliser un critère équivalent au critère de Ward : l'augmentation de la variance. Ce critère de [Ward 1963] est l'un des nombreux critères largement utilisés pour la classification ascendante hiérarchique et il est fréquemment mentionné que cette méthode a surpassé les autres méthodes dans plusieurs études comparatives [Jain 1988].

$EQUAL(KVR1, KVR2) = 1$ si :

$$Variance(KVR1, KVR2) <= \alpha Variance(KVR1) + \beta Variance(KVR2) \quad (9)$$

où α et β sont les poids relatifs des deux KVR ($\alpha + \beta = 1$). α peut être calculé en basant sur le nombre d'images dans chaque KVR. Pour l'instant, dans notre expérimentation, la formule simple suivante est utilisée pour calculer ces poids :

$\alpha = 1$, si $Variance(KVR1) > Variance(KVR2)$ sinon $\alpha = 0$

MERGE L'opérateur MERGE est utilisé pour fusionner 2 KVR similaires en 1 KVR. La fusion de 2 KVR est basée sur le nombre d'éléments dans chaque KVR, critère qui définit leur importance. Nous pouvons trouver que le KVR représenté par le plus grand nombre d'images est plus important que le KVR représenté par le moins d'images :

$$KVR3 = MERGE(KVR1, KVR2)$$

où

$$\overrightarrow{KVR3} = \frac{\left(N1 * \overrightarrow{KVR1} + N2 * \overrightarrow{KVR2}\right)}{N1 + N2} \tag{10}$$

où N_i est le nombre d'images du $KVRi$.

SPLIT L'opérateur SPLIT est utilisé pour scinder 1 KVR en 2 autres KVR ou pour re-calculer un Sac de KVR. On peut faire la division d'un KVR ou d'un Sac de KVR s'il y a modification dans ceux-ci. Pour l'instant nous faisons des expérimentations pour la division d'un KVR seulement. Le principe de la division est similaire à celui des méthodes de regroupement. Un KVR dans le modèle Sac de KVR correspond à un cluster dans les méthodes de regroupement. La division d'un KVR en 2 autres KVR est effectuée après la fusion des 2 KVR (opérateur MERGE) si le KVR fusionné est trop diversi fié/large selon un critère donné. Comme les Sacs de KVR sont construits de façon incrémentale où tous les KVR existants sont déjà véri fiés pour savoir s'il est possible de les scinder, la division est faite seulement après un changement du KVR. En utilisant le même critère que la fusion, la condition de division est basée sur la variance des KVR.

Condition pour diviser KVR3 :

$$Variance(KVR3) > (\alpha Variance(KVR1) + (1 - \alpha)Variance(KVR2))), \tag{10}$$

avec $\alpha = N_1/(N_1 + N_2)$.

Le KVR3 est divisé en 2 autres KVR en utilisant une méthode de regroupement (k-moyennes adaptatif, agglomération compétitive). Dans notre expérimentation, sur nos jeux de données, la méthode d'agglomération compétitive donne une performance qui est légèrement meilleure que la méthode des k-moyennes adaptatif. Mais cependant, nous avons observé que la méthode de regroupement n'influence pas beaucoup la performance. La raison est probablement liée au très petit nombre d'exemples.

$$SPLIT(KVR3) = CompetiveAgglomeration(KVR3) = 2_autres_KVR$$

Les 4 opérateurs sont utilisés par des algorithmes pour l'apprentissage incrémental/par renforcement. Les operateurs ADD, MERGE et SPLIT sont utilisés pour la révision du modèle, tandis que l'operateur EQUAL est utilisé pour l'analyse du modèle. Nous décrivons en détails les algorithmes proposés et l'utilisation des opérateurs dans la partie suivante.

4.3.1.3 Algorithmes proposés

Profitant du retour de pertinence à court terme présenté dans le chapitre 3, un retour de pertinence à long terme avec une stratégie d'apprentissage progressif est également proposé afin de mettre à jour les Sacs de KVR de mots textuels (figure 4.20, étape 4). Dans le retour de pertinence à court terme, après la recherche interactive pour trouver des images intéressantes, les connaissances ne sont pas mémorisées. Les connaissances sont des annotations manuelles et des requêtes idéales à points multiples pour chaque concept (voir chapitre 3). Dans notre retour de pertinence à long terme (apprentissage de connaissances) proposé, ces connaissances sont utilisées pour construire le modèle Sacs de KVR. Ce modèle Sacs de KVR est réactualisé à chaque fois qu'une nouvelle expérience significative (ou une session d'interaction) devient disponible. Le modèle est alors mis à jour en fonction du temps (en continu) comme les expériences arrivent en séquences. Cette caractéristique montre le caractère "incrémental" de l'apprentissage. Du fait que l'apprentissage utilise des informations de renforcement par l'interaction de l'utilisateur (celles-ci sont des images pertinentes/non pertinentes du retour de l'utilisateur), nous considérons que cet apprentissage est un type d'apprentissage par renforcement.

Nous avons identifié deux scénarios possibles d'utilisation de notre système. Dans cette partie, nous proposons deux algorithmes pour l'apprentissage des Sacs de KVR des concepts textuels, correspondant aux deux scénarios que nous avons présentés ci-dessus. Dans le premier scénario, des premiers KVR sont construits en se basant sur des informations fournies par l'utilisateur qui sont des images étiquetées avec un mot. Dans le deuxième scénario, les informations fournies sont des images pertinentes/non pertinentes et des requêtes idéales (calculées par le retour de pertinence à court terme, voir chapitre 3) pour réactualiser les KVR. A chaque expérience, un algorithme d'apprentissage effectue 2 tâches : l'analyse de modèle et la révision de modèle en utilisant les opérateurs présentés dans la partie ci-dessus. La tâche d'analyse est de vérifier des conditions pour les opérateurs et la tâche de révision est d'exécuter les opérateurs sur le modèle Sacs de KVR. Ces algorithmes sont décrits maintenant.

Algorithme 1 Suivant l'hypothèse qu'il n'y a pas de connaissances dans le système au début de sa vie, le système ne peut rien faire d'autre que d'effectuer la recherche d'images par le contenu. En explorant les images, les utilisateurs peuvent trouver des informations intéressantes et ils souhaitent étiqueter des images avec des mots textuels réutilisables par eux-mêmes ou par d'autres utilisateurs (voir scénario 1). Dans ce scénario, les premiers KVR sont construits comme suit.

Tout d'abord, les images étiquetées avec le mot M sont regroupées en K groupes

en utilisant une méthode de regroupement. La requête idéale à points multiples est construite par la technique de retour de pertinence à court terme jusqu'au contentement de l'utilisateur. On considère chaque point de la requête idéale du mot M comme un nouveau candidat KVR ($Nouveau - KVR$). Ce nouveau candidat KVR est alors comparé avec tous les autres KVR du mot textuel M. Si $Nouveau - KVR$ n'est pas significativement différent d'un autre KVR, alors les 2 KVR sont fusionnés pour former un seul KVR et inversement si $Nouveau - KVR$ est très significativement différent de tous les KVR, alors un nouveau KVR pour le mot textuel M est ajouté.

Algorithme 1
Entrée : Une requête visuelle Q
Sortie : Sacs de KVR mis à jour
Début
Etape 1. Recherche d'images par le contenu
Etape 2. Retour de pertinence à court terme
L'utilisateur étiquette N images avec le mot textuel M
Regrouper N images en k groupes
Calculer les points de la requête idéale du mot M où
chaque point est considéré comme un nouveau KVR
Etape 3. Mettre à jour (analyse et revision)
Pour chaque nouveau KVR : $KVR_nouveau$
Pour tous les KVR du mot textuel M : $KVR_existant$
Si $EQUAL(KVR_nouveau, KVR_existant) = 1$
$KVR_merge = MERGE(KVR_nouveau, KVR_existant)$
Si KVR_merge est divisible
$SPLIT(KVR_merge)$
Sinon
$ADD(KVR_nouveau)$
Etape 4. Retour à l'étape 1
Fin

Algorithme 2 Après que les Sacs de KVR des concepts textuels soient appris, le système peut effectuer la recherche d'images en utilisant ces concepts. L'algorithme suivant est utilisé pour la recherche mixte texte/image ou la recherche par mots (voir scénario 2). Par le retour de pertinence à court terme pour la recherche mixte texte/image présentée dans la section 3.4.2, après chaque expérience nous avons (jusqu'à) 4 sous-requêtes de la requête Q, chaque sous-requête représentant le point idéal d'un mot. La limite est à 4 sous-requêtes parce que nous nous sommes basé sur la visualisation en coordonnées polaires présentée dans le chapitre 3. À cette étape, nous supposons qu'un mot k de la requête Q a déjà N KVR. Les Sacs de

KVR des mots sont alors mis à jour en fonction de 2 facteurs (figure 4.19).

En premier lieu, les précisions moyennes de la recherche d'images des 4 sous-requêtes sont utilisées. Une sous-requête (du mot k) dont la précision moyenne dans les 100 premières images de résultat est supérieure à un seuil p est considérée comme un nouveau candidat KVR ($Nouveau - KVR$) pour le mot k. Selon notre expérience, la meilleure valeur de p est 0.3. Un concept/mot est représenté par plusieurs KVR, dont chaque KVR est présent seulement pour un nombre d'images pertinentes mais pas pour toutes les images pertinentes dans la base d'images. Le nombre de 100 premières images est à relier avec notre hypothèse que l'interface ne permet de visualiser au maximum que 100 images et que la précision pour les 100 premières images est plus importante que la précision pour toute la base d'images.

Ensuite, le nouveau candidat KVR ($Nouveau - KVR$) est comparé avec tous les autres KVR du mot k (figure 4.19). Si le $Nouveau - KVR$ n'est pas significativement différent du KVR du mot k, alors ces 2 KVR sont fusionnés pour former un seul KVR ; inversement si le $Nouveau - KVR$ est très significativement différent de tous les différents mots textuels KVR, alors un nouveau KVR du mot k est ajouté.

Algorithme 2 :
Entrée : Une requête textuelle/visuelle Q
Sortie : Sacs de KVR mis à jour
Début
Etape 1. Recherche multimodale d'images par le texte et le contenu
Etape 2. Retour de pertinence pour la recherche mixte (boucle d'interaction)
Calculer les sous-requêtes correspondantes aux mots de la requête Q
$Sous_requete_i = Modifier_requete(Q, technique_Rochhio)$
Calculer un KVR correspondant au mot i
$KVRi =$ la dernière $Sous_requete_i$
Etape 3. Mettre à jour (analyse et révision)
Pour tous les KVR du mot i : $KVR_existant$
Si $EQUAL(KVRi, KVR_existant) = 1$
$KVR_merge = MERGE(KVRi, KVR_existant)$
Si KVR_merge est divisible
$SPLIT(KVR_merge)$
Sinon
$ADD(KVRi)$
Etape 4. Retour à l'étape 1
Fin

Notre système est capable de rechercher des images par le contenu, par une requête mixte texte/contenu ou par le texte seul. Les deux algorithmes ci-dessus

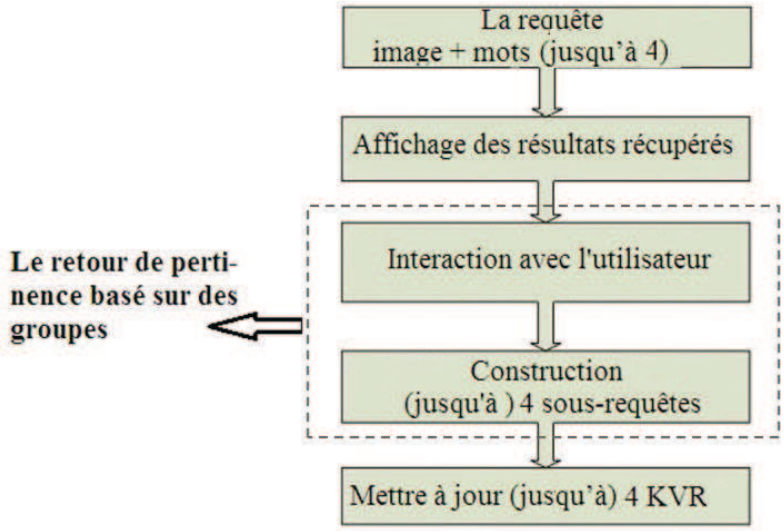

FIGURE 4.19 – Apprentissage incrémental des KVR fondé sur le retour de pertinence basé sur des groupes.

sont utilisés dans deux cas différents d'utilisation du système (deux scenarios présentés au début de cette partie) : l'algorithme 1 est fait pour la recherche par le contenu et l'algorithme 2 est fait à la fois pour la recherche par le texte et pour la recherche mixte texte/contenu. La seule différence entre ces deux algorithmes est l'étape de retour de pertinence qui est utilisée pour construire les candidats de KVR. L'algorithme 1 utilise le retour de pertinence de la recherche par le contenu, tandis que l'algorithme 2 utilise le retour de pertinence de la recherche mixte. Ces deux techniques de retour de pertinence ont été présentées dans le chapitre 3 (Interaction). L'évaluation sur ces 2 algorithmes est donnée dans le chapitre suivant, soit le chapitre 5 (Expérimentation).

4.4 Utilisation de Sacs de KVR

4.4.1 Propagation d'annotations

La propagation d'annotations est utilisée pour transférer des mots pour des images non annotées. Tandis que l'annotation manuelle demande beaucoup d'effort de la part de l'utilisateur, la propagation d'annotations peut être exécutée automatiquement. Dans notre système, la propagation d'annotations est réactualisée lorsque les KVR sont mis à jour (figure 4.20, étape 5). Ainsi, lorsque les KVR d'un mot k sont mis à jour, les similarités entre ce mot k et les images sont recalculées. En analysant des bases d'images avec annotations jointes comme Corel30K [Carneiro 2007] ou Caltech101 [Fei-Fei 2006], nous avons constaté que

chaque image peut être annotée avec plusieurs concepts tout en observant que seuls quelques concepts (jusqu'à 5) ont un poids important en général. Nous sommes donc partis sur l'hypothèse que le nombre de concepts associés à chaque image varie de 1 à 5. Donc si le mot est dans les 5 plus proches de l'image, alors le mot k est attribué à cette image et le 6ème mot est omis pour cette image. Chaque image prend ainsi toujours les 5 mots les plus proches comme annotations. L'annotation est améliorée avec l'évolution des représentations visuelles des mots. Du fait de l'utilisation de la méthode des k plus proches voisins, notre algorithme de propagation d'annotations converge forcément vers un résultat. Par contre, la qualité de ce résultat dépend de la qualité des KVR utilisés.

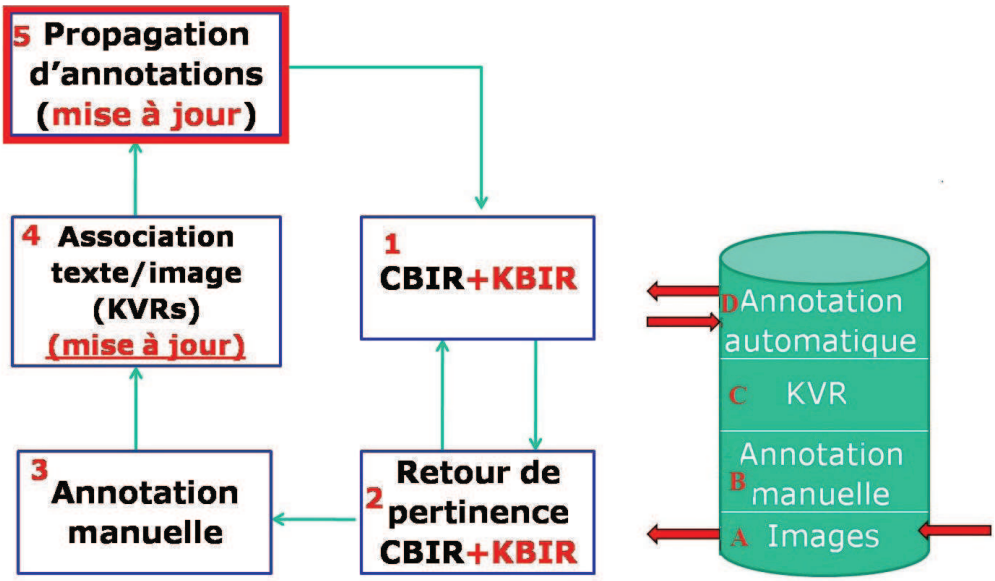

FIGURE 4.20 – La propagation d'annotation d'images par le modèle Sacs de KVR (étape 5).

Il est possible d'améliorer la propagation d'annotations en intégrant les corrélations entre les concepts dans la fonction de similarité des KVR et de l'image. Dans notre cas, la similarité entre un KVR et une image est calculée en fonction de la similarité visuelle. Cette mesure de similarité visuelle s'appuie sur la corrélation entre les concepts. La corrélation entre les 2 concepts k_1 et k_2 est calculée comme la probabilité $p(k_1, k_2)$ pour que ces deux concepts soient présents comme annotation pour la même image. Ces probabilités d'une paire de concepts peuvent être calculées par apprentissage à partir d'un ensemble de données d'apprentissage ou en utilisant une ontologie comme WordNet [1]. Basé sur l'intégration des corrélations entre concepts, la similarité entre un KVR et une image est donc calculée comme suit :

1. WordNet : http ://wordnet.princeton.edu/

1. Le concept (*keyword*) le plus proche de k_1 aura la similarité avec le KVR comme suit :

$$Sim(k_1, KVR) = Sim_visuelle(k_1, KVR)$$

   la similarité visuelle $Sim_visuelle(k_1, KVR)$ est calculée suivant l'équation (8).

2. Le concept suivant k_n aura comme distance :
   $Sim(k_n, KVR) = Sim_visuelle(k_n, KVR) \prod_{i=1}^{n-1} p(k_i, k_n)$

La similarité visuelle *sim_visuelle* entre le concept k_1 (mot textuel) et le KVR $KVR1$ dans les deux équations ci-dessus se calcule en transformant le concept k_1 en sacs de KVR (selon les correspondances déjà apprises, cf section 4.3) et ensuite en utilisant l'équation (8) pour calculer la similarité entre les KVR calculés et $KVR1$.

4.4.2 La recherche d'images par la requête textuelle

Le texte et le contenu visuel correspondent à des niveaux sémantiques différents. Le texte manipule davantage d'informations sémantiques, alors que le contenu visuel est plus perceptif. Ces 2 types d'information sont complémentaires et fournissent des aspects très différents pour la recherche d'images. Cependant, la formation de la requête est plus difficile pour la recherche d'images par le contenu car l'utilisateur doit fournir des images exemples qui ne sont pas toujours disponibles et qui ne sont pas toujours représentatives de toutes intentions de l'utilisateur. C'est d'ailleurs un problème majeur dans les systèmes CBIR. Par contre, à l'opposé un problème de la recherche d'images par le texte est l'indisponibilité des annotations d'images.

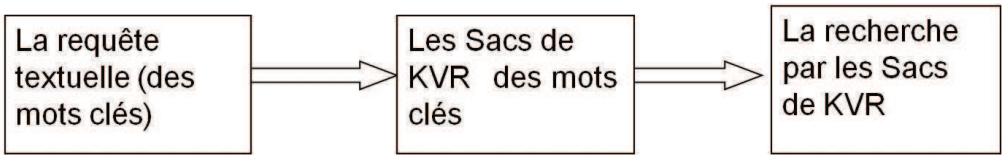

FIGURE 4.21 – La recherche d'images par le contenu dans notre système.

Ces deux problèmes peuvent être surmontés en utilisant la représentation Sac de KVR des concepts. Dans notre système, nous pouvons rechercher des images par le texte et/ou par le contenu en utilisant des requêtes textuelles même si des informations textuelles ne sont pas disponibles initialement dans la base de données. En profitant des interaction des utilisateurs, le système construit ses connaissances, ou

en d'autres termes des annotations à différents niveaux, qui évoluent dans le temps et qui permettent une utilisation continue du système. Nous pouvons transformer automatiquement une requête textuelle en une requête visuelle par la représentation Sac de KVR (voir figure 4.21). Donc, pour une base d'images annotée partiellement ou pour une nouvelle base d'images sans connaissances/annotations jointes nous pouvons utiliser la requête textuelle dès que les premiers KVR sont construits (c'est-à-dire à partir des annotations partielles ou dès les premières interactions). Au départ, pour une base d'images sans annotations, avec peu d'interactions effectuées, et donc peu de KVR construits, les résultats ne sont bien évidemment pas extra-ordinaires, mais ils ont le mérite d'exister et d'offrir la possibilité d'utiliser le système malgré tout ; par ailleurs, ces résultats s'améliorent au fur et à mesure de l'utilisation du système (c'est-à-dire de la construction et du raffinement des KVR).

4.4.3 Comparaison avec les autres modèles

Dans le tableau suivant (tableau 4.1), nous proposons une comparaison de notre modèle avec d'autres modèles d'association texte/image. Les avantages de notre modèle s'appuient sur la possibilité d'utiliser 2 outils différents : 1) la transformation de l'image en texte (annotation d'images) et 2) la transformation du texte en image. Il est possible de rechercher des images par le contenu en utilisant une requête textuelle ou une requête visuelle même s'il y a une partie de la base d'images qui ne comporte pas d'annotations. En d'autres termes nous pouvons rechercher des images non annotées en utilisant des mots textuels (requête textuelle). De plus, notre modèle n'est pas soumis aux difficultés liées à certaines techniques, telles que la segmentation d'images. Toutefois, la performance de l'annotation d'images et de la recherche d'images s'appuie fortement sur la performance de l'apprentissage de Sacs de KVR qui est basé principalement sur le regroupement (clustering). Cela conduit potentiellement à des représentations imparfaites des Sacs de KVR de concepts, du fait de l'imperfection du regroupement. L'imperfection du regroupement est liée à la construction des groupes hétérogènes d'images et aussi à la sélection du nombre de groupes.

Un autre avantage majeur de notre démarche est qu'il n'est pas nécessaire de procéder à un apprentissage hors ligne pour l'annotation d'images. Notre système est en effet capable d'acquérir de manière incrémentale des connaissances sans aucune connaissance au début de vie du système. La grande différence de notre système par rapport à d'autres systèmes est donc la source de connaissances. Les autres systèmes requièrent en effet de la connaissance a priori ou issue de l'internet, alors que la connaissance de notre système vient uniquement de l'interaction avec des utilisateurs/experts. De plus, les connaissances du système sont mises à jour au

Système	**Image->Texte**	**Texte->Image**	**Support la recherche multimodale**	**Source de la connaissance**
Modèle cooccurrence	Oui	Non	Non	a priori
Modèles de traduction automatique	Oui	Non	Non	a priori
LSA (Latent Semantic Analysis)	Oui	Non	Oui	a priori
Modèle de transformation	Oui	Oui	Oui	a priori + WordNet
Notre modèle Sacs de KVR	Oui	Oui	Oui	Interaction

TABLE 4.1 – Comparaison des modèles d'association du texte et du contenu.

fur et à mesure sans demander un apprentissage hors ligne. Cela est différent des autres modèles pour lesquels l'apprentissage hors ligne doit être réitéré à chaque fois que des connaissances supplémentaires arrivent ou que les connaissances changent.

Cependant, en contre-partie, notre modèle demande des interactions de l'utilisateur, ce qui n'est pas toujours nécessaire dans les autres modèles. Cela signifie que la performance des connaissances dans notre système dépend de la durée de l'apprentissage (ou en d'autres termes de l'intensité de l'utilisation du système). De plus dans notre modèle, la visualisation interactive pose potentiellement un problème lié au nombre de concepts utilisés dans la requête. En effet, notre approche ne supporte qu'une requête textuelle de 1 à maximum 4 concepts. Enfin, nos travaux liés à la visualisation ne sont pour l'instant que préliminaires et requièrent d'être largement ré-étudiés et améliorés.

4.5 Conclusion

Dans ce chapitre, nous avons présenté des modèles d'association texte/image et notre modèle basé sur des Sacs de KVR. Le modèle Sacs de KVR est inspiré du modèle Sacs de Mots avec une structure similaire. En profitant du retour de pertinence à court terme, notre modèle construit un dictionnaire des représentations visuelles de concepts textuels. Avec l'hypothèse que chaque concept peut avoir de multiples représentations visuelles (KVR - *Keyword Visual Representation*), nous avons proposé des techniques de fusion/division des KVR pour améliorer la précision et représenter des concepts textuels par des Sacs de KVR. Les Sacs de KVR des concepts textuels sont utilisés pour l'annotation d'images et la recherche d'images.

Notre modèle donne au système CBIR la capacité de l'interrogation par l'image, l'interrogation par mot textuel ou des requêtes hybrides même s'il existe dans la base d'images une partie sans annotation.

Chapitre 5

Expérimentation

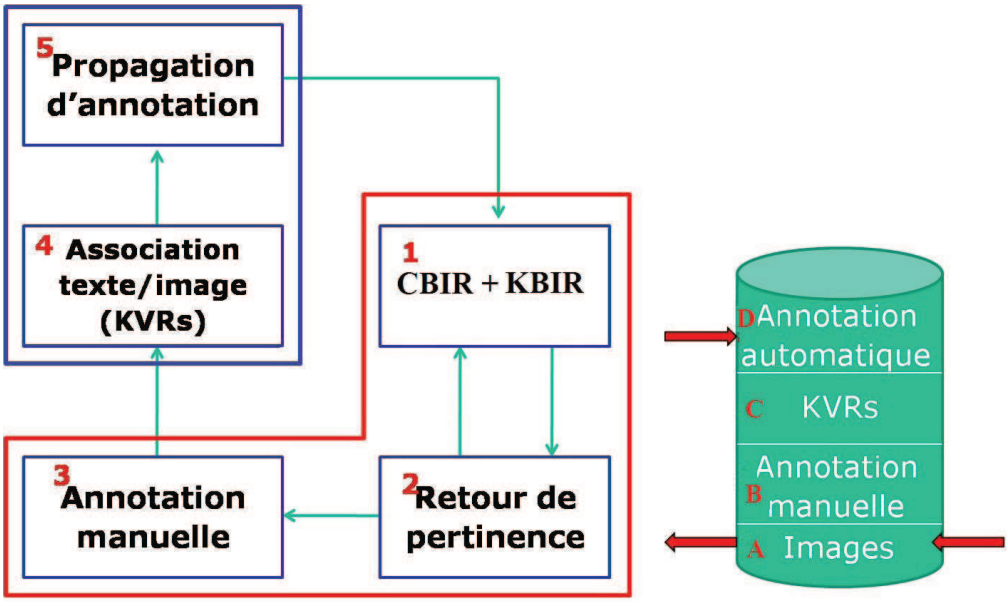

FIGURE 5.1 – Système de recherche d'images

Dans les chapitres précédents, nous avons décrit les différents étages de notre système : La recherche mixte par le texte/contenu d'images dans le chapitre 2 (étapes 1-2, figure 5.1), le retour de pertinence dans le chapitre 3 (étapes 2-3, figure 5.1) et l'apprentissage de la représentation visuelle des concepts - le modèle Sacs de KVR dans le chapitre 4 (étapes 4-5, figure 5.1). Notre système de recherche et d'annotation d'images est construit à partir de ces différentes étapes. Pour rappel, nous résumons tout d'abord ici les hypothèses sur lesquelles nous nous appuyons pour l'élaboration de notre système, ainsi que son contexte d'utilisation :

– Le travail se fait sur une grande base d'images sans connaissance préalable ;

- Le volume d'images n'est pas fixe, et de nouvelles images sont ajoutées dans le temps ;
- La connaissance du système s'appuie sur les annotations des images (images présentées par le texte) et sur les Sacs de KVR de concepts (texte représenté par l'image) ;
- L'apprentissage est interactif et peut se faire par renforcement et/ou de manière incrémentale ; L'interaction entre l'utilisateur, les experts du domaine et le système pour améliorer la connaissance globale de système doit se faire de façon simple en cliquant des images pertinentes/non pertinentes pendant la recherche/l'exploration ;
- Très peu de données d'entraînement (renforcement/incrémental) ; Le nombre d'images cliquées pour chaque interaction doit être très faible (20 maximum) ;
- L'annotation des images se fait en temps réel ;
- Le nombre d'annotations pour chaque image varie de 1 à 5 (cette idée est rediscutée dans les perspectives de cette thèse).

L'objectif de ce chapitre est d'évaluer en détails notre contribution concernant le modèle Sacs de KVR, l'évolution des connaissances, l'annotation d'images et la recherche d'images, afin de donner une évaluation globale pour notre système. Ce chapitre est organisé comme suit. D'abord, dans la section 5.1 nous présentons les difficultés de mise en place d'un protocole objectif d'évaluation de performance, ainsi que la méthodologie retenue pour évaluer notre système sur la base d'expérimentations significatives. Ensuite, la section 5.2 présente la base de données utilisée pour l'évaluation de notre système et la section 5.3 montre le protocole d'expérimentation. Enfin, la section 5.4 présente la performance du modèle Sacs de KVR pour l'apprentissage de connaissances, ainsi que pour l'annotation d'images et la recherche d'images dans notre système.

5.1 Les difficultés de l'évaluation du système

La première difficulté est le caractère "dynamique" de l'apprentissage. Le principe de notre système est l'apprentissage par renforcement via des interactions entre les utilisateurs et la machine. Les connaissances du système sont apprises à partir de celles des utilisateurs/experts pendant la recherche et/ou l'exploration interactive d'images. Afin d'évaluer notre système, nous devons donc déclencher des interactions entre des utilisateurs/experts et la machine. Cette tâche demande un travail conséquent en temps et en effort pour l'expérimentation.

Une autre difficulté concerne la base d'images. Dans notre système, il n'y a pas de frontière entre la base d'apprentissage et la base de tests. Avec l'hypothèse du caractère "dynamique" de la base d'images, son volume n'est pas fixe. De nouvelles

images arrivent constamment, rendant difficile la simulation de la base.

De plus, du fait de son caractère "dynamique", notre apprentissage n'est pas comparable avec l'existant. A notre connaissance, il n'existe pas d'autres travaux de recherche et d'annotation d'images qui ont le même caractère "dynamique" que notre système.

5.1.1 Orientations retenues pour l'expérimentation et l'évaluation

Tout d'abord, afin d'éviter les fastidieuses interactions entre l'utilisateur et la machine, nous avons décidé de mettre en place un protocole permettant de les simuler. Grâce à cette simulation des interactions, notre système est expérimenté de manière automatique sans demander l'intégration des utilisateurs/experts. Il est donc beaucoup plus facile à expérimenter (en plusieurs fois) par rapport à une expérimentation avec des utilisateurs réels. Cette simulation est basée sur des agents qui sont décrits en détails dans le protocole d'expérimentation. Concernant la base d'images, nous la divisons en deux parties : une partie existante et une partie nouvelle, ces deux parties étant discutées en détails dans la section sur le protocole d'expérimentation. La base d'images utilisée est présentée dans la section suivante. Enfin, nous évaluons le caractère "dynamique" de notre système en fonction du temps. Une comparaison de l'annotation et de la recherche d'images de notre approche (avec le simulateur automatique) avec d'autres approches est également proposée dans notre expérimentation.

5.2 Présentation de la base de données pour l'expérimentation

La base de données Corel 30K [Carneiro 2007] est utilisée dans nos expériences. Il s'agit d'un sous-ensemble de la base de données Corel. Une vue générale de la base Corel 30K est présentée dans la figure 5.2. Cette base contient 31695 images divisées en différentes catégories par des experts du domaine, à raison de 100 images par classe. La taille de chaque image est de 384 x 256 pixels. Les images sont annotées en utilisant un total de 5587 mots. Cependant, seuls 950 mots sur les 5587 ont été utilisés comme annotations pour au moins 10 images. Pour notre expérimentation, nous n'utilisons pas des mots qui ont un trop petit nombre d'exemples (<10 images), afin de rester sufisamment générique, puisque ces mots (avec trop peu d'images) ne sont pas assez généraux pour évaluer le résultat.

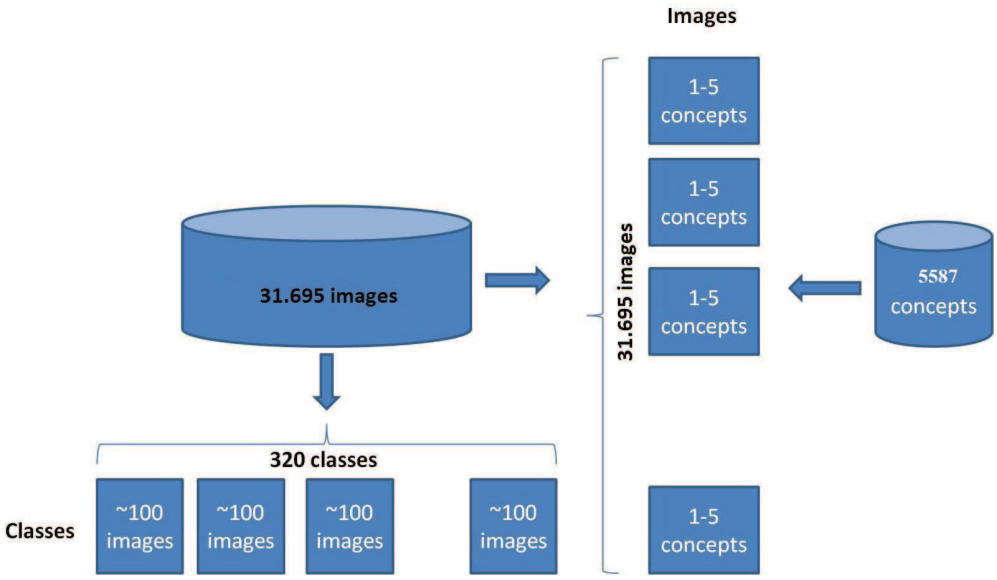

FIGURE 5.2 – Une vue générale de la base Corel 30K.

FIGURE 5.3 – Quelques catégories de la base Corel 30K avec des mots clés.

Pour évaluer le système de recherche et d'annotation d'images, un ensemble de requêtes est déclenché et des jugements de pertinence correspondant sont ana-

lysés. Dans la plupart des travaux de la littérature, les catégories d'images de la base Corel sont utilisées comme vérité terrain. Bien que les catégories de Corel contiennent des images d'un même sujet, de nombreux groupes d'images sont visuellement dissemblables et impactent négativement et de manière significative sur la performance du système. Cette difficulté s'explique par le fait que Corel n'est pas construit pour cette problématique (recherche en traitement d'images / vision par ordinateur) et donc les classes/catégories (voir figure 5.3) de cette base ne sont pas construites dans cette optique. Dans le domaine de la vision par ordinateur, les recherches considèrent parfois que Corel n'est pas une bonne base pour l'expérimentation [Muller 2002]. Toutefois, pour notre système, les annotations manuelles de cette base Corel30K seront utilisées comme vérité terrain pour l'apprentissage de connaissances car nous considérons que cette base est suffisante et raisonnable pour son évaluation. Nous pouvons résumer le protocole de vérité terrain comme suit :

- Pour l'apprentissage de connaissances, ainsi que pour évaluer l'annotation d'images et la recherche mixte d'images en utilisant des Sacs de KVR (avec requête textuelle), les annotations manuelles (mots-clés) de la base Corel30K sont utilisées comme vérité terrain.
- Pour évaluer la recherche mixte d'images par le texte et par l'exemple, la vérité terrain est décrite comme la combinaison des annotations manuelles de la base Corel et des classes/catégories de la base Corel. Les annotations manuelles sont utilisées comme vérité terrain pour des requêtes textuelles, et les classes/catégories sont utilisées pour des requêtes visuelles. Les images sont considérées comme pertinentes si associées avec les mots-clés de la requête textuelle d'une part et si elles sont dans la même classe que l'image d'exemple d'autre part.

Une autre limite importante de la base Corel30K concerne le caractère "vague" ou abstrait de certains concepts dans les annotations. Il est parfois difficile de représenter ces concepts par des caractéristiques visuelles. [Yanai 2005] propose de déterminer dans quelle mesure un concept a des caractéristiques visuelles discriminantes pour l'annotation d'images, ce qu'il appelle le *visualness* (ou caractère visuel) d'un concept. Par exemple, il est impossible de lier le concept "hiver" à des caractéristiques visuelles, ou encore c'est difficile de représenter le concept "animaux" par des caractéristiques visuelles du fait de la grande variabilité de représentation des animaux en termes de types, formes, textures ou couleurs.

5.3 Protocole d'expérimentation

Comme déjà mentionné ci-dessus dans la section sur les difficultés de l'expérimentation, il est difficile d'évaluer notre système du fait du caractère "dynamique". Il n'existe pas de vrai protocole, ce qui nous a conduit à en proposer un pour toute la partie dynamique. Dans les sections suivantes, nous parlons d'abord des validations croisées pour l'expérimentation et des scénarios de recherche d'images dans notre système. Ensuite, nous proposons à la suite l'utilisation d'une méthode d'évaluation pseudo interactive, basée sur des agents. Ces agents sont utilisés pour simuler les interactions entre les utilisateurs et la machine. Enfin, deux types d'évaluation sont discutées, l'évaluation de l'évolution de connaissances (le caractère "dynamique") et l'évaluation de l'utilisation des Sacs de KVR (le caractère "statique").

5.3.1 Validations croisées pour l'expérimentation

Dans notre système, la base d'images est dynamique puisque nous faisons l'hypothèse que de nouvelles images arrivent en temps réel. A un instant t, l'apprentissage de connaissances se fait avec les données existantes dans la base, sur lesquelles s'ajouteront ultérieurement de nouvelles images arrivant après t, divisant ainsi la base en deux parties. Nous appelons la première partie "partie initiale" ou "partie existante" et la deuxième partie "partie nouvelle" ou "partie ajoutée". La partie initiale sert à apprendre des connaissances et la partie nouvelle sert uniquement pour l'évaluation.

Notre système est évalué en effectuant quatre validations croisées pour la base d'images du système, correspondant aux différentes situations que le système est susceptible de rencontrer. Un pourcentage de la base d'images possède des connaissances et un autre n'a aucune connaissance. Pour ces quatre situations, nous définissons arbitrairement la partie initiale comme correspondant à des proportions de 20%, 50%, 75%, 95% de la base Corel 30K. L'apprentissage de connaissances est effectué via l'exploration et la recherche d'images de ces proportions de la base. Les 80%, 50%, 25% et 5% respectivement restants sont retenus comme de nouvelles images arrivées dans le système (la partie nouvelle). Notre expérience s'appuie sur 1000 sessions de recherche pour l'apprentissage de connaissances. Si on considère l'utilisation normale du système par plusieurs utilisateurs, 1000 sessions est un faible nombre dans la vie d'un système qui est utilisé à long terme (par exemple, 1000 sessions pendant quelques jours ou quelques semaines).

Dans chaque cas l'apprentissage est répété 3 fois (avec 1000 différentes sessions de recherche d'images pour chaque fois) et nous retenons la moyenne. Nous définissons alors les 4 validations croisées représentant 4 conditions d'utilisations du

système :

1. **Expérimentation 1** : La partie initiale de la base d'images correspond à 20% de la base Corel 30K. Cette proportion de la base Corel30K est appelée la **partie "initiale" 1**. Les utilisateurs font de l'exploration et de la recherche interactive d'images sur cette partie. Les 80% restants de la base Corel 30K sont considérées comme de nouvelles images arrivant dans le système. Cette proportion est nommée la **partie "nouvelle" 1**. Dans ce schéma d'utilisation du système, l'utilisateur construit des connaissances sur peu d'images, et donc beaucoup images nouvelles arrivent sans aucune connaissance.
2. **Expérimentation 2** : La partie initiale de la base d'images correspond à 50% de la base Corel 30K. Cette proportion de la base Corel30K est appelée la **partie initiale 2**. Les 50% restants de la base Corel 30K sont considérés comme étant la **partie nouvelle 2**.
3. **Expérimentation 3** : La partie initiale de la base d'images correspond à 75% de la base Corel 30K. Cette proportion de la base Corel30K est appelée la **partie initiale 3**. Les 25% restants de la base Corel 30K sont considérés comme étant la **partie nouvelle 3**.
4. **Expérimentation 4** : La partie initiale de la base d'images correspond à 95% de la base Corel 30K. Cette proportion de la base Corel30K est appelée la **partie initiale 4**. Les 5% restant de la base Corel 30K sont considérés comme étant la **partie nouvelle 4**. Dans ce schéma d'utilisation du système, l'utilisateur construit des connaissances sur de nombreuses images et peu d'images nouvelles arrivent sans aucune connaissance.

Dans la réalité, des situations aussi diverses que les expérimentations 1 à 4 sont tout à fait possibles et peuvent même parfois évoluer dans le temps d'une situation à l'autre.

Au début de la vie du système, il n'y a pas d'annotations, ni pour la partie initiale d'images du système, ni pour la partie nouvelle. Autrement dit, il n'y a pas de connaissance a priori. Dans notre système, selon l'hypothèse que les nouvelles images arrivent dans le système en temps réel, la base d'images du système augmente en taille. Cependant, dans le cadre de cette expérimentation nous ne simulons pas pour l'instant l'évolution du nombre d'images. Nous n'utilisons la partie nouvelle que pour regarder la performance de la propagation des annotations et l'utilisation de Sacs de KVR sur les nouvelles images arrivées dans le système.

On utilise les concepts de la base Corel comme la vérité terrain. Il y a 1036 concepts si nous considérons seulement les concepts avec plus de 10 images annotées dans la base Corel 30K. Comme nous ne considérons que les concepts présents à la fois dans les 2 parties initiale et nouvelle, nous avons donc 929 concepts pour

l'**expérimentation 1** ; 957 concepts pour l'**expérimentation 2** ; 951 concepts pour l'**expérimentation 3** ; 935 concepts pour l'**expérimentation 4**.

5.3.2 Scénarios

Dans notre système, les scénarios suivants sont proposés pour l'expérimentation :

1. L'utilisateur prend une image comme requête et interroge le système. C'est une recherche par le contenu (CBIR). Pendant l'exploration des résultats, l'utilisateur observe les images renvoyées et qui sont d'intérêt pour lui et leur affecte des mot-clés (notre interface de visualisation ne supporte qu'un seul mot-clé pour l'instant). Il peut terminer ou poursuivre cette recherche pour trouver d'autres images par retour de pertinence.
2. L'utilisateur prend une image et un ou plusieurs mots-clés pour lancer une requête sur la base. C'est une recherche mixte par requête mixte texte/image. Il continue à interagir avec le système par le retour de pertinence jusqu'à ce qu'il soit satisfait.
3. L'utilisateur prend un ou plusieurs mots-clés pour faire une recherche mixte d'images. C'est alors ici une recherche mixte d'images par la requête textuelle.

Les deuxième et troisième types de requêtes ne peuvent bien sûr être utilisés que dans le cas d'existence d'annotations dans la base de connaissances du système. Il faut distinguer deux bases de connaissances :

1. **La base de connaissances du système** : Ces sont des annotations propagées ou des annotations manuelles qui sont assignées aux images via des interactions avec les utilisateurs (pendant l'apprentissage de connaissances). Il s'agit également de connaissances de niveau "image", c'est-à-dire des représentations Sacs de KVR de concepts. Cette base existe tout le temps pour le système..
2. **La base de connaissances simulées (connaissances des utilisateurs)** : Nous utilisons une méthode d'évaluation pseudo interactive basée sur la simulation, comme il n'y a pas d'utilisateurs réels ici. Dans notre expérimentation, les annotations manuelles (des utilisateurs) sont simulées en utilisant les annotations existantes de la base Corel30K. Cette base n'est pas utilisée par le système, mais seulement dans la simulation des utilisateurs.

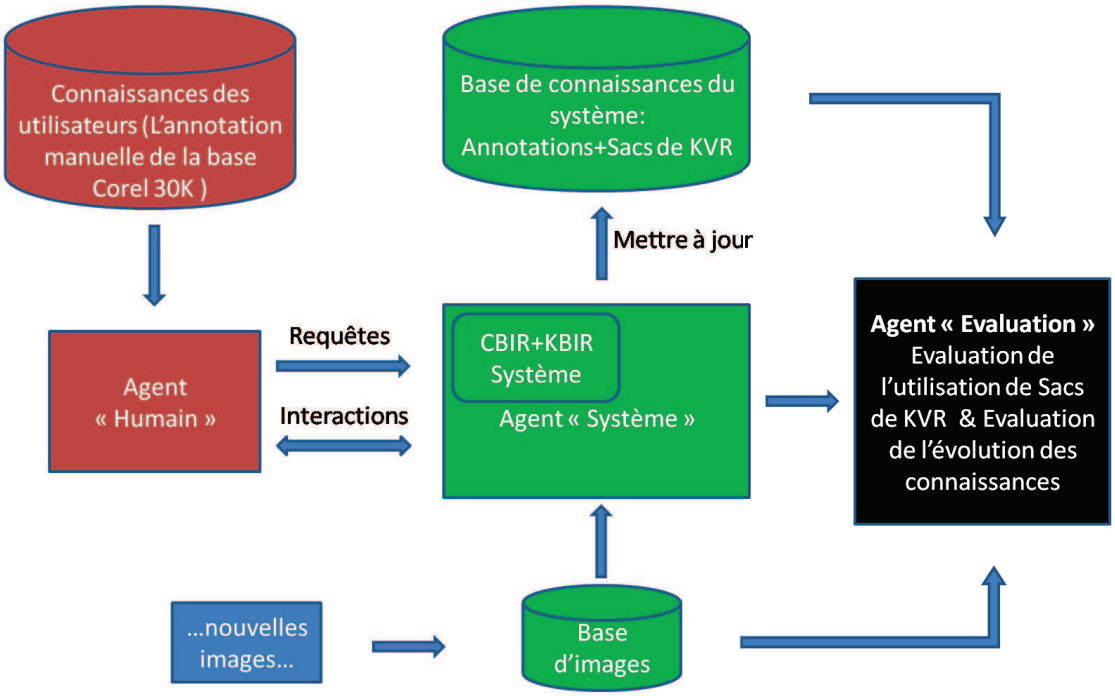

FIGURE 5.4 – Agents pour simuler les interactions. La section en rouge est la simulation des interactions des utilisateurs. La section en vert est le système que l'on veut évaluer. La section en noir représente le processus d'évaluation du système. La base d'images est dynamique avec des nouvelles images arrivant dans le temps.

5.3.3 Méthode d'évaluation pseudo interactive

Pour évaluer notre système, il faut réaliser un grand nombre d'interactions avec l'utilisateur, ce qui est difficile à expérimenter. Afin d'éviter cette intégration de l'utilisateur réel et produire une procédure plus automatisée et surtout répétable d'évaluation, nous proposons l'utilisation d'une méthode d'évaluation pseudo interactive, basée sur des agents, qui est décrite ci-dessous.

Dans l'expérimentation, nous proposons d'utiliser des agents informatiques pour remplacer les utilisateurs et de procéder à des interactions automatisées. Les annotations manuelles de la base Corel 30K sont utilisées comme connaissances des utilisateurs et des experts dans les interactions. Différents types d'agents sont utilisés pour simuler les interactions, mais également pour l'évaluation (figure 5.4) :

1. Un agent "Système" pour effectuer toutes les opérations normales définies dans l'architecture du système (figure 5.1).
2. Un agent "Humain" pour spécifier des requêtes (requêtes sous forme d'images choisies dans la **partie initiale** et/ou requêtes texte en utilisant sa **base de connaissances simulées**) et des interactions automatiques (en se basant

sur sa **base de connaissances simulées**).

3. Un agent "Evaluation" pour collecter les données et produire l'évaluation des résultats. Nous évaluons l'évolution des connaissances du systeme et l'utilisation de Sacs de KVR pour l'annotation et la recherche d'images. Nous évaluons l'utilisation de Sacs de KVR sur deux parties de la base d'images : la **partie initiale** et la **partie nouvelle** pour chaque **expérimentation**. L'évaluation sur la partie initiale nous donne la qualité de connaissances du système au temps présent, tandis que l'évaluation sur la partie nouvelle nous donne l'efficacité des connaissances pour les recherches futures.

Dans les parties suivantes, nous présentons en détails notre méthode de simulation : la simulation de connaissances des utilisateurs/experts, la simulation de la spécification de la requête et la simulation du retour de pertinence.

5.3.3.1 Simulation de connaissances des utilisateurs/experts

Les connaissances des utilisateurs/experts (voir figure 5.4) sont simulées en utilisant la base d'images du système (la base d'images dans les 4 expérimentations que nous avons présentées). Pour la simulation des connaissances des utilisateurs/experts, la vérité terrain de la base Corel 30K est utilisée. Comme exemple de "connaissance" de l'utilisateur, on considère qu'une "image A est liée sémantiquement au concept B" si l'image A est annotée par le concept B dans la vérité terrain. Nous pouvons remarquer donc que la "quantité" de connaissances simulées est différente dans les 4 expérimentations puisque cette grandeur correspond au nombre d'annotations d'images dans la partie initiale de la base d'images pour chaque expérimentation. Notre système est expérimenté avec différentes bases de connaissances des utilisateurs/experts.

5.3.3.2 Simulation de la spécification de la requête

L'agent "Humain" prend aléatoirement 1000 images (dans la base d'images) et procède alors à des recherches. S'il n'y a pas d'annotations (connaissances du système) pour cette image, nous nous trouvons dans le scénario de requête d'exemple. Dans le cas contraire, c'est une requête combinée d'une image et de mots-clés, qui sont des annotations de l'image. Notre expérimentation s'appuie sur 1000 sessions de recherche, avec une hypothèse d'un faible nombre d'interactions entre l'utilisateur et la machine. Après chaque interrogation, les connaissances du système sont mises à jour.

5.3.3.3 Simulation du retour de pertinence

La simulation du retour de pertinence pour les trois scenarios présentés est décrite ci-dessous :

Scénario 1 :

- Selon ce scénario, l'utilisateur affecte un mot-clé pour les images, retournées pour une requête, qu'il trouve intéressantes pour lui. L'agent "Humain" simule ce scénario en prenant aléatoirement un mot-clé M de l'image requête pour affecter les images pertinentes et les retourner au système. Dans notre experimentation, chaque processus de recherche se base sur 5 itérations de retour de pertinence. On considère que cet agent indique les images comme pertinentes si celles-ci ont le mot M dans leurs annotations (selon la vérité terrain dans notre système).
- Après chaque retour de pertinence, de nouveaux KVR sont fusionnés/ajoutés avec les KVR existants. Les annotations propagées sont alors mises à jour en utilisant ces Sacs de KVR.

Scénario 2 :

- L'agent "Humain" simule le scénario 2 en prenant une image requête et ses mots-clés (des connaissances du système) pour faire le retour de pertinence (5 itérations de retour de pertinence, classique ou basé sur des clusters - voir chapitre 3). Comme pour le scénario 1, des images pertinentes/non pertinentes sont sélectionnées en appui de la vérité terrain de la base Corel (les annotations manuelles existantes dans Corel).
- Après chaque retour de pertinence, de nouveaux KVR sont fusionnés/ajoutés avec les KVR existants. Les annotations propagées sont mises à jour en utilisant les Sac de KVR qui sont également mis à jour.
- Ce scénario 2 est réalisé si l'image que l'agent "Humain" prend aléatoirement possède des annotations (des connaissances du système). Ces annotations sont des annotations manuelles données par les utilisateurs via l'interaction, ou des annotations propagées par le système.

Scénario 3 :

- L'agent "Humain" simule le scénario 3 en prenant aléatoirement une requête textuelle de quelques mots-clés (de 1 à 5) pour faire l'interrogation avec des retours de pertinence (5 itérations, classique ou basé sur des clusters). Comme évoqué dans le scénario 1, des images pertinentes/non-pertinentes sont sélectionnées en appui sur la base de la vérité terrain de la base Corel (des annotations manuelles de Corel).

- Apres chaque retour de pertinence, des nouveaux KVR sont fusionnés/ajoutés avec les KVRs existants. Une fois les Sacs de KVR mis à jour, une propagation automatique des annotations est faite dans toute la base.
- Ce scénario est réalisé si l'image que l'agent "Humain" a choisie aléatoirement a des annotations (possède des connaissances dans le système). Ces annotations sont des annotations manuelles données par des utilisateurs via l'interaction, ou des annotations propagées par le système.

Ces trois scénarios sont utilisés alternativement dans notre apprentissage incrémental via des interactions simulées. Par exemple, quand l'agent "Système" prend aléatoirement une image pour faire une interrogation, il existe 3 types de situation correspondant aux différents scénarios : 1) L'image requête n'a pas d'annotations (connaissances du système). Le scénario 1 sera exécuté dans ce cas ; 2) L'image requête a des annotations (des connaissances du système), alors dans ce cas, le sénario 2 ou 3 est choisi.

5.4 Résultats

Dans cette section, nous présentons les résultats de notre expérimentation. L'évaluation est effectuée pour répondre aux questions suivantes :

1. Est-ce que le système est capable de transférer la connaissance depuis les utilisateurs/experts vers le système ?
2. Quels types de connaissances sont bien apprises ?
3. Comment la connaissance évolue en fonction du temps, en termes de quantité et de qualité ?
4. Comment la base de connaissances des utilisateurs influence les connaissances du système, en termes de quantité de connaissances ?
5. Comment les conditions d'utilisation du système influencent ses connaissances ?
6. Est-ce que le système possède des avantages particuliers pour faire de l'annotation d'images et de la recherche d'images ?

5.4.1 L'évolution des connaissances, l'évaluation du caractère "dynamique"

Considérant la dynamique de notre système, il est important de se pencher sur l'évaluation des sacs de KVR et l'évolution de connaissances. En effet, en recherchant des informations et en explorant la base d'images, les utilisateurs effectuent

des interactions avec le système et enrichissent de ce fait sa base de connaissances. Ces connaissances apprises dynamiquement sont de 3 types :

1. Annotations manuelles : Des images se voient affecter des concepts/mots textuels par les utilisateurs/experts pendant l'interaction.
2. Annotations propagées : D'autres images dans la base se voient affecter dynamiquement des concepts textuels en utilisant des Sacs de KVR des concepts.
3. Connaissances de niveau "image" : Les représentations visuelles de concepts ou en d'autre termes, les Sacs de KVR de concepts.

Pour évaluer le modèle de Sacs de KVR et la performance du système, nous proposons d'évaluer l'évolution globale des connaissances. Nous observons la quantité et la qualité des connaissances en fonction du temps, ce qui signifie :

- La quantité : Le nombre d'annotations manuelles, le nombre d'annotations propagées, le nombre de Sacs de KVR. Ces nombres montrent l'évolution du volume de connaissances. Nous pouvons observer la relation entre l'évolution du volume de connaissances et l'évolution de la qualité des connaissances.
- La qualité : Nous tentons d'évaluer la qualité des connaissances liées aux annotations et ainsi que les connaissances de niveau "image", c'est-à-dire les Sacs de KVR des concepts. La précision des annotations propagées est utilisée pour évaluer les connaissances d'annotations et la précision de la recherche d'images par concepts (en utilisant des Sacs de KVR de concepts) est utilisée pour évaluer les connaissances de niveau "image". Nous évaluons sur deux parties de la base d'images : la partie d'images existantes du système (la partie initiale), la partie des nouvelles images arrivées (la partie nouvelle).

Après chaque session de recherche d'images (avec retour de pertinence), les informations ci-dessus sont calculées. Nous allons évaluer l'évolution sur 1000 sessions.

Rappel et précision : Soient *relevant* le nombre total d'images pertinentes et *retrieved* le nombre d'images trouvées. Le *rappel* exprime la proportion d'images correctement trouvées parmi toutes les images pertinentes. Il s'exprime par la formule :

$$rappel = \frac{relevant \cap retrieved}{relevant}$$

La *precision*, quant à elle, représente la proportion d'images correctement trouvées dans le résultat de la recherche et s'exprime par la formule :

$$precision = \frac{relevant \cap retrieved}{retrieved}$$

5.4.2 La quantité de connaissances

5.4.2.1 Le nombre de concepts appris

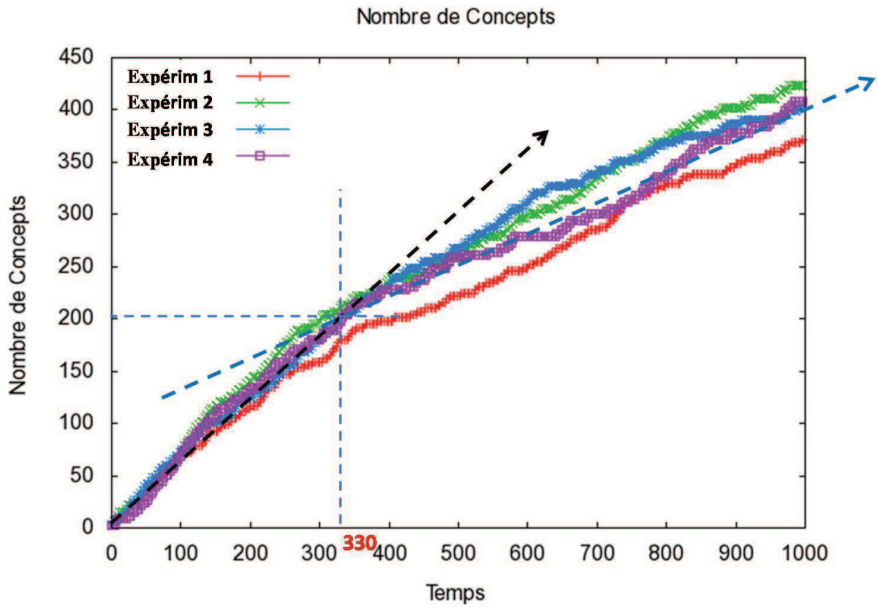

FIGURE 5.5 – Nombre de concepts appris en fonction du temps (itérations). Dans un premier temps, avec le point à $t = 330$, on voit qu'on apprend 200 concepts. Nous observons donc un taux rapide jusqu'à environ la moitié des concepts (200) identifié par la flèche noire en pointillé. Dans un second temps, nous observons un taux beaucoup plus lent identifié par la flèche bleue en pointillé, jusqu'à environ 400 concepts appris à la fin.

Dans les courbes présentées ici (figure 5.5), on remarque une croissant rapide du nombre de concepts appris au début, et passé un certain point, identifié autour de $t = 330$, le taux d'augmentation du nombre de concepts appris diminue. On a une augmentation très rapide au début, identifiée par la flèche noire en pointillé, comme la première phrase. Après un certain point, identifié sur la figure, on observe dans un second temps une augmentation moins rapide du nombre de concepts appris, identifié par la flèche bleue en pointillé.

En fin d'expérimentation (1000 sessions de recherche), on voit sur la figure que le nombre maximum des concepts appris est d'environ 400. Pour rappel, selon les expérimentations, nous avons entre 929 et 957 concepts au total. Le nombre de

concepts appris est donc d'environ 40% du nombre de concepts dans la base Corel 30K. Ces nombres de concepts appris durant les 4 expérimentations ne varient pas trop selon les bases d'images construites pour nos expérimentations (20%, 50%, 75%, 95% de la base Corel 30K). En effet, nous avons effectué 1000 recherches interactives pour faire l'apprentissage pour chaque expérimentation et ces 1000 recherches interactives sont également réparties dans toutes les catégories de la base Corel 30K. Dans notre expérimentation, nous distinguons deux types de concepts : concept "grand" et concept "petit". Nous définissons un concept grand comme étant un concept associé à beaucoup d'images et un concept petit comme étant associé avec peu d'images dans la base. Par exemple dans la figure 5.6, nous trouvons des concepts grands (à la gauche de la figure) comme : "sky", "closeup", "horses" et des concepts petits (à droite de la figure) comme : "glasses", "soldiers", "highway".

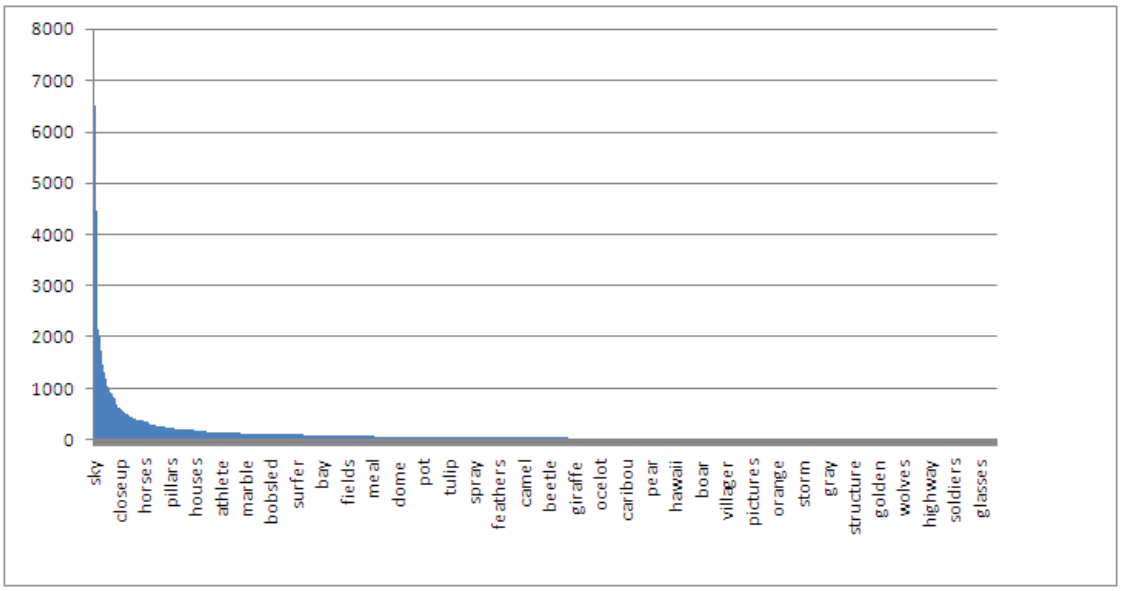

FIGURE 5.6 – La distribution des concepts dans la base Corel 30K. Certains concepts sont associés avec beaucoup d'images (concepts grands - à gauche), tandis que d'autres sont associés avec seulement quelques images (concepts petits - à droite).

Dans la figure 5.5 nous observons que la moitié des concepts sont appris dans le premier tiers du temps d'expérimentation. La plupart du temps restant est utilisé pour améliorer les représentations de ces concepts. Cette amélioration des représentations est illustrée par l'annotation d'images et la recherche d'images dans la section 5.4.3.

Les annotations sont propagées à toutes les images en même temps que les représentations (des Sacs de KVR) des concepts sont apprises. L'évolution du nombre d'annotations et les mesures de rappel/précision de l'annotation caractérisent la

performance de notre système. Nous analysons et évaluons l'évolution de la quantité d'annotations dans notre système dans la partie suivante. La qualité d'annotations est presentée dans la section 5.4.3.

5.4.2.2 Le nombre d'annotations

Nous avons déjà noté que pendant le premier tiers du temps d'expérimentation, la moitié des concepts sont appris (figure 5.5). Avec cela, nous pouvons remarquer que le nombre d'annotations propagées augmente rapidement et beaucoup dans ce premier tiers du temps. Le nombre d'annotations propagées dans le premier tiers est environ 90% du nombre total d'annotations propagées (figure 5.7). Cela s'explique par le fait qu'il existe beaucoup plus d'images qui sont annotées avec les concepts appris dans la première moitié du temps qu'avec les concepts restants (figure 5.5). Cette remarque est confirmée par le fait que les concepts avec plus d'images annotées ont une probabilité plus grande d'apparition dans les premières interactions de l'apprentissage.

Le nombre d'annotations manuelles est plus petit que le nombre d'annotations automatiques (propagées) parce que ces annotations manuelles sont transmises seulement pendant les 1000 recherches interactives d'images par les utilisateurs/experts (simulées dans notre expérimentation). Le nombre d'annotations manuelles pendant l'expérimentation 1 est plus petit parce que le nombre d'images dans la partie initiale est très petit (6000 images). Pendant les 1000 recherches interactives, les utilisateurs affectent des annotations pour seulement un petit nombre d'images. Par contre le nombre d'images dans la partie initiale dans les expérimentations 2, 3 et 4 est beaucoup plus important par rapport aux nombres de recherches interactives pour l'apprentissage. Dans ces expérimentations 2, 3 et 4, pendant 1000 recherches interactives, les utilisateurs peuvent affecter des annotations pour un nombre plus grand d'images que dans l'experimentation 1. Toutefois, ce nombre est encore petit par rapport au nombre d'images total. Notons que nous avons fait l'hypothèse que le nombre d'exemples d'images pertinentes/non pertinentes pour chaque interaction doit être très faible, 20 au maximum. Ainsi le nombre d'annotations manuelles n'est pas trop différent dans les expérimentations 2, 3 et 4.

Le nombre d'annotations propagées dépend du nombre d'images dans la base. Nous pouvons donc trouver 2 ordres réversibles dans les deux graphiques représentant les nombres d'annotations propagées de la partie initiale et la partie nouvelle (figure 5.7, deux graphiques du haut). Le nombre d'annotations propagées de la partie initiale est le plus grand dans l'expérimentation 4, et le plus petit dans l'expérimentation 1. Par contre, le nombre d'annotations propagées de la partie nouvelle est le plus grand dans l'expérimentation 1, et le plus petit dans l'expéri-

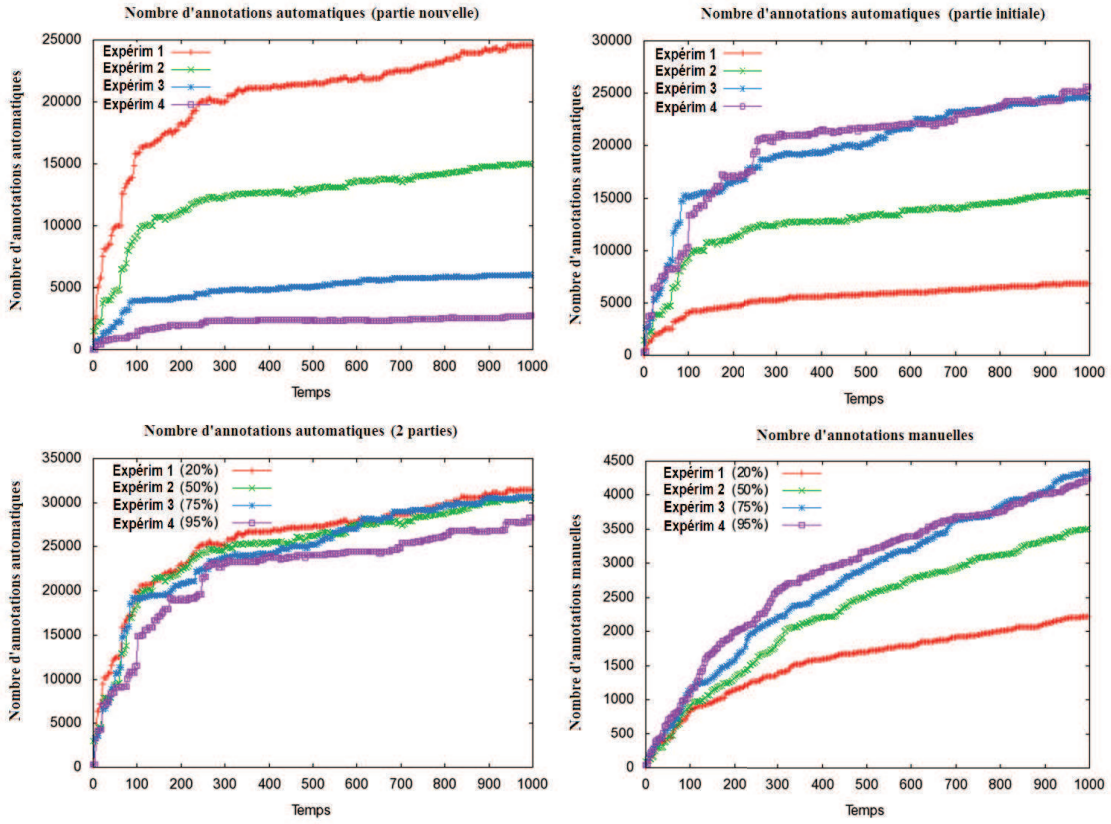

FIGURE 5.7 – Nombre d'annotations propagées dans la partie initiale (en haut à droite), dans la partie nouvelle (en haut à gauche), dans la base totale (en bas à droite) et le nombre d'annotations manuelles (en bas à gauche).

mentation 4. Toutefois la somme des nombres d'annotations propagées de 2 parties (initiale et nouvelle) est presque le même dans les 4 expérimentations.

Par exemple, dans le cas d'une utilisation très intensive du système par l'humain, impliquant ainsi de nombreuses annotations manuelles, beaucoup de connaissances avec peu d'images dynamiques arrivent dans le temps. C'est le cas de l'expérimentation 4. Dans d'autres cas, comme celui de l'expérimentation 1, il peut y avoir peu d'utilisations du système alors que de nombreuses images arrivent. Dans ce cas, nous avons très peu d'annotations manuelles. Dans ces deux cas, le nombre d'annotations propagées est le même parce que le nombre d'images total dans la base d'images est identique.

5.4.2.3 Conclusion sur la quantité de connaissances

La quantité de connaissances augmente plus rapidement dans les premiers temps, mais ce taux d'augmentation diminue au fur et à mesure qu'on utilise le

système. La quantité de connaissances dépend principalement du nombre d'interactions ou en d'autres termes de la fréquence d'utilisation du système. Plus le système est utilisé, plus des connaissances sont apprises. La distribution des concepts sur la base d'apprentissage influence aussi la quantité de connaissances. Si des images dans la base ont un large éventail de concepts, la quantité de connaissances est plus importante que dans le cas d'un petit éventail de concepts.

5.4.3 La qualité de connaissances d'annotations

La propagation d'annotations en utilisant les représentations (Sacs de KVR) des concepts est évaluée par la précision d'annotation et le rappel d'annotation. L'évaluation est basée sur les 4 expérimentations dont nous avons parlé précédemment. La propagation des annotations est effectuée sur les deux parties, initiale et nouvelle, dans chaque expérimentation. L'annotation des images dans la partie nouvelle illustre l'efficacité de l'utilisation des Sacs de KVR pour les nouvelles images arrivées dans le système. La propagation d'annotations d'images sur la partie initiale et sur la partie nouvelle permet d'observer l'influence du nombre d'images initiales dans le système, et l'influence des conditions d'utilisation du système sur la qualité des connaissances en fonction du temps.

5.4.3.1 La partie initiale

	partie initiale	partie nouvelle
Expérimentation 1 (20% - 80%)	6K	25K
Expérimentation 2 (50% - 50%)	15K	16K
Expérimentation 3 (75% - 25%)	23K	8K
Expérimentation 4 (95% - 05%)	29.5K	1.5K

TABLE 5.1 – Nombre d'images dans les bases d'images pour les 4 expérimentations.

La partie initiale contient 6000 (20%), 15000 (50%), 23000 (75%), 29500 (95%) images respectivement pour les expérimentations 1, 2, 3 et 4 (tableau 5.1). Les images de cette base sont reparties dans toutes les catégories de la base Corel 30K. Les 4 conditions d'utilisation différentes du système selon les 4 expérimentations mettent en évidence de différentes performances de la propagation d'annotations en fonction du temps (voir figure 5.8).

La figure 5.8 illustre l'évolution de la précision de la propagation d'annotations dans notre système dans les 4 expérimentations. Dans l'expérimentation 1 (20%-80%) la précision d'annotation est très bonne par rapport aux expérimentations 2, 3 et 4. La raison est que le taux du nombre d'images pertinentes sur le nombre total

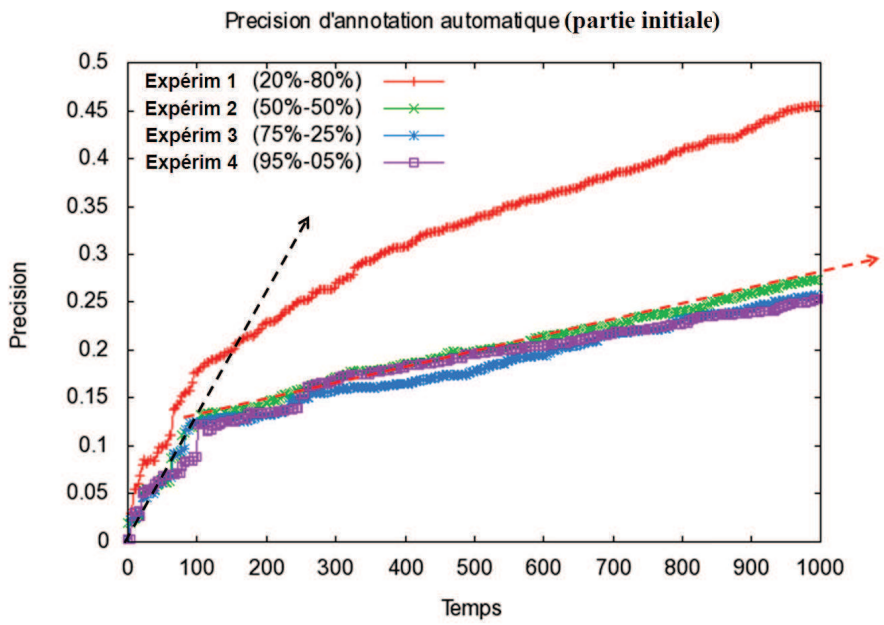

FIGURE 5.8 – L'évolution de la précision pour la propagation d'annotation sur la partie initiale. La précision d'annotation augmente rapidement dans un premier temps (flèche noire) où les Sacs de KVR des concepts grands sont appris (la connaissance augmente rapidement). Dans un second temps (flèche rouge), les Sacs de KVR des concepts grands sont améliorés (la connaissance est améliorée) et les autres concepts petits sont appris (la connaissance augmente moins rapidement que dans le premier temps). Alors la précision d'annotation augmente plus lentement.

d'images dans l'expérimentation 1 est grand. En revanche, ce taux est trop petit dans les autres expérimentations 2, 3 et 4. Nous pouvons imaginer simplement qu'il est facile de chercher une personne dans un groupe de 10 personnes mais que ceci est très difficile dans un groupe de 1000 personnes. Alors, c'est plus facile dans l'expérimentation 1 pour trouver des images pertinentes pour un concept que dans les expérimentations 2, 3 et 4. La propagation d'annotations dans l'expérimentation 3 est meilleure que dans l'expérimentation 2, mais moins bonne que dans l'expérimentation 4. Cependant la différence est petite car le nombre d'images pertinentes pour un concept (dans environ 400 concepts appris) est très petit par rapport au nombre d'images dans la base d'images.

La précision d'annotations augmente rapidement dans un premier temps (la flèche noire dans la figure 5.8) pendant lequel les Sacs de KVR des concepts grands sont appris. Ceci est confirmé dans la figure du nombre d'annotations propagées pour la partie initiale (figure 5.7). Dans ce même premier temps, le nombre d'annotations atteint 80% du nombre total d'annotations. Dans le temps suivant (la flèche en rouge dans la figure 5.8), les Sacs de KVR des concepts grands sont améliorés et les autres concepts plus petits sont appris. Alors la précision d'annotation aug-

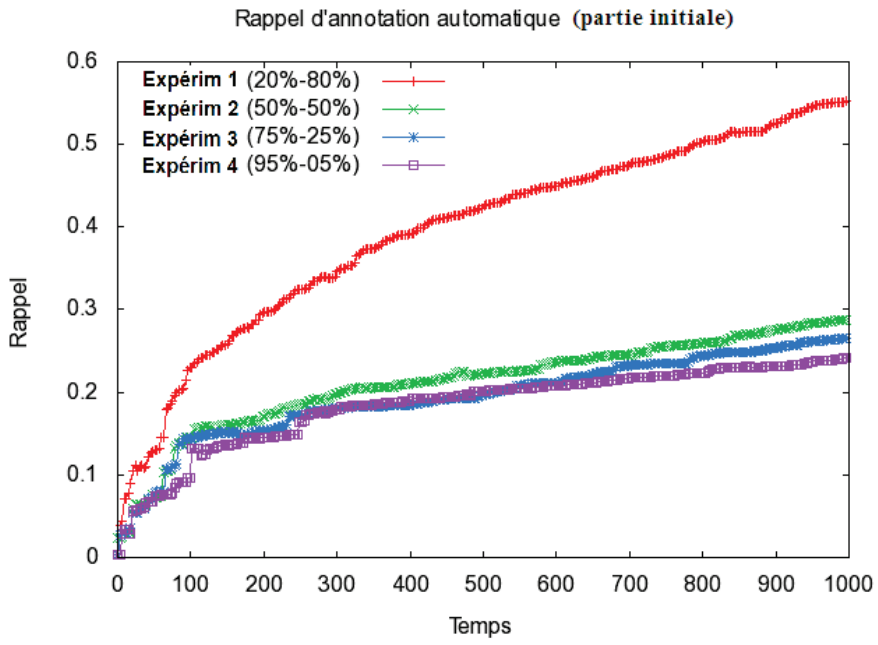

FIGURE 5.9 – L'évolution du rappel de la propagation d'annotations sur la partie initiale.

mente plus lentement. Nous remarquons qu'il est difficile de faire l'annotation des concepts petits parce que le nombre d'images associées est très petit par rapport aux nombre d'images total dans la base. En résumé, la précision d'annotation de la partie initiale s'améliore dans le temps et au fur et à mesure, et ce dans toutes les expérimentations. Cela signifie que les Sacs de KVR deviennent meilleurs au fur et à mesure de l'utilisation du système.

La figure 5.9 illustre le rappel d'annotations de la partie initiale dans les 4 expérimentations. Ces résultats confirment encore que l'annotation de la partie initiale, ou en d'autres termes les connaissances du système, devient meilleure dans le temps au fur et à mesure de l'utilisation du système. Si on utilise le système sur une petite base d'images, la qualité de connaissances est plus fiable que si on utilise le système sur une grand base d'images.

5.4.3.2 La partie nouvelle

La partie nouvelle contient respectivement 24000 (80%), 15000 (50%), 7500(25%), 1500(5%) images pour les expérimentations 1, 2, 3 et 4 (voir tableau 5.1). Les images dans la base sont reparties dans toutes les catégories de la base Corel 30K. Comme dans le cas de la partie initiale, le temps et le nombre d'images de la partie nouvelle influencent la performance de la propagation d'annotations.

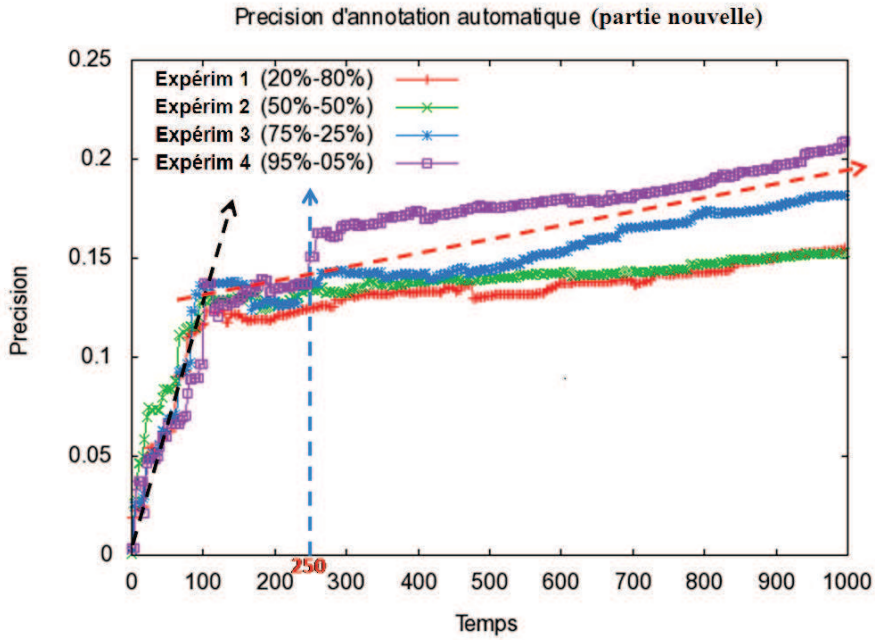

FIGURE 5.10 – L'évolution de la précision de la propagation d'annotations sur la partie nouvelle.

Les figures 5.10 et 5.11 illustrent la performance de la propagation d'annotations dans la partie nouvelle pour les 4 expérimentations. La propagation d'annotations dans la partie nouvelle est plus faible que celle dans la partie initiale parce qu'un certain pourcentage d'annotations manuelles de la partie initiale est utilisé pour propager les annotations. Ce nombre est respectivement de 0.3, 0.2, 0.16, 0.15 dans les expérimentations 1, 2, 3 et 4 (ce sont les ratios entre le nombre d'annotations manuelles et le nombre d'annotations dans la vérité terrain de la partie initiale, voir la figure 5.7).

Comme pour la partie initiale, la propagation d'annotations dans la partie nouvelle augmente rapidement pendant un premier temps où les concepts grands sont appris. Un exemple est illustré au temps t=250 (figures 5.10 et 5.11) où les mesures de précision et de rappel dans l'expérimentation 4 augmentent tout d'un coup. Nous observons dans la figure du nombre d'annotations propagées (figure 5.7) qu'un grand nombre d'annotations est propagé. Cela signifie qu'au temps t=250, la représentation Sacs de KVR d'un concept grand est apprise. Les annotations de ce concept (concept « building » avec 2373 images associées) sont bien propagées aux images de la partie nouvelle.

Dans ce cas, nous pouvons remarquer que dans le cas où on utilise beaucoup le système par rapport au nombre d'images arrivées (c'est le cas de l'expérimentation 4), les annotations propagées sont plus fiables que dans le cas où on utilise peu le système par rapport au nombre d'images arrivées, ce qui est le cas de l'expérimen-

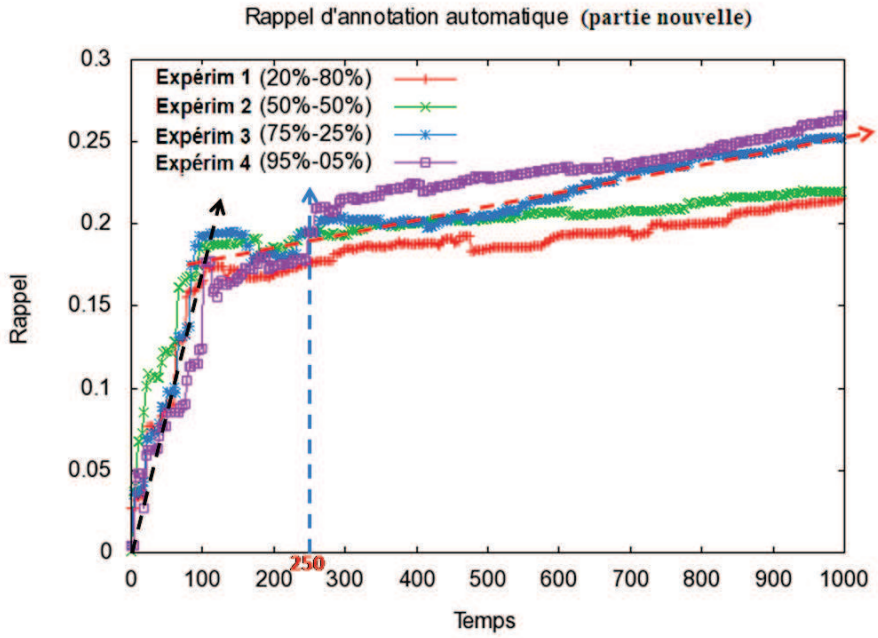

FIGURE 5.11 – Evolution du rappel de la propagation d'annotations sur la partie nouvelle.

tation 1.

5.4.3.3 La propagation d'annotations dans la base Corel 30K

Dans cette partie, nous évaluons la propagation d'annotations dans les 4 expérimentations sur une même base d'images. La base Corel 30K au complet (soit la fusion de la partie initiale et la partie nouvelle pour les 4 expérimentations) est utilisée pour propager des annotations. Dans les expérimentations 1, 2, 3 et 4, les Sacs de KVR des concepts sont appris en simulant les connaissances des utilisateurs par les annotations des bases d'images différentes (20%, 50%, 75%, 95% de la base Corel30K). Cela signifie que les Sacs de KVR des concepts sont appris en se basant sur les différentes sources de connaissances (en termes de quantité de connaissances). Les influences dans les 4 expérimentations sont évaluées par la propagation d'annotations sur la même base d'images Corel 30K.

Les figures 5.12 et 5.13 illustrent la performance de la propagation d'annotations sur la base Corel 30K pour les 4 expérimentations. Dans le premier temps (jusqu'au temps $t = 280$) l'expérimentation 2 donne la meilleure performance, les autres expérimentations fournissant presque les mêmes résultats. La raison est probablement que la connaissance utilisée est encore petite par rapport à la quantité totale de connaissances. Dans le temps restant, la quantité de connaissances commence à influencer la performance de la propagation. Elle est meilleure dans les

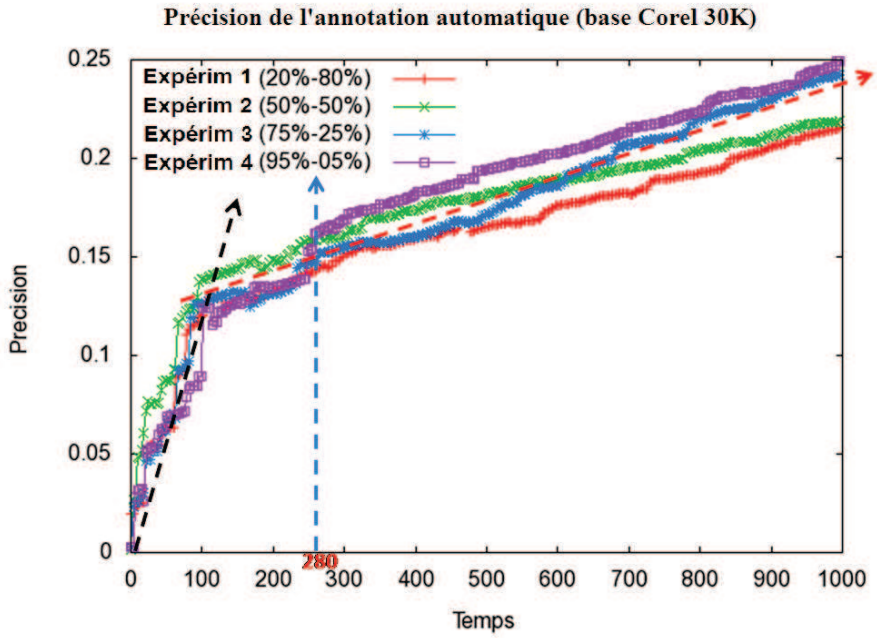

FIGURE 5.12 – L'évolution de la précision de la propagation d'annotations (annotation automatique) sur la base Corel 30K. Dans un premier temps (jusqu'au temps $t = 280$) l'expérimentation 2 donne la meilleure performance, les autres expérimentations ayant presque la même performance. Dans un second temps, la quantité de connaissances commence à influencer la performance de la propagation.

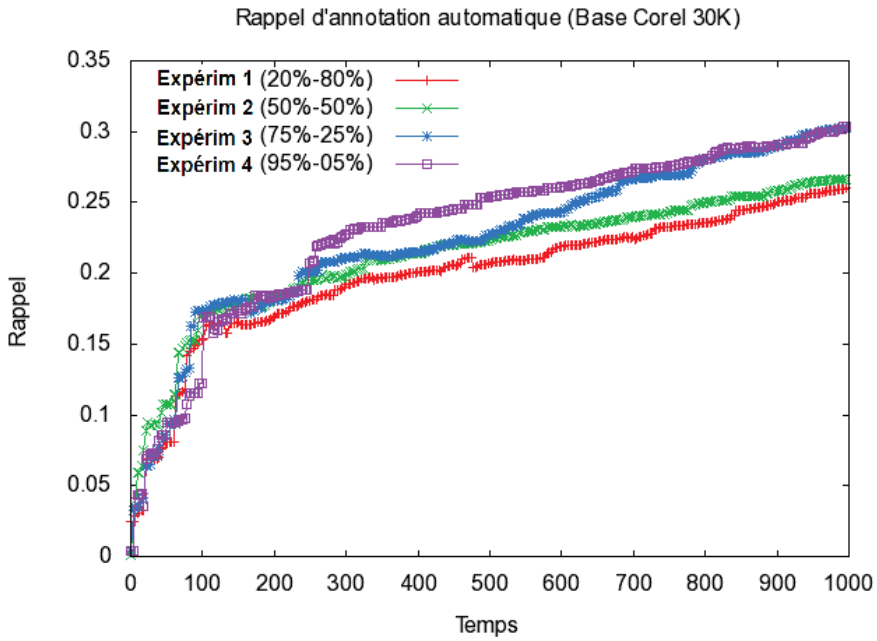

FIGURE 5.13 – Evolution du rappel de la propagation d'annotations sur la base Corel 30K.

expérimentations 1 et 2 que dans les expérimentations 3 et 4. Cependant, nous pouvons remarquer qu'il n'y a pas trop de différences du fait que tous les apprentissages dans les 4 expérimentations sont effectués en simulant un même nombre de 1000 recherches interactives des utilisateurs/experts. En fait, l'influence de la quantité de connaissances sur la propagation peut être confirmée tout au cours du temps. En avançant dans le temps, plus de connaissances sont utilisées pour l'apprentissage, et donc plus la propagation d'annotations est performante.

Nous pouvons observer que différentes conditions d'utilisation du système donnent différents résultats. Dans un cas d'une utilisation intensive du système, c'est-à-dire avec beaucoup d'interactions, le résultat est meilleur. Ceci est confirmé par l'amélioration de la performance de la propagation d'annotations ou de la recherche d'images au cours du temps, ou en d'autres termes, au fur et à mesure du nombre d'interactions. Dans le cas d'utilisation du système sur beaucoup d'images (l'expérimentation 4), le résultat est meilleur que sur peu d'images (l'expérimentation 1). L'expérimentation 4 donne les meilleurs résultats, et cela est certainement dû au fait qu'on se place dans de telles conditions d'utilisation qu'on utilise intensivement le système sur de nombreuses d'images. Mais on voit que dans d'autres conditions d'utilisation, le système réussit quand même bien, et la performance s'améliore avec le temps.

5.4.4 La qualité de connaissances de niveau "image"

La recherche d'images par des Sacs de KVR est utilisée pour évaluer des connaissances de niveau "image". Nous observons si les représentations Sacs de KVR des concepts sont bien apprises en analysant la précision de la recherche d'images.

La figure 5.14 illustre la précision moyenne de la recherche d'images par les Sacs de KVR des concepts. La précision est la moyenne des précisions moyennes de la recherche d'images de tous les concepts.

La précision de l'expérimentation 4 est la meilleure tandis que celle de l'expérimentation 1 est la pire. Plus le nombre d'images dans la base est élevé, plus la précision de la recherche d'images est mauvaise. En général, la précision de recherche augmente dans le temps ou en d'autres termes, les Sacs de KVR s'améliorent dans le temps. Cependant, nous pouvons remarquer qu'il existe des moments où la précision diminue un peu. Cela signifie qu'un Sac de KVR de concept ne s'améliore pas toujours pendant l'apprentissage. Ce problème vient probablement du regroupement d'images pour la construction des KVR. Notre approche de l'apprentissage considère que le regroupement est bon du fait que le nombre d'exemples est faible. Cependant, le regroupement peut être mauvais dans les cas où les exemples sont trop diversifiés.

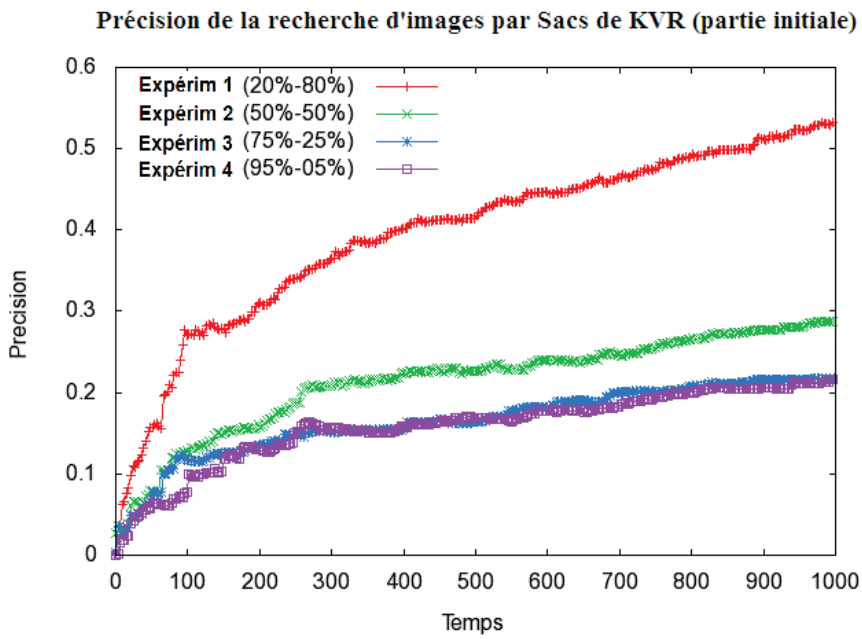

FIGURE 5.14 – Evolution de la précision de la recherche d'images par des Sacs de KVR sur la partie initiale. En général, la précision de recherche est meilleure dans le temps ou en d'autres termes, les Sacs de KVR s'améliorent dans le temps.

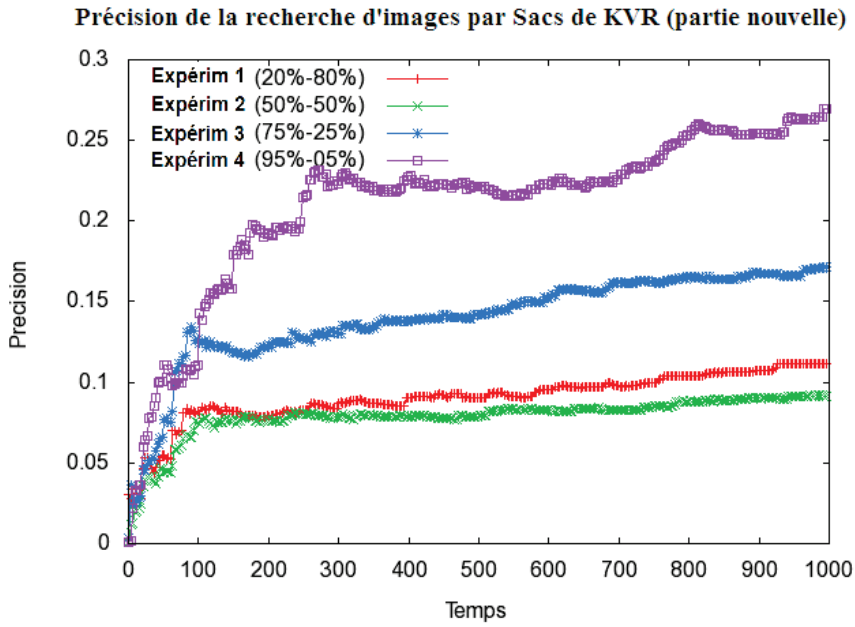

FIGURE 5.15 – Evolution de la précision de la recherche d'images par des Sacs de KVR sur la partie nouvelle.

La figure 5.15 illustre la précision de la recherche par les Sacs de KVR sur la partie nouvelle. Bien que la précision soit meilleure en fonction du temps, elle reste plus faible dans le cas d'une grande base d'images comme pour toutes les évalua-

tions précédentes ci-dessus. Dans les cas où on utilise peu le système par rapport au nombre d'images qui arrivent, on se trouve dans le cas de l'expérimentation 1. Le système s'en sort quand même assez bien dans ces circonstances. Dans l'expérimentation 4, les connaissances évoluent plus vite que le nombre d'images arrivant et on voit que dans ce cas-là, le système fonctionne très bien.

5.4.5 Evaluation sur le retour de pertinence utilisé dans l'apprentissage de connaissances

Nous avons présenté l'apprentissage de connaissances dans le chapitre 4. Dans cet apprentissage, nous utilisons la technique du retour de pertinence que nous avons proposée dans le chapitre 3. Notre retour de pertinence est proposé en se basant sur la combinaison de la technique de mouvement de requête et celle d'extension de requête, aves deux variantes : CR (*Clustering-repeat*) et CNR (*Clustering-non-repeat*). L'évaluation dans le chapitre 3 a montré que la méthode CNR est légèrement meilleure que la méthode CR pour la recherche interactive. Dans cette partie, nous comparons ces deux méthodes pour l'apprentissage de connaissances (Sacs de KVR). La précision d'annotation d'images et la précision de recherche d'images sont utilisées encore ici pour évaluer l'évolution des connaissances dans les deux cas des méthodes CR et CNR. Nous présentons ici seulement le résultat pour l'expérimentation 4 (division 95%-5% des bases d'images). En effet, les autres expérimentations produisent les mêmes résultats.

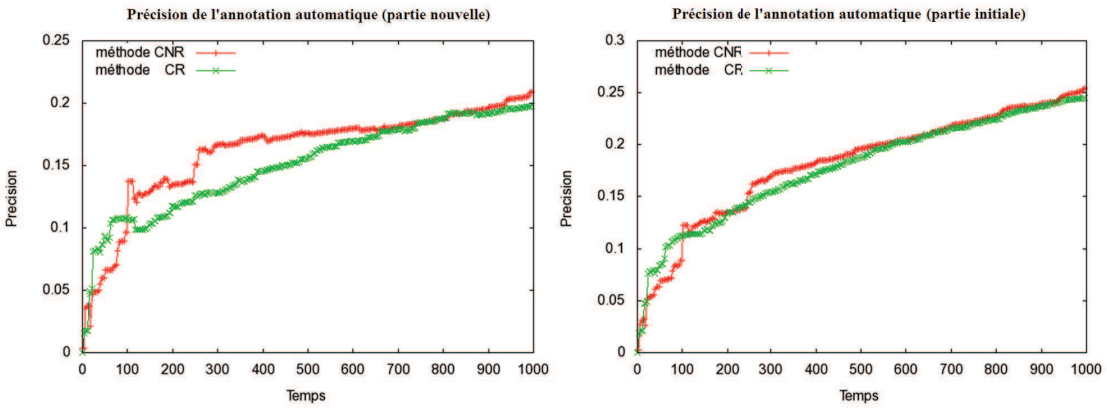

FIGURE 5.16 – Evolution de la précision de l'annotation d'images pour les méthodes CR et CNR.

La figure 5.16 illustre la précision d'annotation et la figure 5.17 illustre la précision de la recherche d'images. Dans la méthode CR, des KVR sont construits en déplaçant des points (KVR) dans l'espace de caractéristiques vers les points

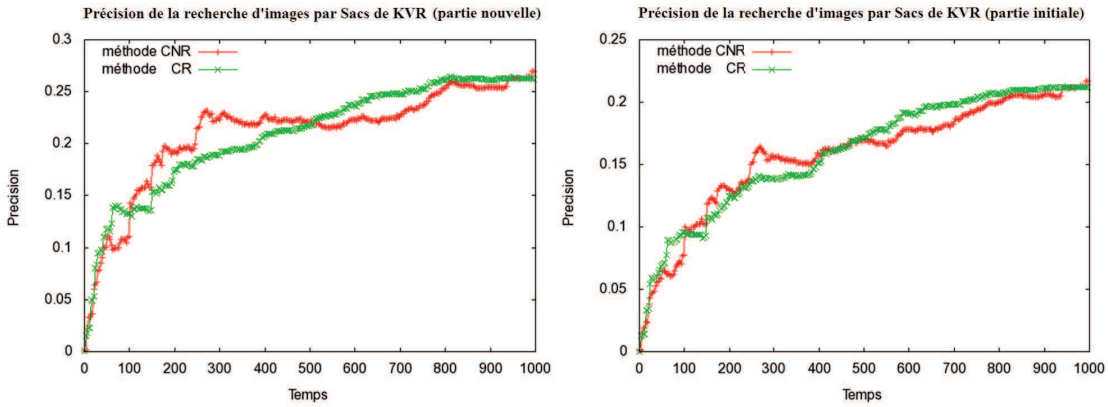

FIGURE 5.17 – Evolution de la précision de la recherche d'images pour les méthodes CR et CNR.

(KVR) idéaux. Par contre, dans la méthode CNR, des KVR sont construits en éliminant des KVR non pertinents et en ajoutant des KVR pertinents. Dans les 2 cas, la méthode CR et la méthode CNR fournissent presque les mêmes résultats, avec toutefois un léger avantage pour la méthode CNR, notamment dans le cas de l'annotation d'images. Ce résultat est confirmé par le fait que l'apprentissage de connaissances est influencé par la recherche interactive (retour de pertinence) pour laquelle nous avons montré dans le chapitre 3 que la meilleure méthode est aussi la méthode CNR.

5.4.6 Evaluation de la méthode de regroupement

Notre modèle Sacs de KVR est basé principalement sur le regroupement (clustering) des exemples d'images. Nous avons présenté notre sélection sur la méthode du regroupement dans le chapitre 4 : la méthode des k-moyennes adaptif et la méthode d'agglomération compétitive. Dans cette partie, ces deux méthodes sont comparées en fonction de la performance de connaissances apprises en analysant la précision de l'annotation et de la recherche d'images. Nous présentons ici seulement le résultat pour l'expérimentation 4 (division 95%-5% des bases d'images). Les autres expérimentations ont les mêmes résultats.

La figure 5.18 illustre la précision de l'annotation automatique et la figure 5.19 la précision de la recherche d'images. Dans un premier temps, le regroupement en utilisant l'agglomération compétitive est meilleur que le regroupement par les k-moyennes, même s'il faut noter que la performance est presque la même pour le temps restant. Nous pouvons constater que le choix de la méthode de regroupement n'influence pas beaucoup le résultat, parce que le nombre total d'échantillons (exemples pertinents / non pertinents) est très faible. A noter que dans notre

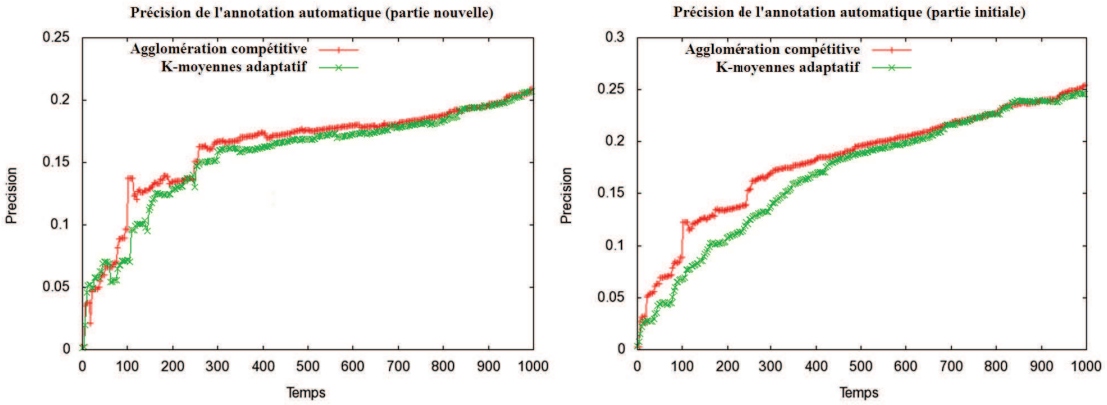

FIGURE 5.18 – L'évolution de la précision de l'annotation d'images selon la methode de regroupement : K-moyennes adaptif et Agglomération compétitive.

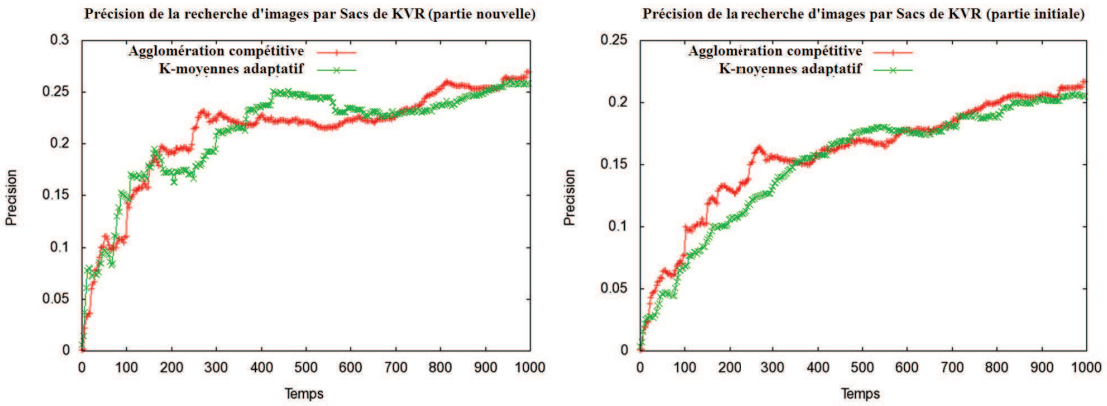

FIGURE 5.19 – L'évolution de la précision de la recherche d'images selon la methode de regroupement : K-moyennes adaptif et Agglomération compétitive.

système, l'utilisateur ne marque que quelques exemples (20 maximum) comme pertinents ou non pertinents pendant l'interaction.

5.4.7 Utilisation de Sacs de KVR : Evaluation du caractère "statique"

Dans cette section, nous évaluons l'utilisation de connaissances. Parmi les connaissances de notre système, les représentations visuelles de concepts ou en d'autres termes, les Sacs de KVR de concepts, sont les plus importantes. Avec ces connaissances de niveau "image", nous pouvons faire de l'annotation d'images ou de la recherche d'images par le contenu en utilisant des requêtes textuelles.

Pour comparer notre système avec les travaux existants nous proposons d'abord d'évaluer deux tâches principales du domaine CBIR : l'annotation d'images d'une part et la recherche d'images d'autre part. Nous proposons ensuite une comparaison globale des avantages et des inconvénients de notre système par rapport aux autres.

5.4.7.1 Annotation d'images

Notre système possède d'excellentes qualités pour la fonction de l'annotation d'images, pour les raisons suivantes.

1. Tout d'abord, il est capable d'annoter dynamiquement une base d'images dépourvue d'annotations au départ de la vie du système, c'est-à-dire qu'il ne demande pas une base d'apprentissage a priori.
2. De plus, il est capable de faire de l'annotation semi-automatique (incrémentale) d'images via l'interaction avec les experts/utilisateurs.
3. Enfin, il est capable de faire une propagation d'annotations automatique pour des nouvelles images arrivées dans la base et cela en temps réel.

Dans ce chapitre nous avons évalué l'évolution de notre système en fonction du temps, ce qui est difficile à comparer avec d'autres travaux. Dans cette partie, nous proposons de comparer notre résultat de propagation d'annotations avec d'autres méthodes d'annotation automatique. La propagation d'annotations est effectuée en se basant sur les représentations Sacs de KVR de concepts. Bien que les Sacs de KVR de concepts soient appris via des interactions avec des experts/utilisateurs, nous considérons nécessaire de comparer notre propagation d'annotations avec d'autres méthodes d'annotation automatique, du fait que les interactions dans notre expérimentation sont simulées automatiquement. Si notre méthode est acceptable par rapport aux méthodes automatiques, nous pouvons confirmer l'efficacité de notre système.

Dans cette section, nous comparons le résultat de notre méthode d'annotation avec d'autres méthodes de la littérature. Pour comparer avec d'autres travaux procédant également à de l'annotation d'images, la base de données Corel 5K [Carneiro 2007], [Lavrenko 2004], [Metzler 2004], [Yavlinsky 2005], [Feng 2004] est utilisée. Corel 5k contient 4500 images d'apprentissage et 500 images de test avec 374 mots textuels. Sur les 374 mots-clés, seulement 260 mots sont présents dans les deux ensembles d'apprentissage et de test.

La base d'apprentissage est utilisée pour faire l'apprentissage incrémental des représentations Sacs de KVR de concepts avec un nombre d'interrogations défini (1000). Comme la simulation des interactions pour l'apprentissage est utilisée, nous considérons que la propagation est une annotation automatique pour comparer

avec les autres méthodes d'annotation automatique. Ensuite, les Sacs de KVR sont utilisés pour procéder à une propagation d'annotations sur la base de test.

Nous comparons notre résultat avec les travaux de [Carneiro 2007], [Lavrenko 2004], [Metzler 2004], [Yavlinsky 2005], [Feng 2004], [Makadia 2010] qui ont également proposé des techniques d'annotation d'images avec la base Corel 5K. Ces travaux sont également discutés dans [Makadia 2010].

Méthode	**Précision**	**Rappel**
CRM ([Lavrenko 2004])	0.16	0.19
InfNet ([Metzler 2004])	0.17	0.24
NPDE ([Yavlinsky 2005])	0.18	0.21
MBRM ([Feng 2004])	0.24	0.25
SML ([Carneiro 2007])	0.23	0.29
JEC ([Makadia 2010])	0.27	0.32
Sacs de KVR (notre approche)	0.23	0.28

TABLE 5.2 – Tableau comparatif des méthodes d'annotation d'images.

Le tableau 5.1 montre la précision et le rappel des méthodes d'annotation d'images. Comme on peut le constater sur ce tableau, notre méthode donne de bonnes performances. Seule la méthode de [Makadia 2010] est meilleure que notre méthode en termes de précision et rappel. Cependant, notons que ce résultat est évalué d'une manière automatique grâce à notre protocole de simulations d'interactions. Nos résultats sont probablement ici en-dessous de la réalité, du fait de l'usage de notre simulateur, dont la performance est, selon nous, bien moins élevée que celle d'un véritable opérateur humain. De plus, au-delà de la seule performance liée à l'annotation, l'avantage majeur de notre approche est la possibilité d'améliorer cette annotation grâce aux interactions des utilisateurs, double qualité qu'aucun système de la littérature n'offre actuellement, du moins à notre connaissance.

Comme évoqué précédemment, un autre avantage de notre méthode est sa capacité d'annoter les nouvelles images en temps réel. Nous avons testé sur un ordinateur PC normal Core 2 Duo CPU, 2 GHz, 2Go de RAM sous Linux Ubuntu 10.04. Le temps moyen pour annoter une image à partir des Sacs de KVR de 374 concepts est d'environ 1 seconde.

5.4.7.2 La recherche d'images

Dans le domaine CBIR, la spécification d'une requête par exemple d'images est toujours un problème difficile. La première raison de cela est que l'image exemple peut être indisponible. La deuxième raison est le fossé sémantique en vertu duquel il est difficile de décrire l'intérêt précis de l'utilisateur par des images. Il est plus

facile de spécifier la requête par le texte. Cependant, la recherche d'images par le texte a un problème concernant la disponibilité des annotations d'images. Au-delà de la capacité liée à l'annotation d'images sans connaissances a priori que nous avons présentée, notre système est capable de rechercher des images par le contenu en utilisant des requêtes textuelles (figure 5.20). La méthode est simplement basée sur le modèle Sacs de KVR. Les mots-clés de la requête textuelle sont d'abord transformés en Sacs de KVR. Ensuite la recherche d'images par le contenu est effectuée en utilisant ces Sacs de KVR qui sont des caractéristiques visuelles.

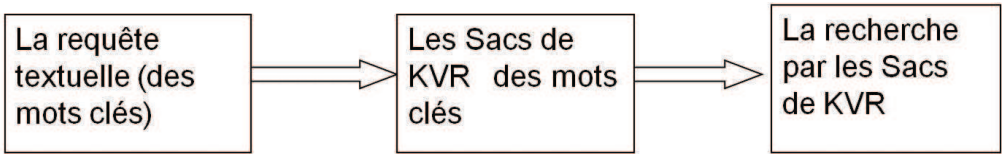

FIGURE 5.20 – La recherche d'images par le contenu dans notre système.

Dans cette partie, nous comparons notre méthode de recherche par le contenu avec les approches classiques (celles qui utilisent l'image exemple comme requête). La recherche d'images par l'exemple est effectuée avec la base Corel5K qui contient 50 classes d'images. Des Sacs de KVR sont appris sur la base d'apprentissage de la base de test de la base Corel 5K qui contient 50 classes de 10 images. La recherche par requête textuelle est effectuée aussi sur cette base de test. La vérité terrain pour la recherche textuelle correspond à l'annotation manuelle de la base Corel 5K. Les images qui partagent un même concept sont considérées pertinentes pour la requête par ce concept. La recherche par requête textuelle a comme avantages la facilité de spécification de la requête par rapport à la recherche par l'exemple. Nous regardons encore si la performance de la recherche par requête textuelle est bonne par rapport à la recherche par l'exemple. La courbe précision/rappel est utilisée pour comparer ces deux méthodes de recherche d'images.

La figure 5.21 montre la performance de la recherche d'images par l'exemple (en noir), de la recherche d'images par requête textuelle pour tous les concepts (en rouge) et pour les concepts avec plus de 10 images associées dans la base de test (en bleue).

Nous n'évaluons pas la performance de la recherche par le contenu mais nous allons comparer avec la performance de la recherche d'images par la requête textuelle en utilisant des représentations Sacs de KVR de concepts. Dans tous les 2 cas, nous utilisons le modèle Sacs de Mots [Sivic 2008] pour représenter des caractéristiques visuelles. Nous pouvons remarquer que la performance est meilleure après élimination des concepts associées avec moins de 10 images. Ceci est normal parce que ces concepts "petits" donnent les plus mauvais résultats parmi tous les concepts à cause de la malédiction de la dimensionnalité.

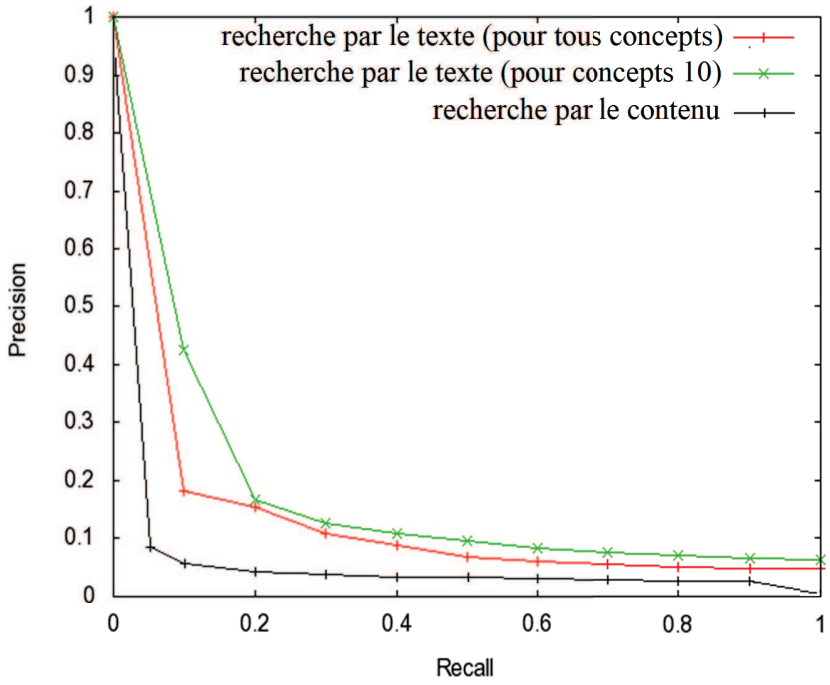

FIGURE 5.21 – Précision/rappel de la recherche d'images par l'exemple (en noir), de la recherche d'images par requête textuelle pour tous les concepts (en rouge) et pour les concepts avec plus de 10 images associées dans la base de test (en bleue).

La figure 5.21 montre la performance de la recherche d'images par requête textuelle comparé avec la recherche d'images par l'exemple. Même si le nombre d'images pertinentes pour chaque concept est encore faible par rapport au volume de la base d'images comme dans le cas de la recherche par l'exemple, la performance de la recherche par la requête textuelle est meilleure. Ceci s'explique par le fait que le modèle Sacs de KVR représente mieux les concepts que la représentation des images pour les classes dans la base d'images.

Des exemples de résultats de la recherche d'images par le contenu et par des requêtes textuelles sont présentées dans les figures 5.22, 5.23 et 5.24, 5.25. La spécification de la requête est plus facile avec des mots textuels. Comme nous l'avons présenté à la fin de la partie 5.2, il est difficile de représenter des concepts avec le *visualness* moins élevé [Yanai 2005], comme par exemple ici avec le concept "street", donc le résultat n'est pas bon. Par contre le résultat est bon dans les cas de concepts d'un grand *visualness*, comme par exemple ici avec le concept "bear".

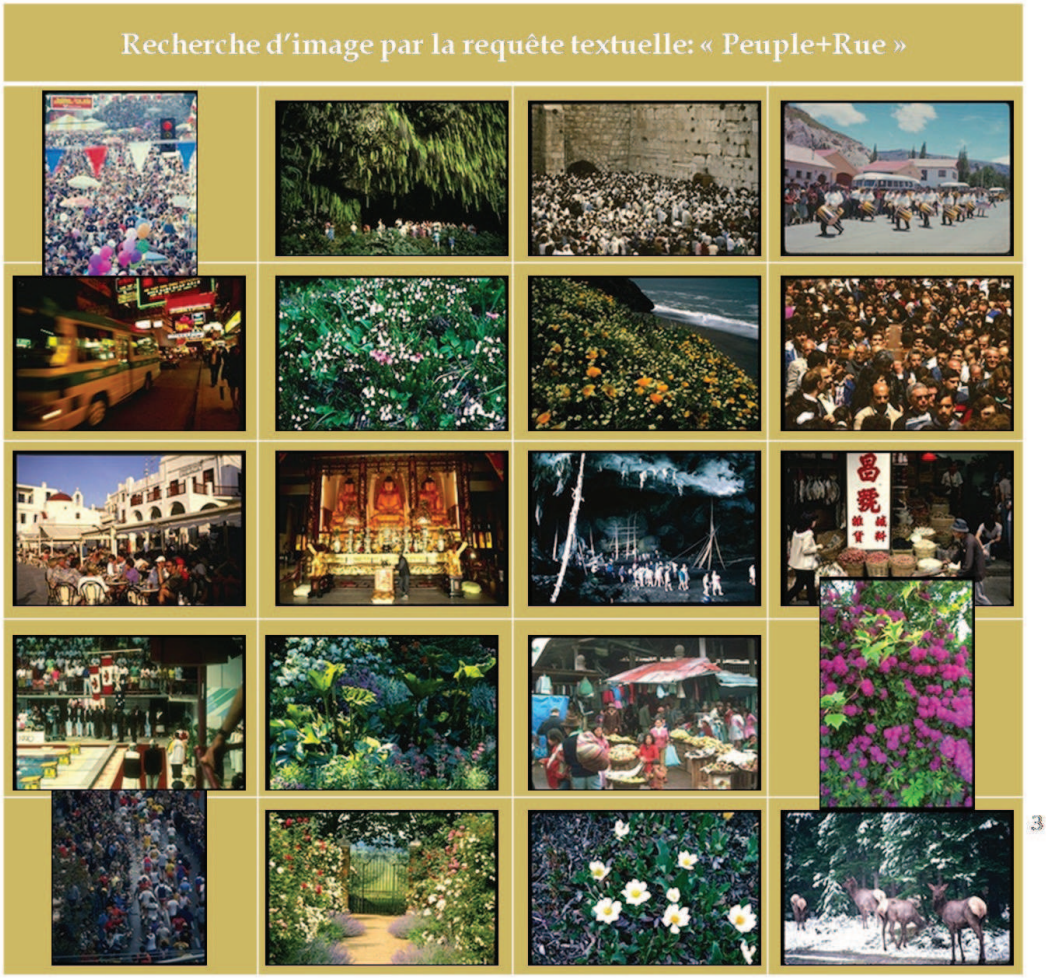

FIGURE 5.22 – Résultat de la recherche d'images par requête textuelle : "people+street".

5.5 Conclusion

Dans ce chapitre nous avons présenté l'évaluation de notre système concernant l'évolution de connaissances, l'utilisation des Sacs de KVR de concepts pour la recherche d'images et l'annotation d'images. Les connaissances du système viennent d'utilisateurs pendant la recherche interactive d'images. L'expérimentation montre que notre apprentissage par renforcement permet d'améliorer incrémentalement les connaissances dans le temps (de l'utilisation du système). De plus, la performance de la recherche d'images par des connaissances au niveau "image" (les Sacs de KVR de concepts) est très prometteuse et la performance de l'annotation dynamique est très bonne en comparaison avec les travaux existants.

Dans notre expérimentation, nous avons séparé la base d'images en deux, cette séparation servant à l'apprentissage initial seulement. Une fois que des Sacs de KVR sont appris, nous pouvons faire des recherches complètes sur toute la base d'images,

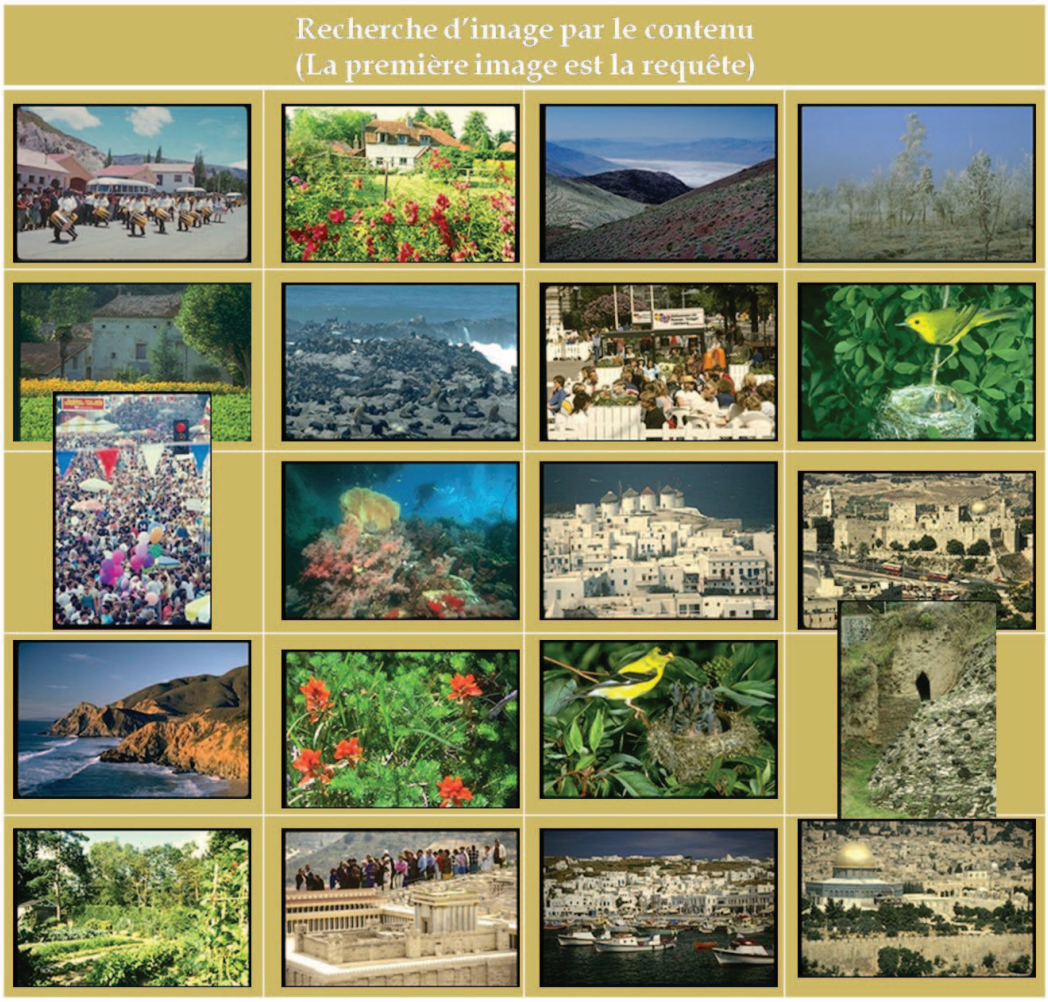

FIGURE 5.23 – Résultat de la recherche d'images par le contenu (pour trouver des gens dans la rue).

sans faire attention si les images sont annotées ou pas annotées. Nous pouvons donc remarquer que notre modèle donne au système la capacité de rechercher des images par le contenu avec une seule requête textuelle sur une base d'images sans annotation et une meilleure performance que la recherche d'images par l'exemple de la littérature. Cela représente un potentiel particulièrement intéressant pour les systèmes CBIR, parce que les utilisateurs pourraient spécialiser la requête plus facilement par le texte que par l'exemple, et recevoir de meilleurs résultats en terme de précision.

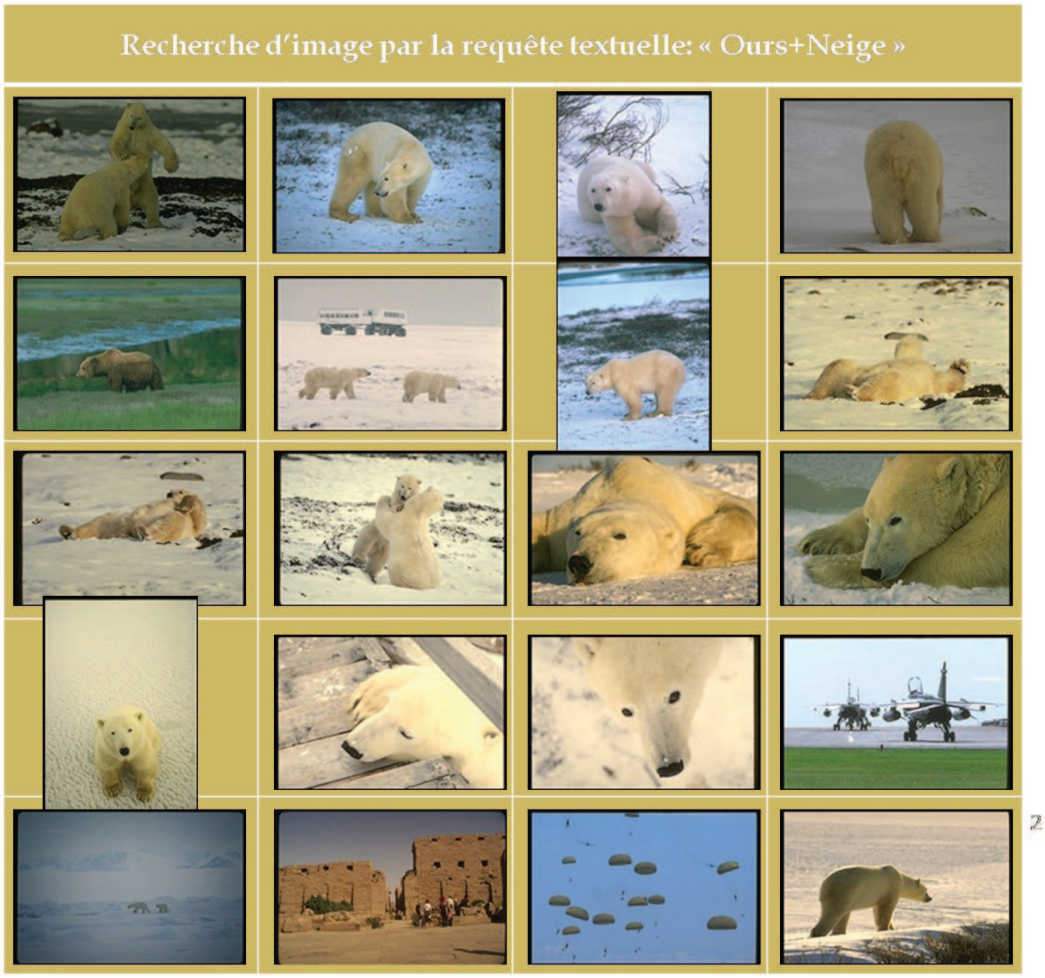

FIGURE 5.24 – Résultat de la recherche d'images par requête textuelle : "bear+snow".

FIGURE 5.25 – Résultat de la recherche d'images par le contenu (pour trouver des ours sur la neige).

Chapitre 6

Conclusion et Perspectives

6.1 Conclusion

La recherche d'images connaît depuis plusieurs années de nombreuses évolutions. Initialement basée sur des critères mono-modaux, celle-ci tente désormais d'intégrer de nombreux aspects couplant à la fois des informations contextuelles, sémantiques, visuelles et même parfois géographiques (avec l'avènement généralisé des GPS notamment). D'autre part, au-delà de ces problématiques contextuelles, notons également les récentes évolutions liées aux connaissances environnant un système de recherche d'images, dont les exigences actuelles portent beaucoup sur leur dynamicité. Ainsi, la majeure partie des applications exigent l'implémentation de systèmes interactifs permettant une évolution incrémentale du capital de connaissances, parfois partant de zéro. Dans cette thèse nous avons abordé ces composants importants de la recherche d'images, en intégrant : la recherche multimodale, la recherche interactive de type retour de pertinence et l'apprentissage de connaissances.

6.1.1 Résumé des contributions

Nous identifions quatre contributions principales dans cette thèse. La première contribution est un système de recherche multimodale d'images qui intègre différentes sources de données, comme le contenu de l'image et le texte. Ce système permet l'interrogation par l'image, l'interrogation par mot-clé ou encore l'utilisation de requêtes hybrides.

La seconde contribution concerne la recherche interactive d'images en appui sur une nouvelle technique de retour de pertinence combinant un mouvement de

requête et une extension de requête. Grâce à cette combinaison de mouvement de requête d'une part, et de l'extension de requête d'autre part, techniques qui sont généralement utilisées isolément, notre système dispose aujourd'hui de facultés de recherche d'images particulièrement rapides et efficaces, comparé aux techniques traditionnelles. Notons également que cette approche combinée est parfaitement combinable avec une recherche mixte texte/images, ce qui représente un atout supplémentaire de notre système.

La troisième contribution est un modèle basé sur des représentations visuelles de mots-clés (KVR : *Keyword Visual Representation*) pour créer des liens entre le texte et le contenu visuel, en s'appuyant sur l'interaction à long terme. Grâce à une stratégie d'apprentissage incrémental, ce modèle fournit une association entre concepts textuels et caractéristiques visuelles qui contribue à améliorer la précision de l'annotation d'images et de la recherche d'images. La représentation visuelle de concepts textuels permet aussi l'annotation rapide des nouvelles images. De plus, ces représentations visuelles donnent à l'utilisateur la possibilité d'interroger le système par des requêtes textuelles ou par des requêtes mixtes texte/images, ceci même si toute la base d'images n'est pas annotée. En particulier, la recherche mixte texte/images peut être réalisée en utilisant seulement une requête textuelle.

Partant de l'hypothèse que la connaissance doit être dynamique et que les systèmes disposent très souvent de faibles quantités d'informations/connaissances au début de leur vie, la quatrième contribution porte sur un mécanisme de construction incrémentale des connaissances à partir de zéro. Nous avons ainsi proposé un mécanisme d'apprentissage interactif de connaissances (entre autres pour l'indexation) grâce à une technique adaptative de type retour de pertinence. La recherche interactive d'images permet de raffiner itérativement l'annotation et la connaissance sur les images. Selon cette stratégie, nous ne séparons pas les phases d'annotation et de recherche, et l'utilisateur peut ainsi faire des requêtes dès la mise en route du système, tout en laissant le système apprendre au fur et à mesure de son utilisation. Nous avons d'autre part montré que notre système a de bonnes performances concernant l'apprentissage de connaissances.

En l'état actuel, notre système intègre seulement l'indexation et la recherche d'images par le texte et le contenu visuel pour l'instant, mais d'autres types d'informations externes à l'image, comme la localisation (GPS) et le temps ont fait l'objet de premières études et sont discutées dans les perspectives.

6.1.2 Problèmes restants

Malgré les bonnes performances obtenues par le système d'indexation et de recherche d'images que nous avons proposé dans cette thèse, un certain nombre

de problématiques scientifiques restent encore très ouvertes. Ces questions portent sur les sujets suivants :

- Tout d'abord, les évolutions concernent la possibilité d'intégrer d'autres informations pour l'indexation : Dans notre système, nous avons réalisé l'indexation et la recherche d'images seulement par le texte et le contenu visuel. D'autres types d'informations externes comme la localisation (GPS) et le temps ne sont pas encore intégrées, même si nous avons conduit de premières études très encourageantes [Lai 2010]. La difficulté ici est liée à la combinaison de différentes sources d'informations pour rechercher des images d'intérêts. Pour l'instant, nous utilisons une fonction linéaire pour combiner la similarité de contenu et la similarité conceptuelle, mais cela ne marche pas si nous intégrons plus de types d'informations.
- D'autre part, des évolutions concernent également la question de la visualisation : Dans notre systèxme, nous avons proposé une technique de visualisation pour la recherche mixte texte/visuel, celle-ci se composant d'images exemples et de mots-clés (concepts textuels). Cette technique permet de voir les relations visuelle et conceptuelle des images avec la requête. Notre visualisation est pour l'instant limitée sur le nombre de mots-clés dans la requête. Le nombre maximal de mots-clés dans une requête est 4, ce qui correspond aux 4 quadrants de l'interface 2D. Il serait nécessaire de faire évoluer cette approche pour étendre le nombre de concepts. Il est difficile de modifier notre approche de représentation en coordonnées polaires pour s'adapter à plus de 4 mots-clés car la relation conceptuelle serait très mal présentée si nous divisions la représentation en plusieurs angles (par rapport aux 4 quadrants pour l'instant).

6.2 Perspectives

6.2.1 A court terme

Les perspectives à court terme de ce travail concernent les points suivants :

La visualisation Dans notre système, nous avons proposé une visualisation pour la recherche mixte texte/image. Cette technique permet de représenter des images avec la requête, en intégrant à la fois la relation visuelle et la relation textuelle. La requête mixte texte/image se compose d'images exemples et des mots-clés (concepts textuels). Notre visualisation est pour l'instant limitée à un certain nombre de mots-clés dans la requête. Le nombre maximal de mots-clés dans une requête est pour l'instant de 4. Dans l'avenir à court terme, nous proposons d'im-

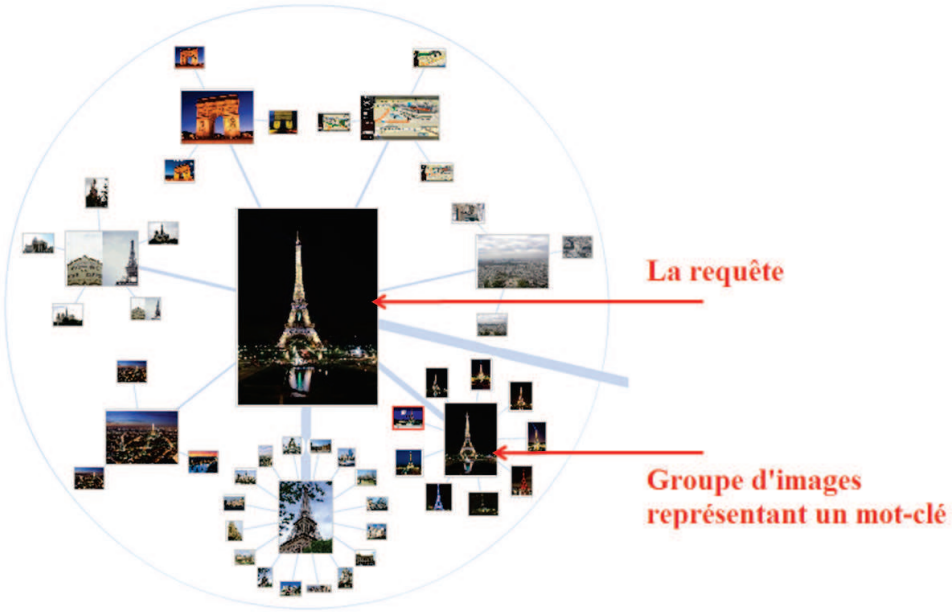

FIGURE 6.1 – Notre proposition basée sur la mise en page radiale utilisée dans Google Swirl.

plémenter une visualisation de graphes conceptuels avec une mise en page radiale (Google Swirl [1]). La mise en page radiale a été pour la première fois utilisée dans [Yee 2001], les résultats d'une requête de recherche d'images étant organisés suivant une structure d'arbre, chaque couche de l'arbre étant disposée radialement autour de son parent. Pour notre système, nous tentons de représenter les mots-clés par des sous-arbres qui contiennent leurs images pertinentes. En utilisant le zoom, des groupes d'images d'un grand nombre de mots-clés peuvent ainsi être tous présentés sur l'écran (voir figure 6.1).

Evaluation sur un groupe d'utilisateurs Pour l'instant, nous avons présenté une technique pour évaluer automatiquement notre système interactif. Cette technique est utilisée pour évaluer le retour de pertinence et aussi l'apprentissage incrémental. Nous aimerions réaliser l'expérimentation sur un groupe d'utilisateurs pour vérifier si notre système est facile à utiliser, et si la performance est confirmée dans des scénarios d'utilisation réels car il existe des différences. Par exemple, la méthode automatique interprète une image d'une seule façon mais les utilisateurs peuvent l'interpréter de différentes façons selon l'utilisateur en question et son contexte. De plus, le fait d'être sur un domaine spécialisé changerait probablement énormément la donne, par rapport à notre contexte d'images de catastrophes naturelles, qui comporte de nombreuses images de catégories différentes.

1. Google Swirl : http ://image-swirl.googlelabs.com/

6.2.2 A long terme

Intégration des information externes A long terme, nous proposons d'étendre notre système en utilisant des informations externes (le temps et la localisation) pour l'indexation et la recherche d'images. Pour la localisation, nous avons tenté une approche avec le stage de fin d'études Master de Lai Hien Phuong [Lai 2010] qui consistait à développer un modèle de recherche d'informations basé sur une double information de contenu visuel de l'image et de localisation géographique pour retrouver des situations d'urgence dans une ville (feux, blessés, bâtiments endommagés, etc.). Il s'agisssait également d'attribuer un niveau d'urgence en fonction de la proximité géographique d'événements similaires. La difficulté d'intégration de la localisation est l'indexation/l'organisation de la double information. Il faut indexer la localisation ensemble avec le contenu pour faciliter la recherche mixte localisation/images. Cela est différent de la recherche mixte texte/images dans notre système parce que nous considérons que l'information de texte est simple et il n'existe pas de relations entre des concepts textuels. Par contre, dans le cas d'images géo-référencées il existe des relations (spatiales, sur un espace continu) liées aux informations de localisation, ce qui augmente la complexité de l'indexation : il est nécessaire d'intégrer la notion de distance topologique dans la mesure de proximité entre images.

Dans ce travail, les images sont représentées par deux descripteurs différents, l'un représentant le contenu visuel de l'image (information interne) et l'autre représentant l'information de localisation géographique (information externe). Dans le cadre du stage de Lai Hien Phuong, nous nous sommes intéressés aux techniques d'indexation multidimensionnelles, telles que le SR-Tree, qui visent à regrouper les descripteurs de base dans des structures unitaires faciles à manipuler (hiérarchie).

Nous avons implémenté la structuration par partitionnement de données SR-Tree [Katayama 1997] pour organiser les descriptions du contenu visuel des images et les descripteurs externes d'informations géographiques afin de faciliter et d'accélérer la recherche des images similaires dans l'espace du contenu d'une part et le calcul de la proximité dans l'espace de la localisation géographique d'autre part. La figure 6.2 illustre des arbres SR-tree que nous avons construits dans le système. Nous avons fait le choix de générer plusieurs arbres SR-tree différents pour organiser les images :

- Des SR-trees structurant l'information du contenu visuel. En utilisant ces SR-trees, nous pouvons organiser des images de types différents et correspondant aux différentes situations d'urgence susceptibles d'être rencontrées (feu, blessé, bâtiment endommagé, route endommagée ou inondation).
- Un SR-tree structurant l'information des monuments dans la ville. Cet arbre est construit en utilisant un fichier *shapefile*. Il est utilisé pour attribuer un

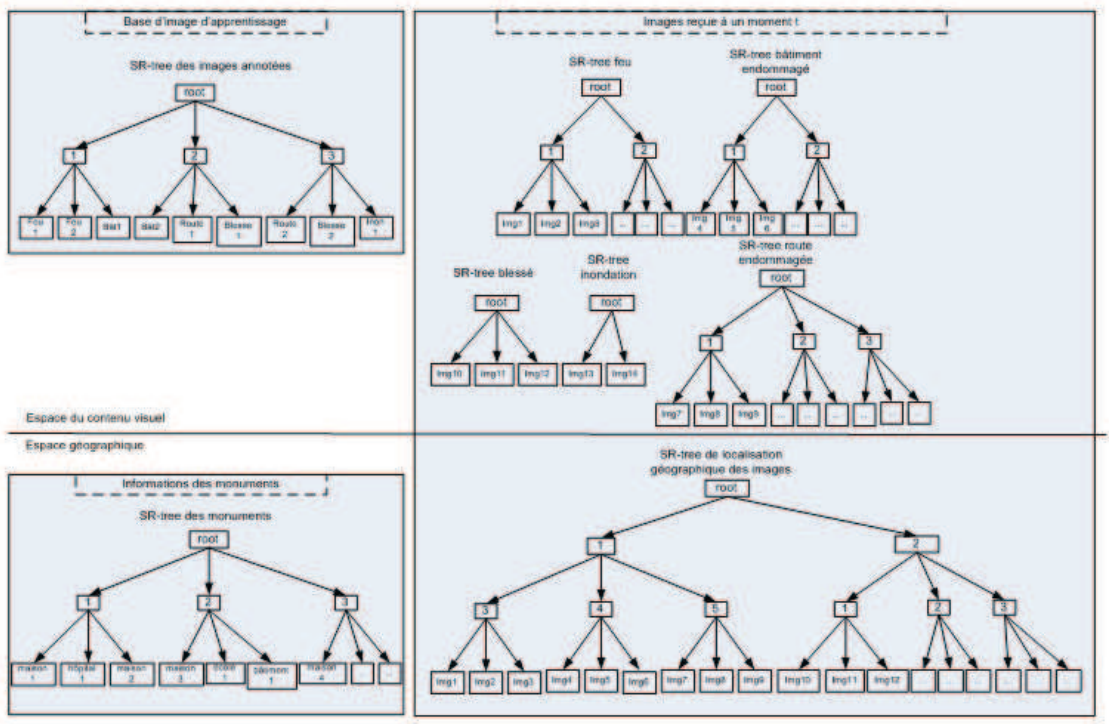

FIGURE 6.2 – Structuration des données par des SR-tree.

niveau d'urgence pour chaque situation (voir figure 6.3).

La structuration SR-tree a été retenue parce qu'elle est bien adaptée pour les applications d'indexation de similarité des images/vidéos d'une part, et pour son potentiel de regroupement toplogique des données qui sont proches par des rectangles englobants et des sphères englobantes correspondant bien à la structuration des données géographiques d'autre part. Ce type de stratégie offre la possibilité d'effectuer une recherche rapide d'images dans les deux espaces, intégrant à la fois des contraintes de contenu et de localisation, tout en évitant des comparaisons exhaustives avec tous les éléments de la base.

Pour l'intégration de la localisation dans notre système, nous avons donc pu expérimenter utiliser la structuration SR-tree pour indexer a la fois le contenu et la localisation. La localisation peut être considérée comme une option pour la recherche d'images par le contenu. Cette idée est donc applicable dans notre système mais nous devrons au préalable approfondir nos études pour bien exploiter la localisation et combiner avec les informations du contenu et du texte.

Visualisation basée sur des agents Une autre possibilité pour la recherche et la visualisation a été proposée dans le stage de Master 2 de Guillaume Chiron [Chiron 2010]. Pendant ce stage, nous avons implémenté un prototype pour la

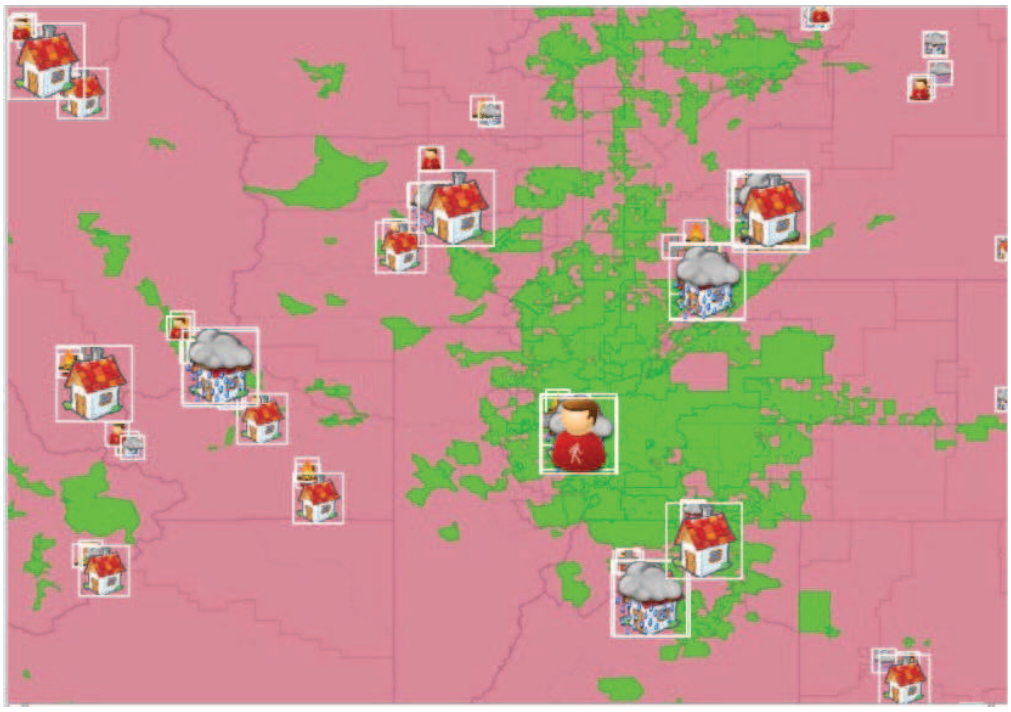

FIGURE 6.3 – Groupes de situations d'urgence dans la ville.

recherche d'images et la visualisation basée sur des agents. La base d'images est représentée sous la forme d'un système multi-agent où chaque image est un agent se déplaçant dans un espace 2D (l'espace de visualisation). Nous avons montré qu'il était possible d'utiliser un modèle à base d'agents réactifs pour faire de la recherche d'images par le contenu et de la visualisation d'images.

Dans la figure 6.4, nous proposons une comparaison entre la méthode traditionnelle et notre méthode à base d'agents. Dans cette méthode, des images sont placées dans l'espace 2D de visualisation en se basant sur un modèle de forces d'attraction et de répulsion, remplaçant la fonction de similarité traditionnelle. Les images similaires sont attirées les unes aux autres et se retrouvent ainsi à proximité, tandis que les images non similaires se repoussent et se retrouvent loin les unes des autres. Chaque image est représentée par un agent, chaque agent (image) interagissant avec ses voisins. Ces interactions se caractérisent par l'application de forces entre ces agents. Un agent subit les forces engendrées par ses voisins, forces qui expriment généralement un désir d'attirer ou de repousser un agent. Ces forces induisent des déplacements des agents-images dans l'espace. Au final, tous les agents se déplacent vers et/ou s'éloignent des autres agents en fonction de leurs similarités. Le système se stabilise enfin lorsqu'un équilibre entre les différentes forces est trouvé.

En fonction de la similarité et de la distance entre des agents-images, différentes règles de forces d'attraction et de répulsion sont définies (figure 6.5). Un seuil de

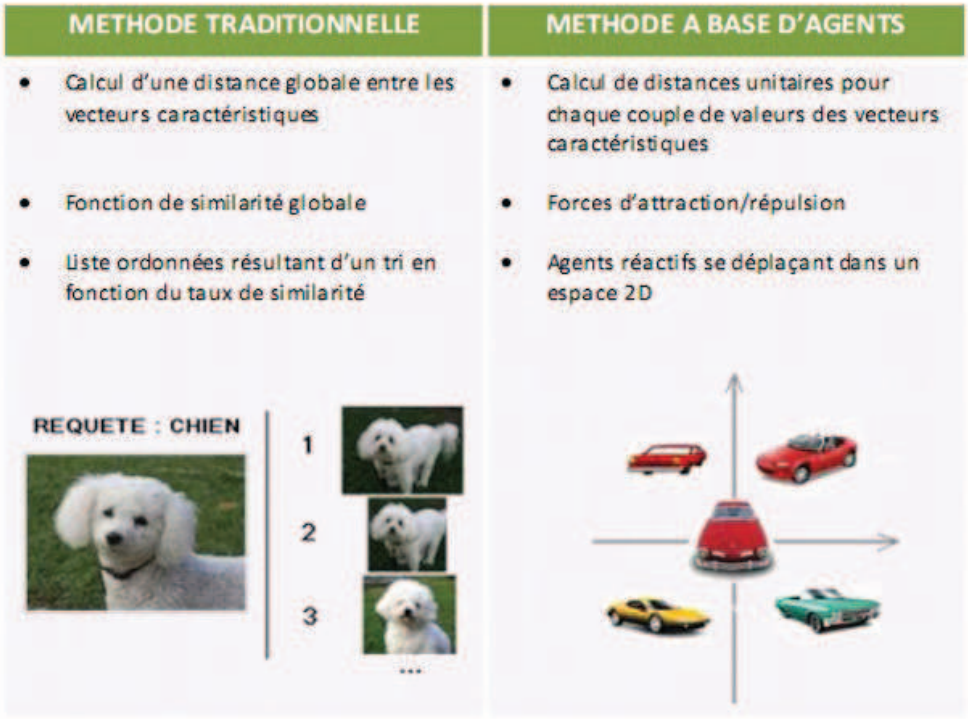

FIGURE 6.4 – Comparatif entre méthode traditionnelle / nouvelle méthode à base d'agents pour la recherche d'images par le contenu.

SIMILARITE	PROXIMITE	FORCE	PUISSANCE
forte	loin	attraction	forte
forte	proche	attraction	faible
faible	loin	répulsion	faible
faible	proche	répulsion	forte

FIGURE 6.5 – Nature et puissance des forces selon la similarité et la proximité des agents.

neutralité est fixé, et si le produit est inférieur à ce seuil, nous sommes en présence d'une force attractive, tandis que si ce dernier est supérieur, nous sommes en présence d'une force répulsive. L'écart au seuil est ensuite pondéré pour fixer la puissance de la force.

Un exemple de ce système de visualisation est illustré dans la figure 6.6. Au milieu se trouve la requête, les images similaires étant placées proches et les autres images se trouvant plus loin du centre. De plus, des images similaires sont situées proches les unes des autres et elles formént des groupes distincts. Ce résultat montre la bonne performance du système.

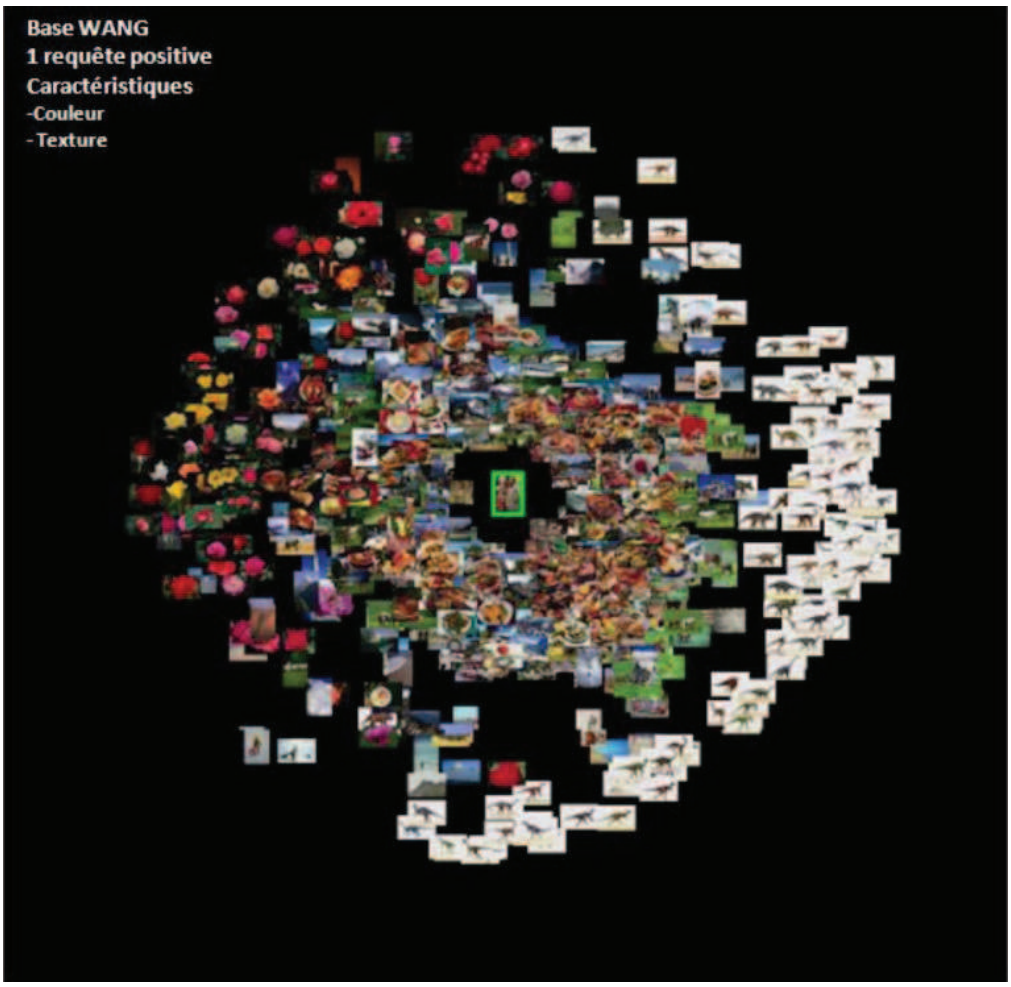

FIGURE 6.6 – Visualisation basée sur des agents.

Dans ce système, il n'existe aucune caractéristique liée à chacun des axes horizontal et vertical, les agents se déplaçant librement dans tout l'espace 2D. L'organisation spatiale est le cœur du système et se fait de façon naturelle comme résultat de toutes les interactions entre tous les agents sans aucune fonction globale venant réguler le système. Grâce aux avantages montrés ci-dessus de cette visualisation

basées sur des agents, elle sera une très bonne visualisation alternative pour notre système. Cependant, pour l'intégrer, il faut re-implémenter des règles de forces d'attraction et de répulsion pour notre information multimodale utilisée, car le travail de Guillaume Chiron [Chiron 2010] utilise principalement le contenu. De plus, il nous faut peut-être étudier d'autres techniques d'interactions pour adapter ce type de visualisation.

6.3 Conclusion finale

Notre système de recherche d'images avec plusieurs aspects est présenté et expérimenté dans cette thèse. Tout d'abord, la recherche mixte texte/images avec une nouvelle technique de retour de pertinence et une nouvelle visualisation a été appliquée et évaluée. Ensuite, nous avons démontré que les connaissances du système peuvent évoluer à partir de zéro en utilisant notre méthode d'apprentissage interactif. L'expérimentation dans le chapitre 5 montre que notre approche donne de bonnes performances dans différentes situations d'utilisation du système

Notre système peut être étendu pour d'autres types d'informations que seulement le texte ou les images comme la localisation et/ou le temps. De même, la visualisation et le retour de pertinence intégré dans le système peuvent être améliorés en utilisant une mise en page radiale ou en se basant sur des agents.

Bien que la méthode d'apprentissage interactif présentée dans cette thèse puisse encore être améliorée, elle peut déjà être utilisée pour des systèmes spécifiques qui n'ont pas de connaissances au début de leur vie et qui demande d'évoluer dans le temps via des interactions avec l'humain.

Enfin, nous croyons que, dans l'avenir, l'apprentissage interactif et l'information multimodale joueront un rôle majeur dans les systèmes, en remplaçant les technologies d'informations unimodales et en perfectionnant l'apprentissage statistique pour profiter des connaissances venant des interactions.

Travaux de l'auteur

[Nguyen 2009] Nhu Van Nguyen and Jean-marc Ogier and Salvatore Tabbone and Alain Boucher. *Text Retrieval Relevance Feedback Techniques for Bag of Words Model in CBIR*. International Conference on Machine Learning and Pattern Recognition (ICMLPR), 2009, Paris, France.

[Nguyen 2009b] Nhu Van Nguyen and Jean-marc Ogier and Salvatore Tabbone and Alain Boucher. *Region-Based Semi-automatic Annotation Using the Bag of Words Representation of the Keywords.* In Proceedings of the 5th International Conference on Image and Graphics (ICIG) 2009), pages 422–427. IEEE Computer Society, 2009, Xi'an, Shanxi, China.

[Lai 2010] Hien Phuong Lai and Nhu Van Nguyen and Alain Boucher and Jean-Marc Ogier. *Using SR-tree in a Content-based and Location-based Image Retrieval System.* In Proceedings of the 5th International Conference on Computer Vision Theory and Applications(VISAPP), pages 491–494. INSTICC Press, 2010, Angers, France.

[Nguyen 2010] Nhu Van Nguyen and Jean-marc Ogier and Salvatore Tabbone and Alain Boucher. *Clusters-Based Relevance Feedback for CBIR : A Combination of Query Movement and Query Expansion.* In Proceedings of the 10th International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), pages 422–427. IEEE Computer Society, 2010, Hanoi, Vietnam.

[Nguyen 2011] Nhu Van Nguyen and Jean-marc Ogier and Salvatore Tabbone and Alain Boucher. *Cluster-based Relevance Feedback for CBIR : A combination of query point movement and query expansion.* Journal of Ambient Intelligence and Humanized Computing. IEEE Computer Society, submitted, 2011

Bibliographie

[Aggarwal 2002] G. Aggarwal, T.V. Ashwin et S. Ghosal. *An image retrieval system with automatic query modification*. IEEE Transactions on Multimedia, vol. 4, no. 2, pages 201 – 214, 2002.

[Agrawal 2006] Rajeev Agrawal, William Grosky et Farshad Fotouhi. *Image Retrieval Using Multimodal Keywords*. International Symposium Multimedia, pages 817–822, 2006.

[Andrews 2007] Keith Andrews, Werner Putz et Alexander Nussbaumer. *The Hierarchical Visualisation System (HVS)*. In IV '07 : Proceedings of the 11th International Conference Information Visualization, pages 257–262, Washington, DC, USA, 2007. IEEE Computer Society.

[Anguera 2008] Xavier Anguera et Nuria Oliver. *MAMI : multimodal annotations on a camera phone*. In Proceedings of the 10th international conference on Human computer interaction with mobile devices and services, MobileHCI '08, pages 379–382, New York, NY, USA, 2008. ACM.

[Aslandogan 1999] Y. Alp Aslandogan et Clement T. Yu. *Techniques and Systems for Image and Video Retrieval*. IEEE Transactions on Knowledge and Data Engineering, vol. 11, pages 56–63, 1999.

[Barnard 2003] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei et Michael I. Jordan. *Matching words and pictures*. The Journal of Machine Learning Research, vol. 3, pages 1107–1135, 2003.

[Barrat 2009] Sabine Barrat et Salvatore Tabbone. *Modeling, Classifying and Annotating Weakly Annotated Images Using Bayesian Network*. In Proceedings of the 10th International Conference on Document Analysis and Recognition, ICDAR '09, pages 1201–1205, Washington, DC, USA, 2009. IEEE Computer Society.

[Bartolini 2010] Ilaria Bartolini et Paolo Ciaccia. *Multi-dimensional keyword-based image annotation and search*. In Proceedings of the 2nd International Workshop on Keyword Search on Structured Data, KEYS '10, New York, NY, USA, 2010. ACM.

[Bay 2006] Herbert Bay, Tinne Tuytelaars et Luc Van Gool. *SURF : Speeded Up Robust Features*. In Aleš Leonardis, Horst Bischof et Axel Pinz, editeurs,

Computer Vision – ECCV 2006, volume 3951 de *Lecture Notes in Computer Science*, pages 404–417. Springer Berlin / Heidelberg, 2006.

[Bederson 2001] Benjamin B. Bederson. *PhotoMesa : a zoomable image browser using quantum treemaps and bubblemaps*. In UIST '01 : Proceedings of the 14th annual ACM symposium on User interface software and technology, pages 71–80, New York, NY, USA, 2001. ACM.

[Belkhatir 2005] Mohammed Belkhatir, Philippe Mulhem et Yves Chiaramella. *A Conceptual Image Retrieval Architecture Combining Keyword-Based Querying with Transparent and Penetrable Query-by-Example*. In Wee-Kheng Leow, Michael Lew, Tat-Seng Chua, Wei-Ying Ma, Lekha Chaisorn et Erwin Bakker, editeurs, Image and Video Retrieval, volume 3568 de *Lecture Notes in Computer Science*, pages 528–539. Springer Berlin / Heidelberg, 2005.

[Berg 2005] Alexander C. Berg, Tamara L. Berg et Jitendra Malik. *Shape Matching and Object Recognition Using Low Distortion Correspondences*. In CVPR '05 : Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1, pages 26–33, Washington, DC, USA, 2005. IEEE Computer Society.

[Camargo 2010] Jorge Camargo, Juan Caicedo et Fabio González. *Multimodal Image Collection Visualization Using Non-negative Matrix Factorization*. In Mounia Lalmas, Joemon Jose, Andreas Rauber, Fabrizio Sebastiani et Ingo Frommholz, editeurs, Research and Advanced Technology for Digital Libraries, volume 6273 de *Lecture Notes in Computer Science*, pages 429–432. Springer Berlin / Heidelberg, 2010.

[Carey 2003] Matthew Carey, Daniel C Heesch et Stefan M Rüger. *Info navigator : A visualization tool for document searching and browsing*. In In Proc. of the Intl. Conf. on Distributed Multimedia Systems (DMS), pages 23–28, 2003.

[Carneiro 2007] Gustavo Carneiro, Antoni B. Chan, Pedro J. Moreno et Nuno Vasconcelos. *Supervised Learning of Semantic Classes for Image Annotation and Retrieval*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 3, pages 394–410, 2007.

[Carson 2002] Chad Carson, Serge Belongie, Hayit Greenspan et Jitendra Malik. *Blobworld : Image Segmentation Using Expectation-Maximization and Its Application to Image Querying*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 8, pages 1026–1038, 2002.

[Chandrika 2010] Pulla Chandrika et C. V. Jawahar. *Multi modal semantic indexing for image retrieval*. In CIVR '10 : Proceedings of the ACM International Conference on Image and Video Retrieval, pages 342–349, New York, NY, USA, 2010. ACM.

[Chang 2003] E. Chang, Kingshy Goh, G. Sychay et Gang Wu. *CBSA : content-based soft annotation for multimodal image retrieval using Bayes point machines*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 1, pages 26–38, 2003.

[Chang 2006] Yih-Chen Chang, Wen-Cheng Lin et Hsin-Hsi Chen. *A Corpus-Based Relevance Feedback Approach to Cross-Language Image Retrieval*. In Carol Peters, Fredric Gey, Julio Gonzalo, Henning Müller, Gareth Jones, Michael Kluck, Bernardo Magnini, Maarten de Rijke et Danilo Giampiccolo, editeurs, Accessing Multilingual Information Repositories, volume 4022 de *Lecture Notes in Computer Science*, pages 592–601. Springer Berlin / Heidelberg, 2006.

[Chang 2008] Yih-Chen Chang et Hsin-Hsi Chen. *Using an Image-Text Parallel Corpus and the Web for Query Expansion in Cross-Language Image Retrieval*. pages 504–511, 2008.

[Chen 2000] J.Y. Chen, C.A. Bouman et J.C. Dalton. *Hierarchical Browsing and Search of Large Image Databases*. IEEE Transactions on Image Processing, vol. 9, no. 3, pages 442–455, March 2000.

[Chen 2005] Y.X. Chen, J.Z. Wang et R. Krovetz. *CLUE : cluster-based retrieval of images by unsupervised learning*. IEEE Transactions on Image Processing, vol. 14, no. 8, pages 1187–1201, August 2005.

[Chevallet 2005] Jean-Pierre Chevallet et Joo-Hwee Lim. *SnapToTell Accès ubiquitaire à de l'information multimédia à partir d'un téléphone portable*. In COnférence en Recherche d'Infomations et Applications - CORIA2005, pages 245–260, 2005.

[Chevallet 2007] Jean-Pierre Chevallet, Joo-Hwee Lim et Mun-Kew Leong. *Object identification and retrieval from efficient image matching. Snap2Tell with the STOIC dataset*. Inf. Process. Manage., vol. 43, no. 2, pages 515–530, 2007.

[Chiron 2010] Guillaume Chiron. *Modèle multi-agent d'attraction/répulsion pour la recherche d'images par le contenu*. In Mémoire de stage Master, Stages de fin etude, Institut de la Francophonie pour l'Informatique, 2010.

[Combs 1999] Tammara T. A. Combs et Benjamin B. Bederson. *Does zooming improve image browsing ?* In DL '99 : Proceedings of the fourth ACM conference on Digital libraries, pages 130–137, New York, NY, USA, 1999. ACM.

[Cox 1992] K. Cox. *Information retrieval by browsing*. In Proceedings of The Fifth International Conference on New Information Technology, pages 69–80, 1992.

[Datta 2008] Ritendra Datta, Dhiraj Joshi, Jia Li et James Z. Wang. *Image retrieval : Ideas, influences, and trends of the new age*. ACM Comput. Surv., vol. 40, no. 2, pages 1–60, 2008.

[Duygulu 2002] P. Duygulu, Kobus Barnard, J. F. G. de Freitas et David A. Forsyth. *Object Recognition as Machine Translation : Learning a Lexicon for a Fixed Image Vocabulary*. In ECCV '02 : Proceedings of the 7th European Conference on Computer Vision-Part IV, pages 97–112, London, UK, 2002. Springer-Verlag.

[Escalante 2008] Hugo Jair Escalante, Carlos A. Hérnadez, Luis Enrique Sucar et Manuel Montes. *Late fusion of heterogeneous methods for multimedia image retrieval*. In Proceeding of the 1st ACM international conference on Multimedia information retrieval, MIR '08, pages 172–179, New York, NY, USA, 2008. ACM.

[Fang 2005] Hui Fang et ChengXiang Zhai. *An exploration of axiomatic approaches to information retrieval*. In SIGIR '05 : Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pages 480–487, New York, NY, USA, 2005. ACM.

[Fei-Fei 2006] Li Fei-Fei, Rob Fergus et Pietro Perona. *One-Shot Learning of Object Categories*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, pages 594–611, April 2006.

[Fei-Fei 2007] Li Fei-Fei. *Tutorial on Bag-of-words models*. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007.

[Feng 2004] S. L. Feng, R. Manmatha et V. Lavrenko. *Multiple Bernoulli Relevance Models for Image and Video Annotation*. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pages 1002–1009, 2004.

[Ferecatu 2008] Marin Ferecatu, Nozha Boujemaa et Michel Crucianu. *Semantic interactive image retrieval combining visual and conceptual content description*. Multimedia Systems, vol. 13, pages 309–322, 2008.

[Flickner 1995] Myron Flickner, Harpreet S. Sawhney, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele et Peter Yanker. *Query by Image and Video Content : The QBIC System*. IEEE Computer, vol. 28, no. 9, pages 23–32, 1995.

[Frigui 1997] Hichem Frigui et Raghu Krishnapuram. *Clustering by competitive agglomeration*. Pattern Recognition, vol. 30, no. 7, pages 1109 – 1119, 1997.

[Fu 2008] Yun Fu, Liangliang Cao, Guodong Guo et Thomas S. Huang. *Multiple feature fusion by subspace learning*. In CIVR '08 : Proceedings of the 2008 international conference on Content-based image and video retrieval, pages 127–134, New York, NY, USA, 2008. ACM.

[Fuhr 1992] Norbert Fuhr. *Probabilistic models in information retrieval.* Comput. J., vol. 35, no. 3, pages 243–255, 1992.

[Furnas 1986] G. W. Furnas. *Generalized fisheye views.* SIGCHI Bull., vol. 17, no. 4, pages 16–23, 1986.

[Haralick 1973] R. M. Haralick, Dinstein et K. Shanmugam. *Textural features for image classification.* IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3, pages 610–621, November 1973.

[Havre 2002] Susan Havre, Elizabeth Hetzler, Paul Whitney et Lucy Nowell. *ThemeRiver : Visualizing Thematic Changes in Large Document Collections.* IEEE Transactions on Visualization and Computer Graphics, vol. 8, no. 1, pages 9–20, 2002.

[He 2004] Jingrui He, Hanghang Tong, Mingjing Li, Hong-Jiang Zhang et Changshui Zhang. *Mean version space : a new active learning method for content-based image retrieval.* In MIR '04 : Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval, pages 15–22, New York, NY, USA, 2004. ACM.

[Heesch 2006] Daniel Heesch, Alexei Yavlinsky et Stefan Rüger. *NNk networks and automated annotation for browsing large image collections from the world wide web.* In MULTIMEDIA '06 : Proceedings of the 14th annual ACM international conference on Multimedia, pages 493–494, New York, NY, USA, 2006. ACM.

[Heesch 2008] Daniel Heesch. *A survey of browsing models for content based image retrieval.* Multimedia Tools Appl., vol. 40, no. 2, pages 261–284, 2008.

[Hofmann 1998] Thomas Hofmann. *Learning and Representing Topic. A Hierarchical Mixture Model for Word Occurrences in Document Databases.* In Proceedings of the Conference for Automated Learning and Discovery (CONALD), Pittsburgh, 1998.

[Hörster 2008] Eva Hörster, Rainer Lienhart et Malcolm Slaney. *Continuous visual vocabulary modelsfor pLSA-based scene recognition.* In CIVR '08 : Proceedings of the 2008 international conference on Content-based image and video retrieval, pages 319–328, New York, NY, USA, 2008. ACM.

[Huang 1997] Jing Huang, S. Ravi Kumar, Mandar Mitra, Wei-Jing Zhu et Ramin Zabih. *Image Indexing Using Color Correlograms.* Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, page 762, 1997.

[Huiskes 2008] Mark J. Huiskes et Michael S. Lew. *Performance evaluation of relevance feedback methods.* In CIVR '08 : Proceedings of the 2008 international conference on Content-based image and video retrieval, pages 239–248, New York, NY, USA, 2008. ACM.

[Huynh 2005] David F. Huynh, Steven M. Drucker, Patrick Baudisch et Curtis Wong. *Time quilt : scaling up zoomable photo browsers for large, unstructured photo collections.* In CHI '05 extended abstracts on Human factors in computing systems, CHI '05, pages 1937–1940, New York, NY, USA, 2005. ACM.

[Ishikawa 1998] Yoshiharu Ishikawa, Ravishankar Subramanya et Christos Faloutsos. *MindReader : Querying Databases Through Multiple Examples.* In VLDB '98 : Proceedings of the 24rd International Conference on Very Large Data Bases, pages 218–227, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[Jaimes 2006] Alejandro Jaimes, Nicu Sebe et Daniel Gatica-Perez. *Human-centered computing : a multimedia perspective.* In MULTIMEDIA '06 : Proceedings of the 14th annual ACM international conference on Multimedia, pages 855–864, New York, NY, USA, 2006. ACM.

[Jain 1988] Anil K. Jain et Richard C. Dubes. Algorithms for clustering data. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[Janecek 2005] Paul Janecek et Pearl Pu. *An evaluation of semantic fisheye views for opportunistic search in an annotated image collection.* International Journal on Digital Libraries, vol. 5, pages 42–56, 2005.

[Jeon 2003] J. Jeon, V. Lavrenko et R. Manmatha. *Automatic image annotation and retrieval using cross-media relevance models.* In SIGIR '03 : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pages 119–126, New York, NY, USA, 2003. ACM.

[Jin 2003a] Sun Jin Li et Hong-Hui. *On interactive browsing of large images.* IEEE Transactions on Multimedia, vol. 5, no. 4, pages 581 – 590, Dec. 2003.

[Jin 2003b] Xiangyu Jin et James C. French. *Improving image retrieval effectiveness via multiple queries.* In MMDB '03 : Proceedings of the 1st ACM international workshop on Multimedia databases, pages 86–93, New York, NY, USA, 2003. ACM.

[Jing 2004] Feng Jing, Mingjing Li, Hong-Jiang Zhang et Bo Zhang. *Relevance Feedback for Keyword and Visual Feature-Based Image Retrieval.* In Peter Enser, Yiannis Kompatsiaris, Noel E. OĆonnor, Alan F. Smeaton et Arnold W. M. Smeulders, editeurs, Image and Video Retrieval, volume 3115 de *Lecture Notes in Computer Science*, pages 648–648. Springer Berlin / Heidelberg, 2004.

[Jones 2000] K. Sparck Jones, S. Walker et S. E. Robertson. *A probabilistic model of information retrieval : development and comparative experiments.* Inf. Process. Manage., vol. 36, no. 6, pages 779–808, 2000.

[Joo-Hwee Lim 2004] Sihem Nouarah Merah Joo-Hwee Lim Jean-Pierre Chevallet. *SnapToTell : Ubiquitous Information Access from Camera*. In Workshop on Mobile and Ubiquitous Information Access (MUIA04), Udine, Italy, 2004. Springer.

[Kadir 2004] T. Kadir, A. Zisserman et J. M. Brady. *An Affine Invariant Salient Region Detector*. In European Conference on Computer Vision. Springer-Verlag, 2004.

[Karthik 2006] S. Karthik et C.V. Jawahar. *Discriminative relevance feedback with virtual textual representation for efficient image retrieval*. IET Conference Publications, vol. 2006, no. CP522, pages 309–314, 2006.

[Katayama 1997] Norio Katayama et Shin'ichi Satoh. *The SR-tree : an index structure for high-dimensional nearest neighbor queries*. In Proceedings of the 1997 ACM SIGMOD international conference on Management of data, SIGMOD '97, pages 369–380, New York, NY, USA, 1997. ACM.

[Kim 2005] Deok-Hwan Kim, Chin-Wan Chung et Kobus Barnard. *Relevance feedback using adaptive clustering for image similarity retrieval*. J. Syst. Softw., vol. 78, no. 1, pages 9–23, 2005.

[Kothari 1999] Ravi Kothari et Dax Pitts. *On finding the number of clusters*. Pattern Recogn. Lett., vol. 20, no. 4, pages 405–416, 1999.

[Lai 2010] Hien Phuong Lai, Nhu Van Nguyen, Alain Boucher et Jean-Marc Ogier. *Using SR-tree in a Content-based and Location-based Image Retrieval System*. In VISAPP-International Conference on Computer Vision Theory and Applications, pages 491–494, Angers, France, 2010. INSTICC Press.

[Lau 2007] C. Lau, D. Tjondronegoro, J. Zhang, S. Geva et Y. Liu. *Fusing Visual and Textual Retrieval Techniques to Effectively Search Large Collections of Wikipedia Images*. In Norbert Fuhr, Mounia Lalmas et Andrew Trotman, editeurs, Comparative Evaluation of XML Information Retrieval Systems, volume 4518 de *Lecture Notes in Computer Science*, pages 345–357. Springer Berlin / Heidelberg, 2007.

[Lavrenko 2004] V. Lavrenko, R. Manmatha et J. Jeon. *A Model for Learning the Semantics of Pictures*. Advances in Neural Information Processing Systems (NIPS), pages 553–560, 2004.

[Lew 2006] Michael S. Lew, Nicu Sebe, Chabane Djeraba et Ramesh Jain. *Content-based multimedia information retrieval : State of the art and challenges*. ACM Trans. Multimedia Comput. Commun. Appl., vol. 2, no. 1, pages 1–19, 2006.

[Li 2005] Fei-Fei Li et Pietro Perona. *A Bayesian Hierarchical Model for Learning Natural Scene Categories*. In CVPR '05 : Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition

(CVPR'05) - Volume 2, pages 524–531, Washington, DC, USA, 2005. IEEE Computer Society.

[Li 2008] Jia Li et James Z. Wang. *Real-Time Computerized Annotation of Pictures.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, pages 985–1002, June 2008.

[Lienhart 2009] Rainer Lienhart, Stefan Romberg et Eva Hörster. *Multilayer pLSA for multimodal image retrieval.* In CIVR '09 : Proceeding of the ACM International Conference on Image and Video Retrieval, pages 1–8, New York, NY, USA, 2009. ACM.

[Lin 2007] Wen-Cheng Lin, Yih-Chen Chang et Hsin-Hsi Chen. *Integrating textual and visual information for cross-language image retrieval : a trans-media dictionary approach.* Inf. Process. Manage., vol. 43, no. 2, pages 488–502, 2007.

[Lindeberg 1998] Tony Lindeberg. *Feature Detection with Automatic Scale Selection.* Int. J. Comput. Vision, vol. 30, no. 2, pages 79–116, 1998.

[Liu 2008] Haiming Liu, Dawei Song, Stefan Rüger, Rui Hu et Victoria Uren. *Comparing dissimilarity measures for content-based image retrieval.* In AIRS'08 : Proceedings of the 4th Asia information retrieval conference on Information retrieval technology, pages 44–50, Berlin, Heidelberg, 2008. Springer-Verlag.

[Liu 2009] Danzhou Liu, Kien A. Hua, Khanh Vu et Ning Yu. *Fast Query Point Movement Techniques for Large CBIR Systems.* IEEE Trans. on Knowl. and Data Eng., vol. 21, no. 5, pages 729–743, 2009.

[Lowe 2004] David G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints.* International Journal of Computer Vision, vol. 60, pages 91–110, 2004.

[Makadia 2010] Ameesh Makadia, Vladimir Pavlovic et Sanjiv Kumar. *Baselines for Image Annotation.* International Journal of Computer Vision, vol. 90, pages 88–105, 2010.

[Mallat 1989] S. G. Mallat. *A Theory for Multiresolution Signal Decomposition : The Wavelet Representation.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 11, no. 7, pages 674–693, 1989.

[Matas 2002] J. Matas, O. Chum, M. Urban et T. Pajdla. *Robust wide baseline stereo from maximally stable extremal regions.* In Proceedings of British Machine Vision Conference, volume 1, pages 384–393, London, UK, 2002.

[Metzler 2004] Donald Metzler et R. Manmatha. *An Inference Network Approach to Image Retrieval.* In CIVR04-Proceedings of the 2004 international conference on Content-based image and video retrieval, pages 42–50, 2004.

[Mikolajczyk 2005a] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir et L. Van Gool. *A Comparison of*

Affine Region Detectors. Int. J. Comput. Vision, vol. 65, no. 1-2, pages 43–72, 2005.

[Mikolajczyk 2005b] Krystian Mikolajczyk et Cordelia Schmid. *A performance evaluation of local descriptors*. IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 27, no. 10, pages 1615–1630, 2005.

[MM 2010] Rahman MM, Antani SK, Long LR, Demner-Fushman D et Thoma GR. Lecture notes in computer science. first miccai international workshop on medical content-based retrieval for clinical decision support (mcbr-cds 2009) ; part of the 12th international conference on medical image computing and computer assisted interventio february 2010, chapitre Multi-Modal Query Expansion Based On Local Analysis For Medical Image Retrieval. National Library of Medicine, NIH, 2010.

[Monay 2003] Florent Monay et Daniel Gatica-Perez. *On image auto-annotation with latent space models*. In MULTIMEDIA '03 : Proceedings of the eleventh ACM international conference on Multimedia, pages 275–278, New York, NY, USA, 2003. ACM.

[Mori 1999] Y. Mori, H. Takahashi et R. Oka. *Image-to-word transformation based on dividing and vector quantizing images with words*. In Proc. First Int'l Workshop Multimedia Intelligent Storage and Retrieval Management, 1999.

[Muller 2002] Henning Muller, Stephane Marchand-Maillet et Thierry Pun. *The Truth about Corel - Evaluation in Image Retrieval*. In Michael Lew, Nicu Sebe et John Eakins, editeurs, Image and Video Retrieval, volume 2383 de *Lecture Notes in Computer Science*, pages 38–49. Springer Berlin / Heidelberg, 2002.

[Natsev 2003] A. P. Natsev et J. R. Smith. *Active selection for multi-example querying by content*. In ICME '03 : Proceedings of the 2003 International Conference on Multimedia and Expo, pages 445–448, Washington, DC, USA, 2003. IEEE Computer Society.

[Natsev 2005] Apostol (Paul) Natsev, Milind R. Naphade et Jelena TešiĆ. *Learning the semantics of multimedia queries and concepts from a small number of examples*. In MULTIMEDIA '05 : Proceedings of the 13th annual ACM international conference on Multimedia, pages 598–607, New York, NY, USA, 2005. ACM.

[Ortega Binderberger 2004] Michael Ortega Binderberger et Sharad Mehrotra. *Relevance feedback techniques in the MARS image retrieval system*. Multimedia Systems, vol. 9, pages 535–547, 2004.

[Pham 2007] Trong-Ton Pham, Nicolas Eric Maillot, Joo-Hwee Lim et Jean-Pierre Chevallet. *Latent semantic fusion model for image retrieval and annotation*. In CIKM '07 : Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pages 439–444, New York, NY, USA, 2007. ACM.

[Pham 2010] Trong-Ton Pham, Philippe Mulhem et Loic Maisonnasse. *Spatial relationships in visual graph modeling for image categorization*. In Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '10, pages 729–730, New York, NY, USA, 2010. ACM.

[Philipp-Foliguet 2009] Sylvie Philipp-Foliguet, Julien Gony et Philippe Henri Gosselin. *FReBIR : An image retrieval system based on fuzzy region matching*. Computer Vision and Image Understanding, vol. 113, pages 693–707, 2009.

[Rocchio 1971] J. Rocchio. Relevance feedback in information retrieval, pages 313–323. Prentice Hall, Englewood, Cliffs, New Jersey, 1971.

[Rodden 2001] Kerry Rodden, Wojciech Basalaj, David Sinclair et Kenneth Wood. *Does organisation by similarity assist image browsing ?* In CHI '01 : Proceedings of the SIGCHI conference on Human factors in computing systems, pages 190–197, New York, NY, USA, 2001. ACM.

[Salton 1975] G. Salton, A. Wong et C. S. Yang. *A vector space model for automatic indexing*. Commun. ACM, vol. 18, no. 11, pages 613–620, 1975.

[Santini 2001] Simone Santini, Amarnath Gupta et Ramesh Jain. *Emergent Semantics through Interaction in Image Databases*. IEEE Trans. on Knowl. and Data Eng., vol. 13, no. 3, pages 337–351, 2001.

[Schvaneveldt 1990] Roger W. Schvaneveldt, editeur. Pathfinder associative networks : studies in knowledge organization. Ablex Publishing Corp., Norwood, NJ, USA, 1990.

[Sclaroff 1999] Stan Sclaroff, Marco La Cascia et Saratendu Sethi. *Unifying textual and visual cues for content-based image retrieval on the World Wide Web*. Comput. Vis. Image Underst., vol. 75, no. 1-2, pages 86–98, 1999.

[Shechtman 2007] E. Shechtman et M. Irani. *Matching Local Self-Similarities across Images and Videos*. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE Computer Society, 2007.

[Shneiderman 1996] Ben Shneiderman. *The Eyes Have It : A Task by Data Type Taxonomy for Information Visualizations*. In VL '96 : Proceedings of the 1996 IEEE Symposium on Visual Languages, page 336, Washington, DC, USA, 1996. IEEE Computer Society.

[Singhal 2001] Amit Singhal. *Modern Information Retrieval : A Brief Overview*. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, vol. 24, no. 4, pages 35–42, 2001.

[Sivic 2008] J. Sivic et A. Zisserman. *Efficient Visual Search for Objects in Videos*. Proceedings of the IEEE, vol. 96, no. 4, pages 548–566, 2008.

[Smeulders 2000] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta et Ramesh Jain. *Content-Based Image Retrieval at the End of the Early Years.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, pages 1349–1380, December 2000.

[Snoek 2005] Cees G. M. Snoek et Marcel Worring. *Multimodal Video Indexing : A Review of the State-of-the-art.* Multimedia Tools Appl., vol. 25, no. 1, pages 5–35, 2005.

[Stan 2003] Daniela Stan et Ishwar K. Sethi. *eID : a system for exploration of image databases.* Inf. Process. Manage., vol. 39, no. 3, pages 335–361, 2003.

[Tahaghoghi 2002] Seyed M. M. Tahaghoghi, James A. Thom et Hugh E. Williams. *Multiple Example Queries in Content-Based Image Retrieval.* In SPIRE 2002 : Proceedings of the 9th International Symposium on String Processing and Information Retrieval, pages 227–240, London, UK, 2002. Springer-Verlag.

[Tamura 1978] H. Tamura, T. Mori et T. Yamawaki. *Textural Features Corresponding to Visual Perception.* Systems, Man and Cybernetics, IEEE Transactions on, vol. 8, pages 460–473, June 1978.

[Tao 2006] Dacheng Tao, Xiaoou Tang, Xuelong Li et Xindong Wu. *Asymmetric Bagging and Random Subspace for Support Vector Machines-Based Relevance Feedback in Image Retrieval.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 7, pages 1088–1099, 2006.

[Theobald 2002] Anja Theobald et Gerhard Weikum. *The Index-Based XXL Search Engine for Querying XML Data with Relevance Ranking.* In EDBT '02 : Proceedings of the 8th International Conference on Extending Database Technology, pages 477–495, London, UK, 2002. Springer-Verlag.

[Tian 2001] Qi Tian, Baback Moghaddam et Thomas S. Huang. *Display Optimization for Image Browsing.* In MDIC '01 : Proceedings of the Second International Workshop on Multimedia Databases and Image Communication, pages 167–178, London, UK, 2001. Springer-Verlag.

[Tieu 2004] Kinh Tieu et Paul Viola. *Boosting Image Retrieval.* International Journal of Computer Vision, vol. 56, pages 17–36, 2004.

[Tirilly 2008] Pierre Tirilly, Vincent Claveau et Patrick Gros. *Language modeling for bag-of-visual words image categorization.* In CIVR '08 : Proceedings of the 2008 international conference on Content-based image and video retrieval, pages 249–258, New York, NY, USA, 2008. ACM.

[Tollari 2008] Sabrina Tollari, Philippe Mulhem, Marin Ferecatu, Hervé Glotin, Marcin Detyniecki, Patrick Gallinari, Hichem Sahbi et Zhong-Qiu Zhao. *A Comparative Study of Diversity Methods for Hybrid Text and Image Retrieval Approaches.* In CLEF, pages 585–592, Berlin, Heidelberg, 2008. Springer-Verlag.

[Torres 2003] Ricardo S. Torres, Celmar G. Silva, Claudia B. Medeiros et Heloisa V. Rocha. *Visual structures for image browsing.* In CIKM '03 : Proceedings of the twelfth international conference on Information and knowledge management, pages 49–55, New York, NY, USA, 2003. ACM.

[Torres 2006] Ricardo Da Silva Torres et Alexandre Xavier Falc ao. *Content-Based Image Retrieval : Theory and Applications.* Revista de Informática Teórica e Aplicada, vol. 13, pages 161–185, 2006.

[Tuytelaars 2004] Tinne Tuytelaars et Luc Van Gool. *Matching Widely Separated Views Based on Affine Invariant Regions.* Int. J. Comput. Vision, vol. 59, no. 1, pages 61–85, 2004.

[Urban 2004] Jana Urban et Joemon M. Jose. *Evidence Combination for Multi-Point Query Learning in Content-Based Image Retrieval.* In ISMSE '04 : Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering, pages 583–586, Washington, DC, USA, 2004. IEEE Computer Society.

[Vasconcelos 2000] Nuno Vasconcelos et Andrew Lippman. *A Probabilistic Architecture for Content-Based Image Retrieval.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, page 1216, 2000.

[Vasconcelos 2004] N. Vasconcelos. *On the efficient evaluation of probabilistic similarity functions for image retrieval*, 2004.

[Veltkamp 2001] Remco C. Veltkamp et Michiel Hagedoorn. *State of the art in shape matching.* pages 87–119, 2001.

[Vidal-Naquet 2003] Michel Vidal-Naquet et Shimon Ullman. *Object Recognition with Informative Features and Linear Classification.* In ICCV '03 : Proceedings of the Ninth IEEE International Conference on Computer Vision, page 281, Washington, DC, USA, 2003. IEEE Computer Society.

[Villena-Román 2008] Julio Villena-Román, Sara Lana-Serrano et José González-Cristóbal. *MIRACLE at ImageCLEFmed 2007 : Merging Textual and Visual Strategies to Improve Medical Image Retrieval.* In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas Oard, Anselmo Penas, Vivien Petras et Diana Santos, editeurs, Advances in Multilingual and Multimodal Information Retrieval, volume 5152 de *Lecture Notes in Computer Science*, pages 593–596. Springer Berlin / Heidelberg, 2008.

[Vinokourov 2003] Alexei Vinokourov, David R. Hardoon et John Shawe-Taylor. *Learning the semantics of multimedia content with application to web image retrieval and classification.* In Fourth International Symposium on Independent Component Analysis and Blind Source Separation. Elsevier Publisher, Amsterdam, 2003.

[Vogel 2006] J. Vogel et B. Schiele. *On Performance Characterization and Optimization for Image Retrieval.* In Anders Heyden, Gunnar Sparr, Mads

Nielsen et Peter Johansen, editeurs, Computer Vision ECCV 2002, volume 2353 de *Lecture Notes in Computer Science*, pages 51–55. Springer Berlin / Heidelberg, 2006.

[Wang 2000] James Wang, Jia Li et Gio Wiederholdy. *SIMPLIcity : Semantics-sensitive Integrated Matching for Picture LIbraries.* In Robert Laurini, editeur, Advances in Visual Information Systems, volume 1929 de *Lecture Notes in Computer Science*, pages 171–193. Springer Berlin / Heidelberg, 2000.

[Wang 2004] Lei Wang, Li Liu et Latifur Khan. *Automatic image annotation and retrieval using subspace clustering algorithm.* In MMDB '04 : Proceedings of the 2nd ACM international workshop on Multimedia databases, pages 100–108, New York, NY, USA, 2004. ACM.

[Wang 2008] Xuanhui Wang, Hui Fang et ChengXiang Zhai. *A study of methods for negative relevance feedback.* In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '08, pages 219–226, New York, NY, USA, 2008. ACM.

[Ward 1963] Jr. Ward. *Hierarchical grouping to optimize an objective function.* Journal of the American Statistical Association, vol. 58, pages 236–244, 1963.

[Wenyin 2001] Liu Wenyin, Susan Dumais, Yanfeng Sun, Hongjiang Zhang, Mary Czerwinski et Brent Field. *Semi-Automatic Image Annotation.* In In INTERACT2001, 8th IFIP TC.13 Conference on Human-Computer Interaction, pages 326–333, 2001.

[Westerveld 2000] T. Westerveld. *Image retrieval : Content versus context.* In Content-Based Multimedia Information Access, RIAO, Paris, France, 2000.

[Westerveld 2003] Thijs Westerveld et Arjen P. de Vries. *Experimental result analysis for a generative probabilistic image retrieval model.* In SIGIR '03 : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pages 135–142, New York, NY, USA, 2003. ACM.

[Westerveld 2004] Thijs Westerveld et Arjen P. de Vries. *Multimedia Retrieval Using Multiple Examples.* In Image and Video Retrieval, volume 3115 de *Lecture Notes in Computer Science*, pages 2048–2049. Springer Berlin / Heidelberg, 2004.

[Wu 2004] Yimin Wu et Aidong Zhang. *Interactive pattern analysis for relevance feedback in multimedia information retrieval.* Multimedia Systems, vol. 10, pages 41–55, 2004.

[Yanai 2005] Keiji Yanai et Kobus Barnard. *Image region entropy : a measure of "visualness" of web images associated with one concept.* In Proceedings of

the 13th annual ACM international conference on Multimedia, MULTIMEDIA '05, pages 419–422, New York, NY, USA, 2005. ACM.

[Yang 2005] Changbo Yang, Ming Dong et Farshad Fotouhi. *Semantic feedback for interactive image retrieval.* In MULTIMEDIA '05 : Proceedings of the 13th annual ACM international conference on Multimedia, pages 415–418, New York, NY, USA, 2005. ACM.

[Yang 2007] Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann et Chong-Wah Ngo. *Evaluating bag-of-visual-words representations in scene classification.* In MIR '07 : Proceedings of the international workshop on Workshop on multimedia information retrieval, pages 197–206, New York, NY, USA, 2007. ACM.

[Yavlinsky 2005] Alexei Yavlinsky, Edward Schofield et Stefan Rüger. *Automated Image Annotation Using Global Features and Robust Nonparametric Density Estimation.* In Wee-Kheng Leow, Michael Lew, Tat-Seng Chua, Wei-Ying Ma, Lekha Chaisorn et Erwin Bakker, editeurs, Image and Video Retrieval, volume 3568 de *Lecture Notes in Computer Science*, pages 507–517. Springer Berlin / Heidelberg, 2005.

[Yee 2001] Ka-Ping Yee, Danyel Fisher, Rachna Dhamija et Marti Hearst. *Animated Exploration of Dynamic Graphs with Radial Layout.* In Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01), Washington, DC, USA, 2001. IEEE Computer Society.

[Yee 2003] Ka-Ping Yee, Kirsten Swearingen, Kevin Li et Marti Hearst. *Faceted metadata for image search and browsing.* In CHI '03 : Proceedings of the SIGCHI conference on Human factors in computing systems, pages 401–408, New York, NY, USA, 2003. ACM.

[Zhang 2005] Ruofei Zhang, Zhongfei (Mark) Zhang, Mingjing Li, Wei-Ying Ma et Hong-Jiang Zhang. *A Probabilistic Semantic Model for Image Annotation and Multi-Modal Image Retrieva.* In ICCV '05 : Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, pages 846–851, Washington, DC, USA, 2005. IEEE Computer Society.

[Zhao 2002] Rong Zhao et W. I. Grosky. *Narrowing the semantic gap - improved text-based web document retrieval using visual features.* IEEE Transactions on Multimedia, vol. 4, no. 2, pages 189–200, August 2002.