Contributions à l'analyse d'images par catégorisation pour la description et la reconnaissance

Muriel Visani

Contributions à l'analyse d'images par catégorisation pour la description et la reconnaissance

—

Vers l'exploitation interactive d'images

Habilitation à diriger des recherches

présentée devant

l'Université de La Rochelle

(Spécialité Informatique)

Par

Muriel VISANI

Soutenue publiquement le 02 décembre 2014 devant le jury :

Président :

Pr. Matthieu CORD	-	Université de Paris VI Pierre et Marie Curie - LIP6

Rapporteurs :

Pr. Philippe BOLON	-	Polytech Annecy-Chambéry - LISTIC
Pr. Rolf INGOLD	-	Université de Fribourg, Suisse - DIVA
Pr. Florence SÈDES	-	Université de Toulouse III Paul Sabatier - IRIT

Examinateurs :

Pr. Isabelle BLOCH	-	Télécom ParisTech (ENST) - LTCI
Pr. Jean-Michel JOLION	-	INSA de Lyon - LIRIS
Pr. Jean-Marc OGIER	-	Université de La Rochelle - L3i
Pr. Karl TOMBRE	-	Université de Lorraine - LORIA

Remerciements

En premier lieu, je tiens à remercier mes rapporteurs, Philippe Bolon, Professeur à Polytech Annecy-Chambéry, Rolf Ingold, Professeur à l'université de Fribourg et Florence Sèdes, Professeur à l'université de Toulouse III, pour le temps et l'attention qu'ils ont consacré à mon travail. Je remercie également Matthieu Cord, professeur à l'université Pierre et Marie Curie (Paris VI) pour avoir accepté de présider ce jury. Un grand merci également à Isabelle Bloch, professeur à Télécom ParisTech (ENST), à Jean-Michel Jolion, professeur à l'INSA de Lyon et à Karl Tombre, professeur à l'université de Lorraine, qui ont accepté de participer à ce jury.

Je suis enfin très reconnaissante à Jean-Marc Ogier, professeur à l'université de La Rochelle, de m'avoir toujours encouragée et soutenue depuis mon recrutement en tant que maître de conférences en 2006.

Les contributions et les résultats présentés dans ce mémoire sont avant tout le fruit d'un travail collectif, notamment au travers des thèses que j'ai co-encadrées. Je souhaite donc remercier Lai Hien Phuong, Sophea Prum, Stéphanie Guillas, Nathalie Girard, Kieu Van Cuong et Bui Quang Anh et pour les moments d'échanges et de réflexion scientifiques partagés dans le cadre de leurs travaux de thèse.

Je tiens également à remercier les chercheurs, en France et à l'international, avec lesquels j'ai co-encadré ces doctorants et des stagiaires de master 2, ou bien qui m'ont fait confiance notamment au travers de projets menés en collaboration : Karell Bertet, Jean-Marc Ogier, Rémy Mullot, Arnaud Revel et Michel Ménard de l'université de La Rochelle (L3i), mais aussi Nicholas Journet et Jean-Philippe Domenger de l'université de Bordeaux (LaBRI), Thierry Urruty de l'université de Poitiers (XLIM), Antoine Tabbone de l'université de Nancy (LORIA), Josep Lladós, Alicia Fornés et Oriol Ramos-Terrades de l'Université Autonome de Barcelone (CVC), Andreas Fischer de l'université de Concordia à Montréal (CENPARMI), Tran Cao Dê et Huynh Xuan Hiep de l'université de Can Tho (Vietnam), Nguyen Dung Duc et Luong Chi Mai de la Vietnamese Academy of Science and Technology (IOIT), et enfin Alain Boucher, Alexis Drogoul et Jean-Daniel Zucker (IRD) de l'équipe MSI de l'UMI UMMISCO à Hanoï.

Merci encore et surtout à tous les membres du laboratoire L3i : stagiaires, doctorants, ingénieurs, assistantes, enseignants-chercheurs, avec lesquels j'ai eu le plaisir de travailler depuis 2006.

Enfin, je tiens à remercier mes proches, amis et parents, de leurs encouragements et de leur patience. Un merci tout particulier à mon père pour sa relecture détaillée de ce manuscrit.

Résumé

Ce document présente une synthèse de mes principaux travaux de recherche depuis une dizaine d'années. Ces travaux portent sur l'analyse d'images, et plus précisément sur la description et la reconnaissance d'images de niveaux de structuration variables (images de scènes naturelles, images de documents, etc.). En l'absence de connaissance du domaine formalisée explicitement, les approches que j'ai privilégiées reposent sur une catégorisation des images, ou des objets/motifs contenus dans les images. Pour cela, j'ai engagé et encadré des travaux de recherche dans le contexte de différentes thèses, souvent en collaboration avec d'autres chercheurs en France et à l'international. Ces travaux ont abouti à des contributions d'ordre méthodologique et applicatif, valorisées d'un point de vue scientifique et au travers de cas d'étude concrets, souvent en lien avec une activité contractuelle.

Dans le cas de la description, la catégorisation doit généralement être menée sans connaître *a priori* les catégories perçues par un humain dans la collection. Nous avons donc choisi, après avoir confié à la machine la tâche de découvrir des groupes d'images similaires selon leur description de bas niveau sémantique, de faire interagir l'utilisateur avec ces groupes d'images, de manière à guider un rapprochement de ces groupes vers les concepts de plus haut niveau perçus par l'utilisateur. Dans le cas de la reconnaissance, nous nous sommes focalisés sur la recherche d'approches adéquates pour intégrer dans le processus l'ensemble des informations disponibles (ces dernières étant parfois de natures hétérogènes et, dans certains cas, très partielles), et pour restituer les informations apprises à l'utilisateur d'une manière qui soit intelligible pour ce dernier.

D'un point de vue très général, on peut considérer que l'ensemble de ces travaux vise à concevoir des outils pouvant servir *in fine* à indexer des images en vue d'une exploitation ultérieure par l'humain. Forte de l'expérience acquise au travers de ces travaux, la direction de recherche que je souhaite privilégier dans les années à venir consiste à me pencher sur les différents moyens pour amener la machine à assister un utilisateur humain dans l'exploitation de corpus d'images de contenus variés, en accordant un rôle-clé à l'utilisateur final du système. Afin d'engager ces recherches, j'ai d'ores et déjà obtenu le financement de projets de recherche et/ou de thèses qui débuteront prochainement.

Table des matières

Chapitre 1

Introduction générale

Sommaire

1.1 Avant-propos : présentation du domaine de recherche de l'analyse d'images

1.1.1 Contexte scientifique

Depuis une vingtaine d'années, le volume d'images disponibles pour le grand public a été accru de manière exponentielle en raison notamment de la démocratisation d'appareils d'acquisition numériques et de l'essor d'internet. On peut citer comme exemple le partage intensif de photographies personnelles sur des sites *web* tels que Flickr ou Facebook : en 2014, en une minute ce sont en moyenne 3000 photographies qui sont transférées et 20 millions de photographies qui sont visualisées vers/depuis Flickr. D'autres services, comme par exemple Google Street View, participent à l'accès de tout un chacun à d'immenses bases d'images. Et ce phénomène est toujours en pleine expansion. On estime qu'en 2015 le nombre d'appareils connectés à internet atteindra le double de la population globale mondiale et selon Google, en 2016, 96%

du volume de données transitant par internet sera sous forme pixellique (images ou vidéos). Cela explique en partie le grandissant intérêt des chercheurs et des entreprises pour des outils relevant des domaines de la **vision par ordinateur** [Chen 2009] et de la **recherche d'information multimedia** au sens large [Sèdes 2002, Lew 2006], outils permettant respectivement d'acquérir, de traiter, d'analyser, d'interpréter automatiquement ou semi-automatiquement les images, et de naviguer/rechercher efficacement dans des masses d'images.

Les techniques d'**analyse d'images** [Yoo 2004, Russ 2010] qui consistent à extraire, à partir du contenu généralement peu structuré des images (éventuellement additionné de quelques informations ou connaissances supplémentaires), une information d'un niveau sémantique intermédiaire, sont très souvent mises à contribution. En fonction de l'objectif visé et de la nature des images, les applications relevant de l'analyse d'images couvrent un large spectre. Ce dernier s'étend par exemple de la segmentation d'images de scènes naturelles en régions homogènes à la reconnaissance d'écriture dans des images de manuscrits anciens.

Évidemment, les systèmes d'analyse d'images nécessitent souvent des phases de traitement d'images. Néanmoins, l'analyse d'images diffère du traitement d'images : si tous deux prennent en entrée des images, l'analyse renvoie de l'information extraite de ces images, tandis que le traitement renvoie les images traitées. À la différence des méthodes d'interprétation d'images qui renvoient de l'information d'un niveau sémantique élevé, les techniques d'analyse retournent le plus souvent une information d'un niveau sémantique intermédiaire. Ainsi, l'interprétation d'images peut aller jusqu'à la compréhension de scène, pouvant résulter en des assertions du type « Un homme portant un pull-over rouge est en train de promener son chien dans un environnement arboré par beau temps », tandis que l'analyse d'image se cantonne à décrire certaines propriétés de l'image (« Les couleurs dominantes dans l'image sont le vert, le bleu et le rouge »), voire à détecter/reconnaître certains objets présents dans une image (« Cette image contient un homme, un chien, et six arbres »). La frontière entre les processus de traitement, d'analyse et d'interprétation d'images est cependant parfois floue, et la terminologie employée par les chercheurs dépend beaucoup de leur culture scientifique.

On peut considérer que l'analyse et l'interprétation d'images cherchent à imiter le fonctionnement du cortex visuel humain, qui excelle dans le fait d'extraire, à partir d'un simple stimulus visuel, des informations de plus haut niveau sémantique. Elles reposent sur des algorithmes complexes, relevant souvent de la modélisation et de l'**apprentissage artificiel** (*machine learning*) [Alpaydin 2004, Cord 2008], sous-discipline de l'intelligence artificielle [Bloch 2014] à la croisée des chemins avec la statistique [Vapnik 1998] et la fouille de données [Witten 2011].

Le contenu des images est souvent analysé au travers de la reconnaissance d'objets, de motifs ou de structures spécifiques dans les images. L'analyse d'images fait donc naturellement souvent appel à des techniques de **reconnaissance de formes** [Bishop 2006]. Certaines des approches utilisées en analyse d'images sont inspirées de méthodes qui ont fait leurs preuves dans des disciplines connexes telles que la recherche d'information textuelle ou encore l'analyse de la parole.

Durant la dernière décennie, en parallèle de l'accroissement exponentiel du volume d'images disponibles, de nouveaux dispositifs permettant aux humains d'interagir de manière intuitive avec les contenus numériques se sont répandus auprès du grand public. De nombreux travaux ont émergé, visant à intégrer l'humain dans l'analyse d'images. Du fait notamment de la popularisation de ces nouveaux dispositifs et de l'émergence de tels travaux, l'analyse d'images s'est progressivement liée à des problématiques de visualisation de données [Chen 2007] et d'**interaction homme-machine** [Xie 2013].

1.1.2 Diffusion scientifique

En raison notamment de la trans-disciplinarité évoquée ci-dessus, la diffusion scientifique des travaux relevant de l'analyse d'images se fait au travers de nombreux journaux couvrant un très large spectre thématique, dont je propose ci-après une liste organisée par thématiques. Du fait de la multiplicité des thématiques concernées, il ne s'agit pas ici d'en présenter un inventaire exhaustif. Dans cette liste, je mets particulièrement l'accent sur les thématiques les plus en lien avec mes activités de recherche.

Parmi les **revues internationales** concernées, on retrouve évidemment des journaux spécialisés dans la vision par ordinateur : *International Journal of Computer Vision*, *Foundations and Trends in Computer Graphics and Vision*, *Computer Vision and Image Understanding*, *IEEE Transactions on Image Processing*, etc. On trouve également de nombreux travaux relevant de l'analyse d'images dans des revues traitant de l'apprentissage artificiel et de la reconnaissance de formes : *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *Foundations and Trends in Machine Learning*, *Machine Learning*, *Pattern Recognition*, *Pattern Recognition Letters*, *Pattern Analysis and Applications*, *International Journal of Pattern Recognition and Artificial Intelligence*, etc. Des articles relevant de l'analyse d'image sont également publiés dans des journaux plus proches du domaine de l'interaction homme-machine : *IEEE Transactions on Systems, Man, and Cybernetics*, *IEEE Transactions on Affective Computing*, etc. On note depuis plusieurs décennies l'émergence de revues permettant aux chercheurs de diffuser des travaux concernant des images d'un type spécifique, comme par exemple *IEEE Transactions on Medical Imaging*, dont la première édition a été publiée en 1982, ou encore l'*International Journal on Document Analysis and Recognition*, paru pour la première fois en 1998.

Parmi les **revues francophones** concernées par des publications en analyse d'images, on peut citer par exemple Traitement du Signal (TS) ou les Revues des Sciences et Technologies de l'Information (RSTI) (en particulier la revue Document Numérique, la Revue d'Intelligence Artificielle, ou encore la Revue de Technique et Science Informatiques).

La communauté scientifique internationale et des industriels en quête de transfert technologique se retrouvent régulièrement lors de **conférences** reconnues telles que *international conference on Computer Vision and Pattern Recognition* (CVPR), *International Conference on Computer Vision* (ICCV), *European Conference on Computer Vision* (ECCV), *International Conference on Pattern Recognition* (ICPR), *International Conference on Machine Learning* (ICML), etc. Récemment, on a assisté à l'émergence de conférences dédiées à un type spécifique d'applications, comme par exemple la *IEEE International Conference on Automatic Face and Gesture Recognition* (FGR), l'*International Conference on Medical Image Computing and Computer Assisted Intervention* (MICCAI), l'*International Conference on Document Analysis and Recognition* (ICDAR) ou encore l'*International Conference on Frontiers in Handwriting Recognition* (ICFHR). Ces conférences, souvent associées à des événements satellites encore plus spécialisés (comme par exemple les *workshops Graphic RECognition* (GREC) ou *Historical Image Processing* (HIP) organisés en marge d'ICDAR), permettent aux spécialistes concernés de se retrouver régulièrement afin d'échanger sur les dernières nouveautés du domaine. Certains *workshops* très spécialisés se tiennent même indépendamment de plus gros événements, et jouissent néanmoins d'une bonne renommée, comme par exemple l'*international workshop on Document Analysis Systems* (DAS) [1]. On retrouve également des communications relevant

1. Selon les années, DAS (qui vient de connaître en 2014 sa onzième édition) est noté A ou B dans le

de l'analyse d'images dans certaines conférences ou *workshops* internationaux focalisés sur un type particulier d'approches, comme par exemple l'*International Joint Conference on Neural Networks* (IJCNN), *Concept Lattices and Applications* (CLA) ou *Graph-based Representations in Pattern Recognition* (GbR).

Au niveau francophone, on peut citer les congrès du Groupement de Recherche en Traitement du Signal et des Images (GRETSI), les conférences Reconnaissance de Formes et Intelligence Artificielle (RFIA), les rencontres de la Société Francophone de Classification (SFC), les journées francophones des jeunes chercheurs en vision par ordinateur (ORASIS), l'école d'été en traitement du signal et des images (PEYRESQ), ou encore des congrès plus focalisés sur des applications particulières très diverses, allant par exemple du Colloque International Francophone sur l'Écrit et le Document (CIFED) au colloque francophone de visualisation et de traitement d'images en mécanique des fluides (FLUVISU).

1.1.3 Communauté scientifique

En France, l'Association Française pour la Reconnaissance et l'Interprétation des Formes (**AFRIF**) est une association à but non lucratif qui œuvre pour soutenir la recherche dans ces domaines, en France comme au niveau international. Cette association est fédérée au sein de la fédération des Associations françaises des Sciences et Technologies de l'Information (ASTI) avec d'autres associations ayant des thématiques connexes, telles que l'Association Française pour l'Intelligence Artificielle (**AFIA**), en collaboration avec laquelle l'AFRIF organise la conférence francophone RFIA tous les deux ans, ou encore le Groupement de Recherche en Traitement du Signal et des Images (**GRETSI**), sous l'égide duquel est publiée la revue francophone Traitement du Signal. Au sein de l'ASTI, on retrouve également des groupes de recherche touchant une communauté plus restreinte, tels que le Groupe de Recherche en Communication Écrite (**GRCE**), qui co-organise tous les deux ans avec l'Association francophone de Recherche d'Information et Applications (**ARIA**) leurs conférences respectives CIFED et CORIA lors de la Semaine du Document Numérique et de la Recherche d'Information. La communauté scientifique française se retrouve également autour de deux groupes de recherche d'animation de l'INS2I du CNRS : le **GdR I3** (Information-Interaction-Intelligence), et le **GdR ISIS** (Information, Signal, Image et viSion).

Au niveau international, la communauté scientifique est fédérée en particulier par l'**IAPR** (*International Association for Pattern Recognition*) – dont l'AFRIF est la branche française – et ses différents comités techniques. Comme on a pu le voir en section précédente, beaucoup de revues et de conférences du domaine sont également soutenues par l'association IEEE (*Institute of Electrical and Electronics Engineers*) et ses différentes sociétés.

classement Core : http://core.edu.au/index.php/conference-rankings.

1.2 Positionnement de mes travaux de recherche

1.2.1 Positionnement dans la communauté scientifique

Le positionnement de mes activités de recherche dans la communauté scientifique peut être caractérisé au niveau local, national et international par mes responsabilités dans les différentes instances d'animation de la recherche et mes publications.

Au niveau local du laboratoire L3i, j'ai été co-responsable dès 2007 des équipes[2] de recherche auxquelles j'ai appartenu (IMEDOC puis IDDC). IDDC a été évaluée A par l'AERES en 2011 ; en 2013 elle comptait 5 professeurs et 11 maîtres de conférences. Comme expliqué plus tard au travers de mon cheminement scientifique, je mène mon travail de recherche en collaboration avec plusieurs des professeurs et maîtres de conférences du L3i (ce qui a donné lieu à des co-publications et/ou co-encadrements avec 3 PR et 1 MCF du laboratoire). Mes activités contribuent aux coopérations nationales et internationales du L3i.

Au niveau national, j'appartiens à la communauté scientifique structurée autour de la reconnaissance des formes et de la perception et à celle, plus restreinte, de l'analyse et de l'interprétation d'images de documents. Ainsi, je suis notamment membre des conseils d'administration de l'AFRIF depuis Mars 2010 et du GRCE depuis Novembre 2012. Je participe – ou ai participé – à plusieurs projets (dont deux financés par l'ANR), en collaboration avec plusieurs laboratoires renommés dans mon domaine d'expertise, comme par exemple le LaBRI (Bordeaux), le LORIA (Nancy), le LITIS (Rouen), le XLIM/SIC (Poitiers), etc. J'ai pris part à plusieurs reprises à des journées organisées par le GdR ISIS, qui finance par ailleurs le projet exploratoire CINÉDI dont je suis porteuse. Je participe très régulièrement à la conférence RFIA, durant laquelle j'ai revêtu à deux occasions le rôle de *chairman*, aux conférences CIFED-CORIA et j'ai publié des articles dans les revues Traitement du Signal et la Revue de Technique et Science Informatiques.

Au niveau international, je publie très régulièrement dans les conférences ICPR et ICDAR[3]. Concernant les revues internationales, certains de mes articles sont parus dans des journaux aux spectres thématiques assez larges (tels que *Pattern Recognition Letters, IEEE Transactions on Systems, Man, and Cybernetics, Pattern Analysis and Applications* et *International Journal of Pattern Recognition and Artificial Intelligence*), voire très large (*Fundamenta Informaticae*). J'ai également publié dans des journaux aux thématiques plus ciblées, comme l'*International Journal on Document Analysis and Recognition.*

1.2.2 Objectifs généraux visés

Mes travaux de recherche se placent dans le domaine de l'analyse du contenu de collections d'images. Le processus d'analyse permet d'extraire, généralement à partir du contenu pixellaire des images de la collection voire d'informations/connaissances additionnelles, une information de niveau sémantique intermédiaire. Cette information peut être extraite dans un but exploratoire ou inférentiel. Dans un but exploratoire, on peut par exemple chercher à obtenir un aperçu des données (souvent volumineuses) voire à en dégager certaines tendances, à détecter

2. Terminologie employée par l'AERES dans son rapport de 2011 sur le L3i (le terme de « projets scientifiques »étant préféré en interne).

3. Avec respectivement 4 et 6 communications depuis 2010.

des anomalies ou à formuler des hypothèses sur les images. Je m'intéresse en particulier à l'objectif de description qui consiste, selon les cas, à rechercher une représentation simplifiée d'une image en particulier, ou de la collection dans son ensemble. Dans un but inférentiel en revanche, on cherche à inférer certaines propriétés d'une image, en fonction de ses autres propriétés (observées) et, le cas échéant, d'informations/connaissances additionnelles ; les objectifs possibles sont multiples, mais je m'intéresse particulièrement dans mes recherches à l'objectif de reconnaissance. La suite de cette section présente les deux objectifs généraux vers lesquels tendent mes efforts de recherche, à savoir la description et la reconnaissance d'images.

Dans un **objectif de description d'images**, nous cherchons à construire une représentation simplifiée qui permette de synthétiser le contenu d'images ou de collections d'images. Cette représentation simplifiée peut être véhiculée sous une forme statistique ou structurelle [4] par le biais de « descripteurs » [5] d'images. Les descripteurs peuvent décrire une région de l'image, une image en particulier, ou plus globalement une collection d'images. Ils peuvent être de niveaux sémantiques variés, selon les processus employés pour les obtenir et l'information dont on dispose en entrée. À un bas niveau sémantique, on cherchera à décrire une région de l'image par exemple par sa texture ou sa forme, tandis que la représentation d'une image pourra, entre autres, être basée sur une description des régions qui la composent (et le cas échéant de leur agencement spatial) ou bien sur une représentation plus globale. Pour la description d'une collection d'images, on pourra par exemple s'appuyer sur des similarités entre les images de la collection pour en proposer une représentation simplifiée. Des descripteurs de plus haut niveau sémantique peuvent être obtenus grâce à des informations ou des connaissances collectées le plus souvent auprès (ou sous la supervision) d'un humain, et éventuellement de la représentation de bas niveau sémantique.

Dans un **objectif de reconnaissance d'images**, on cherche généralement à inférer des propriétés de plus haut niveau sémantique. Cette inférence se fait souvent à partir de descriptions de bas niveau sémantique des images et, éventuellement, d'informations additionnelles voire de connaissances souvent spécifiques au domaine d'application (imagerie médicale, image de document, etc.). Les propriétés inférées portent le plus souvent sur les types et/ou la structure d'éléments d'intérêt (objets/motifs) composant les images. Tiré vers le haut par de très nombreuses applications, le domaine de la reconnaissance d'images a connu de grandes avancées durant les dernières décennies. Malgré cela, les résultats obtenus sont, à l'heure actuelle, loin d'être parfaits. Ceci peut être expliqué en partie par l'insuffisance des outils de représentation des informations et/ou des connaissances sur lesquelles elle se base. En outre, la multiplicité des tâches et des applications visées, chacune avec ses fortes spécificités et ses situations exceptionnelles, rend difficile une formalisation unifiée du problème. Ces difficultés conceptuelles s'accompagnent d'obstacles techniques liés en particulier à la qualité et à la résolution des images, aux conditions d'acquisition plus ou moins contrôlées et, le cas échéant, à un manque de données.

D'un point de vue scientifique, les deux objectifs de description et de reconnaissance d'images se rejoignent (de manière non exhaustive) sur les thématiques de la modélisation de l'information, de l'extraction et de la sélection d'indices visuels caractéristiques du contenu de l'image, de l'apprentissage artificiel et de la combinaison/fusion d'informations. Lorsque l'humain est

4. Voir annexe A.
5. *Cf.* section 2.1 pour une définition des notions de descripteurs, de caractéristiques et de signatures d'images.

impliqué, peuvent se poser des problèmes liés à la visualisation de données et aux interfaces homme-machine.

D'un point de vue applicatif, la description et la reconnaissance d'images sont très souvent entremêlées ; les modalités possibles pour la collaboration entre modules de description et de reconnaissance sont multiples. Ainsi, les descripteurs extraits des images sont très fréquemment utilisés pour alimenter un processus de reconnaissance d'images. C'est la chaîne de traitement traditionnelle, enchaînant description puis reconnaissance. Plusieurs approches permettent également de mener à bien conjointement la description et la reconnaissance ; les descripteurs sont alors souvent calculés de manière à optimiser les performances du moteur de reconnaissance. Inversement, il arrive qu'un objet/motif particulier soit détecté ou reconnu dans une image, et que cette information soit utilisée pour la description de cette image, soit directement en utilisant la présence ou l'absence de cet objet comme descripteur de cette image, soit pour déclencher le calcul de descripteurs spécifiques de ce type d'objets. De plus, l'ensemble des informations extraites des images (que ce soit au terme d'un processus à visée plutôt descriptive ou de reconnaissance) peuvent conjointement être utilisées par un même processus, par exemple en vue d'indexer les images par leur contenu.

1.3 Problématique de recherche

1.3.1 Problématique de la catégorisation d'images

Les processus d'analyse d'images s'appuient fréquemment sur des groupes d'images. Si l'on se place dans un objectif descriptif où l'on cherche à donner à l'humain un aperçu rapide d'une collection d'images qui est trop volumineuse pour qu'il puisse la visualiser rapidement dans son intégralité, on pourra par exemple la lui présenter sous la forme de quelques images représentatives de chacun des groupes d'images similaires dans la collection. La notion de similarité est ici à entendre au sens large, et peut être définie à des niveaux de sémantique divers. Par exemple, on peut considérer que deux images sont similaires dès lors que leurs textures et/ou couleurs dominantes sont proches, ou bien dès lors qu'elles représentent un même concept pour l'homme. Une application de reconnaissance d'images peut (entre autres) reposer sur l'identification de leur(s) groupe(s) d'appartenance, ou sur la détection d'instances de groupes pré-établis dans la collection d'images.

Quel que soit l'objectif visé, le procédé qui consiste à affecter une image (ou un objet/motif présent dans cette image) à un groupe est appelé catégorisation, les groupes étant alors appelés catégories. C'est à cette problématique que je m'intéresse principalement dans le cadre de mes travaux.

Dans le cas de la catégorisation d'objets ou de motifs, que ceux-ci soient préalablement détectés/localisés ou extraits par nos soins, on se ramène dans tous les cas à une catégorisation des imagettes contenant ces objets/motifs. Pour plus de simplicité dans le propos, l'expression **« catégorisation d'images » est une expression à entendre dans un sens large**, qui couvre dans la suite de ce manuscrit la catégorisation des images entières ou des imagettes contenant les objets/motifs présents dans les images.

L'information extraite à l'issue de la catégorisation des images pourra *in fine* être retournée à un humain, soit pour information, soit pour l'assister dans une prise de décision (cas de la biométrie dans des applications de vidéo-surveillance par exemple). Cet humain, selon les cas, peut être un expert du domaine d'application (expert judiciaire, photographe, archiviste, etc.), un chercheur en vision par ordinateur, ou encore un simple utilisateur du système (non-expert).

Cette information pourra également être retournée à la machine afin d'alimenter un processus ultérieur. Par exemple, l'information de catégorisation pourra être utilisée pour indexer les images en vue d'une recherche ultérieure [Smeulders 2000, Fournier 2001, Datta 2008]. Dans un objectif purement descriptif, les catégories d'images pourront être mises à profit – souvent conjointement avec des indices visuels de ces images – par le biais d'un moteur de visualisation/navigation par similarité dans les bases d'images [Ma 1999, Borth 2008, Jing 2012]. L'information de catégorisation extraite grâce à un processus de reconnaissance pourra être utilisée par la machine pour alimenter des processus de reconnaissance plus élaborés, voire d'interprétation d'images. Ainsi, à l'issue d'une première phase de catégorisation pour détecter un objet dans une image (p. ex. un oiseau), on peut chercher à appliquer une catégorisation de grain plus fin (p. ex. l'espèce de l'oiseau préalablement détecté) ; on parle dans ce dernier cas de « reconnaissance de grain fin ».

L'objectif visé et la destination de l'information de sortie de la catégorisation conditionnent en particulier le type et le niveau de structuration désiré pour cette information, et influe sur les méthodes de catégorisation à privilégier, comme expliqué dans la section suivante.

1.3.1.1 Types de méthodes de catégorisation

Lorsque l'on cherche à mettre en œuvre une procédure de catégorisation d'images, on dispose traditionnellement en entrée d'un ensemble d'indices visuels (caractéristiques) extraits des images, éventuellement additionné d'informations concernant leurs catégories [6]. Les techniques utilisées pour la catégorisation sont donc typiquement basées sur un apprentissage, et largement utilisées – bien au-delà du cadre des images – pour l'analyse de données. L'apprentissage se fait dans un **espace de représentation** (vectoriel ou graphique), souvent défini grâce aux caractéristiques, et muni d'une mesure de similarité. Dans tous les cas, le choix de l'espace de représentation (qui peut être fait soit de manière explicite, soit de manière implicite par l'usage d'un noyau par exemple) influe énormément sur la solution de catégorisation produite en sortie. Le type de méthodes à mettre en œuvre est très largement conditionné par l'objectif visé au final, les spécificités du problème ainsi que l'information disponible en entrée et celle désirée en sortie, comme expliqué ci-après.

Dans un objectif de description, on dispose en entrée d'un ensemble d'images (ou imagettes) à catégoriser. Le plus souvent, on ne dispose d'aucune information concernant les groupes d'images recherchés, ou bien d'une information très partielle concernant ces groupes (par exemple sous la forme de quelques exemples d'images appartenant à un même groupe, ou bien de quelques images appartenant à des groupes différents). La sémantique éventuellement associée à ces groupes est le plus souvent inconnue. Les techniques mises en œuvre sont typiquement des méthodes de « ***clustering*** [7] » [Jain 2010]. Elles permettent de produire des groupes d'images appelés *clusters*. Si l'on ne dispose en entrée d'aucune information concernant les *clusters*, on parle de *clustering* non supervisé, tandis que si l'on dispose d'une information partielle concernant les *clusters* (voir ci-avant), alors cette information peut (dans une certaine mesure) guider l'apprentissage et l'on parle alors de *clustering* semi-supervisé.

Dans un objectif de reconnaissance, on dispose généralement d'une information plus complète concernant les catégories d'images recherchées. On parle alors de classes (à différencier des *clusters* évoqués ci-avant). Ces classes sont le plus souvent définies par (ou sous la supervision de) l'humain, ce qui leur confère un niveau sémantique plus élevé que les *clusters*. L'information disponible se présente généralement sous la forme d'un ensemble d'exemples d'images étiquetées par leurs classes d'appartenance (vérité-terrain). À la différence de l'objectif de description où l'on cherche à découvrir des *clusters* dans la collection à décrire, l'objectif est ici avant tout d'inférer, à partir des exemples étiquetés, la (les) classe(s) d'appartenance de nouvelles images (non encore étiquetées) qui seraient présentées au système. Selon l'exhaustivité de l'information de classe disponible à propos de la base d'apprentissage, les méthodes mises en œuvre sont typiquement basées sur des techniques de **classification** [Duda 2012] semi-supervisées ou supervisées. À noter que ces techniques sont désignées par le terme de « techniques de classement » dans la communauté statistique. Dans la suite, nous utiliserons la terminologie informatique, et l'outil servant à la classification sera désigné par le terme de « classifieur ».

Le foisonnement des méthodes de catégorisation (qu'il s'agisse de *clustering* ou de classification) dans la littérature rend difficile la présentation synthétique d'une taxinomie exhaustive. Très grossièrement, les méthodes de catégorisation peuvent être rangées selon deux typologies : suivant la fonction de coût minimisée lors de l'apprentissage, on distingue les méthodes généra-

6. Notons que, parfois, on dispose d'informations, voire de connaissances, du domaine (*cf.* section 1.3.1.2).

7. À noter que ces techniques sont communément désignées par le terme de « classification » dans la communauté statistique. Ce dernier terme pouvant être source d'ambiguïtés avec la terminologie informatique (voir ci-après), je lui préfèrerai dans la suite de ce mémoire le terme anglophone de *clustering*.

tives des méthodes discriminatives, tandis que selon la manière dont l'information apprise est restituée, on distingue les méthodes symboliques des méthodes numériques.

La différence entre les méthodes génératives et discriminatives peut être résumée ainsi : tandis que les méthodes génératives cherchent avant tout à modéliser les données à l'intérieur des catégories, les méthodes discriminatives, elles, cherchent à modéliser les frontières entre ces catégories. Si l'usage de méthodes discriminatives pour la classification semble assez naturelle [Bouchard 2005], dans le cas du *clustering* il requiert généralement la présence d'informations partielles sous la forme de quelques exemples appartenant à des catégories différentes. Une distinction stricte entre ces deux types de méthodes est néanmoins un peu dépassée aujourd'hui, puisque l'une des tendances actuelles consiste à les mêler (voir annexe D).

Les approches symboliques, issues originellement du domaine de l'intelligence artificielle, sont en général non paramétriques et permettent de représenter l'information apprise depuis les données d'entrée sous la forme d'expressions (en général propositionnelles). Cela se fait souvent au prix d'une discrétisation des données numériques en données symboliques qui peut engendrer une perte d'information difficilement réversible. À l'inverse, les méthodes numériques (issues du domaine des statistiques ou de l'intelligence artificielle), qui ne nécessitent généralement pas de discrétisation des données numériques, peuvent être paramétriques ou non-paramétriques et ont souvent un aspect « boîte noire ». Elles sont souvent entraînées par le biais de méthodes d'optimisation numérique (p. ex. pour l'ajustement des poids synaptiques des réseaux de neurones ou, dans le cas des SVM, la recherche des frontières entre catégories).

Selon le niveau de structuration désiré pour l'information de sortie et les méthodes employées, le résultat de la catégorisation peut être présenté « à plat » ou sous une forme hiérarchique (c'est-à-dire où les catégories se déclinent en sous-catégories). Les techniques floues permettent d'obtenir en sortie des degrés d'appartenance aux catégories, tandis que d'autres techniques permettent d'obtenir des probabilités d'appartenance, ou encore ou des scores parfois plus difficilement interprétables. Assez récemment, les techniques multi-étiquettes [Tsoumakas 2010], où une même image peut appartenir à plusieurs catégories, ont attiré une certaine attention de la part des chercheurs, motivés par de nouvelles applications telles que l'annotation automatique d'images [Zhang 2012].

Chacun de ces grands types d'approches se décline en une multiplicité de méthodes de *clustering* ou de classification, dont chacune peut être implémentée selon divers algorithmes (et souvent avec divers paramétrages). Chacune de ces méthodes renvoie un résultat de catégorisation différent, dont il est parfois difficile d'évaluer la qualité (particulièrement dans le cas du *clustering*, à visée exploratoire). Cela rend difficile le choix *a priori* de l'algorithme à appliquer dans un contexte donné, d'autant plus qu'il arrive que différentes méthodes produisent des solutions de catégorisation toutes intéressantes, mais localement différentes, voire complémentaires. Dès lors que le problème de catégorisation est trop complexe pour être résolu à l'aide d'un unique algorithme, une tendance consiste à **combiner/fusionner un ensemble de techniques de catégorisation** (possiblement obtenues par différents algorithmes et/ou avec différentes valeurs de paramètres, dans des espaces de représentation pouvant être différents). L'idée générale est d'obtenir en sortie une unique solution de catégorisation, tirant avantage des spécificités de chacune des méthodes utilisées et qui, dans certains cas, peut même surpasser la meilleure des méthodes considérée individuellement (*cf.* [Strehl 2003, Ghosh 2011] pour le *clustering* et [Kittler 1998, Lam 2000] pour la classification).

1.3.1.2 Images considérées

Les travaux décrits dans ce manuscrit concernent des images en niveaux de gris ou en couleurs. Plus particulièrement, les images à analyser sont, soit des images de scènes naturelles (« images naturelles »), soit des images de documents « symboliques », c'est-à-dire contenant du texte et/ou du graphique [Trupin 2005] (« images de documents »). Les images que nous considérons sont acquises soit à partir d'un appareil photographique, soit à partir d'un scanner, soit reconstruites à partir du signal en-ligne donné par un matériel d'acquisition de type tablette électromagnétique [8].

Selon les collections d'images considérées et les conditions de l'étude, on dispose de plus ou moins d'informations à propos de ce que représentent ces images (en plus des informations extraites automatiquement concernant leur contenu visuel). En grande partie, il s'agit d'**informations du domaine** concernant le contenu des images ou leurs catégories.

Outre les éventuelles métadonnées associées aux images et dont la nature dépend du domaine d'application, une part des informations du domaine concernant les images provient de leur éventuelle structuration. Les images sont composées simplement de tableaux de valeurs de pixels associées à quelques métadonnées (résolution, taille, espace couleur utilisé, etc.) et sont, par conséquent, généralement peu structurées. Les bases d'images diffèrent donc fondamentalement des bases de texte par exemple, où les mots ont déjà été structurés logiquement par l'auteur. Cependant, il existe souvent une **structuration sous-jacente à l'image à analyser**, dont le degré varie en fonction de son contenu. Par exemple, on comprend aisément que les images de documents symboliques (produits par l'homme et pour l'homme) sont en général plus structurées que des images de scènes naturelles. C'est par exemple le cas des symboles graphiques, généralement composés de primitives appartenant à un lexique fermé (segments de droite, arcs de cercle, etc.) dont les paramètres d'apparence et l'agencement spatial sont spécifiques à chaque symbole. Certaines images naturelles peuvent néanmoins être qualifiées par leur niveau de structuration. Par exemple, les images de visages peuvent être considérées comme relativement structurées, dans le sens où les visages sont tous composés de caractéristiques faciales (coin des yeux, iris, nez, bouche, etc.), dont l'apparence varie selon l'individu, mais qui sont agencées spatialement de manière similaire quel que soit l'individu.

Dans certains cas d'étude bien spécifiques, la structure sous-jacente est si bien connue qu'il est possible (le cas échéant avec l'aide d'un expert) de la formaliser sous la forme de connaissance du domaine, par exemple par le biais d'un atlas, d'une nomenclature ou d'une ontologie. Ces connaissances sont alors souvent organisées sous une forme hiérarchique permettant de décrire les liens entre éléments d'intérêt de l'image. C'est entre autres souvent le cas en imagerie médicale, où l'on peut adjoindre à chaque élément d'intérêt extrait de l'image à analyser des informations fonctionnelles et/ou pathologiques qui peuvent aider à guider le processus d'analyse (voir par exemple l'approche de segmentation/reconnaissance d'images de cerveaux présentée dans [Fouquier 2012]). Dans le cas d'une image de document, cette connaissance peut permettre de guider des processus d'analyse de sa structure physique et de sa structure logique [9] [André 1990].

Dans le cadre des travaux présentés dans ce manuscrit où **nous ne disposons pas de connaissance formalisée** de manière explicite à propos du contenu des images à analyser, nous adoptons un point de vue de plus bas niveau. Dans la suite de ce mémoire, l'aspect

8. Un exemple de ce cas spécifique est détaillé en section 3.3.1.1.
9. Les structures physique et logique d'un document sont définies en section 3.1.1.

« structuré » ou « non structuré » des images est avant tout lié à la présence ou à l'absence d'information *a priori* sur l'existence, la nature, la variabilité d'apparence et/ou l'agencement spatial d'éléments d'intérêt dans les images considérées. Dans notre contexte, cette notion est donc intimement liée à celle d'homogénéité du contenu de la collection à analyser. Le cas échéant, le niveau de structuration des images pourra être pris en compte implicitement ou explicitement lors de la conception d'approches de description ou de reconnaissance.

Comme évoqué dans la section précédente, selon les collections d'images considérées et les conditions de l'étude, la nature et l'exhaustivité de l'information dont on dispose concernant les catégories recherchées est variable. Si, dans un objectif de description d'images, on ne dispose souvent pratiquement d'aucune information *a priori* concernant les catégories recherchées, dans un objectif de reconnaissance au contraire, la collection d'images est en général associée à un ensemble de métadonnées décrivant (au moins partiellement) les catégories à reconnaître, incluant souvent des pointeurs vers quelques images représentatives de ces catégories.

Lorsqu'elle est disponible, l'**information sur les catégories** peut être de nature hétérogène. Elle peut par exemple porter sur les catégories directement, ou bien sur des sous-catégories. En plus d'informations sur l'apparence visuelle de ces catégories (ou sous-catégories), on peut disposer d'informations de nature plus sémantique les concernant.

1.3.2 Verrous scientifiques et techniques généraux

Les obstacles auxquels se trouve confrontée l'analyse d'images sont nombreux. Ils sont principalement liés à la difficulté de traduire sous la forme de problèmes mathématiques bien posés des tâches complexes, dont le système visuel humain s'accommode pourtant très facilement. J'ai choisi de focaliser mes efforts sur les verrous suivants :

- Un premier verrou scientifique, très général, est lié à la différence de niveau entre l'information binaire manipulée par les machines, et les concepts de plus haut niveau sémantique attendus par l'humain dans un contexte donné. Sans nous attarder sur le terme quelque peu galvaudé du **fossé sémantique/numérique**[10], nous pouvons noter que ce fossé engendre un certain nombre de difficultés en pratique. D'une part, il est ardu de concevoir des méthodes automatiques qui satisfassent les attentes de l'utilisateur humain. D'autre part, se pose la question de comment restituer de manière intelligible à l'humain l'information retournée par la machine à l'issue du processus d'analyse ;
- Un deuxième verrou scientifique peut être formalisé de la manière suivante : « **comment tirer au mieux parti des informations dont on dispose sur les images**, sachant que ces informations sont potentiellement de nature hétérogène et/ou très partielles ? ». Le problème est en fait double.
  Premièrement, la variabilité dans la nature des informations d'entrée pose des problèmes liés à l'intégration conjointe des divers types d'information dans le système. Ces problèmes sont, en partie, liés à la différence de niveau dans les traitements de ces informations.
  Deuxièmement, il peut arriver que l'information dont on dispose soit insuffisante au regard de la complexité du problème posé. À titre d'illustration, lorsque l'on cherche à mettre en œuvre un apprentissage à partir de données en nombre trop limité par rapport au volume de l'espace de représentation, cela suppose de mener à bien une inférence à partir

10. « *The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation* » [Smeulders 2000].

d'un nombre réduit d'exemples, dans un espace de possibilités trop vaste pour que le problème soit bien posé mathématiquement. Ce problème, originellement identifié par Richard Bellman pour des problèmes d'optimisation dynamique [Bellman 1961], est connu sous le nom de « malédiction de la dimensionnalité » ;

– Un troisième verrou auquel nous cherchons à nous attaquer est celui de la **non-modularité**, qui est un verrou à la fois d'ordre scientifique et technique. La diversité des problèmes pratiques soulevés dans un domaine tel que l'analyse d'images fait que les chercheurs en arrivent souvent à déployer des solutions spécifiquement conçues pour répondre à un problème particulier, ces solutions étant alors difficilement transposables vers un autre problème ou un autre contexte applicatif.

Les verrous généraux relatifs à l'analyse d'images, listés ci-dessus, se déclinent en un certain nombre de questions de recherche auxquelles on est confronté dès lors que l'on s'intéresse plus spécifiquement à la catégorisation d'images. C'est l'objet de la section suivante.

1.3.3 Questions abordées

Concernant le verrou du fossé sémantique, une approche typique pour s'y attaquer consiste à formaliser la connaissance d'un expert humain pour guider le processus de manière automatique. Ce type d'approches est illustré plus haut avec l'exemple d'images structurées pour lesquelles le processus de catégorisation peut bénéficier de connaissances en provenance d'un expert, formalisées de manière explicite. Dans ce cas, les informations récoltées en sortie du processus sont généralement de haut niveau sémantique et donc, pour la plupart, **ces approches relèvent plutôt de l'interprétation d'images que de leur analyse**. Dans notre contexte où nous ne disposons d'aucune connaissance formalisée explicitement, nous nous sommes demandé dans quelle mesure il serait possible/efficace de travailler à partir de descripteurs visuels des images d'un bas niveau sémantique, tout en cherchant à y adjoindre une information (partielle) de plus haut niveau sémantique concernant les catégories de certaines images. Nous nous intéressons en particulier au cas où **l'information serait apportée interactivement** par l'utilisateur du système dans le processus de catégorisation, rendant ce dernier semi-automatique. Cette information viendrait alors guider les traitements de données effectués dans l'espace de représentation de bas niveau sémantique pour rapprocher les catégories découvertes/inférées par la machine des concepts de plus haut niveau perçus par l'utilisateur dans ces images. Se pose alors le problème de choisir des modes d'interaction adaptés à la fois à l'utilisateur et aux processus mis en œuvre par la machine.

En ce qui concerne la présentation des résultats de la catégorisation à l'humain à l'issue de l'apprentissage, on peut s'interroger sur les possibilités et les limites rencontrées dès lors qu'il s'agit de **restituer les mécanismes de catégorisation** appris par la machine à un humain, sous une forme qui soit intelligible pour ce dernier.

Le deuxième verrou auquel nous cherchons à nous attaquer concerne la manière d'intégrer efficacement dans le système les informations dont on dispose sur les images, sachant que ces informations sont potentiellement de nature variable et/ou partielles. Dans les contextes auxquels je m'intéresse, les informations dont on dispose peuvent provenir du contenu visuel des images ou d'informations du domaine concernant les images ou leurs catégories (ces dernières étant, selon les cas, connues au travers d'une vérité-terrain, ou collectées interactivement auprès de l'utilisateur). Ces informations sont de **natures et de niveaux sémantiques hétérogènes**.

Nous nous intéressons en particulier à la manière de faire coopérer de la manière la plus adéquate possible les traitements de ces informations hétérogènes.

Il est en outre courant que certaines de ces informations soient très partielles. C'est en particulier souvent le cas des informations concernant les catégories à découvrir ou à inférer. En effet, d'une part, selon les contextes applicatifs, la vérité-terrain peut se révéler coûteuse ou difficile à collecter et, d'autre part, dans un contexte interactif, on cherche généralement à minimiser l'effort de l'utilisateur. Dans ce contexte, nous nous sommes intéressés aux façons de **répercuter de la manière la plus efficace possible ces informations partielles** dans le processus de catégorisation. Quitte à déduire, à partir de cette information partielle, de nouvelles informations pour alimenter l'apprentissage. Et ce, avec les risques que cela suppose si les mécanismes de déduction sont insuffisamment adaptés au contexte.

Le troisième verrou auquel nous nous intéressons est celui de la non-modularité. Les spécificités d'un contexte applicatif donné sont parfois telles que les efforts en termes de recherche et développement portent souvent sur la conception de solutions dédiées à ce contexte applicatif. Si ces solutions dédiées sont en général plus performantes que des techniques plus génériques pour ce contexte en particulier, se posent des difficultés liées à leur portabilité vers un autre contexte applicatif. Dans certains cas extrêmes, le processus de conception est repris à zéro, sans tirer bénéfice de l'expérience acquise lors de la conception de la première approche. Sans aller jusque-là, il est souvent impératif de passer par une phase fastidieuse de ré-apprentissage des paramètres (parfois nombreux) de la méthode, voire par un changement d'architecture globale du système afin d'envisager la portabilité de l'approche.

Sur ce point, nous nous sommes interrogés sur la possibilité de s'appuyer sur l'utilisateur pour ajuster ces paramètres. Plus précisément, nous avons envisagé deux possibilités. La première est que, dans un contexte interactif, le système s'appuie sur les retours donnés par l'utilisateur pour ajuster ses paramètres. La seconde est de laisser à un utilisateur expert le soin d'adapter les paramètres du système, le cas échéant sous la supervision de la machine. Bien entendu, cela a un impact sur la manière de concevoir l'approche, afin que sa portabilité soit rendue possible par l'ajustement de **paramètres réglables grâce à une intervention humaine** (directement ou indirectement).

Dans un domaine de recherche à visée applicative tel que celui de l'analyse d'images, un problème supplémentaire auquel on est systématiquement confronté est celui de la **caractérisation des performances**. La caractérisation de performances est ici à entendre au sens large et va bien plus loin que l'évaluation, en fin de chaîne, de la solution de catégorisation retournée par la machine. Elle couvre une évaluation à la fois quantitative et qualitative, qui doit être mise en œuvre sous des formes variées et à des étapes diverses du processus global de catégorisation. Les questions posées sont par exemple : « Selon quel procédé choisir les descripteurs les plus adaptés aux images à analyser ? », « Comment sélectionner les images à présenter à l'humain (dans un contexte interactif) ? », « Comment caractériser la qualité des informations déduites à partir des informations limitées dont on disposait initialement ? », « Comment évaluer/comparer de manière systématique la robustesse des approches vis-à-vis de certaines sources de dégradation dans les images ? », etc.

1.4 Fil conducteur et principaux jalons de mon cheminement scientifique

Cette section présente l'ensemble de mon parcours scientifique, émaillé de ses principaux jalons en termes notamment d'encadrements, de projets, et de collaborations. Elle est structurée autour des deux objectifs généraux poursuivis, à savoir la description et la reconnaissance d'images. Afin d'éviter d'alourdir le contenu par de multiples notes de bas de page, j'utilise dans la suite les renvois suivants :

- * Cette responsabilité d'équipe ou de laboratoire est décrite en section 2.3.1 du tome I (Curriculum-Vitæ) ;
- ** Plus de détails sur cette thèse sont donnés en section 2.3.2 du tome I ;
- *** Cette collaboration scientifique est détaillée en section 2.4 du tome I ;
- **** Ce projet de recherche est décrit en section 2.5 du tome I.

1.4.1 Dans un objectif de description d'images

Dans un objectif de description d'images, nous cherchons à extraire, à partir du contenu pixellaire des images, éventuellement associé à des informations du domaine, une représentation simplifiée qui permette de synthétiser ce contenu. Cette représentation repose sur des descripteurs, qui peuvent être extraits à plusieurs niveaux, notamment en fonction du type d'images à décrire et de l'information dont on dispose sur ces images.

Dans le cas d'une collection constituée d'un type donné d'images, il est possible de concevoir des **descripteurs spécifiques** (de nature statistique et/ou structurelle) de ce type d'images. C'est ce que j'appellerai dans la suite des « descripteurs de grain fin ». Lorsque les images à décrire sont structurées, ces descripteurs peuvent être construits de manière à prendre en compte (implicitement ou explicitement) les informations du domaine dont on dispose concernant la structuration des images d'entrée, afin de véhiculer la description la plus adaptée possible au type d'images considéré. C'est le travail que nous avons mené dans le contexte de **ma thèse** [Visani 2005a] pour des visages puis, peu après mon recrutement au L3i, pour des symboles graphiques.

Force est de constater que la littérature regorge de descripteurs d'images parfois redondants entre eux. Pour les chercheurs, ce foisonnement rend difficile le choix de descripteurs adaptés à leur problématique applicative. Peu après mon arrivée au L3i, j'ai souhaité m'appuyer sur mes compétences en statistiques pour m'intégrer dans la communauté de recherche sur les images de documents (thème central au L3i), au travers de la problématique de la caractérisation de performances des descripteurs. Je me suis pour cela rapprochée de certains partenaires du L3i, et en particulier du Computer Vision Center (**CVC, Barcelone**)*** et du **LORIA***** (partenaire du projet ANR Navidomass****). Avec Oriol Ramos Terrades et Antoine Tabbone, nous avons proposé un protocole pour la **caractérisation des performances de descripteurs** de formes. Ce protocole repose sur des mesures quantitatives et qualitatives permettant de caractériser le pouvoir descriptif individuel et la complémentarité des descripteurs [11].

Lorsque les types d'images dans la collection sont hétérogènes et/ou que l'on ne dispose d'aucune information du domaine, nous avons choisi de travailler au niveau plus grossier de la description de la collection d'images. En particulier, nous nous sommes intéressés à la manière

11. Ce travail a donné lieu à une co-publication dans une revue internationale.

d'organiser la collection d'images pour pouvoir en fournir une description synthétique, cette description pouvant servir de base à une visualisation ou à une navigation. À ces fins, nous nous sommes focalisés sur la conception d'approches d'extraction de descripteurs par catégorisation des images. Afin de réduire le fossé sémantique (sans pour autant prétendre le combler) entre les catégories perçues par l'humain et celles proposées par la machine, nous avons cherché à intégrer efficacement dans le processus de catégorisation une information partielle concernant les catégories et collectée interactivement auprès de l'humain (typiquement un expert du domaine, par exemple un archiviste). La description devient alors semi-automatique, puisqu'assistée par l'humain. Forte de mon expérience de l'analyse de données de type image, je me suis rapprochée d'Alain Boucher, spécialiste de l'interaction avec l'humain, à l'époque affilié à l'**Unité Mixte Internationale UMMISCO** (UPMC / IRD) et basé à Hanoï. Ce rapprochement a bien sûr été facilité par mes **collaborations internationales avec le Vietnam*****. C'est sous la direction de Jean-Marc Ogier (PR au L3i), expert de la description d'images et très impliqué dans les collaborations avec l'Asie, que nous avons proposé en 2009 une thèse « ministérielle » sur la problématique peu investiguée du *clustering* interactif d'images. Il s'agit de la **thèse de Lai Hien Phuong****, soutenue en 2013 [Lai 2013a], qui a débouché sur la proposition d'une approche de ***clustering* hiérarchique semi-supervisé et interactif de bases d'images** naturelles tout-venant. L'information collectée auprès de l'utilisateur est intégrée incrémentalement dans le processus de *clustering* en tirant bénéfice de la structure hiérarchique sous-jacente. On obtient au final une hiérarchie de groupes d'images plus représentatifs des catégories perçues par l'humain dans la collection. Malgré l'émergence de travaux très récents, ce point de vue reste assez original dans la communauté, où l'interaction avec l'utilisateur est le plus souvent reléguée ultérieurement, typiquement lors d'une phase de recherche dans les bases d'images. Il était d'autant plus novateur en 2009, à l'heure où nous avons initié cette étude.

À la fin de ce travail, je me suis interrogée sur la possibilité de répercuter les retours fournis par l'humain sur les couches plus basses du système. L'idée est d'apprendre, grâce à l'information apportée incrémentalement par l'humain, non plus une hiérarchie de groupes d'images, mais plutôt une **hiérarchie de descripteurs visuels**. Une telle hiérarchie serait plus générique, au sens où elle serait par exemple applicable à une nouvelle collection d'images tout en tirant parti des retours donnés par l'utilisateur sur une première collection. Et ce, même si les groupes présents dans la nouvelle collection sont différents. Je me suis alors rapprochée de la **fédération de recherche MIRES** [12] et de Thierry Urruty, jeune MCF spécialiste au laboratoire XLIM-SIC des descripteurs visuels locaux. En 2013, nous avons co-encadré un stage de M2 pour une étude préalable sur ce sujet. Nous pourrons poursuivre nos recherches communes grâce à un **financement du GdR ISIS** que nous venons d'obtenir (projet CINÉDI, voir section 1.4.3).

Toujours dans un objectif de description d'images, j'ai souhaité tisser un lien entre nos travaux concernant, d'une part, la conception de descripteurs « de grain fin » et, d'autre part, la catégorisation interactive. C'est au travers de l'application à la description de documents textuels écrits dans des scripts/langages anciens ou rares, que je me suis intéressée à ce problème, initialement dans le cadre d'un **projet PCSI** [13] **sur la valorisation du patrimoine khmer******. L'idée est de décrire le texte de ces documents sans aucune information *a priori* sur le script et/ou le langage utilisé, au travers d'« invariants » (primitives composant le texte, découvertes depuis la collection de documents à décrire). En collaboration avec Rémy Mullot (PR

12. Fédération de laboratoires du PRES Limousin-Poitou-Charentes (FR CNRS 3423).
13. Projet de Coopération Scientifique Inter-universitaires (PCSI), financé par l'Agence Universitaire pour la Francophonie (AUF).

au L3i), spécialiste du traitement d'images de documents et impliqué dans ce projet, nous avons proposé une thèse sur ce sujet en 2011. Il s'agit de la **thèse de Bui Quang Anh****. Afin de pallier l'absence d'information *a priori*, les invariants sont découverts par ***clustering* interactif des traits d'écriture automatiquement segmentés** dans les documents de la collection. L'utilisateur visé est un expert du domaine (archiviste par exemple) ; l'interaction se fait selon des modalités permettant au système de corriger localement les invariants et/ou les traits d'écriture selon les interventions de l'humain. Les invariants ainsi extraits peuvent par exemple être utiles pour des applications de navigation dans la collection de documents par *word spotting*.

1.4.2 Dans un objectif de reconnaissance d'images

Dans un objectif de reconnaissance, nous nous intéressons particulièrement aux images de documents, par nature structurées. Lorsqu'il s'agit non plus de description mais de reconnaissance d'images, on dispose d'informations concernant les catégories (ici les classes) recherchées, et le cas échéant d'informations du domaine. Ces informations se présentent parfois sous des formes hétérogènes, et peuvent dans certains cas se révéler très limitées. Au travers des trois cas d'étude développés dans les trois paragraphes ci-après, nous avons cherché à concevoir des approches tirant le meilleur parti de ces informations, en termes d'apprentissage par la machine et/ou de restitution à l'humain.

Dès mon recrutement en 2006, j'ai été sollicitée, en raison de mon expérience de l'analyse de données, pour **participer à la supervision scientifique de la thèse de Stéphanie Guillas**[14], qui portait sur la conception d'une méthode symbolique de classification supervisée basée sur un treillis des concepts. Cette thèse a été menée sous la supervision Karell Bertet (MCF, L3i) et de Jean-Marc Ogier (PR, L3i) et soutenue en 2007 [Guillas 2007]. La thématique applicative principalement visée était au cœur des activités de recherche du L3i à l'époque ; il s'agit de la reconnaissance de symboles graphiques. L'idée était de chercher une méthode de classification symbolique qui permette de restituer à l'expert en sortie du processus de classification non seulement une décision, mais aussi l'ensemble des règles ayant mené à cette décision. Cette restitution se fait de manière lisible sous forme graphique *via* le treillis. La principale originalité de la méthode que nous avons proposée (à savoir Navigala) par rapport aux approches existantes de classification basées sur des treillis des concepts est que, dans Navigala, le treillis n'est pas utilisé comme un outil de sélection des concepts/règles contextuelles les plus pertinents, mais plutôt comme un outil de navigation dans les données (à la manière d'un arbre de décision). Ces travaux nous ont permis de mettre en lumière les liens structurels existants (sous certaines conditions) entre la famille des treillis manipulés dans Navigala et les arbres de classification.

Forts de ce résultat théorique, nous avons souhaité aller plus loin dans le rapprochement entre l'usage de treillis en classification supervisée et les arbres de décision, afin de bénéficier des avantages des deux structures (robustesse du treillis et faible complexité des arbres). Nous avons donc proposé et obtenu, en collaboration avec Karell Bertet, une thèse « ministérielle » sur la **conception d'une approche hybride entre treillis des concepts et arbre de classification** pour la classification supervisée. L'objectif est d'obtenir une méthode assez générique pour être utilisée bien au-delà de la reconnaissance d'objets graphiques, dans de multiples problèmes d'analyse de données. Il s'agit de la **thèse de Nathalie Girard****, qui a pu bénéficier à la fois des compétences de Karell Bertet concernant les structures de treillis [Bertet 2011] et de

14. Participation « informelle » (soutenance 1 an après mon recrutement), qui s'est soldée par 4 co-publications.

mon expérience de la classification supervisée. Cette thèse a été soutenue en 2013 [Girard 2013]. Grâce à l'encadrement de deux stages de master, **nous avons constitué autour de cette thématique une petite équipe de recherche** qui a intégré son code dans une bibliothèque commune [15].

Ce travail de recherche m'a ouvert les portes de la communauté de l'Analyse Formelle des Concepts. En particulier, nous avons pu identifier des pistes de recherche communes avec Karell Bertet et Rokia Missaoui (PR, Université du Québec en Outaouais) [16].

Dans un deuxième temps, je me suis intéressée au cas d'images de documents de structure moins contrôlée, et pour lesquelles les informations d'entrée sont de nature plus hétérogène. J'ai mené ces travaux dans le contexte applicatif de la reconnaissance de mots manuscrits cursifs à partir d'un signal en-ligne acquis par le biais d'une tablette électromagnétique. L'une des principales difficultés rencontrées réside dans la manière de tirer le meilleur bénéfice de l'information hétérogène disponible (informations concernant l'apparence visuelle des caractères, la dynamique du tracé et le lexique) lors de la reconnaissance des mots. Dans le cadre du **projet Eurêka Reconomad******, nous avons encadré en collaboration avec Jean-Marc Ogier la **thèse de Sophea Prum** sur ce sujet**. Cette thèse, soutenue en 2013 [Prum 2013a], était environnée grâce au projet de deux ingénieurs, et nous avons co-encadré trois stages de Master sur la personnalisation du moteur de reconnaissance vis-à-vis du scripteur, ce qui a permis la **création d'une petite équipe de recherche**. Ces travaux ont débouché sur l'introduction d'un système de reconnaissance de mots manuscrits en-ligne basé sur une segmentation semi-explicite du mot en caractères et une classification à deux niveaux d'entités plus élémentaires que le mot.

Cette étude, initiée en 2009, a constitué à l'échelle du L3i des travaux exploratoires sur la reconnaissance d'écriture manuscrite. Afin de bénéficier de l'expérience d'un chercheur confirmé dans ce domaine très spécifique et riche, nous avons organisé en 2012 un séjour de recherche au L3i pour Andreas Fischer, à l'époque post-doctorant dans le laboratoire **CENPARMI** de l'université de Concordia*** au Canada et préalablement dirigé par Horst Bunke lors de sa thèse sur la reconnaissance d'écriture manuscrite. Ce séjour s'est soldé par la co-publication d'un article de conférence internationale. Nous sommes en train d'investiguer conjointement diverses pistes de recherche concernant la reconnaissance de styles d'écriture en vue de la personnalisation de moteurs de reconnaissance d'écriture. Ce type d'approches de personnalisation peut être vu comme un moyen de tirer au mieux parti des informations d'entrée, en enchaînant deux processus de catégorisation (de style d'écriture puis de texte), le premier permettant de spécialiser le second afin d'en améliorer les performances finales.

Dans les deux premiers cas d'étude évoqués ci-dessus, je me suis particulièrement intéressée à la manière d'intégrer et de restituer des informations de nature variable au travers d'un processus de reconnaissance d'images de documents. Mais, que faire lorsque les exemples annotés sont tout simplement en nombre insuffisant pour que l'apprentissage et/ou la caractérisation des performances des systèmes de reconnaissance soit fiable ? C'est un cas de figure assez fréquent, en particulier dès lors que l'on s'intéresse à l'analyse d'images de documents anciens. En collaboration avec Nicholas Journet, MCF spécialiste au **LaBRI (Bordeaux)***** du traitement et de l'analyse d'images de documents anciens, nous sommes co-responsables dans le cadre du **projet ANR DIGIDOC****** d'un *workpackage* sur la génération d'images

15. La bibliothèque « *Lattice* », mise à disposition par Karell Bertet sous licence LGPL [Bertet 2011].
16. R. Missaoui a effectué un séjour de recherche au L3i en juin 2014 (financement ULR porté par K. Bertet).

de documents semi-synthétiques. L'idée est de générer, à partir d'un faible nombre d'images annotées, des documents semi-synthétiques et, le cas échéant, la vérité-terrain associée. Ces nouvelles données pourront être utilisées, soit pour enrichir l'apprentissage (ré-apprentissage), soit pour caractériser finement les performances d'algorithmes d'analyse d'images. Sur ce sujet, je co-supervise la **thèse de Kieu Van Cuong**** avec Nicholas Journet, Jean-Philippe Domenger (PR au LaBRI et spécialiste d'analyse d'images) et Rémy Mullot (PR, L3i, directeur de thèse). Cette thèse, initiée en 2011 et financée par DIGIDOC, a débouché sur la proposition de méthodes/modèles de dégradation spécifiquement conçus pour imiter certains des défauts fréquemment rencontrés dans des images de documents anciens. Par application de ces dégradations sur les quelques images annotées dont on dispose, on peut **générer une grande variété d'images semi-synthétiques visuellement réalistes** (et, le cas échéant, la vérité-terrain associée). Ces méthodes/modèles sont conçus de manière à ce que l'expert humain puisse contrôler (au moins partiellement) la nature et le niveau de dégradation des images générées. Notre travail apporte des éléments de solution à des problèmes récurrents pour la communauté scientifique de l'analyse d'images de documents ; il a donc logiquement suscité un fort intérêt de la part de cette communauté, et a mené à plusieurs co-publications avec des partenaires impliqués dans DIGIDOC et, au-delà de DIGIDOC, avec des partenaires internationaux. Concernant la caractérisation de performances, nous avons entre autres co-organisé avec le CVC de Barcelone une **compétition internationale** en marge de la conférence ICDAR 2013 sur la détection de lignes de portées musicales [17]. Au-delà de cette compétition, la base que nous avons publiée à cette occasion est d'ores et déjà réutilisée par des chercheurs comme base de référence. En ce qui concerne le ré-apprentissage, nous avons montré la pertinence de notre approche par le biais de diverses applications, notamment pour la reconnaissance de texte manuscrit ancien (en collaboration avec Andreas Fischer, actuellement à Montréal, voir plus haut) [18].

Suite à un séjour de recherche que nous avons organisé pour Cuong à l'**université de Fribourg** (équipe de recherche DIVA) en Mars 2014, nous sommes en train de travailler en collaboration avec les chercheurs de cette équipe sur d'autres applications telles que le ré-apprentissage d'outils de segmentation de zones/lignes de texte.

Au terme de la présentation du cheminement scientifique que j'ai parcouru jusqu'à aujourd'hui, je donne ci-après des éléments plus prospectifs liés à des projets ou programmes de recherche, dont le financement est acquis, et qui débuteront dans les prochains mois.

1.4.3 Projets ou programmes de recherche à venir

1.4.3.1 CINÉDI et financements complémentaires

Comme évoqué plus haut, le projet CINÉDI porte sur l'extraction d'une hiérarchie sémantique de descripteurs visuels d'images. Il sera financé par le GdR ISIS à compter de la fin 2014 et durera 24 mois au total. Il nous permettra notamment de recruter deux stagiaires de M2 que nous co-encadrerons, avec Thierry Urruty, spécialiste des descripteurs locaux d'images (XLIM).

Sur une thématique similaire, j'ai obtenu du gouvernement vietnamien *via* l'université des Sciences et Technologies de Hanoï (USTH)*** le financement d'une **thèse qui devrait commencer au début de l'année 2015**. La co-supervision scientifique se ferait avec Thierry Urruty, Arnaud Revel (PR du L3i), que j'ai sollicité pour son expérience dans les domaines de l'apprentissage actif et des interactions avec l'humain, et Nguyen Dung Duc, chercheur spécia-

17. Ce travail a donné lieu à un article de conférence internationale et à un chapitre de livre.
18. Ce travail a donné lieu à la co-publication d'un article de conférence internationale.

liste de l'apprentissage semi-supervisé et supervisé au sein de la **Vietnamese Academy of Science and Technology (VAST)***** [19] à Hanoï.

1.4.3.2 ARCHIVES et RELISH

En 2013, lors d'un séjour de recherche de deux mois à Hanoï, j'ai travaillé activement à la conception du **projet ARCHIVES******, financé essentiellement par la Banque Asiatique pour le Développement (BAD) et le Ministères des Affaires Étrangères *via* l'USTH. Ce projet, qui fédère autour de lui plusieurs partenaires français et vietnamiens sur la thématique des humanités digitales (ou humanités numériques), porte sur l'analyse et la reconstitution d'événements catastrophiques par modélisation géo-historique. Certains de ces événements sont décrits au travers de documents écrits en sino-vietnamien, en français (durant la période coloniale) et en vietnamien moderne. Ce projet (et plus particulièrement le *workpackage* « analyse d'images » dont je suis responsable) intègre donc naturellement un volet sur l'analyse d'écriture multi-script et multi-langage, dans lequel l'expérience acquise par le biais des travaux réalisés dans le cadre des thèses de Quang Anh [20] et de Sophea nous sera bien sûr utile. Les évènements catastrophiques sont également relatés au travers, d'une part, de nombreuses cartes et, d'autre part, de documents contenant des éléments graphiques (tampons, sceaux, etc.) cruciaux pour déterminer leur origine. Certains des travaux que nous avons menés, en particulier sur l'organisation de collections d'images naturelles et sur la reconnaissance de graphiques, pourraient être ré-utilisés et étendus dans le contexte de ce projet.

Forte de mon expérience de co-responsable des équipes de recherche IMEDOC (2007-2010) et IDDC (2010-2014)*, et souhaitant m'impliquer plus activement dans la structuration de la recherche à l'USTH, j'ai répondu en 2014 à l'**appel de l'Ambassade de France à Hanoï « objectif labos »**. Ce programme vise à encourager la création et le développement de nouveaux laboratoires mixtes entre l'USTH et ses partenaires vietnamiens et français. L'un des partenaires de recherche vietnamiens majeurs de l'USTH est la VAST (voir plus haut). Du côté français, l'USTH s'appuie pour l'informatique sur un consortium composé de 11 universités françaises (dont l'université de La Rochelle), et sur un fort investissement humain de la part de l'Institut de Recherche pour le Développement (IRD).

Suite à l'acceptation de notre dossier intitulé RELISH par l'ambassade (dont je suis porteuse et Alexis Drogoul, DR à l'IRD, est co-porteur), je deviendrai début 2015 ***acting director* du laboratoire *Vietnam-France ICT Lab****, durant sa phase de création. Les thématiques scientifiques de ce laboratoire (modélisation, traitement/analyse d'images, apprentissage artificiel, interaction utilisateur, SIG et capteurs) sont en appui sur les trois spécialités du master d'informatique auquel il est adossé, et couvrent certains de mes domaines d'expertise. Dans cette phase de création d'un laboratoire, nous avons choisi d'adopter un point de vue pragmatique par le biais d'une « approche projet ». La création de ce laboratoire s'appuie donc en particulier sur le projet ARCHIVES. L'objectif de ce programme à moyen terme est de déboucher sur la création d'un laboratoire comprenant au minimum une équipe internationale dont le L3i serait partenaire, et qui pourrait à plus long terme se placer sous la tutelle de l'IRD et/ou du CNRS.

19. Institut de recherche public (équivalent à l'échelle vietnamienne du CNRS pour les sciences et technologies).

20. D'un point de vue de ressources humaines, il est utile de préciser ici que la thèse de Quang Anh est financée par le gouvernement vietnamien *via* l'USTH. Le contrat régissant cette bourse prévoit qu'à l'issue de son doctorat, le bénéficiaire retournera travailler comme enseignant-chercheur à l'USTH. Si toutefois il le souhaite, ce projet permettra à Quang Anh de s'intégrer scientifiquement à l'USTH lorsqu'il y sera recruté à l'issue de sa thèse, tout en favorisant la pérennisation des collaborations de recherche entre le L3i et l'USTH.

1.5 Organisation du reste du mémoire

J'ai choisi d'organiser le reste de ce mémoire autour des deux grands objectifs d'analyse d'image visés dans le cadre de nos travaux : la description d'images et la reconnaissance d'images. J'ai fait ce choix pour des raisons didactiques, mais aussi parce que, comme nous venons de le voir au travers de mon cheminement scientifique, les faits marquants liés à nos contributions dans ces deux objectifs sont relativement équilibrés (en termes notamment d'encadrement scientifique, de publications, d'activité contractuelle et de collaborations).

Le **chapitre 2** traite donc de la description d'images, tandis que le **chapitre 3** est consacré à la reconnaissance d'images. Dans chacun de ces deux chapitres, je commence par placer nos travaux dans leur contexte scientifique et applicatif, avant de présenter nos principales contributions. Ces contributions sont systématiquement discutées par rapport, d'une part, aux questions de recherche abordées dans ce manuscrit (*cf.* section 1.3.3) et, d'autre part, vis-à-vis de la littérature. Ces chapitres se terminent par la présentation de nos principales perspectives de recherche, et par un récapitulatif des principaux faits marquants liés à nos contributions.

Ce mémoire s'achève avec le **chapitre 4** qui en dresse les conclusions générales et situe les principales perspectives de recherche évoquées au fil du manuscrit dans le cadre unificateur de l'exploitation interactive de collections d'images.

Ce manuscrit comporte cinq annexes, dont la dernière (annexe E) est composée d'un recueil de trois articles parus dans des revues internationales.

Chapitre 2

Contributions dans le domaine de la description d'images

Description d'images de niveaux de structuration variables par clustering interactif

---

Sommaire

Tables

Figures

2.1 Introduction

Chaque image est composée d'un nombre potentiellement important de valeurs de pixels peu structurées (en comparaison avec des données texte par exemple). Notre objectif est ici d'extraire, à partir de ce contenu pixellaire éventuellement additionné d'informations du domaine, une représentation simplifiée qui permette de synthétiser ce contenu. Concrètement, il s'agit de calculer des descripteurs, plus synthétiques que le contenu initial, mais représentatifs de ce contenu.

Il convient ici de définir plus précisément l'usage que nous ferons dans la suite des termes « caractéristique », « descripteur » et « signature », souvent utilisés de manière indifférenciée dans la littérature :

- Une caractéristique (ou caractéristique visuelle) est un indice visuel extrait de l'image, souvent obtenu par un filtre ou une transformation mathématique appliquée localement ou globalement sur les pixels de l'image.
- Un descripteur vise également à décrire le contenu de l'image. Il peut être, selon les cas, composé uniquement de caractéristiques visuelles (on parle alors de « descripteur visuel »), ou comporter des caractéristiques de plus haut niveau sémantique associées à l'image (type, mots-clés, etc.). Comme expliqué en annexe A, les éléments du descripteur peuvent être organisés sous une forme vectorielle (descripteurs statistiques) ou bien séquentielle ou structurée (descripteurs structurels) ;
- Les signatures des images sont constituées de l'ensemble des descripteurs retenus au final pour décrire une image. Elles permettent de définir l'espace de représentation des images (vectoriel ou graphique). Une signature est en général associée à une mesure de similarité adaptée [Jolion 2001], qui permet de comparer différentes images dans l'espace de représentation. Une signature peut être statistique, structurelle, ou statistico-structurelle (c.-à-d. composée de descripteurs structurels ramenés directement ou indirectement dans un espace de représentation vectoriel, généralement dans une optique d'efficacité de l'analyse).

Selon les cas, les descripteurs pourront être retournés soit à un humain (expert du domaine d'application, chercheur en vision par ordinateur, ou un simple utilisateur du système), soit à la machine afin d'alimenter un processus ultérieur. Dans ce dernier cas de figure, les descripteurs seront typiquement réutilisés pour indexer les images en vue d'une recherche ultérieure [Smeulders 2000, Fournier 2001, Datta 2008], ou par un moteur de reconnaissance. Il arrive également qu'ils soient mis à profit par un moteur de visualisation/navigation par similarité dans les bases d'images [Ma 1999, Borth 2008, Jing 2012].

La destination de ces descripteurs conditionne en particulier leur niveau de structuration, et le niveau auquel on les extrait. Les descripteurs peuvent en effet être extraits au niveau d'une région de l'image, d'une image donnée ou au niveau plus global de la collection d'images. Au niveau de l'image ou de ses régions, on cherche le plus souvent des descripteurs visuels représentant le contenu de l'image, en général pour alimenter des processus ultérieurs d'indexation/recherche ou de reconnaissance. C'est l'objet de la section 2.1.1. Au niveau plus global de la collection d'images, on cherche des descripteurs qui permettent de représenter de manière synthétique le contenu souvent volumineux de la collection ; ces descripteurs pourront typiquement être utilisés par un moteur de visualisation/navigation par similarité dans les bases d'images. Une manière de décrire la collection d'images passe par une catégorisation des images de la collection ou de leurs éléments d'intérêt ; c'est l'objet de la section 2.1.2.

2.1.1 Extraction et caractérisation de performances de descripteurs visuels

On se place ici au niveau de l'image ou de ses régions, où l'on cherche le plus souvent des descripteurs visuels, en général pour alimenter des processus ultérieurs d'indexation/recherche ou de reconnaissance d'images. Les nombreux descripteurs visuels proposés dans la littérature sont traditionnellement regroupés selon des taxinomies basées sur leur nature (p. ex. statistique et/ou structurelle), sur la nature de ce qu'ils cherchent à décrire (p. ex. couleur, forme, texture), sur la manière dont ils sont extraits depuis l'image (localement, globalement, « spatialement [1] ») ou, le cas échéant, selon leurs bonnes propriétés et éventuellement leurs invariances. Je les amène ici d'une manière un peu plus originale, en lien avec le fil conducteur de mes travaux, à savoir en fonction de l'information dont on dispose *a priori* sur les images.

Dès lors que l'**on dispose *a priori* d'informations concernant le(s) type(s) des images** à décrire, ou des objets/motifs présents dans ces images, il est possible d'extraire des descripteurs tirant parti de cette information.

En présence d'un type donné d'images structurées [2], il peut être avantageux d'utiliser implicitement ou explicitement les informations sur leur structure pour le calcul de descripteurs dédiés à ce type d'images (le cas échéant au travers d'un apprentissage). Je parle dans ce cas de « descripteurs de grain fin », que l'on retrouve très fréquemment dans des contextes applicatifs tels que la biométrie [Turk 1991], l'imagerie médicale [Cachier 2001] ou encore l'analyse d'images de documents [Bunke 2011]. C'est l'objet de certains des travaux réalisés dans ma thèse [Visani 2005a] et, peu après mon recrutement au L3i, dans le cadre d'un stage de M2 que j'ai co-encadré et qui a donné lieu par la suite à plusieurs publications. Ces travaux portent respectivement sur la proposition de descripteurs statistiques pour des visages, et d'une signature statistico-structurelle pour des symboles graphiques (les visages ou les symboles graphiques

1. Souvent au travers d'un graphe des relations spatiales entre les objets visuels composant l'image.
2. C-à-d. lorsque l'on dispose implicitement ou explicitement d'informations du domaine sur la présence, la nature et/ou l'agencement d'éléments d'intérêt dans les images, *cf.* section 1.3.1.2 « Images considérées ».

étant préalablement détectés/localisés dans l'image). Le lecteur intéressé pourra se référer à l'annexe A pour plus de détails concernant ces travaux, ayant donné lieu à des contributions à la fois méthodologiques et applicatives. S'ils s'avèrent particulièrement efficaces dès lors qu'on les extrait depuis le type d'images pour lequel ils ont été conçus, ce type de descripteurs dédiés ne sauraient néanmoins décrire efficacement des images de contenu moins contrôlé. Autrement dit, ils souffrent d'une importante non-généricité.

En présence d'images peu structurées et/ou appartenant à des types hétérogènes, il reste possible de tirer parti de l'information concernant le type des images (ou objets/motifs dans ces images), en entraînant de manière supervisée des outils d'extraction de descripteurs. La supervision confère aux descripteurs ainsi appris un certain caractère sémantique, ce qui explique en partie leurs excellentes performances en pratique. Les descripteurs extraits par un apprentissage profond – en plein essor – sont réputés parmi les plus efficaces [Goh 2013, Iandola 2014, Krause 2014, Donahue 2014], mais requièrent le plus souvent de très nombreux exemples. Plus généralement, l'entraînement de tels descripteurs est souvent guidé par un processus de détection et/ou de reconnaissance d'objets (parfois basée sur la détection/reconnaissance de certaines de leurs parties). Même si, en théorie, certains de ces descripteurs peuvent être utilisés pour représenter une grande variété de contenus, dans la pratique il est souvent nécessaire que les bases d'apprentissage et de test soient suffisamment similaires. Pour illustrer cette assertion, on peut citer les travaux menés dans l'équipe LEAR de l'INRIA [Paulin 2014] qui montrent une baisse sensible des performances des descripteurs DeCAF [Donahue 2014] lorsqu'ils sont appris sur les images du challenge ImageNet 2012 et utilisés pour décrire le contenu d'ImageNet 2010, et ce, malgré le fort recouvrement entre les deux bases.

Dans les contextes applicatifs que je considère dans la suite de ce chapitre, nous ne disposons *a priori* d'aucune information concernant les catégories des images à décrire. Donc, nous ne considérerons pas ces types de descripteurs.

En l'absence d'information *a priori* concernant le type des images ou des objets qu'ils contiennent, il est courant d'utiliser des descripteurs visuels génériques (généralement statistiques) capables de décrire au mieux les variétés de formes, textures et couleurs présentes dans la nature (comme par exemple l'un des nombreux descripteurs décrivant l'image localement [Li 2008] largement adoptés par la communauté [3], tels que SIFT [Lowe 2004] ou l'histogramme de gradient orienté (HoG) [Dalal 2005]), ou encore des descripteurs basés sur un apprentissage non-supervisé [Ranzato 2007].

Ces descripteurs sont le plus souvent d'un **niveau sémantique trop faible** pour permettre à la machine de caractériser précisément le contenu de l'image en adéquation avec les concepts de plus haut niveau sémantique perçus par un humain. On peut citer pour appuyer ce propos l'initiative menée par des chercheurs du MIT dans un contexte applicatif de détection d'objets dans des images [Vondrick 2013] : partant du constat que le système visuel humain nous permet d'appréhender très efficacement le contenu d'une image, ils se sont demandés comment les humains se débrouilleraient si, au lieu de visualiser la totalité des pixels de l'image, ils n'en visualisaient que ce qu'en perçoivent généralement les machines, à savoir leur signature. Ils ont donc pré-traité un ensemble d'images, en ont extrait des signatures basées sur HoG, et ont transformé ces signatures en une représentation visuelle susceptible d'être appréhendée facilement par un humain. Leurs conclusions montrent que les humains commettent de nombreuses erreurs de détection sur la foi des images ainsi constituées, ces erreurs rejoignant dans leur

3. Nous avons introduit un de ces descripteurs locaux, basé sur des ondelettes couleur, dans [Laurent 2003].

grande majorité les erreurs commises par la machine. Il arrive par exemple que la signature d'une région de l'image correspondant à des vaguelettes dans l'eau ressemble à s'y méprendre à une voiture, trompant à la fois l'humain et la machine.

En raison de la très grande variété des contenus possibles et vu le relativement faible pouvoir expressif de ces descripteurs, des **signatures volumineuses** sont (dans l'immense majorité des cas) requises pour décrire les images. Et ce, malgré les efforts entrepris pour en réduire la taille : réduction de dimension [Van der Maaten 2009], sélection de caractéristiques [Jain 1997], discrétisation des descripteurs locaux par des approches basées sur des sacs/chaînes de mots visuels [Sivic 2003, Ros 2009, Perronnin 2010, Avila 2013], etc. Cela peut poser un certain nombre de difficultés dans les traitements ultérieurs qui en sont faits.

Derrière cette brève taxinomie de descripteurs que je viens de donner en fonction de l'information dont on dispose *a priori* concernant les types d'images à décrire, se cache en réalité un foisonnement de descripteurs visuels plus ou moins redondants. Cette redondance rend difficile le choix *a priori* de la signature la plus adaptée à un problème donné. Souvent, les chercheurs tendent donc à sélectionner les descripteurs constituant la signature à utiliser dans leur cas à l'aide d'une comparaison expérimentale de leurs performances [Deselaers 2008]. Mais, les conclusions de ces études diffèrent souvent selon le processus appliqué en aval de l'extraction de la signature, la base d'images considérée, etc. Au final, de nombreux auteurs choisissent d'appliquer une sélection automatique des caractéristiques ou des descripteurs composant la signature [Jain 1997]. Dans [Visani 2011b], nous avons proposé un protocole permettant de caractériser de manière explicite les performances d'un ensemble de descripteurs. L'objectif visé au final est de pouvoir sélectionner, pour une collection d'images donnée et sur la base de cette étude, une signature (composée éventuellement de plusieurs descripteurs) présentant de bonnes propriétés [4].

À l'issue de cette brève présentation des descripteurs visuels, nous allons introduire dans la section suivante le cœur du sujet du présent chapitre, à savoir la description d'images par *clustering* interactif.

2.1.2 Pourquoi extraire des descripteurs par *clustering* interactif ?

La question posée dans le titre de cette section est en fait double : « Pourquoi extraire des descripteurs par catégorisation, et plus spécifiquement par *clustering* ? » d'une part, et « Pourquoi le faire de manière interactive ? » d'autre part.

Tâchons d'abord de répondre à la première question. Nous nous plaçons ici dans le contexte où l'on ne dispose d'aucune information *a priori* concernant le type des images ou des objets représentés dans la collection d'images à décrire. Dans ce contexte, on utilise généralement des signatures visuelles basées uniquement sur des descripteurs génériques (voir section précédente) pour décrire individuellement chacune des images. Dès lors qu'on recherche une représentation synthétique non plus des images mais de la base d'images, il peut être utile de travailler à un niveau plus global, en cherchant des similarités entre les images de la

4. Ces bonnes propriétés concernent l'unicité, le pouvoir discriminant et la robustesse vis-à-vis du bruit de chaque descripteur considéré indépendamment, ainsi que la complémentarité des descripteurs composant la signature. Elles sont évaluées à l'aide de mesures d'évaluation qualitatives et quantitatives que nous avons proposées, paramétrables par la mesure de similarité ou de dissimilarité la mieux adaptée à chaque descripteur.

base (ou certains des éléments extraits depuis ces images) ; l'idée à terme est d'alimenter un moteur de visualisation/navigation par similarité dans les bases d'images. Lorsque les bases d'images sont trop volumineuses pour que l'humain puisse en avoir rapidement un aperçu, **une représentation naturelle passe par des groupes d'images similaires** (le cas échéant, seules certaines images représentatives de chacun des groupes peuvent être présentées à l'humain). C'est dans cette optique que nous avons choisi d'extraire des descripteurs par catégorisation, et plus spécifiquement par *clustering*, puisque nous ne disposons pas *a priori* d'information sur les catégories d'images dans la collection à décrire.

La deuxième question est : « Pourquoi le faire de manière interactive ? ». Vu les applications que nous visons au final, nous souhaitons obtenir à l'issue du *clustering* des groupes d'images qui soient similaires au sens de certains critères humains. Mais, comme souligné ci-avant, les signatures visuelles génériques sur lesquelles nous devrons nous baser pour mener à bien le *clustering* sont de très bas niveau sémantique, et il y a donc peu de chances qu'ils permettent de découvrir des catégories qui satisfassent l'humain, comme nous le détaillerons ci-après.

En fait, à l'heure actuelle, seul un humain est capable de maîtriser le niveau de sémantique requis pour intégrer les éléments de contexte permettant de lever les ambiguïtés dans l'analyse du contenu de ces images. Ce fait est illustré par l'existence de nombreuses recherches basées sur l'utilisation de connaissances humaines [Maillot 2008, Chuan 2011, Depeursinge 2014]. Dans ces travaux, c'est en effet l'humain, souvent expert (par exemple un médecin dans le cas d'images médicales [Depeursinge 2014]), qui fournit les connaissances qui sont ensuite formalisées (typiquement sous la forme d'ontologies), utilisées *in fine* pour l'analyse d'images.

Dans notre contexte où nous ne disposons d'aucune connaissance formalisée de manière explicite sur les collections d'images à décrire, nous ne pouvons utiliser le type d'approches citées ci-dessus. Nous avons donc choisi d'**intégrer l'humain dans le processus de *clustering*** de manière à ce qu'il fournisse incrémentalement une information partielle concernant la pertinence des *clusters*, information qui est utilisée par le système pour corriger itérativement ces derniers. Nous nous focalisons donc ici sur une tâche de *clustering* interactif d'images ou de motifs extraits des images, qui nous permet d'obtenir une organisation de la collection en groupes qui soient plus conformes aux concepts perçus par l'utilisateur dans la collection.

L'information extraite du *clustering* peut être agencée sous la forme d'un descripteur. Selon la méthode de *clustering* considérée, le descripteur obtenu sera de nature statistique (vectorielle) ou structurelle (le plus souvent de forme graphique ou arborescente). Comme mentionné ci-dessus, ce descripteur pourra être utilisé (le cas échéant conjointement avec des descripteurs visuels) pour permettre à l'humain d'avoir un aperçu rapide du contenu de la collection d'images, ou d'y naviguer intuitivement. Ce qui n'empêche pas qu'il pourra en outre être mis à profit par des processus d'indexation, de recherche ou de reconnaissance d'images.

2.1.3 Contenu et organisation du reste du chapitre

La suite de ce chapitre est dédiée à la description de collections d'images de niveaux de structuration divers.

La section 2.2 s'intéresse à l'organisation de collections d'images naturelles pour lesquelles on ne dispose d'aucune information du domaine (en particulier concernant leur éventuelle structuration). Il s'agit d'être capable de décrire la collection au travers de catégories qui font

sens pour un humain.

La section 2.3 est focalisée sur la description de collections d'images plus structurées, à savoir des collections de documents textuels anciens. En contrepartie de cette structuration, on ne dispose d'aucune information *a priori* concernant les scripts ou les langages utilisés. On cherche à décrire ces documents au travers de la découverte de motifs (« invariants ») revenant fréquemment dans les images et d'un niveau sémantique suffisant pour être réutilisés par un expert humain souhaitant naviguer, ou rechercher de l'information, dans la collection.

Ce chapitre se poursuit avec la section 2.4, qui propose une discussion autour des avancées réalisées grâce à nos travaux. Cette discussion met en perspective ces avancées avec le chemin qu'il nous reste à parcourir au regard des questions de recherche abordées, et le cas échéant avec les tendances qui se sont dégagées entre-temps dans la communauté scientifique pour s'attaquer à ces questions.

Les principales perspectives qui se dégagent des travaux présentés dans ce chapitre sont abordées en section 2.5.

Enfin, la section 2.6 permet de synthétiser les faits marquants liés à ces contributions en termes notamment d'encadrements, de publications, de projets et de collaborations nationales et internationales.

2.2 Proposition d'une approche de *clustering* semi-supervisé et interactif de bases d'images tout-venant

2.2.1 Positionnement de l'étude

2.2.1.1 Objectif visé

L'objectif visé dans cette section est l'organisation de bases d'images potentiellement volumineuses et de contenus très hétérogènes, ici des images naturelles tout-venant dont on ne connaît pas *a priori* le type. Ces images sont donc peu structurées, ou bien on ne dispose d'aucune information concernant leur structure sous-jacente. Plus précisément, le problème posé ici est d'organiser ces bases d'images d'une manière qui se rapproche le plus possible des concepts perçus dans la collection par un utilisateur humain, en cherchant à créer des groupes d'images similaires selon des critères humains. En fonction du contexte applicatif, l'utilisateur peut être simplement un particulier souhaitant organiser sa collection de photographies personnelles, ou un expert du domaine d'application (archiviste, journaliste, etc.). Nous nous intéressons en particulier à l'utilisateur expert. Parmi les applications finales possibles, nous sommes en train d'étudier dans le cadre du projet PCSI sur la valorisation du patrimoine khmer la possibilité d'embarquer notre système dans un outil d'assistance à l'organisation de bases d'images pour des archivistes du centre de ressources audio-visuelles Bophana situé à Phnom Penh, au Cambodge. Nous sommes en train d'initier des discussions avec la BnF, avec laquelle nous travaillons déjà dans le cadre du projet DIGIDOC, et qui devrait participer au prochain *workshop* du projet, qui se tiendra à Phnom Penh à l'automne 2014.

Malgré des similitudes certaines avec l'indexation, et l'existence de quelques travaux utilisant le résultat d'un *clustering* pour de la recherche d'images [Rubner 2000, Käster 2003, Chen 2005], **notre objectif diffère de celui de l'indexation traditionnelle** (p. ex. par arbres ou par tables de hachage) sur plusieurs points. Dans son acception traditionnelle, un index est défini comme une structure (le plus souvent multi-dimensionnelle) permettant de regrouper les images ayant des signatures similaires, et de renvoyer aux images originales (par exemple en utilisant des fichiers inversés) [Eakins 1999, Ai 2013]. L'objectif principal de l'indexation traditionnelle est d'optimiser le temps de recherche ultérieure des images de la base[5]. Plus précisément, la plupart des méthodes d'indexation récemment proposées visent à partitionner l'espace de représentation (espace des signatures) en cellules disjointes (appelées en anglais *buckets* dans le cas des fonctions de hachage). La phase de recherche qui s'ensuit se fait alors sur la base des valeurs de l'index avant de porter, le cas échéant, sur les signatures des images des *buckets* concernés.

Les méthodes d'indexation traditionnelles ont deux désavantages majeurs dans le cas qui nous intéresse :

– Premièrement, les *buckets* qu'elles produisent sont généralement de tailles similaires (en termes de nombre d'images). C'est logique, puisque l'objectif est d'optimiser le temps de recherche en utilisant des structures de données équilibrées. Mais, ces *buckets* ne peuvent être conformes aux catégories d'images perçues par l'utilisateur que lorsque ces catégories sont elles-mêmes équilibrées, ce qui est très rarement le cas en pratique. Certaines stratégies visant à regrouper les *buckets* par similarité dans une phase ultérieure de *clustering* peuvent permettre de corriger ce défaut ;

5. Par extension, certains auteurs considèrent que l'indexation consiste à assigner une signature à une image.

– Deuxièmement, les méthodes d'indexation traditionnelles sont adaptées à une recherche d'images de contenu quasi-similaire. En effet, les signatures et les index qui en découlent sont typiquement calculés en utilisant exclusivement des descripteurs visuels génériques décrivant l'apparence visuelle des images (principalement leur couleur et leur texture). Mais, dans la plupart des cas, les souhaits des utilisateurs sont d'un plus haut niveau sémantique et les images retournées par le système ne peuvent donc satisfaire ces souhaits, comme illustré en Figure 2.1 avec un moteur de recherche commercial (ici Google Images) ;

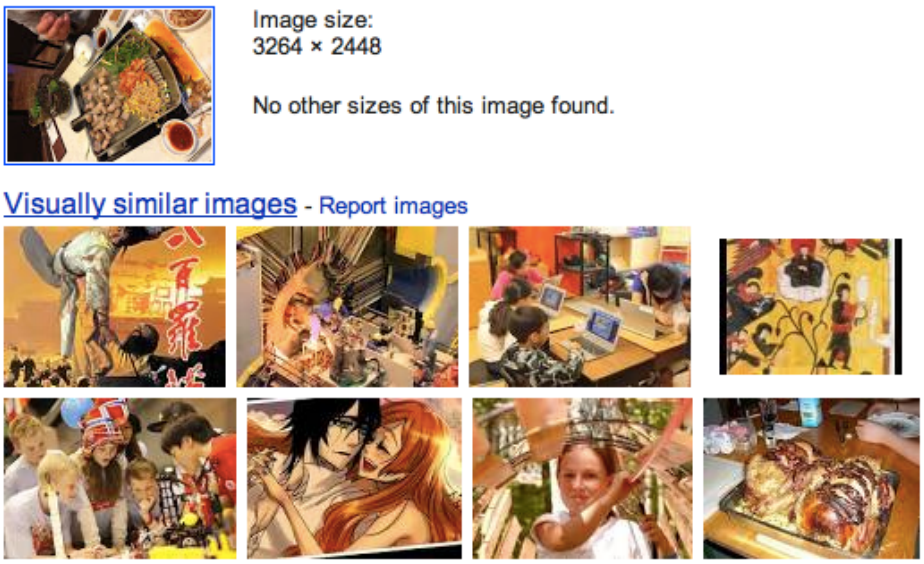


FIGURE 2.1 – Le moteur de recherche de Google Image retourne majoritairement des images non pertinentes pour l'image de requête (en haut à gauche). Ici, l'unique image pertinente (image de nourriture) retournée par le moteur de recherche se trouve en huitième position (seconde ligne, quatrième colonne).

Avec un moteur de recherche *web*, l'utilisateur peut généralement tenter d'améliorer les résultats en ajoutant des mots-clés descriptifs dans le champ de recherche. La recherche est alors raffinée à l'aide du texte associé aux pages *web* contenant les images retournées. Dans le cas où l'on ne dispose pas *a priori* de données textuelles associées aux images (cas très courant dans la pratique), un tel raffinement par une recherche textuelle nécessiterait une phase d'annotation manuelle coûteuse et potentiellement subjective, et/ou une phase d'annotation automatique dont les résultats restent à l'heure actuelle assujettis à la qualité et à l'exhaustivité de la base d'apprentissage [Zhang 2012].

Le « bouclage de pertinence » permet également de réduire partiellement ces désavantages en impliquant l'utilisateur dans la phase de recherche, rendant ainsi la recherche interactive. À chaque itération, l'utilisateur étiquette les images retournées par le système comme pertinentes ou non-pertinentes, et le système adapte ses paramètres à ce retour de l'utilisateur. Les stratégies les plus courantes pour l'adaptation des paramètres du système consistent à modifier la signature de l'image de requête et/ou la mesure de similarité utilisée pour la recherche [Fournier 2001, Zhou 2003]. Le bouclage de pertinence reste donc généralement cantonné à la phase de recherche et ne permet pas d'améliorer l'organisation des données de manière à ce qu'elle convienne mieux aux attentes de l'utilisateur. Il n'est donc pas directement utilisable dans notre cas.

Plus récemment, des techniques visant à rendre l'indexation par table de hachage supervisée [Kulis 2009b, Norouzi 2011, Strecha 2012] ou semi-supervisée [Wang 2012] ont été proposées. L'idée est d'intégrer dès le début de la procédure d'indexation une information de plus haut niveau sémantique, à prendre en compte par le système lors de la création de la structure d'index. Tandis que les méthodes supervisées souffrent rapidement de sur-apprentissage en présence de données annotées bruitées ou en faible nombre, seules peu de méthodes d'indexation par table de hachage sont interactives [Grewe 1995]. Plus généralement, peu de travaux s'intéressent à la problématique de l'indexation interactive, et restent généralement cantonnés à des applications très particulières comme par exemple la télédétection [Schroder 2000].

C'est dans ce contexte que nous avons proposé la **thèse de Lai Hien Phuong**[6] [Lai 2013a, Lai 2014a], portant sur la conception d'une approche capable d'**adapter l'organisation d'une collection d'images naturelles tout-venant aux catégories perçues par l'utilisateur** en impliquant ce dernier dès la phase de description de la collection, et non lors d'une éventuelle phase ultérieure de recherche. Afin de s'adapter à l'organisation des données, naturellement non équilibrée dans la plupart des cas, nous avons choisi de baser notre système sur une méthode de *clustering*.

2.2.1.2 Approches existantes

Dans cette section, je commence par donner une typologie grossière des approches de *clustering*, avant de passer en revue les principales approches de *clustering* non-supervisé et leurs applications à la catégorisation d'images. Puis, je présente diverses applications d'approches de *clustering* semi-supervisé dans le domaine de l'analyse d'images), avant d'aborder la notion de *clustering* semi-supervisé interactif.

Jusqu'à aujourd'hui, une multitude de méthodes de ***clustering*** ont été proposées, dans des domaines très variés allant bien au-delà de l'analyse d'images. Cela rend difficile l'établissement d'une typologie exhaustive et claire des méthodes de *clustering* [Jain 2010]. Très grossièrement, on distingue les méthodes qui produisent des *clusters* « à plat » de celles qui présentent les *clusters* sous une forme hiérarchique (arborescente dans la plupart des cas). Parmi les méthodes produisant des *clusters* à plat, certaines sont basées sur la recherche de *clusters* compacts, c'est-à-dire pour lesquels la similarité intra-*cluster* est supérieure à la similarité inter-*cluster*. C'est le cas des méthodes parfois dites « par partitionnement ». Dans certains cas, ces approches tendent à définir dans l'espace de représentation des *clusters* de formes prédéfinies (p. ex. hypersphériques dans le cas des k-moyennes avec une distance Euclidienne). D'autres approches « à plat » cherchent plutôt à s'appuyer sur la distribution des données dans l'espace de représentation afin d'obtenir des *clusters* de formes variées, séparés par des régions de faible densité. Les approches produisant une structure hiérarchique de données le font soit de manière ascendante, soit de manière descendante (par fusions ou divisions successives de groupes d'images selon des critères de similarité).

Dans le contexte de l'analyse d'images, même si la plupart des travaux existants se sont focalisés sur l'indexation d'images dans un but de recherche, il existe quand même des travaux sur le ***clustering* non-supervisé d'images**. Les résultats du *clustering* sont le plus souvent utilisés pour la visualisation [Schaefer 2010, Jing 2012] et/ou la navigation

6. Doctorante en co-supervision entre le L3i et l'équipe MSI du laboratoire UMMISCO (UPMC / IRD).

[Ma 1999, Krishnamachari 1999, Chen 2000, Borth 2008] dans des collections d'images. Ces outils sont typiquement basés sur un *clustering* hiérarchique qui calcule une structure représentative des similarités entre images. Cela permet à l'utilisateur de visualiser la collection de manière dynamique au travers d'opérations telles que le *zoom* dans la base d'images. Certains travaux utilisent également le *clustering* pour l'indexation en vue d'une recherche ultérieure [Rubner 2000, Käster 2003, Chen 2005].

La principale limitation de ces approches de *clustering* entièrement non-supervisé est que les groupes sont constitués sur la foi de signatures visuelles génériques extraites automatiquement des images. Ces signatures véhiculent donc généralement une information de bas niveau sémantique et ne correspondent donc que rarement aux critères de similarité humains. Afin de résoudre ce fossé sémantique, plusieurs stratégies sont possibles. La première stratégie possible est, comme évoqué en section précédente pour l'indexation, de prendre en compte une information textuelle de plus haut niveau sémantique lors du *clustering* [Gao 2005, Agrawal 2007]. Cette information textuelle peut être par exemple issue soit d'une annotation (manuelle ou automatique [Zhang 2012]), soit du texte entourant l'image dans des pages *web*. Ces approches souffrent des mêmes limitations que dans le cas de l'indexation (voir plus haut).

Une autre piste pour réduire ce fossé sémantique est d'intégrer en entrée du *clustering* des descripteurs de plus haut niveau sémantique, décrivant par exemple la nature et l'agencement spatial de certains objets présents dans les images[7] [Du 2009, Chen 2012], préalablement localisés à l'aide de détecteurs d'objets [Felzenszwalb 2010]. Parce que la détection d'objets est appliquée sur toutes les images de la base, le processus de *clustering* est alourdi en termes de temps de calcul, sans réelle garantie que le résultat du *clustering* soit plus proche des attentes de l'utilisateur au final. D'autant plus que, dans notre contexte applicatif, nous n'avons aucune information *a priori* concernant les objets que l'on peut s'attendre à trouver dans les images.

La mise en œuvre du *clustering* de manière semi-supervisée peut partiellement résoudre ce problème. La plupart des algorithmes de ***clustering* semi-supervisé** sont issus d'algorithmes de *clustering* usuels (c'est-à-dire non supervisé), dans lesquels on intègre une certaine part d'information supervisée, généralement collectée en amont de l'apprentissage, le plus souvent auprès (ou sous la supervision) de l'humain. L'algorithme cherche à apprendre de manière complètement automatique une solution de *clustering* qui tire parti conjointement de la distribution des données dans l'espace des signatures, et de cette information supervisée partielle.

Il existe plusieurs manière d'intégrer de l'information supervisée dans le processus de *clustering*. L'information supervisée peut porter sur l'étiquette (numéro du *cluster* par exemple) de certaines images de la base. Peu de travaux utilisant directement ce type d'information dans du *clustering* d'images [Dubey 2010] ou de vidéos [Kinoshita 2005] ont été présentés dans la littérature. Pour des raisons de performance essentiellement, les chercheurs préfèrent généralement intégrer l'information supervisée au travers de contraintes entre paires d'images [Wagstaff 2001, Klein 2002, Basu 2004, Davidson 2005, Grira 2005, Kulis 2009a, Lelis 2009]. Certaines méthodes, comme COP-kmeans [Wagstaff 2001], n'autorisent aucune violation de contraintes tandis que d'autres, comme par exemple HMRF-kmeans [Basu 2004], les pénalisent. De manière à minimiser le nombre de contraintes à inclure dans le *clustering* semi-supervisé (et ainsi la complexité calculatoire tout comme l'effort d'annotation de l'humain), certains travaux se focalisent sur la manière de sélectionner, en amont de l'apprentissage, les exemples à proposer à l'utilisateur pour étiquetage. Cette sélection peut par exemple se faire en utilisant

7. On parle parfois de descripteurs « spatiaux ».

des méthodes d'apprentissage actif [Grira 2008].

En parallèle de ces travaux, des chercheurs se sont intéressés à rendre interactif le *clustering* semi-supervisé sous contraintes, en présentant itérativement à l'utilisateur des exemples bien choisis. Le domaine d'application phare est la recherche d'information et en particulier le *clustering* de documents textuels, avec la méthode Scatter/Gather de [Cutting 1992] ou encore la méthode présentée au chapitre 2 de [Basu 2008], complémentaires en termes d'applications. Si les applications au domaine de l'image sont encore très rares, on peut citer les travaux présentés très récemment dans [Biswas 2012][8], qui reposent sur une extension semi-supervisée interactive de l'algorithme de *clustering* par Classification Ascendante Hiérarchique (CAH). C'est dans ce contexte scientifique peu investigué du ***clustering* semi-supervisé interactif d'images** que nous avons initié en 2009 les travaux décrits dans la suite de cette section.

Les deux sections suivantes visent à présenter l'approche que nous avons adoptée dans le cadre de ces travaux, et les principales contributions liées à ces derniers.

2.2.2 Aperçu de l'approche proposée

Le fonctionnement de l'approche que nous avons proposée est illustré en Figure 2.2, et détaillé ci-après.

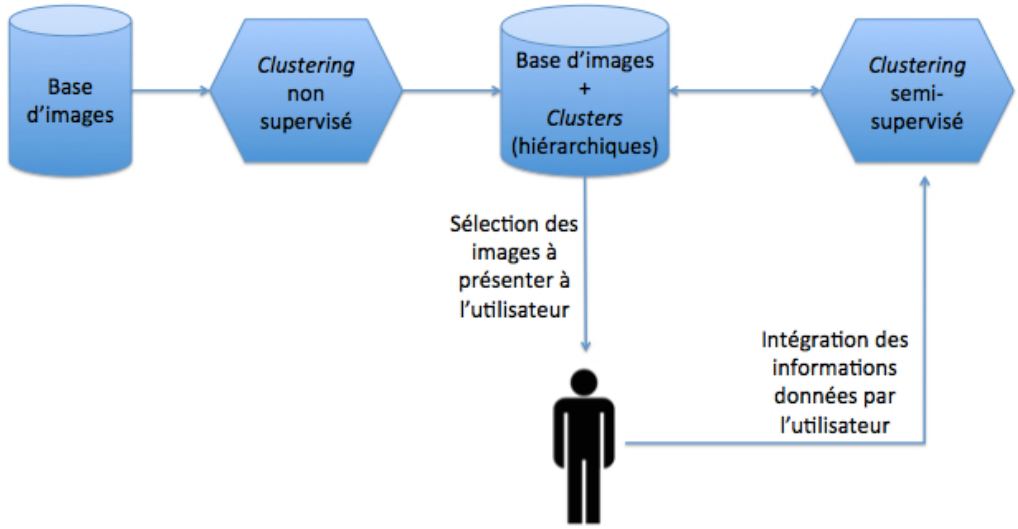


FIGURE 2.2 – Vue globale de l'approche de *clustering* semi-supervisée et interactive proposée.

À partir de la base d'images tout-venant à décrire et sans information *a priori* concernant son contenu, un *clustering* non-supervisé est appliqué. Puis, on entre dans une **boucle itérative d'interaction** avec l'utilisateur, où ce dernier apporte incrémentalement des informations au système afin de corriger la solution initiale de *clustering*.

Plus précisément, lors de chaque itération interactive, seules certaines images (sélectionnées en fonction de la distribution des *clusters* correspondants et des interactions précédentes) sont présentées à l'utilisateur, de manière à minimiser l'effort de ce dernier. L'utilisateur fournit alors

8. Nous reviendrons sur ces travaux en section 2.2.5.

un retour de pertinence concernant quelques-unes de ces images grâce à l'interface montrée en Figure 2.3. Grâce à cette interface, l'humain peut confirmer qu'une image a bien été rangée dans le groupe qu'il souhaite (c'est le cas des images détourées de bleu dans la Figure 2.3). Si ce n'est pas le cas, l'humain peut également déclarer qu'une image a été rangée dans un mauvais groupe (images détourées de rouge), ou même glisser et déposer des images d'un *cluster* vers un autre *cluster*. Pour prendre ces décisions, l'humain peut inférer la sémantique associée à un *cluster* donné en se basant sur les prototypes et les images du *cluster* qui lui sont présentées dans l'interface. L'itération interactive se poursuit jusqu'à ce que l'utilisateur pense qu'il a donné suffisamment d'information au système, voire qu'il n'ait plus aucun changement à apporter dans les étiquettes des images qui lui sont présentées.

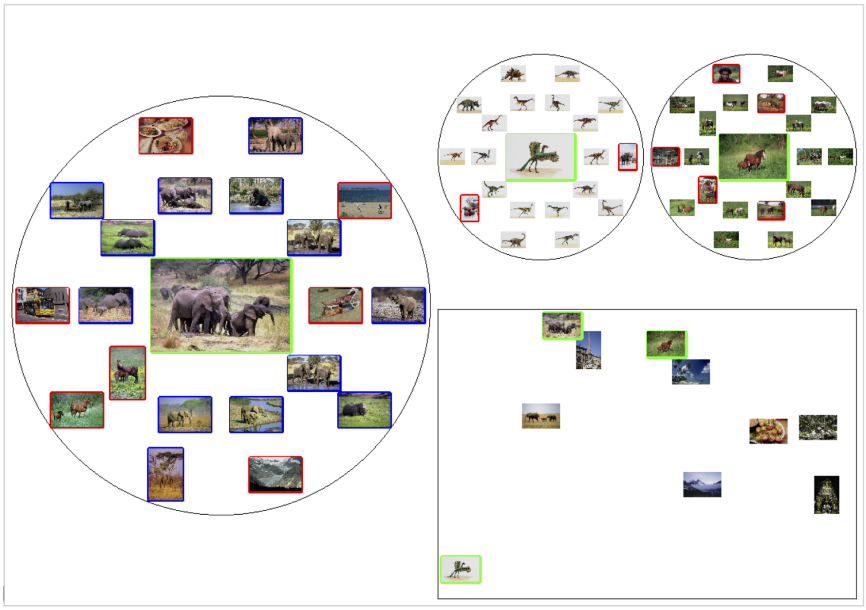


FIGURE 2.3 – Interface permettant à l'utilisateur d'interagir avec les images. Le rectangle en bas à droite est le plan principal composé des deux premiers axes principaux obtenus par ACP, dans lequel les prototypes (images les plus représentatives) de chaque *cluster* sont représentés. Les cercles correspondent aux détails des *clusters* sélectionnés par l'utilisateur (par simple clic sur leur prototype dans le plan principal). *Figure extraite de [Lai 2013a].*

À l'issue de chaque itération interactive, le système déduit des retours de l'utilisateur un ensemble de contraintes qui sont intégrées dans une phase de re-*clustering* semi-supervisé. Nous avons pour cela conçu une méthode de *clustering* semi-supervisé interactif, qui sera détaillée plus loin. Puis, les résultats de ce re-*clustering* sont présentés à l'utilisateur pour interaction.

Cette boucle interactive est réitérée jusqu'à ce que la solution de *clustering* soit assez stable et/ou que ses résultats satisfassent l'humain.

2.2.3 Principales contributions

Les **contributions liées à ce travail sont principalement d'ordre méthodologique** et reposent essentiellement sur deux points. Premièrement, nous avons mené une comparaison théorique et expérimentale des différents algorithmes de *clustering* non supervisé de l'état de l'art, combinés avec différentes signatures d'images génériques, afin de sélectionner la configu-

ration la mieux adaptée à notre contexte. Deuxièmement, nous avons proposé une extension semi-supervisée interactive de l'algorithme de *clustering* non-supervisé sélectionné à l'issue de cette comparaison (à savoir BIRCH). Ces contributions principales font l'objet des deux sections suivantes.

2.2.3.1 Comparaison formelle et expérimentale de différentes méthodes de *clustering* non supervisé dans le cas d'images tout-venant

a) Comparaison formelle

Nous avons dressé dans [Lai 2012a] une comparaison formelle dont l'objectif est de sélectionner un ensemble de méthodes de *clustering* non supervisé qui soient adaptables au contexte semi-supervisé et interactif auquel on s'intéresse. Cette comparaison porte au total sur 19 méthodes de *clustering* non-supervisé, parmi lesquelles 10 méthodes « à plat » – plus précisément 6 méthodes par partitionnement (k-moyennes, k-médoïdes, CLARA, CLARANS, ISODATA et SOM) et 4 méthodes basées sur la densité des données (EM, DBSCAN, OPTICS, CLIQUE) – et 9 méthodes hiérarchiques (STING, DIANA, *Minimum Spanning Tree*, CAH, BIRCH, CURE, R-tree, SS-tree, SR-tree).

Afin de pouvoir aisément découper ou regrouper des *clusters* sur la foi des retours de l'utilisateur, et aussi afin de pouvoir facilement nous adapter à des applications finales de navigation ou d'indexation, il est souhaitable de sélectionner des méthodes produisant une structure hiérarchique de données. Il faut également que les méthodes choisies soient *a priori* adaptées à des grands volumes de données (en particulier parce que les signatures génériques issues des images sont généralement volumineuses), peu sensibles aux valeurs aberrantes ou du moins qui permette leur détection, et peu dépendantes de leurs paramètres pour pouvoir s'adapter à des bases d'images de contenus variés. Afin de permettre l'interactivité avec l'utilisateur, il faut également que les méthodes sélectionnées soient incrémentales et peu complexes en temps de calcul. En effet, les phases de re-*clustering* sont exécutées en-ligne avec l'utilisateur.

Cette étude formelle nous a permis de sélectionner quatre méthodes qui vérifient (au moins en grande partie) ces conditions. Il s'agit des méthodes SR-tree, R-tree, CAH et BIRCH. Pour plus de détails sur cette comparaison formelle, merci de se référer à [Lai 2012a].

b) Comparaison expérimentale

Toujours dans dans [Lai 2012a], nous avons comparé expérimentalement les quatre méthodes sélectionnées avec une variante de la méthode par partitionnement très prisée de la communauté scientifique k-moyennes (*global k-means* [Likas 2003]). Afin d'évaluer (dans une certaine mesure) leur capacité de passage à l'échelle, nous avons mené cette comparaison expérimentale sur 4 bases d'images tout-venant de taille croissante, et plus précisément dont le nombre d'images varie de 1000 à 30000 (Wang, Pascal-VOC 2006, Caltech101 et Corel30k). Nous avons dans un premier temps mesuré leurs performances de manière non supervisée, avec la mesure de performance interne *Silhouette-Width* (SW). Puis, afin de comparer les solutions de *clustering* obtenues avec la vérité-terrain associée (catégories d'appartenance des images, préalablement définies par – ou sous la supervision de – l'humain), nous avons utilisé différents types de mesures d'évaluation externes. Plus d'informations concernant ces mesures d'évaluation sont données en section 2.2.4.

Outre la présence d'une comparaison formelle préalable, une originalité de cette comparaison expérimentale de méthodes de *clustering* d'images au regard de celles précédemment publiées (par exemple [Käster 2003]) repose sur la combinaison de ces méthodes avec diverses signatures génériques. Cela permet de réduire le biais dans la comparaison expérimentale des méthodes de *clustering*. Nous avons pour cela choisi d'utiliser des signatures variées, à savoir, d'une part, une signature globale basée à la fois sur des histogrammes couleur, des filtres de Gabor (descripteur de texture) et les moments de Hu (descripteur de forme) et, d'autre part, plusieurs signatures issues d'une description locale discrétisée par « sacs de mots visuels » [Sivic 2003]. Pour la description locale, nous avons expérimenté SIFT [Lowe 2004] et quelques-unes de ses variantes couleur : rgSIFT, CSIFT, RGB SIFT et Opponent-SIFT [Van de Sande 2010].

La principale conclusion de cette comparaison expérimentale est que l'**algorithme BIRCH** (*Balanced Iterative Reducing and Clustering using Hierarchies*) [Zhang 1996], utilisé conjointement avec une signature **rgSIFT**[9], donne les meilleurs résultats selon les mesures externes. C'est-à-dire que la combinaison rgSIFT+BIRCH fournit la solution de *clustering* la plus proche des catégories de la vérité-terrain (fournie par l'humain).

2.2.3.2 Proposition d'une approche de *clustering* semi-supervisé interactif

En raison de ses bonnes propriétés et de ses excellentes performances en pratique, c'est donc l'algorithme BIRCH que nous avons choisi d'adapter à notre contexte semi-supervisé et interactif. Cet algorithme est décrit en annexe B (section B.2) et un exemple de structure arborescente produite par BIRCH est donné en Figure B.1, page 161.

Très grossièrement, cet algorithme construit de manière descendante un « arbre CF » dont les « feuilles CF » contiennent les enregistrements d'entrée. Ces feuilles sont équilibrées en termes de nombre d'enregistrements, sauf quelques feuilles qui contiennent des valeurs isolées/aberrantes écartées des autres par l'algorithme (ce qui nous permet le cas échéant d'appliquer un post-traitement particulier à ces valeurs isolées/aberrantes). Afin de déduire, à partir des feuilles CF, des *clusters* d'images similaires – qui, eux, ne sont pas forcément équilibrés, comme souligné plus haut –, un algorithme de *clustering* « à plat » peut être employé (p. ex. celui des k-moyennes). Ce *clustering* s'applique directement sur les feuilles CF, décrites au travers d'un « vecteur CF » caractéristique de son contenu dans l'espace initial de représentation. Si l'on note p la taille de l'espace initial de représentation, alors le vecteur CF caractéristique de chaque feuille est de taille $p + 2$.

L'un des questionnements principaux auxquels nous avons dû faire face dans la conception de notre approche concerne la manière d'intégrer efficacement dans le processus de catégorisation les quelques informations retournées par l'utilisateur à chaque itération. Le problème est double. D'une part, les informations retournées par l'utilisateur sont d'un niveau sémantique supérieur à celui des signatures visuelles de bas niveau utilisées pour mettre en œuvre le *clustering* initial. D'autre part, vu que l'on souhaite minimiser l'effort à fournir par l'utilisateur afin d'éviter de le lasser, ces informations sont forcément en nombre limité au regard du volume d'images de la collection.

La première question à laquelle nous avons cherché à apporter une réponse est donc : « Com-

9. Avec un histogramme normalisé de mots visuels extraits en appliquant les k-moyennes sur les descripteurs ; le nombre de mots (ici 200) a été sélectionné par compromis entre performance et temps de calcul.

ment intégrer dans le processus de catégorisation semi-supervisée les informations fournies par l'utilisateur à chaque itération interactive ? ». Afin de répondre à cette question, nous nous sommes dans un premier temps tournés vers la littérature. Pour cela, nous avons comparé dans [Lai 2012b] différentes méthodes de *clustering* semi-supervisé. Cette étude nous a permis de déterminer que les techniques de *clustering* semi-supervisé où l'information supervisée est intégrée dans l'algorithme sous la forme de contraintes *must-link* ou *cannot-link* entre paires d'images (voir section 2.2.1.2) sont les plus efficaces. En effet, un seul clic de l'utilisateur peut déclencher de très multiples contraintes entre paires d'images. Dans les approches de la littérature, les contraintes entre paires d'images sont généralement prises en compte directement dans l'optimisation de la fonction objective de l'algorithme de re-*clustering* semi-supervisé.

Dans notre contexte où nous disposons d'une structure hiérarchique de *clustering*, nous avons cherché à **tirer parti de cette structure hiérarchique** afin d'intégrer plus efficacement les informations données par l'utilisateur dans le système. Cela constitue une originalité de notre approche.

Pour cela, nous avons proposé dans [Lai 2014b] une méthode de *clustering* semi-supervisé interactif où, à chaque itération interactive, l'intégration des retours de l'utilisateur (sous la forme de contraintes entre paires d'images) déclenche une redéfinition des feuilles de l'arbre BIRCH (au sens des images qu'elles contiennent), et la définition de contraintes *must-link* ML_{CF} (respectivement *cannot-link* CL_{CF}) entre ces feuilles. À l'issue de chaque itération interactive, ce sont ces nouvelles feuilles (et non les images en elles-mêmes) que l'algorithme de re-*clustering* final cherchera à réorganiser, en pénalisant dans sa fonction objective les éventuelles violations de contraintes ML_{CF} et CL_{CF}.

L'idée de la méthode est d'« inciter » le processus de re-*clustering* à placer dans un même *cluster* des groupes d'images que l'humain considère comme similaires, mais qui ne sont pas suffisamment similaires dans l'espace de représentation défini par les signatures visuelles de bas niveau sémantique[10] pour que l'algorithme BIRCH original les range d'emblée dans le même *cluster*. La même idée s'applique bien sûr pour les groupes d'images que l'utilisateur considère comme dissimilaires (le système étant alors incité à les ranger dans des *clusters* différents).

Dans un souci d'efficacité, la méthode s'appuie sur deux types de regroupements d'images de taille intermédiaire, respectivement les « noyaux » et les « voisinages » pour, respectivement, diviser certaines feuilles CF afin d'obtenir de nouvelles feuilles, et calculer les contraintes par paires ML_{CF} et CL_{CF} entre ces nouvelles feuilles à partir des contraintes ML et CL entre paires d'images.

En annexe B (section B.3), je décris la méthode que nous avons proposée sous une forme différente de nos publications précédentes, plus synthétique mais néanmoins exhaustive et soulignant, à chaque étape du processus, la finalité de cette étape dans le processus global.

On en arrive à la deuxième question à laquelle nous avons cherché à apporter une réponse, et qui peut se formuler ainsi : « Comment répercuter au mieux au niveau de la solution de *clustering* les retours fournis par l'utilisateur à chaque itération interactive, sachant que ces retours sont par essence en nombre limité ? ». Notre choix pour solutionner ce problème a été, à partir de la première liste de contraintes « primaires » entre paires d'images (que l'on peut déduire des retours positifs et négatifs de l'utilisateur à chaque itération interactive), de **déduire un ensemble de contraintes « secondaires »** qui seront également prises en compte pour la re-définition des feuilles CF et la définition des contraintes entre paires de feuilles. Pour

10. Ici l'histogramme normalisé de 200 mots visuels issu du descripteur rgSIFT, voir plus haut.

cela, nous avons dans un premier temps tenté d'« accumuler » les contraintes sur l'ensemble des itérations $1, \ldots, t$: par exemple, si lors de l'itération interactive courante t, l'utilisateur étiquette l'image X comme appartenant au *cluster* C_1 et qu'il avait préalablement étiqueté l'image Y comme appartenant à C_1 (lors d'une itération interactive précédente $t' < t$), alors l'image X devient liée par une contrainte *must-link* à Y. *Idem* pour les contraintes *cannot-link*. Nos expérimentations ont montré que, dans ce cas, la solution de *clustering* s'approche plus de la solution de catégorisation donnée dans la vérité-terrain que si l'on n'utilise que les contraintes primaires, et ce, en un nombre d'itérations interactives assez faible. Mais, chaque itération devient très lente en raison de la prise en compte de ces multiples contraintes secondaires dans l'algorithme de re-*clustering*, qui accroît la complexité calculatoire de la phase de re-*clustering*. Dans notre contexte applicatif où chaque itération interactive est menée en-ligne avec l'utilisateur, cette solution n'est pas réaliste.

Nous avons donc expérimenté quatre autres stratégies de sélection des contraintes primaires et secondaires, afin de trouver un bon compromis entre amélioration des performances et complexité en temps de calcul. Il s'est avéré que, dans nos expérimentations, la stratégie offrant le meilleur compromis est une stratégie qui sélectionne à la fois certaines des contraintes déduites des retours de l'utilisateur lors de l'itération courante, et certaines des contraintes secondaires obtenues en accumulant les retours reçus lors des itérations $1, \ldots, t$. L'idée de cette stratégie de sélection est de ne conserver que les contraintes dont on peut penser qu'elles auront un fort impact sur la re-définition des feuilles CF et la définition des contraintes entre paires de feuilles ; à savoir les contraintes ML entre les images les plus distantes dans l'espace de représentation, et certaines des contraintes CL entre les images les plus proches dans l'espace de représentation (les contraintes CL secondaires étant de plus filtrées de manière à ne garder que celles qui sont le moins redondantes avec les contraintes précédemment données). On peut noter ici des similitudes avec l'idée sous-jacente aux techniques d'**apprentissage actif** [Cohn 1996, Cord 2008], qu'il conviendrait d'approfondir. Plus de détails sur les stratégies de sélection/déduction de contraintes étudiées sont données dans la section 5.2. de [Lai 2013a], et ce point sera rediscuté dans la suite de ce chapitre.

Dans la section suivante, nous nous intéressons aux protocoles et aux mesures que nous avons proposés afin de caractériser les performances de méthodes de *clustering* semi-supervisé interactif de bases d'images.

2.2.4 Protocoles/mesures proposés pour la caractérisation de performances

Dans la littérature, deux types de mesures ont été présentées pour l'évaluation d'un résultat de *clustering* d'images [He 2004, Jain 2010] : les mesures internes et les mesures externes. Une définition très résumée de ces deux types de mesures est donnée ci-après. Pour plus de détails, et les références correspondantes, le lecteur est invité à se référer à la section 3.4. de [Lai 2013a].

Dans le cas où l'on ne dispose pas de la vérité-terrain (c'est-à-dire des groupes d'images annotés par – ou sous la supervision de – l'humain), on utilise généralement des **mesures d'évaluation « internes »** telles que la compacité (aussi appelée homogénéité), la séparation, et des mesures telles que celle de He *et al.*, *Silhouette Width* (SW), la statistique Γ de Hubert ou sa version normalisée. Les mesures internes permettent de vérifier que les images dotées de signatures similaires sont rangées dans le même groupe, tandis que les images ayant des signatures dissimilaires sont rangées dans des groupes différents. Mais, lorsque les signatures sont de bas niveau sémantique (ce qui est notre cas), les mesures internes ne permettent pas de

déterminer si les groupes d'images constitués font sens d'un point de vue sémantique, ou non.

Les **mesures d'évaluation « externes »**, en revanche, permettent d'évaluer la pertinence sémantique des *clusters* produits. En effet, elles comparent les *clusters* avec les catégories de la vérité-terrain (par la suite appelées « classes »). La comparaison entre les *clusters* et les classes se fait le plus souvent selon deux critères : l'homogénéité (les objets d'un même *cluster* doivent provenir d'une même classe) et la complétude (les objets en provenance d'une même classe doivent être assignés à un même *cluster*). Ces critères peuvent être mesurés soit de manière globale à l'échelle du *cluster* ou de la classe (mesure de pureté, mesure d'entropie et ses variantes, $F-$mesure, $V-$mesure, statistique Γ, etc.), soit de manière combinatoire, c'est-à-dire en étudiant le nombre de paires d'objets qui sont rangés dans le même groupe à la fois dans la solution de *clustering* et dans la vérité-terrain (*rand-index*, indice de Jaccard, indice de Fowlkes-Mallows, métrique de Mirkin, etc.).

Dans notre approche de *clustering* semi-supervisé interactif, nous utilisons une caractérisation de la qualité du *clustering* en deux occasions :

- À chaque itération interactive, pour sélectionner les images à présenter à l'utilisateur afin qu'il interagisse avec elles ;
- Pour évaluer l'adéquation d'une solution de *clustering* donnée avec la vérité-terrain (en fin de chaîne ou à la fin de chaque itération interactive, par exemple si l'on souhaite évaluer expérimentalement la convergence de la méthode).

À chaque itération interactive, notre objectif est de sélectionner les images à proposer à l'utilisateur pour interaction dans l'interface (Figure 2.3, p. 35). En effet, le choix des c *clusters* et des p images par *cluster* à présenter à l'utilisateur à chaque itération interactive est déterminant dans les changements apportés (le cas échéant) à la solution de *clustering* lors du re-*clustering*.

Plusieurs stratégies ont été étudiées et évaluées expérimentalement. Concernant les *clusters*, en l'absence de choix de l'utilisateur, ceux-ci sont sélectionnés au hasard (stratégie choisie en raison de ses meilleures performances que diverses stratégies basées sur les dissimilarités entre *clusters* dans l'espace de représentation). Concernant les images, la stratégie la plus performante parmi celles que nous avons testées consiste à présenter les $p + 1$ images les plus représentatives (parmi lesquelles l'image de prototype est la plus représentative) et les p images les moins représentatives de chaque *cluster* sélectionné. L'idée est de sélectionner les images pour lesquelles un éventuel retour de l'utilisateur aurait potentiellement les plus grandes répercussions sur la solution de *clustering*, afin de minimiser l'effort à fournir par l'utilisateur. Pour évaluer la représentativité d'une image au sein de son *cluster* d'appartenance, nous utilisons sa contribution à la mesure interne SW [Rousseeuw 1987]. Ce point sera ré-évoqué en section suivante.

En fin de chaîne (voire à la fin de chaque itération interactive), l'objectif est une évaluation de haut niveau sémantique de l'adéquation entre la solution de *clustering* proposée et les catégories de la vérité-terrain. Nous pouvons donc, dans ce cas, utiliser des mesures d'évaluation externes. Afin d'automatiser notre campagne intensive d'évaluation, nous avons implémenté un agent logiciel qui simule le comportement de l'utilisateur humain en se basant sur la vérité-terrain pour fournir ses retours au système (à la manière d'un « oracle »), comme détaillé dans [Lai 2013b]. Outre le fait d'automatiser les tests, nous avons choisi un protocole basé sur un agent afin de pouvoir comparer équitablement, selon une même vérité-terrain, notre approche vis-à-vis d'autres méthodes (et notamment de la méthode HMRF-kmeans [Basu 2004] que nous

avons adaptée à notre contexte interactif). Les conditions d'interaction de l'agent utilisateur sont les mêmes que celles de l'humain. C'est-à-dire que l'agent ne peut interagir à chaque itération interactive qu'avec un nombre c fixé de *clusters* (choisis au hasard), et un nombre fixé p d'images par *cluster*, images sélectionnées selon la procédure détaillée plus haut [11]. Nous supposons que l'agent utilisateur infère la classe d'un *cluster* donné comme étant la classe majoritaire parmi les images qui lui sont présentées pour ce *cluster*.

2.2.5 Bilan et améliorations possibles

Nous avons proposé une approche de *clustering* semi-supervisé et interactif originale, qui repose sur une extension semi-supervisée de l'algorithme de *clustering* hiérarchique BIRCH. Elle repose sur des modifications locales de la structure hiérarchique guidées, à chaque itération interactive, par les retours de l'utilisateur. Ces modifications portent sur des divisions des feuilles existantes afin de s'adapter aux retours de l'utilisateur, et un re-*clustering* mené directement au niveau des nouvelles feuilles. Afin de répercuter au mieux au niveau de la solution de *clustering* les quelques retours fournis par l'utilisateur à chaque itération interactive, nous avons étudié diverses stratégies pour déduire des contraintes secondaires à partir des contraintes entre paires d'images calculées selon ces retours, et nous avons pu observer que celle qui donnait le meilleur compromis entre temps de calcul et performance [12] tient compte à la fois de la distribution des données dans l'espace de représentation initial (celui des signatures visuelles de bas niveau sémantique), et des retours précédemment fournis par l'utilisateur.

Nous avons en outre proposé un protocole expérimental basé sur un agent utilisateur, qui nous permet de comparer équitablement, à partir d'une vérité-terrain figée, notre approche à d'autres méthodes. Nos expérimentations montrent que, selon ce protocole expérimental, notre approche produit des groupes d'images plus sémantiques que d'autres méthodes de la littérature. C'est-à-dire qu'ils correspondent mieux à la vérité-terrain annotée par – ou sous la supervision de – l'humain que les groupes formés avec du *clustering* non-supervisé, ou même que les groupes découverts avec d'autres approches semi-supervisées (notamment une variante de HMRF-kmeans [Basu 2004] que nous avons adaptée à notre contexte interactif). Une analyse quantitative des résultats est disponible en annexe E (article [Lai 2013b]) et dans [Lai 2013a].

Ces travaux nous ont néanmoins permis de mettre en évidence plusieurs pistes d'amélioration, parmi lesquelles les trois principales sont détaillées ci-après.

Une première piste qui mériterait d'être explorée est celle de l'**apprentissage de distance** [Yang 2006]. Ce type d'approches permet d'apprendre une distance (dans l'espace initial de représentation) en fonction des interactions de l'utilisateur. La distance ainsi apprise est donc *a priori* plus en adéquation avec les dissimilarités perçues par l'utilisateur que la distance Euclidienne que nous utilisons jusqu'à présent. Une telle distance pourrait être utilisée à de nombreuses occasions dans notre approche. Dans la fonction objective de l'algorithme de re-*clustering*, elle pourrait permettre de répercuter de manière plus globale dans la solution de *clustering* les effets localement induits par les contraintes entre paires de feuilles. En ce qui concerne l'espace de présentation des prototypes d'images à l'utilisateur (*cf.* Figure 2.3), nous pourrions substituer au plan 2D composé des deux premiers axes principaux (obtenus par ACP), actuellement utilisé, un espace de présentation obtenu par *Multi Dimensional Scaling* (MDS) calculé à partir de cette distance. Un autre exemple où une telle distance pourrait être utile est

11. Sur la foi de nos expérimentations, nous avons fixé les valeurs de ces paramètres à $c = 10$ et $p = 10$.
12. En termes d'adéquation entre la solution de *clustering* proposée par le système et celle de la vérité-terrain.

celui de la sélection des images à présenter à l'utilisateur, actuellement basée sur des mesures de distances *intra-cluster* et *inter-cluster*. Enfin, elle pourrait être utilisée pour caractériser les similarités entre de nouvelles images et les groupes d'images constitués interactivement, dans le cadre d'une application finale à la recherche d'images par le contenu (dans l'optique de retourner à l'utilisateur des images plus en adéquation avec ses attentes).

Le choix de la manière de mener l'apprentissage de distance n'est cependant pas évident *a priori*. S'il est bien sûr possible de le mettre en œuvre lors de la deuxième étape de l'algorithme de re-*clustering* des feuilles, à la manière de l'algorithme MPCK-Means présenté dans [Bilenko 2004], nous pouvons nous interroger sur la capacité d'une distance apprise globalement à capturer les subtilités des dissimilarités entre catégories perçues par l'utilisateur. Des réponses à ces questionnement ne pourront être apportées qu'au travers d'une étude fouillée de la littérature, suivie d'expérimentations pratiques.

Une deuxième piste d'amélioration concerne l'interaction avec l'utilisateur. Même si ses performances sont très satisfaisantes selon notre protocole expérimental, dans sa forme actuelle, notre approche ne gère pas réellement les éventuelles incohérences dans les retours donnés par l'utilisateur. Cela peut se justifier dans une certaine mesure étant données les applications visées, qui relèvent des humanités digitales où l'organisation des images souhaitée par les utilisateurs (experts du domaine tels que des archivistes) suit généralement une typologie précise. Néanmoins, le cas d'un utilisateur dont les critères de catégorisation varieraient avec le temps se présente fréquemment dans la pratique, dès lors que l'on sort de ces applications, et mérite donc d'être abordé.

Les raisons possibles d'une évolution dans les critères de catégorisation de l'humain sont multiples. Elle peut par exemple provenir d'une évolution du contenu de la base, ou d'une inconsistance dans le comportement de l'utilisateur. Le cas de l'évolution du contenu de la base sera discuté dans les perspectives générales de ce chapitre (section 2.5). Ici, nous nous focalisons sur le cas où l'évolution provient d'un comportement inconsistant de l'utilisateur (par exemple, après avoir considéré pendant quelques itérations interactives que les éléphants étaient à ranger avec les chevaux car ce sont des animaux, il choisit finalement de les séparer en deux catégories distinctes). Dans la version actuelle de notre approche, lors de la phase de re-*clustering*, la violation de contraintes n'est pas interdite, elle est simplement pénalisée. Donc, dans ce type de cas, le système continuerait à fonctionner, mais risquerait d'engendrer des résultats qui ne satisfassent pas l'utilisateur (ici par exemple la solution de *clustering* pourrait comporter un *cluster* contenant – de manière non exhaustive – des éléphants et des chevaux, un *cluster* contenant des éléphants mais pas de chevaux, et un *cluster* contenant des chevaux, mais pas d'éléphant). Le problème est alors de supprimer de la liste des contraintes actives les contraintes *must-link* entre images de chevaux et d'éléphants précédemment intégrées dans le système, afin d'inciter au final le système à redéployer les images du premier *cluster* mêlant chevaux et éléphants vers les deux autres *clusters*.

Parmi les six stratégies de déduction/sélection des contraintes entre paires d'images à partir des retours de l'utilisateur que nous avons étudiées (voir section 2.2.3.2), certaines reposent sur un oubli progressif des contraintes les plus anciennes (elles n'ont cependant pas été retenues dans notre contexte applicatif où l'on considère que la catégorisation souhaitée par l'utilisateur est figée). Ces stratégies d'oubli pourraient partiellement résoudre ce problème, mais potentiellement avec un délai assez long, vu que la totalité des *clusters* ne peut être présentée à l'utilisateur à chaque itération interactive. À noter que, du fait de leur caractère systématique, elles pourraient engendrer des effets de bord, comme par exemple des contraintes qui resteraient

valides mais seraient quand même oubliées. On peut donc envisager **des stratégies d'oubli ciblé**, qui pourraient être soumises à la validation de l'utilisateur, et qui ne concerneraient que les contraintes les plus anciennes dont la satisfaction entraînerait le plus de changements dans la structure arborescente, voire dans la solution de *clustering* proposée au final. La sélection de ces contraintes pourrait se faire automatiquement au niveau des contraintes entre images ou au niveau des contraintes entre feuilles. Les paires d'images ou de feuilles concernées seraient alors présentées à l'utilisateur, pour que celui-ci valide ou infirme ses retours précédents.

Afin de pouvoir mener des expérimentations avec des utilisateurs humains de divers horizons (experts ou simples utilisateurs), d'une manière qui soit la moins biaisée possible, nous sommes en train de développer une application *web* que nous souhaitons mettre à disposition du grand public. Ces expérimentations devraient nous permettre (entre autres) de choisir la meilleure stratégie d'oubli ciblé.

Passons maintenant à la troisième principale amélioration possible de ce travail, qui concerne la sélection des images à présenter à l'utilisateur lors de chaque itération interactive. L'idée est de présenter en priorité à l'utilisateur les images pour lesquelles une interaction de sa part aurait potentiellement le plus grand impact sur la solution de *clustering*. Tout comme pour la sélection des contraintes entre paires d'images à prendre en compte dans le processus de *clustering* semi-supervisé, l'idée sous-jacente est proche de celle de l'**apprentissage actif** [Cohn 1996, Cord 2008]. Dans le cas des rares travaux sur le *clustering* semi-supervisé d'images basé sur un apprentissage actif (si l'on exclut le cas du *clustering* flou), nous pouvons citer la stratégie introduite dans [Biswas 2012] et qui est basée sur la sélection des exemples dont un éventuel changement d'étiquette aurait le plus grand impact sur la fonction objective du *clustering* semi-supervisé. Dans notre cas, la fonction objective (*cf.* annexe B) pénalise la violation d'éventuelles contraintes entre feuilles CF, et non entre images directement. Ce qui rend difficile la transposition directe de ce genre d'approches dans notre cas, puisqu'il n'est pas trivial de quantifier de manière individuelle l'impact du changement d'étiquette d'une image donnée sur la fonction objective du re-*clustering*. Notre stratégie actuelle de sélection des images à présenter à l'utilisateur, basée sur l'étude de la contribution de chaque image à la compacité et à la séparabilité de son *cluster* d'appartenance courant (*via* la mesure interne SW), est cependant d'un principe proche de celle utilisée dans [Biswas 2012].

Néanmoins, notre méthode de sélection des images à présenter à l'utilisateur, tout comme celle introduite dans [Biswas 2012], comporte un désavantage étant donnés les modes d'interaction que nous avons choisi de mettre en place. En effet, puisqu'il nous faut choisir dès le début de chaque étape itérative l'ensemble des exemples à présenter à l'utilisateur lors de cette étape itérative, nous devons procéder à cette sélection « en bloc ». Il est donc *a priori* possible que deux images pour lesquelles une interaction avec l'utilisateur apporterait une information redondante soient sélectionnées au cours d'une même itération interactive. C'est pour éviter ce genre de cas qu'un critère de non-redondance a été introduit dans [Grira 2008] dans un contexte d'apprentissage actif pour le *clustering* semi-supervisé flou (difficilement transposable dans notre cas). Dans notre cas en revanche, nous pourrions tirer avantage des regroupements d'images intermédiaires (noyaux, voisinages, feuilles) pour introduire dans notre algorithme d'apprentissage actif un critère de non-redondance des images à présenter à l'utilisateur (par exemple en n'autorisant pas la sélection de plus de quelques images représentatives par voisinage ou par feuille).

La section suivante vise à détailler nos travaux relevant de la description de collections

d'images plus structurées (en l'occurrence de documents textuels) par *clustering* interactif. Nous reviendrons sur les travaux présentés ci-dessus à l'occasion, d'une part, d'une discussion plus globale sur ce chapitre (section 2.4) et, d'autre part, de la présentation de leurs perspectives à plus long terme (section 2.5).

2.3 Extraction d'invariants dans des documents textuels par *clustering* interactif

2.3.1 Positionnement de l'étude

Dans le cadre de **travaux en cours**, nous nous intéressons au cas d'images de documents textuels (imprimés ou manuscrits) appartenant à des collections anciennes (typiquement datant du Moyen-Âge). Étant donné qu'un ouvrage ancien est généralement écrit en utilisant une unique police (cas imprimé) ou en respectant un style d'écriture très standardisé (cas manuscrit), le texte extrait d'une collection donnée est caractérisé par une variabilité intra-catégorie limitée, et ce, quel que soit le niveau auquel on place la catégorie (graphème, caractère ou mot). De plus, leur structure est généralement assez stéréotypée. Une grande part des difficultés rencontrées lorsque l'on cherche à analyser ce type de documents provient plutôt des dégradations liées à leur âge ou à leurs conditions de conservation, ou bien de l'usage fréquent d'un vocabulaire non standardisé. Dans notre contexte applicatif, ces difficultés sont largement majorées par le fait que nous ne disposons d'aucune information *a priori* sur le script ou le langage utilisé.

2.3.1.1 Objectif visé

L'objectif visé est de pouvoir **extraire**, à partir d'une collection de documents donnée, des « invariants ». Ces invariants sont définis comme **des formes revenant de manière récurrente dans la collection** (il peut s'agir par exemple de graphèmes, de caractères ou de bouts de mots) et peuvent être vus comme des primitives constituant l'écriture. Ils devront pouvoir être extraits même en présence d'un langage ancien ou rare, à propos duquel nous n'avons aucune information *a priori*, et pour lequel aucun moteur de reconnaissance d'écriture n'existe.

Les invariants seront par la suite utilisés pour décrire l'apparence visuelle du texte. Ils pourront par exemple être réutilisés par des moteurs de navigation dans la collection de documents. Cette navigation pourra se baser entre autres sur la détection (*spotting*) de mots ou de bouts de mots dans des documents textuels en provenance d'une même collection et décrits à l'aide des invariants. C'est le sujet de la **thèse de Bui Quang Anh** [13], commencée en 2011. Les travaux présentés ci-après ont été menés essentiellement dans le contexte de cette thèse.

L'une des applications visées est qu'un humain puisse se servir des invariants pour générer lui-même ses propres requêtes dans une application de recherche de mots, comme expliqué dans [Bui 2012]. Pour cela, on peut envisager que chaque invariant soit associé à un code ASCII et que l'utilisateur compose le mot qu'il recherche de manière textuelle, ou bien que les invariants soient présentés à l'utilisateur au travers d'une interface tactile et qu'il compose lui-même son image de requête grâce aux imagettes d'invariants. Il est donc primordial que les invariants soient en nombre limité, mais suffisamment exhaustifs pour permettre à l'utilisateur final du système de composer facilement ses requêtes.

Nous avons donc choisi de faire intervenir l'humain pour raffiner les invariants automatiquement découverts par la machine. Cette interaction se fait de manière personnalisée avec un expert du domaine (archiviste par exemple), qui connaît suffisamment le script utilisé pour aider le système à corriger les invariants.

13. Doctorant effectuant sa thèse au L3i sous la direction de Rémy Mullot et sous ma supervision scientifique.

Notre proposition est de **mettre en œuvre un *clustering* interactif** de formes élémentaires segmentées dans l'image, pour découvrir les invariants composant le texte.

2.3.1.2 Travaux existants connexes

Si l'apparence du texte à l'intérieur de la collection de documents considérée est suffisamment homogène (ce qui est généralement le cas à l'intérieur d'un ouvrage ancien), alors les invariants peuvent être considérés comme des primitives de l'image, en nombre fini.

À la différence du cas des symboles traité en annexe A cependant, on ne connaît pas la liste des invariants possibles, puisque l'on ne dispose d'aucune information *a priori* concernant le script utilisé. Cela rend bien évidemment plus complexe leur extraction depuis l'image.

En termes d'exploitation, si les primitives à rechercher sont connues, alors il suffit de concevoir ou d'utiliser un détecteur/moteur de reconnaissance adapté à ces primitives pour les exploiter. On peut citer par exemple les approches de *word spotting* permettant de générer des images de requêtes à partir d'un texte tapé au clavier et d'un alphabet codant des primitives prédéfinies. Les images ainsi obtenues sont utilisées pour la recherche par le contenu de mots similaires dans la collection. L'avantage de ces approches est que l'utilisateur n'a pas besoin de repérer dans le document une occurrence du mot qu'il recherche ; il lui suffit de le taper au clavier. Parmi ces dernières approches, on peut citer celle proposée dans [Marinai 2006], qui génère une image de requête à partir d'une requête textuelle, en utilisant une police spécifique. On peut également citer l'approche introduite dans [Konidaris 2007], où chaque caractère textuel (code ASCII) est associé manuellement à une image de caractère et où une procédure d'alignement des caractères est utilisée. Ces deux méthodes ne sont applicables qu'aux documents imprimés. L'approche proposée dans [Leydier 2009], qui consiste à construire semi-automatiquement un dictionnaire de glyphes et une grammaire contenant l'ensemble des règles d'édition dirigeant le positionnement spatial des glyphes, est applicable aux documents imprimés comme aux documents manuscrits.

Néanmoins, toutes ces méthodes nécessitent des informations *a priori* sur le script (voire la police) utilisé(e), passant par une phase d'annotation préalable. Cette phase d'annotation peut être manuelle ou semi-automatique. Dans le cas semi-automatique, l'annotation repose typiquement sur un OCR ou un moteur de reconnaissance d'écriture manuscrite dédié au langage considéré. Cela suppose l'existence d'un tel OCR ou moteur de reconnaissance, ce qui n'est pas acquis dans le cas d'un langage ancien ou rare. Or, dans notre contexte applicatif, nous ne disposons d'aucune information *a priori* sur le script ou le langage utilisé, ni *a fortiori* d'annotations (manuelles ou semi-automatiques) du texte présent dans les documents de la collection. Ce type d'approches est donc inapplicable dans notre contexte.

Dans notre contexte où **nous n'avons aucune information *a priori*** concernant les primitives (invariants) composant le texte, nous cherchons à les découvrir automatiquement. L'usage ultérieur qui pourra être fait, à terme, des invariants découverts, est similaire aux exemples ci-dessus. Pour extraire les invariants, nous avons choisi de nous baser sur l'analyse des formes élémentaires revenant fréquemment dans la collection de documents. Plus précisément, les formes auxquelles nous nous intéressons plus particulièrement sont les « *strokes* ». Le terme *stroke* est emprunté au vocabulaire de l'analyse d'écriture manuscrite en-ligne (voir section 3.3), où il s'agit d'un trait d'écriture. Dans le contexte hors-ligne, il peut être étendu au cas imprimé pour désigner un motif élémentaire du texte, indivisible et fréquent.

Plusieurs méthodes visent à extraire des *strokes* à partir du signal hors-ligne, le plus souvent dans le but de reconstruire le signal en-ligne à partir du signal hors-ligne afin d'améliorer la reconnaissance d'écriture [Lallican 2000]. Certaines de ces méthodes sont basées sur une squelettisation des composantes connexes de l'image binarisée (et, le plus souvent, sur une recherche des points d'extrémité ou de jonction entre différentes branches du squelette) ; d'autres sont basées sur une analyse des contours des composantes connexes (passant généralement par une recherche des « points dominants » qui correspondent à des changements brusques de la courbure du contour). Pour un état de l'art plus exhaustif des méthodes d'extraction de *strokes* depuis des documents hors-ligne, merci de se référer à [Bui 2013].

Maintenant que nous avons défini les objectifs de nos travaux et que nous les avons situés dans leur contexte scientifique, nous allons les replacer dans le cadre des questions abordées dans ce manuscrit et vis-à-vis de l'étude précédente, avant de décrire plus précisément leurs principales originalités.

2.3.2 Fil conducteur des questions abordées

Jusqu'à présent, dans ce chapitre dédié à la description d'images, nous avons brièvement évoqué l'extraction de descripteurs visuels depuis les images, avant de nous focaliser sur l'organisation de collections d'images par *clustering* interactif.

Les travaux présentés dans cette section font le lien entre ces deux types d'approches. En effet, nous cherchons ici à décrire les motifs présents dans une image donnée de manière individuelle, à partir d'éléments d'intérêt découverts par *clustering* interactif à l'échelle de la collection entière.

Comme dans les travaux évoqués ci-avant, nous avons fait le choix de l'interaction avec l'utilisateur dans le but de rapprocher la description retournée par la machine des concepts de plus haut niveau manipulés par un humain. Mais, l'expérience acquise au travers de ces travaux précédents nous a montré que l'utilisateur humain a généralement tendance à donner plus de retours négatifs que positifs. Autrement dit, un humain a naturellement tendance à s'intéresser en priorité au fait de corriger les résultats renvoyés par la machine. Nous nous sommes donc focalisés ici sur la manière de le faire interagir cet humain de manière à lui permettre de corriger le plus efficacement possible les résultats retournés par la machine. De manière plus générale, les spécificités du problème traité nous ont poussé à définir des **modes d'interaction différents** du cas précédent de l'organisation de collections d'images tout-venant.

Tout d'abord, nous avons choisi de faire interagir l'utilisateur au niveau plus global des *clusters* de *strokes* (à la différence du cas précédent où l'interaction se faisait au niveau individuel des images). À cela, deux raisons. Premièrement, d'un point de vue très pratique, les éléments que nous cherchons à regrouper ici ne sont que des *strokes*, c'est-à-dire des traits d'écriture, avec lesquels il est fastidieux pour l'utilisateur d'interagir de manière individuelle. Deuxièmement, le fait de mettre en œuvre l'interaction au niveau des *clusters* nous permet d'imaginer diverses solutions pour répercuter chacune des interventions de l'utilisateur de la manière la plus collective possible dans les *clusters* concernés.

Nous cherchons en outre ici à concevoir un système qui soit le plus **générique** possible (omni-script, omni-langage). Le cas échéant, nous cherchons donc à faire en sorte que les interventions de l'utilisateur permettent au système d'ajuster automatiquement les paramètres du système proposé vis-à-vis du script/langage courant.

La section suivante décrit le système que nous avons conçu pour apporter des éléments de réponse à ces questions.

2.3.3 Aperçu du système proposé

Le système d'extraction interactive d'invariants que nous proposons, dont un aperçu est donné en Figure 2.4, repose sur trois phases : extraction des *strokes*, *clustering* de ces *strokes* afin d'obtenir des invariants (définis comme les *clusters* de *strokes*), et raffinement des invariants lors d'une phase d'interaction avec l'expert. Ces trois phases sont détaillées dans les paragraphes ci-après.

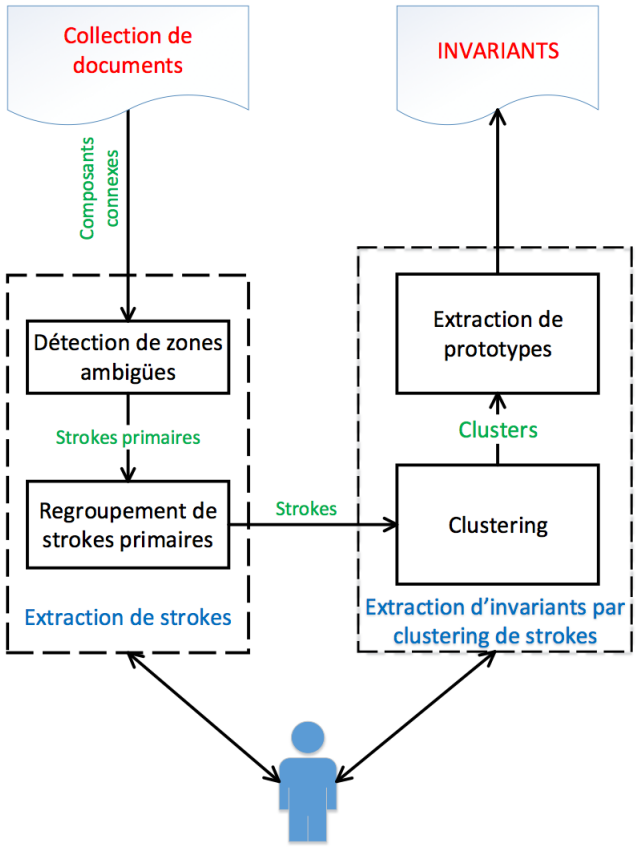


FIGURE 2.4 – Vue globale du système d'extraction interactive d'invariants.

Extraction de *strokes*. À partir des composantes connexes extraites depuis le texte, on extrait les *strokes*. Cette extraction se fait en deux étapes menées séquentiellement :

1. À l'intérieur de chaque composante connexe extraite du document, on détecte les « zones ambiguës » de l'écriture avec la méthode proposée dans [Su 2009] et basée sur l'analyse des points de jonction dans les squelettes des composantes connexes. Chaque fragment d'une composante connexe relié à une zone ambiguë constitue un « *stroke* primaire » (c'est-à-dire une partie de *stroke*), tandis que chaque composante connexe ne présentant pas de zone ambiguë est directement considérée comme un *stroke* ;

2. Afin de regrouper les *strokes* primaires en *strokes*, nous utilisons une méthode de segmentation que nous avons introduite dans [Bui 2013] et qui est basée sur l'étude de la continuité visuelle entre couples de *strokes* primaires reliés par une zone ambiguë.

Extraction d'invariants par *clustering* des *strokes*. Une fois les *strokes* extraits, on constitue de manière complètement automatique des groupes de *strokes* semblables en taille et en forme. Plus précisément, l'étape d'extraction d'invariants prend en entrée des signatures décrivant la taille et la forme des *strokes* (nous utilisons pour cela des descripteurs classiques de la littérature). En sortie, nous obtenons un ensemble de *clusters* qui constituent les invariants.

L'une des principales difficultés ici réside dans le choix de l'algorithme de *clustering* (et le cas échéant de son paramétrage), ainsi que du nombre k de *clusters* le mieux adapté au problème. En effet, on obtient une solution de *clustering* différente (avec le cas échéant un nombre k de *clusters* différent) pour chaque méthode de *clustering* et chaque paramétrage. Or, vu la nature du problème posé, nous ne disposons d'aucune vérité-terrain et ne pouvons donc avoir recours aux mesures externes [14] pour comparer la qualité des différentes solutions de *clustering*. Les mesures internes usuelles, elles, ne sont généralement pas normalisées vis-à-vis du nombre k de *clusters* et donnent parfois des résultats contradictoires. Enfin, les mesures de stabilité [Lange 2004], sont plutôt conçues pour fixer le nombre de *clusters* k et éventuellement les paramètres d'une méthode donnée, plutôt que pour sélectionner l'algorithme le mieux adapté.

Nous avons donc choisi d'appliquer une approche de *clustering* par consensus [Strehl 2003], qui consiste à unifier un ensemble de solutions de *clustering* (possiblement obtenues par différents algorithmes et/ou avec différentes valeurs de paramètres) en une unique solution. Ici nous appliquons le consensus à partir de solutions de *clustering* fournies par les algorithmes *global k-means*, DBSCAN, CAH et SOM, avec différents nombres de *clusters* k. La solution optimale de *clustering* (ainsi que son paramétrage optimal) sont déterminés automatiquement en utilisant la mesure basée sur la théorie de l'information NMI_{max} [Kvalseth 1987], qui présente les avantages d'être métrique et normalisée de manière à ne pas favoriser les solutions de *clustering* avec une grande valeur de k [Nguyen 2010]. Il est coûteux en temps d'appliquer un tel algorithme, mais ce n'est pas forcément gênant dans notre contexte, puisque cette phase est menée de manière complètement automatique (sans intervention de l'utilisateur), et donc hors-ligne.

Raffinement interactif des invariants. Une fois les *clusters* (invariants) calculés, ils sont présentés à l'utilisateur, qui peut les raffiner itérativement en interagissant (en-ligne) avec le système de deux manières :

- Par fusion/division des *clusters* dans l'espace de représentation (espace des signatures) ;
- Par regroupement ou découpage spatial des invariants. Ces raffinements spatiaux peuvent être automatiquement répercutés au niveau de l'ensemble des *strokes* du *cluster*.

2.3.4 Principales originalités du système proposé

Les principales originalités du système proposé reposent, d'une part, sur les modalités de l'interaction avec l'utilisateur humain (expert du domaine) pour raffiner les invariants et, d'autre part, sur le fait que notre approche puisse être déployée même en l'absence de toute information *a priori* concernant le script ou le langage utilisé. Cela lui confère en théorie une certaine généricité

14. Voir section 2.2.4.

qu'il convient de discuter en pratique. C'est donc sur ces deux points que je focaliserai mon propos dans les deux sections ci-après.

2.3.4.1 Raffinements interactifs des invariants

Comme illustré en Figure 2.5, l'interaction avec l'utilisateur pour raffiner les invariants se fait grâce à une interface très proche de celle que nous utilisions précédemment pour les images naturelles. Les prototypes des invariants sont présentés à l'utilisateur dans le plan des deux premières composantes principales calculées à partir des descripteurs de forme extraits des *strokes*. L'utilisateur peut cliquer sur un prototype d'invariant pour voir les *strokes* les plus représentatifs et les moins représentatifs du *cluster* correspondant apparaître dans deux cercles de la fenêtre 2 (en haut à droite de l'interface). En survolant l'un de ces *strokes* avec la souris, l'utilisateur peut le faire apparaître (en bleu) dans son contexte, dans la fenêtre juste en-dessous de la fenêtre 2.

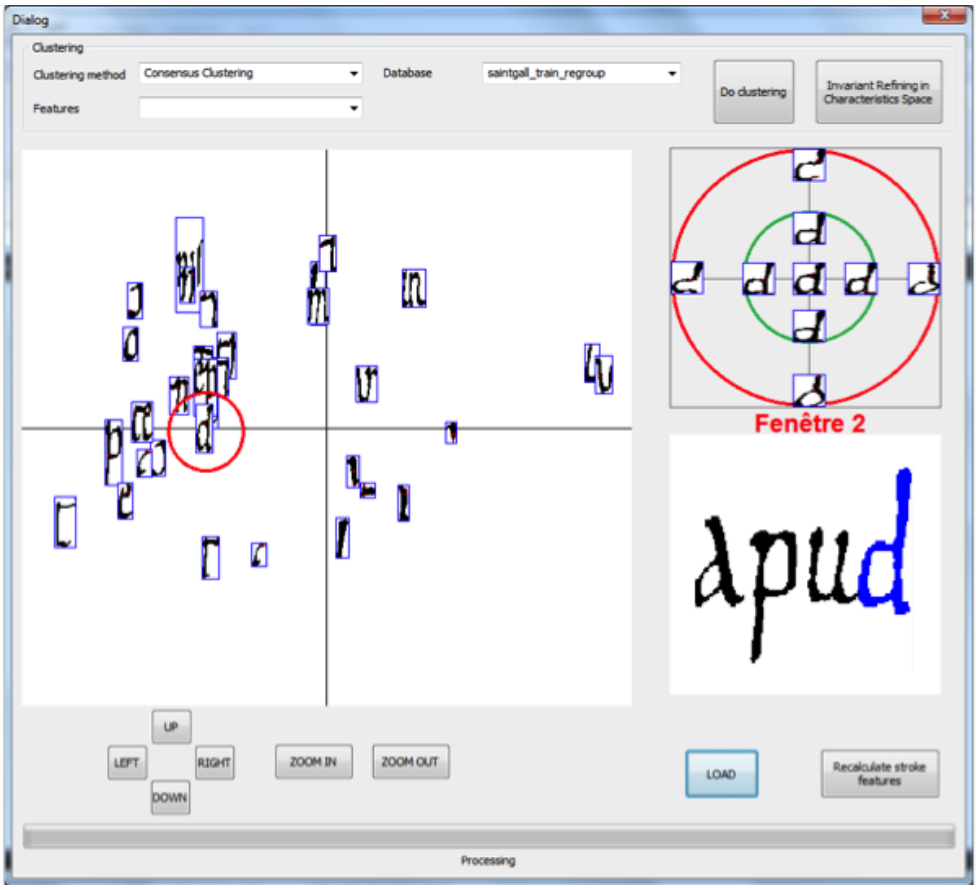


Figure 2.5 – Interface permettant à l'humain d'interagir avec les invariants découverts par le système. Ici, les invariants sont ceux extraits complètement automatiquement (avant le raffinement interactif) sur dix pages de la base Saint Gall [15] (extraite d'un ouvrage écrit par un unique scripteur au 9ème siècle, dans un script carolingien et un langage latin).

Nous avons choisi deux modes d'interaction de l'utilisateur avec les *clusters*. Le premier mode d'interaction se déroule dans l'espace de représentation et vise à permettre à l'utilisateur de

15. www.iam.unibe.ch/fki/databases/iam-historical-document-database/saint-gall-database

commander des opérations de division/fusion de certains *clusters*, en fonction de leur contenu visuel. Le second mode d'interaction se déroule dans le plan (x, y) de l'image et permet de répercuter les interventions de l'humain au plus bas niveau de l'extraction des strokes, en imposant le regroupement ou le découpage spatial de certains des *strokes* d'un *cluster* donné.

Ces interactions entre le système et l'utilisateur sont effectuées itérativement. À chaque itération interactive, l'interaction se fait d'abord dans l'espace de représentation, puis spatialement. Les itérations interactives sont répétées jusqu'à ce que l'utilisateur soit satisfait du résultat. Plus de détails sur ces deux modes d'interaction sont donnés ci-après.

La fusion ou la division de *clusters* se fait dans l'espace de représentation des *strokes*. Suivant les cas, le système pré-sélectionne les *clusters* à présenter à l'utilisateur, ou bien l'utilisateur les choisit directement depuis l'interface montrée en Figure 2.5.

La **division de *clusters*** consiste à découper un *cluster* trop hétérogène en k sous-*clusters*. Soit l'utilisateur choisit à l'aide de l'interface les *clusters* qu'il souhaite diviser, soit le système présente automatiquement à l'utilisateur les *clusters* avec la plus faible valeur de la mesure d'évaluation interne SW. L'utilisateur est alors libre de choisir de diviser ces *clusters* ou non, en précisant le cas échéant la valeur de k (par défaut fixée à 2). Le *clustering* des *strokes* de ce *cluster* initial en k sous-*clusters* est alors mené automatiquement avec la méthode des k-moyennes globale proposée dans [Likas 2003].

Dans le cas d'une **fusion**, l'utilisateur sélectionne directement depuis l'interface les invariants qui lui semblent très proches (par simple clic sur leurs images de prototypes), visualise les *strokes* composant les *clusters* correspondants, et le cas échéant peut commander au système de les fusionner en un seul *cluster*.

À la différence du cas des images naturelles traité précédemment, on peut également mettre en œuvre des opérations de découpage ou de regroupement de *strokes* dans l'espace spatial (c'est-à-dire dans le plan (x, y) de l'image). À chaque itération, ces opérations sont déclenchées lorsque l'utilisateur décide que ses interactions dans l'espace de représentation sont terminées.

En ce qui concerne le **découpage spatial des *strokes***, notre système fonctionne de la manière suivante. Il propose à l'utilisateur, à l'aide de l'interface montrée en Figure 2.5, les *clusters* pour lesquels la taille des *strokes* qu'ils contiennent est en moyenne largement plus élevée que celle des autres *clusters*. L'utilisateur doit alors décider si les *strokes* de ce *clusters* mériteraient d'être découpés, ou non. Si c'est le cas, le système revient alors sur la phase d'extraction des *strokes* de ce *cluster*. Selon le cas considéré, cela peut se faire en modifiant les paramètres soit du module d'extraction de zones ambiguës, soit de regroupement des *strokes* primaires (le système détermine automatiquement le réglage le mieux adapté). Cela débouche sur le découpage spatial de certains des plus grands *strokes* du *cluster* courant, ces nouveaux *strokes* étant alors affectés au *cluster* existant le plus proche dans l'espace de représentation de bas niveau sémantique.

Intéressons-nous maintenant au **regroupement spatial de *strokes*** en provenance de deux *clusters* différents. Selon la configuration du système, soit ces regroupements se font de manière complètement automatique, soit en utilisant l'interface de la Figure 2.6. À chaque itération interactive, le système mène de manière automatique une étude de la distribution des positionnements relatifs des *strokes* des différents *clusters*. Pour chaque couple de *clusters*, le système recherche la configuration spatiale la plus fréquente : p. ex. les *strokes* du *cluster* K_i sont souvent localisés en haut à gauche des *strokes* du *cluster* K_j dans les images de documents de la collection. Notons $K_i^* \subset K_i$ et $K_j^* \subset K_j$ les sous-ensembles de *strokes* vérifiant cette confi-

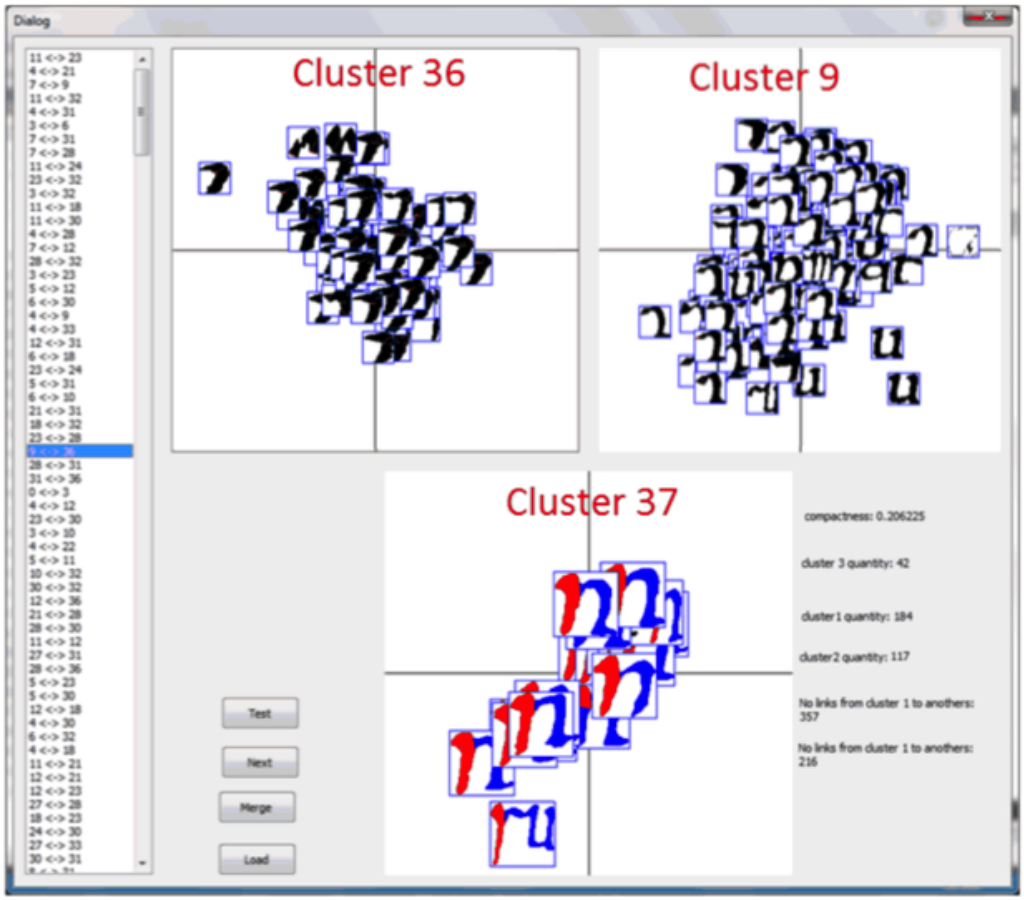


FIGURE 2.6 – Interface permettant à l'utilisateur de décider d'un éventuel regroupement spatial de *strokes*. Ici, les *strokes* des *clusters* 9 et 36 sont fréquemment reliés par une zone ambiguë et les *strokes* du *cluster* 36 sont le plus fréquemment localisés à gauche de *strokes* du *cluster* 9 dans le document. Puisque, de plus, la qualité du *cluster* 37 issu du regroupement spatial des paires de *strokes* des *clusters* 9 et 36 vérifiant cette configuration spatiale est satisfaisante, le système propose à l'utilisateur de procéder au regroupement spatial. S'il le souhaite, l'utilisateur n'a alors plus qu'à cliquer sur le bouton « *Merge* ».

guration spatiale. Le système évalue automatiquement (avec des mesures internes) la qualité du *cluster* K_l qui serait produit par un regroupement spatial (partiel) des *clusters* K_i et K_j : $K_l = K_i^* \cup K_j^*$. Si sa qualité est satisfaisante alors, selon la configuration du système, soit le système procède directement au regroupement spatial, soit l'utilisateur est sollicité afin de déterminer s'il juge ce regroupement pertinent. Le cas échéant, seuls les *strokes* vérifiant la configuration spatiale la plus fréquente seront regroupés spatialement. Le regroupement spatial des *clusters* K_i et K_j résulterait donc en trois *clusters* : $K_i - K_i^*$, $K_j - K_j^*$ et $K_l = K_i^* \cup K_j^*$.

Le processus est alors réitéré (en alternant des phases de raffinement dans l'espace de représentation puis dans l'espace spatial), jusqu'à ce que l'utilisateur soit satisfait du résultat.

2.3.4.2 Discussion sur la généricité du système proposé

Ces travaux en cours visent à proposer un système capable d'extraire des invariants sans information *a priori* concernant le script/langage utilisé (système omni-script/omni-langage).

Or, chaque étape de notre système (extraction de *strokes* primaires, regroupement de *strokes*, *clustering*) repose sur un certain nombre de paramètres (en particulier des seuils). Autant que faire se peut, nous avons cherché à automatiser l'initialisation de ces paramètres, le plus souvent par apprentissage non supervisé en fonction de descripteurs visuels extraits de la collection.

Une fois cette initialisation faite automatiquement, les raffinements interactifs permettent d'ajuster (localement) les paramètres, à la fois de l'extraction de *strokes* et du *clustering*, sur lesquels repose l'extraction d'invariants. Nos expérimentations préliminaires montrent cependant que, **dans la pratique, l'initialisation de ces paramètres a un fort impact** sur le volume d'effort qui devra être demandé à l'utilisateur pour les raffinements interactifs. En effet, les paramètres initiaux optimaux dépendent fortement de certaines spécificités du script considéré, entre autres du nombre de traits d'écriture que l'on retrouve typiquement dans les éléments composant l'écriture (relativement faible dans le cas du système alphabétique latin mais potentiellement très élevé dans les sinogrammes par exemple [16]).

D'un point de vue pratique, on peut dès lors se demander s'il ne serait pas bénéfique de prévoir différents jeux de paramètres initiaux, qui pourraient être ajustés par exemple en fonction du type de script (alphabétique, logographique, etc.). Vu qu'il ne s'agit que d'une initialisation qui pourra être affinée par la suite, ils pourraient être appris le cas échéant sur des documents correspondant au même type de script, mais pas au même langage, si ce dernier est rare.

2.3.5 Bilan et applications visées

Dans le cadre des travaux en cours présentés ci-dessus, nous nous intéressons au cas d'images de documents textuels (imprimés ou manuscrits) appartenant à des collections anciennes. Le contenu de tels documents est généralement très homogène (en termes notamment d'apparence visuelle du texte). Mais, du fait que l'on s'intéresse en particulier à des langages anciens ou rares, nous ne disposons d'aucune information *a priori* sur le script ou le langage utilisé, et donc *a fortiori* d'aucun moteur de reconnaissance automatique de ce texte. Afin d'en décrire le contenu en vue d'une exploitation ultérieure (par *word spotting* par exemple), nous avons proposé un système permettant d'extraire interactivement des « invariants », c'est-à-dire des formes revenant de manière récurrente dans la collection, pouvant être vus comme des primitives constituant l'écriture.

Le fonctionnement du système que nous avons proposé pour l'extraction d'invariants peut être résumé comme suit. Dans une première étape, le système découvre automatiquement dans la collection un premier jeu d'invariants, par extraction de traits d'écriture (*strokes*), suivie du *clustering* non supervisé des *strokes*. Puis, on entre dans une phase d'interaction avec l'utilisateur, qui est itérative et permet au système de corriger localement certains *clusters*, et le cas échéant certains des *strokes* qui les composent, sur la foi des interventions de l'utilisateur. Plus précisément, une itération interactive comprend une phase de raffinement dans l'espace de représentation, et une phase de raffinement dans l'espace spatial, qui peuvent entraîner toutes les deux des modifications locales du jeu d'invariants.

Mais, une fois ces raffinements terminés, le système ne reboucle pas sur la phase de *clustering*. C'est en ce sens qu'il s'agit de *clustering* interactif, mais pas de *clustering* semi-supervisé interactif (comme c'était le cas de l'approche que nous avons proposée pour des images naturelles). Rendre le *clustering* semi-supervisé est possible techniquement et pourrait

16. Le nombre de traits d'écriture dans un sinogramme peut atteindre 36.

permettre de répercuter de manière plus globale les actions de l'utilisateur. Mais cela soulève un certain nombre de questions, notamment liées à l'incrémentalité (du fait que le système procède à des modifications de *clusters* et/ou de *strokes* à chaque retour de l'utilisateur), auxquelles nous n'avons pas eu le temps de nous attaquer jusqu'à présent.

On peut noter que, à la différence de nos travaux précédents sur l'organisation de collections d'images tout-venant, nous n'avons pas introduit une méthode, mais plutôt un système dont les principales originalités reposent sur des modalités d'interaction (entre l'utilisateur et la machine) qui sont assez spécifiques à l'application visée (extraction d'invariants depuis des documents textuels). En ce sens, **les principales contributions liées à ce travail sont plutôt applicatives** que méthodologiques.

Nous sommes en train d'évaluer les performances de notre système d'extraction d'invariants. Nous devons pour cela faire face à une difficulté, qui est que nous ne disposons d'aucune vérité-terrain au niveau des invariants. En effet, chaque utilisateur est libre de guider le système vers une solution de *clustering* qui peut différer d'un utilisateur à l'autre (en fonction de sa connaissance du script par exemple). Cela rend difficile l'évaluation de la pertinence des invariants retournés, et en particulier de l'intérêt de la phase de raffinements interactifs.

Afin d'évaluer les performances de notre système, nous sommes donc en train de nous tourner vers le contexte applicatif final de **la détection de mots** (*word spotting*). Pour cela, nous avons conçu une signature structurelle simple permettant de décrire un mot au travers d'un graphe représentant l'agencement spatial des invariants qui le composent. Nous procédons actuellement à une comparaison avec d'autres approches de la littérature dans le cadre d'une application au *word spotting*. Nos expérimentations préliminaires menées sur la base Georges Washington [17] (qui date du 18ème siècle et contient 4894 images de mots manuscrits segmentés en langue anglaise) montrent que notre approche donne (après quelques itérations interactives) une précision moyenne qui est similaire aux meilleures approches non-supervisées de la littérature (comme par exemple la méthode décrite dans [Lladós 2012]), mais qui reste évidemment nettement inférieure à celles des approches basées sur un apprentissage supervisé, comme par exemple la méthode présentée dans [Frinken 2012].

On compte deux **avantages principaux** de notre approche par rapport à ces méthodes. Premièrement, notre approche ne nécessite aucune annotation *a priori* d'une base d'entraînement, à la différence des approches supervisées. Deuxièmement, dans une application au *word spotting*, il n'est pas nécessaire que l'utilisateur ait préalablement détecté une occurrence du mot à rechercher pour le retrouver dans la collection, comme c'est généralement le cas dans les approches non-supervisées. En effet, à terme, les invariants pourraient être utilisés pour composer visuellement la requête, à la manière des travaux présentés en section 2.3.1.2, sauf que dans notre cas nous ne disposons d'aucune information *a priori* sur le langage/script du document d'entrée.

Ce travail de thèse a été entrepris initialement pour analyser des documents écrits selon différentes variantes anciennes du langage cambodgien, dans le cadre d'un projet PCSI sur la valorisation du patrimoine khmer (financé par l'AUF, *cf.* section 1.4). En raison de difficultés à établir un accord de propriété entre le consortium du projet et l'ONG qui détient ces documents écrits, nous n'avons pas été en mesure, jusqu'à ce jour, d'obtenir ces documents.

17. www.iam.unibe.ch/fki/databases/iam-historical-document-database/washington-database/

Les résultats issus de ces travaux pourront néanmoins être valorisés et étendus dans le contexte du **projet ARCHIVES**, qui commencera en 2015 (voir section 1.4.3.2). Ce projet porte sur l'analyse et la reconstitution d'événements catastrophiques décrits en particulier au travers de documents écrits en sino-vietnamien (langage ancien pour lequel il n'existe pas de moteur de reconnaissance), en français (durant la période coloniale) et en vietnamien moderne. L'adaptabilité de notre système vis-à-vis de différents types de langages sans nécessiter d'apprentissage supervisé est un atout pour l'analyse des documents manuscrits écrits en sino-vietnamien, qui respectent un style d'écriture très standardisé, et une structure très stéréotypée. Et ce, en particulier pour des applications de *word spotting*.

La sélection des images de documents écrits en sino-vietnamien dans le corpus (par ailleurs très hétérogène) d'ARCHIVES pourra être facilitée par la présence systématique de sceaux dans ces documents. Plus précisément, nous pourrons utiliser pour cette sélection un outil de catégorisation automatique de documents basés sur la détection/reconnaissance d'éléments graphiques dans les images, comme par exemple ceux que nous avons proposés dans [Le 2012, Le 2013, Le 2014]. Dans le cas des langages modernes (français ou vietnamien), des méthodes de *word spotting* supervisées telles que celle présentée dans [Frinken 2012] sont plus adaptées.

2.4 Discussion

Au terme de la présentation de l'ensemble de nos travaux concernant la catégorisation interactive d'images dans un objectif de description, il convient de mettre en perspective ces avancées avec le chemin qu'il nous reste à parcourir, au regard des verrous scientifiques auxquels nous avons cherché à nous attaquer. Avec le recul dont nous disposons aujourd'hui, il est également intéressant de replacer ces travaux, initiés en 2009, dans le contexte des tendances qui se sont dégagées entre-temps dans la communauté scientifique. La discussion qui suit est axée autour des verrous scientifiques et techniques auxquels nous cherchons à nous attaquer, et des questions de recherche qui en découlent dans notre contexte.

Le premier verrou auquel nous avons cherché à nous attaquer au travers de ces travaux est celui de la différence de niveau sémantique entre l'information binaire manipulée par les machines et les concepts perçus par l'humain. Afin de tenter de réduire ce fossé sémantique, nous nous sommes appuyés sur une interaction avec l'utilisateur. Les informations obtenues grâce aux interventions de l'utilisateur permettent de guider itérativement le processus de catégorisation vers une solution qui soit plus conforme aux concepts de haut niveau manipulés par l'humain. Au travers des deux cas d'étude abordés, **nous avons étudié différents modes d'interaction** entre l'utilisateur et la solution de catégorisation proposée, à chaque étape, par la machine. Dans le premier cas (catégorisation d'images naturelles), l'intervention de l'utilisateur se fait au niveau des images sous la forme de retours positifs ou négatifs concernant leurs groupes d'appartenance. Dans le deuxième cas où l'on cherche à catégoriser des primitives (« invariants ») automatiquement extraites du texte, nous avons étudié des modes d'interaction plus globaux où l'humain intervient directement au niveau des *clusters*.

Néanmoins, nous sommes restés cantonnés à des **contextes applicatifs qui ne reflètent pas de manière exhaustive les usages variés** qui pourraient être faits des outils que nous avons conçus. En particulier, nous avons pu constater qu'il n'était pas trivial d'adapter le système aux comportements parfois inconsistants des utilisateurs humains, que l'on peut observer dès lors que l'on sort des applications relevant des humanités digitales, où l'organisation des images souhaitée par les utilisateurs suit généralement une typologie précise.

De même, les groupes d'images perçus par l'utilisateur ne sont pas forcément de frontières perméables (particulièrement dans le cas d'images tout-venant), tandis que nos approches font l'hypothèse qu'une image donnée appartient à un seul groupe. Dans ce cas de figure, il est généralement plus pertinent de chercher à annoter les images avec des mots-clés, plutôt que de leur associer un groupe. Les approches contemporaines d'annotation automatique [Zhang 2012] sont généralement supervisées ; elles doivent plutôt être vues comme un complément de notre approche plutôt que comme un remplacement direct. En effet, il est par exemple possible d'utiliser les catégories [18] retournées par notre système pour annoter les images collectivement, au lieu de les annoter individuellement. J'aborderai à nouveau le cas de l'annotation d'images dans les perspectives du chapitre 3 (section 3.6).

Dans le même ordre d'idée, de par leur nature même, les approches proposées sont très peu adaptées au cas où de multiples utilisateurs ne partageant pas les mêmes souhaits cherchent à interagir (de manière synchrone ou asynchrone) avec le système. Dans ce type de cas applicatifs, la tendance actuelle qui consiste à adopter des cadres coopératifs, s'appuyant par exemple sur des système de recommandation [Brut 2011], est *a priori* plus pertinente que nos approches.

En ce qui concerne la manière de présenter l'information à l'utilisateur pour l'interaction,

18. Ou des regroupements intermédiaires (p. ex. les noyaux/feuilles du *clustering* semi-supervisé interactif).

les interfaces que nous avons proposées dans nos deux cas d'étude sont relativement semblables. Si elles permettent au système de collecter les informations dont il a besoin pour améliorer ses résultats, grâce à des opérations assez faciles à réaliser pour l'utilisateur (clic ou glisser-déposer), elles restent néanmoins assez basiques et peu ergonomiques. Vu la démocratisation de dispositifs tactiles et les progrès réalisés dans le domaine de la visualisation 3D ces dernières années, il serait certainement possible de concevoir des interfaces dynamiques et plus intuitives pour l'utilisateur. Outre le confort d'usage amélioré pour l'humain, de telles interfaces pourraient accroître l'information exploitable par le système, à effort constant pour l'utilisateur.

Cela nous amène à la deuxième question de recherche à laquelle nous avons cherché à nous attaquer, et qui peut être formulée ainsi : « Comment tirer au mieux parti des informations dont on dispose sur les images ? ». Le problème est double. D'une part, ces informations sont de niveau sémantique variable (signatures visuelles de bas niveau sémantique et informations d'un niveau sémantique supérieur retournées par l'utilisateur). D'autre part, étant donné que l'on cherche à minimiser l'effort à fournir par l'utilisateur, les informations de plus haut niveau sémantique sont forcément partielles au regard du volume de données à traiter.

Concernant la manière de prendre en compte l'information de plus haut niveau sémantique donnée par l'utilisateur au système, dans le premier cas d'étude, nous avons conçu un système semi-supervisé à l'aide de cette information. C'est-à-dire qu'à l'issue de chaque itération interactive, l'ensemble de la solution de *clustering* est recalculée dans l'espace de représentation de bas niveau sémantique, de manière à pénaliser les contraintes que le système a déduit (en s'appuyant sur la structure hiérarchique de la solution de *clustering*) des retours de l'utilisateur. Dans le second cas applicatif où l'interaction se fait directement au niveau des *clusters* et non des images, l'information apportée par l'utilisateur est intégrée en fin de chaîne, pour raffiner les catégories *a priori* découvertes par le système. Dans les deux cas, la solution de catégorisation proposée par la machine est **modifiée localement en fonction des interventions de l'utilisateur**. Nous avons évoqué au fil de ce chapitre des stratégies possibles [19] afin de rendre ces modifications plus globales. Cependant, vu les variétés des formes (dans l'espace de représentation de bas niveau sémantique) des catégories perçues par l'utilisateur, on peut se demander dans quelle mesure tout processus automatique visant à répercuter, de manière très globale, une information fournie par l'utilisateur sur une image (ou une catégorie) donnée, ne risquerait pas d'introduire un biais dans le système. Il nous faudra donc, le cas échéant, être particulièrement vigilants dans le choix de ces stratégies.

On en arrive maintenant à l'autre facette de cette deuxième question de recherche, qui peut se résumer ainsi : « Comment tirer au mieux parti des retours fournis par l'utilisateur à chaque itération interactive, sachant que ces retours sont par essence en nombre limité ? ».

Nous avons tout d'abord travaillé en amont de la collecte. Afin de minimiser l'effort de l'utilisateur, nous avons cherché à ne le solliciter qu'« à bon escient », c'est-à-dire de manière à ce que chacune de ses interventions ait la plus grande répercussion possible sur le ré-arrangement de la solution de catégorisation. Concrètement, nous avons choisi de **sélectionner les images (voire les *clusters*) à présenter à l'utilisateur** en fonction de leur distribution dans leurs *clusters* d'appartenance et, le cas échéant, des interactions précédentes de l'utilisateur. Comme nous l'avons évoqué précédemment cependant, le problème de cette sélection, qui pourrait être

19. Telles que l'apprentissage de distance par exemple.

formalisé comme une question d'apprentissage actif, reste assez largement ouvert.

En aval de la collecte, nous avons cherché à **déduire un certain nombre d'informations** supplémentaires (à partir des informations données par l'utilisateur) pour alimenter l'apprentissage. En particulier, dans le cas de collections d'images tout-venant, nous avons proposé des stratégies de déduction de « contraintes secondaires » à partir des contraintes données par l'utilisateur à l'itération courante. Ces stratégies, en s'appuyant sur les interventions de l'utilisateur aux itérations interactives précédentes, s'accommodent mal d'éventuels comportements inconsistants de ce dernier. Pourtant, ce genre de comportements se retrouve fréquemment dans la pratique, dès lors que l'on sort du cadre des applications visées dans nos travaux. Nous avons évoqué des stratégies d'oubli ciblé afin de limiter ce risque, mais ces stratégies mériteraient d'être évaluées dans la pratique (c'est l'objet de certains de nos travaux en cours).

Concernant la caractérisation de performances, nos efforts ont porté sur la définition de protocoles et, le cas échéant, de mesures, permettant d'évaluer les performances des approches proposées ou concurrentes. La caractérisation de performances est ici à entendre au sens large (qualitative ou quantitative) et a été déployée à différentes étapes du processus : sélection des images à présenter à l'utilisateur et, bien sûr, évaluation finale du résultat de la catégorisation avec comparaison vis-à-vis d'autres approches. Dans un souci de valider expérimentalement les approches proposées par une comparaison équitable avec les autres approches de la littérature, nos évaluations reposent bien souvent sur une **vérité-terrain figée *a priori*, ce qui constitue une faiblesse** de nos travaux basées sur l'interactivité, particulièrement dans le cas d'étude de l'organisation de collections d'images tout-venant. Dans ce dernier cas, nous sommes en train de tenter d'y pallier par le développement d'une application *web* qui permettrait au système de travailler avec des utilisateurs finaux humains, comme expliqué plus haut. Dans le cas de l'extraction d'invariants où, en l'absence de toute information *a priori* concernant le script ou le langage utilisé, mais surtout en l'absence de définition exhaustive des invariants souhaités par l'utilisateur (qui dépend entre autres de sa connaissance du script/langage utilisé), nous avons cherché à attaquer le problème de la caractérisation de performances au travers d'applications telles que le *word spotting*. L'objectif est avant tout d'évaluer l'intérêt pratique de l'approche proposée. Le problème de l'évaluation de la qualité des invariants reste néanmoins un problème largement ouvert. Au-delà de cet exemple, l'évaluation d'une solution de *clustering* reste l'objet de nombreux débats au sein de la communauté, certains clamant même que toute évaluation d'une solution de *clustering* en dehors d'une application pratique spécifique est vouée à l'échec [von Luxburg 2012], car elle va à l'encontre même du but exploratoire sous-jacent.

Un autre verrou scientifique auquel nous nous sommes attaqué est celui de la non-modularité, qui recouvre l'ensemble des difficultés que l'on rencontre dès lors que l'on cherche à transposer une solution développée dans un contexte applicatif donné vers un autre contexte. Dans le cadre de nos travaux sur la description d'images, nous avons franchi un premier pas en ce sens lorsque nous avons choisi en 2009 de suspendre nos travaux sur la conception de signatures visuelles dédiées à un type d'images spécifiques (visages, symboles, etc.) pour nous tourner vers des descripteurs extraits par catégorisation, depuis des collections d'images pour lesquelles on dispose de moins d'informations du domaine.

Nous nous sommes donc par la suite intéressés à la recherche d'approches plus génériques de description d'images, en cherchant le cas échéant à nous appuyer sur l'humain pour ajus-

ter (directement ou indirectement) les paramètres du système. Nous avons cependant récemment atteint **certaines limites dans cette généricité**, au travers de nos travaux en cours sur l'extraction d'invariants omni-script/omni-langage sans information *a priori* concernant le script/langage utilisé (voir section 2.3.4.2). De manière plus générale, les approches proposées dans ce chapitre ne permettent pas dans leur état actuel de conserver le bénéfice des informations déduites des retours de l'utilisateur, dès lors que l'on cherche à les transposer dans des contextes différents de ceux sur lesquels l'interaction a porté dans un premier temps. Nous avons réfléchi à différentes pistes pour pallier ce problème dans le cadre de l'organisation de bases d'images naturelles ; la piste que nous souhaitons privilégier est présentée en section suivante.

2.5 Perspectives

Les principales perspectives de recherche que nous avons identifiées à l'issue des travaux présentés dans ce chapitre concernent l'organisation de collection d'images naturelles. Dans son état actuel, l'approche de *clustering* semi-supervisé interactif que nous avons proposée nous permet d'obtenir des groupes d'images dotés d'un certain caractère sémantique (c'est-à-dire plus similaires au sens de l'utilisateur que s'ils étaient découverts sans supervision). Mais, de par la nature même du système, le travail fourni par l'utilisateur n'est finalement que partiellement généralisable dans un autre contexte (autre base d'images par exemple), ou dans le cas où la base d'images considérée varie avec le temps (corpus ouvert). Or, ce genre de cas se présente très fréquemment dans une grande variété d'applications.

Si, plutôt que d'obtenir en sortie une hiérarchie de groupes sémantiques d'images, nous obtenions une **hiérarchie sémantique de descripteurs visuels**, cette dernière pourrait plus facilement être transposée pour différentes collections d'images, ou pour des bases en évolution. Cependant, si l'interaction de l'utilisateur avec les images est assez naturelle, il n'en est pas de même de l'interaction directe de l'utilisateur avec les descripteurs extraits automatiquement des images.

Ce n'est que très récemment que certains chercheurs ont commencé à s'intéresser à la sélection des descripteurs d'images sur la foi de retours de l'utilisateur. On peut par exemple citer les travaux décrits dans [Sun 2012] dans un contexte d'indexation et de recherche d'images (CBIR). L'objectif est ici de se baser sur les retours de l'utilisateur concernant une image de requête donnée pour ne considérer que les caractéristiques (issues de la signature) les plus adaptées à cette requête lors de la recherche (et le cas échéant lors de recherches similaires). La méthode présentée consiste à intégrer, dans la fonction objective d'un algorithme de sélection de caractéristiques enveloppant [20], un critère de cohérence entre une caractéristique donnée et les retours exprimés par l'utilisateur lors de cette phase de « bouclage de pertinence ». Cette approche peut permettre d'obtenir des résultats localement satisfaisants (pour une image de requête donnée), mais la sélection ne se fait pas de manière globale pour toute la base. De plus, l'objectif visé ici se restreint à la sélection de caractéristiques, alors que nous souhaitons aller plus loin en concevant une structure hiérarchique de descripteurs.

Dans la littérature, on trouve quelques travaux qui s'attèlent à construire une hiérarchie de descripteurs visuels. Dans la plupart de ces travaux, la notion de hiérarchie porte sur l'aspect spatial [Lazebnik 2006, Ranzato 2007, Boureau 2012, Goh 2013].

Notre objectif est de concevoir une méthode permettant de construire une structure hiérarchique de descripteurs bénéficiant à la fois des informations de bas niveau sur le contenu de l'image, d'informations partielles apportées par les interactions avec l'utilisateur et, le cas échéant, d'informations spatiales sur le contenu de l'image. Bien évidemment, la méthode que nous concevrons au final dépendra grandement des modalités retenues pour l'interaction avec l'utilisateur, et du type de caractéristiques visuelles considérées (voir section 2.1.1). Pour spécifier notre objectif, nous nous plaçons ci-après dans le cas où l'interaction est mise en œuvre au niveau des groupes d'images (à la manière de notre système de *clustering* semi-supervisé

20. Ces techniques [Jain 1997] recherchent le sous-ensemble minimal de caractéristiques permettant d'optimiser une fonction objective liée à la tâche finale (ici la recherche d'images).

et interactif présenté en section 2.2), et où l'on utilise des descripteurs basés sur des « mots visuels » [Sivic 2003, Chatfield 2011, Avila 2013].

Dans la pratique, on s'aperçoit souvent que les lexiques de sacs de mots visuels construits « à plat » contiennent naturellement de nombreux synonymes et/ou polysèmes (en fonction notamment de la taille du lexique considéré), ce qui est source d'ambiguïtés et d'incertitude dans la description. Quelques travaux se sont intéressés à générer un lexique de niveau plus global dans l'image, en agrégeant non plus simplement des mots, mais des chaînes de mots [Ros 2009] ou des phrases, constituées de groupes de mots visuels significatifs que l'on retrouve fréquemment ensemble [Yuan 2007], voire dans une même configuration spatiale [Zhang 2011] dans les images de la collection.

Dans notre cas, en plus de ces informations spatiales, nous pourrions nous aider des retours de l'utilisateur pour constituer une hiérarchie sémantique de phrases visuelles. Cette hiérarchie pourrait être construite en se basant sur un algorithme d'apprentissage semi-supervisé multi-niveaux qui nous permettrait de répercuter les interactions de l'utilisateur, obtenues au niveau des images, sur le plus bas niveau des descripteurs, tout en prenant en compte des informations spatiales (par le biais par exemple de phrases visuelles). Il s'agit donc d'être capable de mener un *clustering* à plusieurs niveaux d'analyse tout en permettant aux groupes des niveaux supérieurs de partager les caractéristiques que l'on retrouve à des niveaux inférieurs de la hiérarchie. Il s'agit d'un problème de modélisation que l'on retrouve, sous différentes formes, dans des domaines variés de traitement du signal [Wulsin 2012], mais qui à notre connaissance est encore peu traité dans le domaine de l'image.

Les principales questions auxquelles il nous faudra faire face afin d'atteindre cet objectif sont détaillées ci-après.

Le **choix des descripteurs locaux** à utiliser en entrée de notre système. Jusqu'à présent dans nos études initiées en 2009, nous ne sommes pas allés plus loin que la sélection de la variante couleur du très traditionnel SIFT qui nous donnait les meilleurs résultats expérimentaux dans notre contexte (en l'occurrence il s'agissait de rgSIFT). Et ce, à la fois pour la phase de détection des points d'intérêt (basée sur un détecteur de Harris), et pour la phase de description. Or, des études plus récentes montrent que les descripteurs échantillonnés depuis une grille dense sont plus performants pour des tâches de catégorisation d'objet [Chatfield 2011]. En ce qui concerne la phase de description proprement dite, la complexité calculatoire du descripteur rgSIFT est un frein dans notre contexte interactif. Il convient donc d'investiguer d'autres solutions, qui soient plus efficaces d'un point de vue calculatoire, tout en gardant une capacité de description suffisante [Burghouts 2009].

Le **choix de la représentation par mots visuels** de ces descriptions locales est également crucial. Dans nos travaux précédents, nous nous sommes cantonnés à utiliser la représentation par sacs de mots visuels de [Sivic 2003], qui repose sur une quantification « en dur », alors que des travaux plus récents montrent les bénéfices de la prise en compte de l'incertitude lors de l'assignation des descripteurs aux mots du lexique pour améliorer la stabilité des mots visuels [Philbin 2008]. Plusieurs alternatives à l'agrégation moyenne traditionnellement utilisée, visant à injecter des contraintes de proximité entre les descripteurs, ont également été proposées [Avila 2013]. Le choix de la représentation par mots visuels devra nécessairement s'appuyer sur une étude en profondeur de la littérature, qui est riche sur ce sujet. Spécifiquement dans notre contexte interactif qui impose des contraintes importantes en termes de temps de calcul, il nous

faudra être vigilant dans le choix d'une solution offrant un bon compromis entre compacité et capacité de représentation.

Mais, les principales questions liées à ce travail concerneront l'**apprentissage semi-supervisé multi-niveaux**, et plus spécifiquement la manière de prendre en compte conjointement dans cet apprentissage les contraintes spatiales et celles, de plus haut niveau sémantique, apportées par l'utilisateur, pour les répercuter sur la structuration des descripteurs visuels. Et ce, d'une manière qui soit incrémentale et efficace, étant donnée l'interactivité avec l'utilisateur. La prise en compte de l'incertitude dans la hiérarchie pourrait améliorer sa stabilité. Une question corollaire à ce traitement multi-niveaux touche à la possibilité de déterminer, à chaque étape interactive, quelles sont les images qui, si elles étaient annotées par l'utilisateur, pourraient induire le plus de changements dans la hiérarchie de descripteurs (et non plus dans les groupes d'images).

S'agissant de perspectives de recherche s'attaquant à de nombreux verrous scientifiques liés notamment à l'apprentissage, à la théorie de l'incertitude et aux interactions homme-machine, celles-ci ne pourront être déployées qu'à moyen, voire à plus long terme. Comme détaillé en section 1.4 (page 15), nous avons obtenu deux financements pour poursuivre nos recherches dans cette direction. Premièrement, **une thèse** devrait commencer en 2015, en co-supervision entre le L3i, le laboratoire XLIM-SIC de Poitiers et la VAST (Hanoï). Afin de nous permettre de constituer une petite équipe de recherche sur cette thématique, nous venons en outre d'obtenir un financement de type « projet exploratoire » de la part du GdR ISIS (**projet CINÉDI**). Par ailleurs, nous avons déposé auprès de l'ANR un projet « Jeunes Chercheurs » (dont je suis porteuse) sur ce thème, en octobre 2014.

2.6 Faits marquants liés à ces contributions

2.6.1 Synthèse des faits marquants

Section	Type de contribution majoritaire	Encadrement	Projets	Collaborations (co-publication ou co-encadrement)	Publications
Section 2.1 + annexe A	Méthodologique et applicatif	2 stages [ST-Chir09] [ST-Cous07]	Projet ANR NAVIDOMASS	CVC, Barcelone LORIA, Nancy	2 revues internationales [Coustaty 2011] [Visani 2011b] 1 chapitre de livre [Coustaty 2008] 6 conf. inter. [Laurent 2003, Visani 2004a] [Visani 2004b, Visani 2005b, Visani 2005c] [Visani 2005d] 1 conf. francophone [Visani 2005e] 1 brevet international PCT [Visani 2006]
Section 2.2	Méthodologique	1 thèse [TH-Lai13] 2 stages [ST-Bui14] [ST-Feli13]	Projet GdR ISIS CINÉDI	UMMISCO, Hanoï XLIM/SIC, Poitiers VAST/IOIT, Hanoï	1 thèse [Lai 2013a] 4 revues internationales [Lai 2012a] [Lai 2013b, Lai 2014a, Lai 2014b] 1 conf. internationale [Lai 2012b]
Section 2.3	Applicatif	1 thèse [TH-BuiXX]	Projet PCSI AUF Projet ARCHIVES	Univ. Can Tho, Vietnam	4 conf. internationales [Le 2012] [Bui 2013, Le 2013, Le 2014] 1 conf. nationale [Bui 2012]

TABLE 2.1 – Faits marquants liés aux contributions présentées dans le chapitre 2.

2.6.2 Encadrements en lien avec ces contributions [21]

[TH-Lai13] Lai H.P. *Vers un système interactif de structuration des index pour une recherche par le contenu dans des grandes bases d'images*. Thèse de doctorat co-supervisée par moi-même et Alain Boucher, laboratoire UMMISCO (UMI UPMC/IRD), Hanoï, Vietnam. Directeur de thèse : OGIER Jean-Marc, L3i/université de La Rochelle. Thèse soutenue à l'université de La Rochelle le 2 Octobre 2013. Manuscrit en anglais [Lai 2013a].

[TH-BuiXX] Bui Q.A. *Extraction d'invariants par clustering interactif pour la navigation dans des bases de documents anciens omni-script*. Thèse de doctorat co-supervisée par moi-même et Rémy Mullot, directeur de thèse, L3i/université de La Rochelle. Thèse commencée en Avril 2011, en cours de rédaction.

[ST-Bui14] Bui D.C. Stage de Master 1. *Review of metric learning techniques for semi-supervised clustering*. 2014.

[ST-Feli13] J.L. Félix. Stage de Master 2. *Amélioration de la description visuelle d'une image*. 2013.

[ST-Chir09] G. Chiron. Stage de Master 1. *Étude préliminaire sur la comparaison de la robustesse de différents descripteurs de forme vis-à-vis du bruit de Kanungo pour des symboles techniques*. 2009.

[ST-Cous07] M. Coustaty. Stage de Master 2. *Conception d'une signature structurelle dédiée aux symboles*. 2007.

21. Pour plus de détails sur ces encadrements, merci de se référer à la section 2.3 du Tome I.

Chapitre 3

Contributions dans le domaine de la reconnaissance d'images

Reconnaissance d'images structurées par classification

Sommaire

Tables

Figures

3.1 Introduction

La reconnaissance d'images fait partie des traitements très couramment rencontrés dans l'analyse d'images. Elle vise à inférer, à partir d'informations de bas niveau sémantique

extraites des images et, éventuellement, d'informations ou de connaissances souvent spécifiques au domaine d'application (imagerie médicale, image de document, etc.), des propriétés. Ces propriétés portent le plus souvent sur le(s) type(s) et/ou la structure du contenu des images. En fonction de la tâche et de l'application visée, la reconnaissance peut se décomposer en une **multitude de problèmes, chacun avec ses spécificités et ses situations exceptionnelles**.

Les efforts de recherche déployés depuis plusieurs décennies ont fait que certains problèmes très spécifiques de reconnaissance sont aujourd'hui considérés comme étant pratiquement résolus, comme par exemple l'identification de chiffres isolés (*digits*) ou encore la vérification d'empreintes digitales [1].

Cependant, dans le cas général, une très grande variété de problèmes de reconnaissance d'images restent à l'heure actuelle non résolus. En attestent les nombreux concours organisés par la communauté scientifique, qui permettent d'évaluer à intervalles réguliers les avancées réalisées pour une tâche particulière de reconnaissance. Les tâches visées sont très diverses et vont par exemple de la détection et/ou reconnaissance d'objets dans des images de scènes naturelles [2] à la reconnaissance de formules mathématiques manuscrites dans des images de documents [3]. Dans des conditions d'application réelles, les résultats obtenus sont souvent plus imparfaits que ceux annoncés lors de ces compétitions.

L'incapacité des approches actuelles à résoudre complètement ces problèmes peut être expliquée par la complexité de ces derniers et par les difficultés liées à leur formalisation, mais aussi – en partie – par l'insuffisance des outils de description des informations et/ou des connaissances sur lesquels ils reposent. À ces difficultés d'ordre scientifique viennent s'ajouter de nombreuses difficultés techniques fréquemment rencontrées en pratique (mauvaise qualité d'images ou faible résolution, conditions d'acquisition peu contrôlées, pré-traitements mal ajustés, etc.).

Dans ce chapitre, je m'intéresse plus spécifiquement à la reconnaissance d'images structurées, et en particulier à la **reconnaissance d'images de documents**. La section suivante vise à donner un aperçu de ce domaine de recherche, avant de se focaliser sur la manière dont la classification peut être mise à profit dans ce cadre applicatif.

3.1.1 Reconnaissance d'images de documents par classification supervisée

La reconnaissance d'images de documents [Nagy 2000, Ingold 2002, Trupin 2005] est un domaine de recherche très ancien. En effet, la lecture optique de caractères a connu ses balbutiements dès le 19$^{\text{ème}}$ siècle avec le dépôt de brevets concernant par exemple des systèmes d'assistance à la lecture pour des non-voyants. La première machine à trier le courrier basée sur la reconnaissance d'adresses postales (dactylographiées) fut inaugurée aux États-Unis en 1965. Mais c'est dans les années 1980 que ce domaine a connu un véritable essor, avec la démocratisation des ordinateurs personnels et de la bureautique. Cet essor n'a fait que s'amplifier dans les trente dernières années, avec le développement de campagnes de dématérialisation massives menées dans le domaine privé, les services publics (en général dans un souci d'efficacité du traitement de l'information) et dans les bibliothèques ou centres d'archives (souvent dans une

1. Arrêt des compétitions Fingerprint Verification Competition en 2006.
2. On peut citer les compétitions Pascal VOC (de 2005 à 2012) et IMAGENET ILSVRC (depuis 2010).
3. Par exemple, la compétition CROHME organisée en marge de la conférence ICFHR.

optique de préservation ou de sauvegarde, voire de diffusion du patrimoine documentaire).

La notion de document est très générale et peut, dans sa terminologie la plus large, s'étendre à différents types d'objets multimedia. Dans la suite de ce chapitre, le terme « document » concerne les documents à prédominance textuelle et/ou graphique (ceux qu'Éric Trupin qualifiait dans [Trupin 2005] de « symboliques »). Ces documents peuvent être statiques ou dynamiques (digitalisés à partir de documents papier ou acquis par le biais d'un dispositif tactile ou électromagnétique par exemple).

Les tâches relevant de la reconnaissance d'images de documents sont très nombreuses. Elles permettent généralement d'**extraire des propriétés de l'image pouvant contribuer à une meilleure compréhension** de son contenu. Ces propriétés pourront être mises à profit, le cas échéant, par un processus d'interprétation d'images. Les tâches de reconnaissance peuvent être guidées par l'analyse de la structure physique et/ou logique des documents. La structure physique d'un document peut être décrite par la localisation, la description et l'agencement spatial relatif des différentes zones de texte (resp. de graphique) composant l'image originale. Sa structure logique consiste en la description de la nature et du rôle de chaque élément (titre de chapitre, titre de section, illustration, légende, etc.), ainsi que de l'ensemble des liens (d'ordre hiérarchique ou logique) entre éléments [André 1990]. L'analyse des structures physiques et logiques peut s'appuyer sur des traitements menés dans l'espace de représentation de bas niveau sémantique des descripteurs visuels (détection ou segmentation de motifs dans l'image par exemple) et/ou sur des informations voire des connaissances souvent spécifiques au domaine d'application.

Quelle que soit la chaîne de processus embarquée dans le système de reconnaissance, on est très fréquemment confronté au problème de la classification semi-supervisée ou supervisée, relevant de la fouille de données. Très généralement, la classification d'images consiste souvent à détecter, authentifier (vérifier) ou identifier un objet ou un motif dans l'image à reconnaître [Javidi 2002]. Dans le cas de la reconnaissance d'images de documents, il peut par exemple s'agir de détecter un logo dans un document, de vérifier la validité d'une signature ou d'identifier un caractère dans une case de formulaire. Dans la suite de ce chapitre, je m'intéresse plus spécifiquement aux tâches d'identification et de vérification.

Comme évoqué précédemment, la reconnaissance d'images de documents, tout comme la reconnaissance d'images, se décline en une multitude d'applications ayant chacune de fortes spécificités, rendant toute tentative d'unification du formalisme difficile. Du point de vue des systèmes, en fonction du type d'application visé, les processus de reconnaissance peuvent être soit basés sur une chaîne de traitement assez traditionnelle [Ingold 2002], soit sur des chaînes de traitement plus spécifiques. Si l'on se focalise sur la reconnaissance par classification (semi-supervisée ou supervisée), un aperçu d'une chaîne de traitement traditionnelle est illustré en Figure 3.1. Cet aperçu ne prétend pas être assez exhaustif pour couvrir l'ensemble des applications possibles ; il vise simplement à décrire les processus typiquement embarqués dans un système de reconnaissance d'images reposant sur une classification. Il nous servira de fil conducteur dans la suite de ce chapitre, où nous replacerons systématiquement les travaux présentés dans cette chaîne de traitement. À cette occasion, des exemples concrets d'applications finales seront détaillés, et nous passerons en revue un certain nombre d'approches typiquement utilisées dans ces applications.

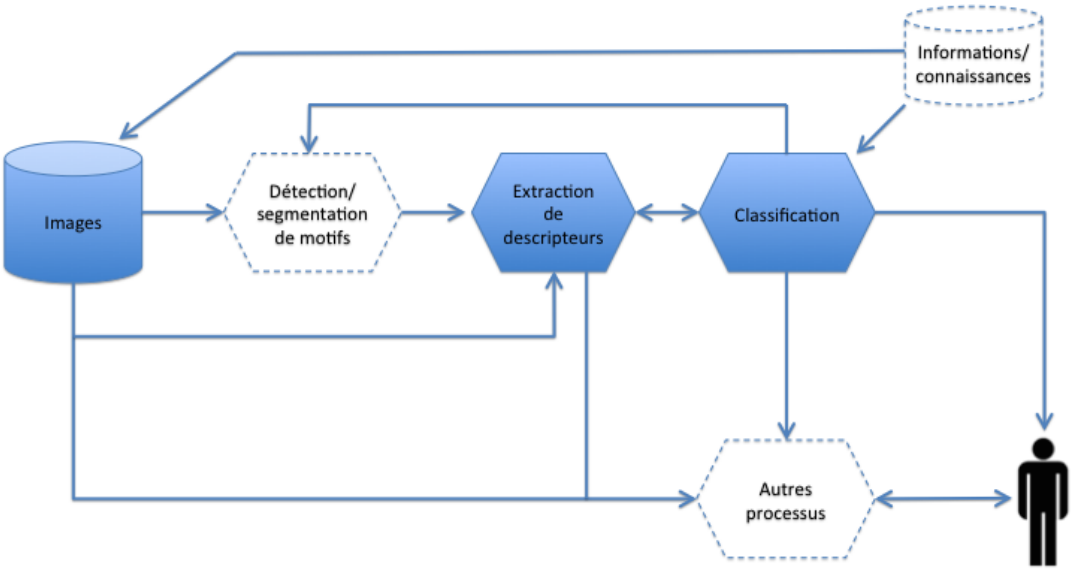


FIGURE 3.1 – Aperçu de la chaîne de traitement traditionnelle de classification pour la reconnaissance d'images de documents. Les étapes en pointillés sont optionnelles.

Très généralement, cette chaîne de traitement vise, à partir des images (éventuellement additionnées de quelques informations ou connaissances supplémentaires), à classer le contenu de ces images dans des catégories connues. Les techniques de classification mises en œuvre sont typiquement basées sur un apprentissage discriminatif et/ou génératif (voir annexe D).

La classification peut s'appliquer au niveau des images directement, par exemple pour « dégrossir » le contenu de la collection en rangeant les images dans des types tels que des documents à prédominance graphique ou textuelle. Elle peut également s'appliquer au niveau plus fin de motifs tels que des objets graphiques, des mots, des caractères, des symboles mathématiques ou des graphèmes détectés et/ou segmentés dans l'image. Elle passe par une phase d'extraction de descripteurs, sur laquelle nous ne nous étendrons pas outre mesure ici, puisqu'il s'agit de l'objet du chapitre précédent.

Ces différentes étapes (détection/segmentation, extraction de descripteurs et classification) peuvent être menées de manière séquentielle ou coopérative. Par exemple, certaines approches visent à raffiner la segmentation grâce aux sorties du classifieur, ou encore à apprendre conjointement les descripteurs et le classifieur (*cf.* section 2.1.1).

L'information extraite à l'issue de la catégorisation pourra alors être retournée, sous une forme plus ou moins structurée, soit à l'humain directement, soit à la machine pour la mise en œuvre d'autres processus (qui sont en fait le plus souvent hors de la chaîne de traitement traditionnelle, mais représentés sur le même schéma pour placer cette chaîne dans un contexte plus large). Ces autres processus peuvent par exemple consister en une indexation pour la recherche ultérieure des documents, en une analyse plus poussée de ses structures physiques et logiques, voire en de l'interprétation de plus haut niveau du contenu des images.

Il est important de noter ici que, de par la nature même du problème, on dispose *a priori* d'**informations supplémentaires par rapport au cas de la description d'images** étudié au chapitre précédent. En effet, dans le chapitre 2, l'interaction avec l'utilisateur sert essentiel-

lement à aider la machine à découvrir les catégories perçues par l'utilisateur. Dans les études menées dans le présent chapitre au contraire, on dispose en entrée de l'apprentissage d'un ensemble de documents associés à une vérité-terrain obtenue le plus souvent grâce à une annotation menée par (ou sous la supervision de) l'humain. Nous nous intéressons en particulier au cas de la classification supervisée, où l'ensemble des exemples d'apprentissage sont annotés par leur vérité-terrain. Dans les contextes applicatifs visés, on considère que pour chacune des catégories (ou sous-catégories) à reconnaître, on dispose d'un ensemble d'exemples d'images lui appartenant.

En revanche, comme expliqué en section 1.3.1.2 « Images considérées » (page 11), nous ne disposons pas dans le contexte de nos études de connaissances du domaine formalisées explicitement.

3.1.2 Contenu et organisation du reste du chapitre

La suite de ce chapitre vise à étudier différents tronçons de la chaîne de traitement présentée en Figure 3.1.

La section 3.2 s'intéresse en particulier à la phase de classification supervisée, et à la manière de restituer les informations apprises lors de la catégorisation à un humain. La section 3.3 est focalisée sur les étapes de segmentation de motifs et de classification, et sur leurs modes de coopération, dans un contexte où l'on dispose d'informations supplémentaires de natures variées. La section 3.4, elle, porte plutôt sur les manières de pallier un éventuel manque de données annotées pour l'apprentissage ou la caractérisation de performances. Deux de ces trois études, menées dans un contexte contractuel, portent sur un cas d'étude ciblé : respectivement, la reconnaissance d'écriture manuscrite en-ligne (section 3.3) et l'analyse de documents anciens (section 3.4).

Chacune de ces trois sections se termine par une discussion critique sur les avancées liées à nos travaux. Il s'agit de mesurer le chemin qu'il nous reste à parcourir au regard des questions de recherche concernées et, le cas échéant, des tendances qui se sont dégagées entre-temps dans la communauté scientifique.

La section 3.5 conclut ce chapitre, tandis que la section 3.6 en dresse les perspectives.

Enfin, la section 3.7 permet de synthétiser les faits marquants liés à ces contributions en termes notamment d'encadrements, de publications, de projets et de collaborations nationales et internationales.

3.2 Proposition d'une approche de classification basée sur un treillis des concepts

3.2.1 Introduction

3.2.1.1 Objectifs visés

Dans cette section, nous nous intéressons aux phases de classification supervisée et de restitution des informations de classification apprises à l'utilisateur humain. C'est-à-dire que nous nous focalisons sur la portion de la chaîne de traitement traditionnelle de la reconnaissance d'images mise en évidence dans la Figure 3.2.

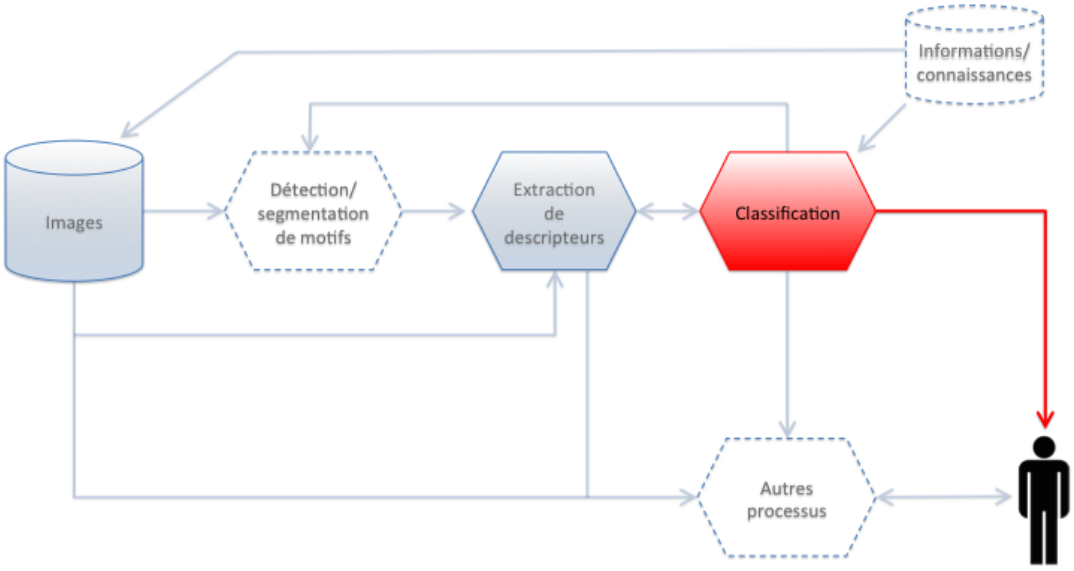


FIGURE 3.2 – Portion de la chaîne de traitement traditionnelle de classification pour la reconnaissance d'images de documents à laquelle on s'intéresse dans cette section.

On cherche à concevoir des méthodes de classification supervisée (basées sur un apprentissage discriminatif), qui soient capables de restituer – le cas échéant de manière ciblée – à un humain expert du domaine d'application (ici l'analyse de documents) les informations apprises lors de l'entraînement du classifieur, d'une manière qui soit intelligible pour cet expert.

Dans ces travaux, nous avons également souhaité nous intéresser à la question de la généricité, au travers de la conception d'une approche qui soit facilement transposable d'un contexte applicatif à un autre. Concrètement, nous cherchons à concevoir notre approche de manière à ce que ses éventuels paramètres ne soient pas dépendants du domaine d'application visé (par exemple l'analyse d'images de documents).

Notre proposition (formulée dans le cadre des thèses de Stéphanie Guillas[4] [Guillas 2007], puis de Nathalie Girard[5]) [Girard 2013], est de concevoir à ces fins **une approche non paramétrique, et plus précisément symbolique, inspirée de l'Analyse Formelle des**

4. Dont j'ai participé informellement à l'encadrement (soutenance 1 an après mon recrutement).
5. Doctorante que j'ai supervisée scientifiquement sous la direction de Karell Bertet (L3i).

Concepts (AFC).

L'AFC est une théorie de la représentation de connaissances, de la gestion de l'information et de l'analyse de données qui vise à identifier, depuis un jeu de données éventuellement associé à des informations ou des connaissances additionnelles, des structures conceptuelles agencées de manière hiérarchique sous la forme de treillis. L'AFC fournit des outils très puissants qui sont encore peu utilisés dans les domaines de l'analyse et de l'interprétation d'images, à l'exception notable et récente de [Atif 2014] qui s'attaque au problème de la compréhension de scènes en proposant une approche basée sur un treillis des concepts construit en prenant en compte à la fois des informations extraites des images, et des connaissances *a priori* apportées par des ontologies.

Dans notre cas, nous nous intéressons à l'application de cette théorie à la classification de données pour lesquelles on ne dispose pas d'informations formalisées explicitement sous la forme de connaissances. Le niveau de sémantique du traitement que l'on applique aux données est donc bien en-deçà des possibilités offertes par les treillis. Dans ce contexte, nous avons choisi d'utiliser l'AFC principalement parce que la structure de treillis permet de décrire les règles de classification apprises sous une forme graphique lisible pour un humain, tout en conservant leur diversité.

Comme nous venons de le voir, l'un de nos objectifs principaux est de définir une méthode générique qui puisse s'appliquer sur une variété de problèmes d'analyse de données (quantitatives ou qualitatives). Nous avons entre autres étudié une application relevant de la reconnaissance d'images de documents, à laquelle nous accorderons une importance particulière dans la suite, et qui est présentée dans la section suivante.

3.2.1.2 Contexte applicatif concernant la reconnaissance d'images de documents

La principale application à la reconnaissance d'images de documents visée par ces travaux concerne l'identification de symboles graphiques dans des documents techniques. Il peut s'agir par exemple des symboles architecturaux étudiés en annexe A au travers de leur description, et dont un aperçu est donné en Figure 3.3.

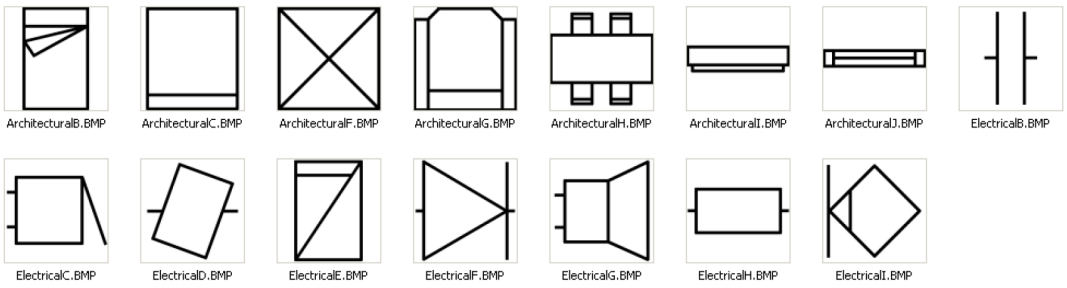


Figure 3.3 – Exemples de symboles graphiques extraits de la base GREC 2003 [6].

Il convient ici de distinguer la reconnaissance de symboles graphiques « en contexte », où les symboles peuvent être connectés à d'autres éléments graphiques ou textuels du document, de la reconnaissance de symboles graphiques isolés [Tombre 2006]. Dans le cas « en contexte », plus complexe à traiter, la segmentation des symboles s'aidera généralement d'une phase préalable de

6. http://www.cvc.uab.es/grec2003/SymRecContest/

segmentation du document en texte et graphique, voire de la localisation d'autres éléments d'intérêt dans l'image (par exemple des murs dans le cas de symboles architecturaux) [Dosch 2000]. Dans tous les cas, il n'est pas rare que la segmentation et la classification se fassent de manière coopérative, le cas échéant *via* une phase de vectorisation du symbole en primitives élémentaires (traits, arcs de cercle, etc.).

Dans nos travaux, l'application que nous avons privilégiée, au cœur des activités du laboratoire L3i lorsqu'ils ont été initiés en 2004, concerne l'**identification de symboles graphiques isolés**. On considère que les symboles ont préalablement été détectés dans les images de documents et qu'ils sont décrits au travers de signatures visuelles (de bas niveau sémantique). Les signatures visuelles utilisées pour ce genre de symboles graphiques visent généralement à décrire les formes présentes dans le symbole. Elles peuvent être de nature statistique [Tabbone 2006] et/ou structurelle [Llados 2003, Coustaty 2011, Bunke 2011, Li 2013]. Nous nous intéressons ici en particulier au cas des signatures de forme de nature statistique ou statistico-structurelle. Ces dernières nous permettent d'alimenter, pour chaque image, le processus de classification supervisée par un vecteur de caractéristiques visuelles du contenu de cette image.

3.2.2 Usages des treillis des concepts pour la classification supervisée

Dans la suite de cette section, je m'attache à définir la structure de treillis des concepts et à donner une brève typologie des approches de classification supervisée de données utilisant cette structure.

3.2.2.1 Définition de la structure de treillis des concepts

Les treillis des concepts (aussi appelés treillis de Galois) ont été introduits pour la première fois de manière formelle dans la théorie des graphes et des structures [Birkhoff 1967, Davey 1991]. Ce n'est que plus tard que cette théorie a été développée dans le domaine de l'AFC [Ganter 1999].

Un treillis des concepts est un graphe permettant de représenter la relation R entre un ensemble d'enregistrements (aussi appelés objets ou individus) O et un ensemble d'attributs discrets I. Ce graphe est composé d'un ensemble de « concepts formels » ordonnés par une relation vérifiant la **propriété de treillis**, à savoir qu'il s'agit d'une relation d'ordre (transitive, réflexive et antisymétrique) telle que, pour chaque paire de concepts dans le graphe, il existe à la fois une borne supérieure et une borne inférieure (voir ci-après). La relation R peut également se décrire au travers de deux relations f et g qui forment une connexion de Galois entre les objets et les attributs, c.-à-d. que f et g sont antitones, et leurs compositions fog et gof sont extensives :

$$\forall O' \subseteq O, \quad f(O') = \{i \in I \mid oRi \ \forall \ o \in O'\} \tag{3.1}$$

$$\forall I' \subseteq I, \quad g(I') = \{o \in O \mid oRi \ \forall \ i \in I'\} \tag{3.2}$$

Un concept formel représente les correspondances maximales entre objets $O' \subseteq O$ et attributs $I' \subseteq I$ qui vérifient $f(O') = I'$ et $g(I') = O'$. Deux concepts formels (O_1, I_1) et (O_2, I_2) sont en relation dans le treillis dès lors qu'ils vérifient la propriété d'inclusion suivante :

$$(O_1, I_1) \leq (O_2, I_2) \Leftrightarrow \left\| \begin{array}{l} O_2 \subseteq O_1 \\ \text{(équivalent à } I_1 \subseteq I_2\text{)} \end{array} \right. \tag{3.3}$$

L'ensemble des concepts formels (plus simplement désignés par le terme « concepts » dans le reste de la section 3.2), ordonnées par $\leq$, constitue le treillis des concepts. En effet, il vérifie la propriété de treillis : la relation $\leq$ est clairement une relation d'ordre, et pour chaque paire de concepts (O_1, I_1) et (O_2, I_2), il existe une borne inférieure (resp. une borne supérieure) appelée *meet* (resp. *join*) et notée $(O_1, I_1) \wedge (O_2, I_2)$ (resp. $(O_1, I_1) \vee (O_2, I_2)$) définie par :

$$(O_1, I_1) \wedge (O_2, I_2) = (g(I_1 \cap I_2), (I_1 \cap I_2)) \tag{3.4}$$

$$(O_1, I_1) \vee (O_2, I_2) = ((O_1 \cap O_2), f(O_1 \cap O_2)) \tag{3.5}$$

En conséquence, un treillis contient un concept minimum et un concept maximum selon la relation $\leq$, appelés respectivement *bottom* et *top* du treillis, et notés $\perp = (O, f(O))$ (resp. $\top = (g(I), I)$). De plus, les compositions $\beta o\alpha$ et $\alpha o\beta$ sont des opérateurs de fermeture (c.-à-d. des opérateurs isotones, extensifs et idempotents) définis respectivement sur O et I, et la restriction des concepts aux seuls objets (respectivement aux seuls attributs) forme un treillis des fermés sur I (resp. O). Donc, si (O', I') est un concept formel du treillis des concepts, alors O' est un fermé de O (puisque $\beta o\alpha(O') = O'$) et I' est un fermé de I (puisque $\beta o\alpha(I') = I'$).

Le treillis est généré à partir d'un « contexte formel » résumant les informations en provenance du triplet (O, I, R) (ou $(O, I, (f, g))$). L'ensemble I peut être constitué d'attributs originellement qualitatifs, ou bien d'attributs qualitatifs dérivés d'attributs quantitatifs. Dans le cas applicatif de la reconnaissance de symboles par exemple, nous considérons que chaque enregistrement est constitué d'une image de symbole, et que chaque attribut d'entrée est constitué d'une des variables composant les vecteurs des signatures d'images. Ces attributs d'entrée sont généralement quantitatifs ; nous avons choisi de les transformer en un ensemble I d'attributs correspondant aux modalités d'attributs qualitatifs ordinaux obtenus par « discrétisation » des variables d'entrée en intervalles disjoints. Nous reviendrons plus en détails sur cette notion de discrétisation par la suite. La relation R est décrite à partir de toutes les correspondances observées entre objets et attributs. Un exemple de contexte formel ainsi obtenu est donné dans la Table 3.1.

Le treillis peut donc être vu comme une **structure hiérarchique de représentation de données**, tout comme l'arbre. Néanmoins, dans les structures arborescentes, chaque nœud possède un unique père et donc le chemin menant de la racine à un nœud donné est unique, tandis que dans les treillis des concepts il y a plusieurs chemins possibles entre le *bottom* (à rapprocher de la racine dans le cas des arbres), et un concept donné du treillis (à rapprocher des nœuds internes ou des feuilles dans les arbres). Le diagramme de Hasse (représentation simplifiée par la suppression des arcs de transitivité et de réflexivité) du treillis construit à partir du contexte formel de la Table 3.1 est montré à titre d'exemple en Figure 3.4.

Classe	Objet	Attributs					
		A		*B*		*C*	
		$\mathbf{a_1}$ *[0-3]*	$\mathbf{a_2}$ *[6-20]*	$\mathbf{b_1}$ *[0-4]*	$\mathbf{b_2}$ *[12-20]*	$\mathbf{c_1}$ *[0-2]*	$\mathbf{c_2}$ *[11-20]*
1	1	X		X			X
	2	X		X			X
2	3	X			X		X
	4	X			X		X
	5	X			X		X
3	6		X		X		X
	7		X		X		X
	8		X		X		X
4	9		X	X		X	
	10		X		X	X	

TABLE 3.1 – Exemple de contexte formel agrémenté (en première colonne) de la classe des objets. Chacun des trois attributs A, B et C ont été discrétisés en deux intervalles (respectivement a_1 et a_2, b_1 et b_2 et c_1 et c_2) couvrant l'ensemble des valeurs observées dans la base d'apprentissage. À noter que les attributs ne sont pas forcément binaires, même si c'est le cas dans cet exemple. *Table adaptée de [Girard 2013].*

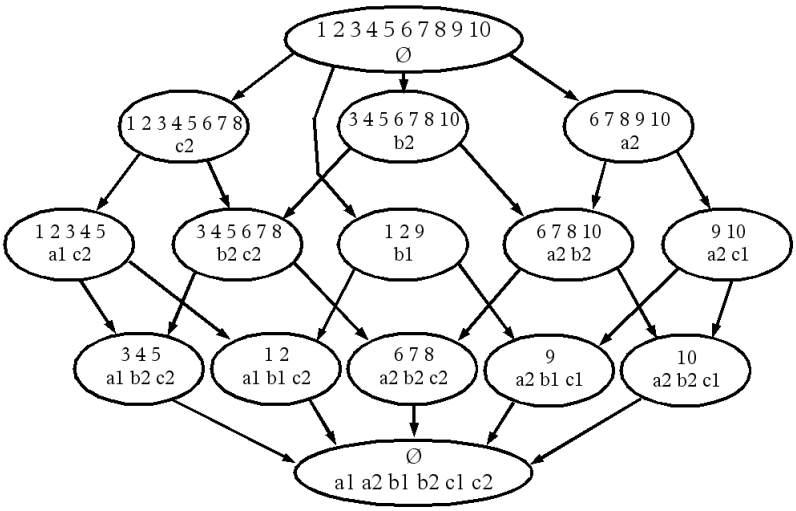


FIGURE 3.4 – Diagramme de Hasse du treillis des concepts généré à partir du contexte formel de la Table 3.1. Afin de se rapprocher du mode de représentation des arbres, le diagramme est inversé par rapport au diagramme de Hasse traditionnel, à savoir que nous avons placé le *bottom* (équivalent de la racine de l'arbre) en haut et non en bas. *Figure extraite de [Girard 2013].*

3.2.2.2 Approches traditionnelles de classification supervisées basées sur un treillis des concepts

Les avancées technologiques des dernières décennies permettent l'utilisation des treillis des concepts pour des problèmes de fouille de données, malgré le fait que leurs complexités

de construction et de stockage soient souvent exponentielles (en fonction de la taille des données) dans le pire des cas. On peut cependant noter qu'en pratique, selon les spécificités du problème, la taille et le temps de construction du treillis peuvent rester raisonnables, surtout si les attributs d'entrée I sont judicieusement pré-sélectionnés [Jain 1997] ou transformés [Van der Maaten 2009].

La section 3.4. de la thèse de Nathalie Girard [Girard 2013] détaille les principales méthodes traditionnelles de classification supervisée existantes basées sur les treillis des concepts. Ces approches peuvent être rangées en deux types. Le premier type regroupe les méthodes qui sont basées sur une sélection de certains concepts dans le treillis, soit directement (p. ex. LEGAL et LEGAL-E [Mephu-Nguifo 1993], GALOIS [Carpineto 1993] et la méthode présentée dans [Zenou 2004]), soit pour en déduire une sélection de règles de classification (p. ex. GRAND [Oosthuizen 1988] et RULEARNER [Sahami 1995]), soit encore pour en déduire une sélection des objets les plus représentatifs du jeu de données à analyser (p. ex. CIBLe [Njiwoua 1999]). Le second type d'approches repose sur l'extraction/sélection de règles contextuelles à partir du treillis : c'est le cas des méthodes CLNN et CLNB [Xie 2002] par exemple.

La classification se fait le plus souvent grâce à un processus externe, à partir des informations sélectionnées dans le treillis, typiquement avec une règle des k plus proches voisins ou un classifieur Bayésien naïf. Le cas échéant, un vote majoritaire est appliqué.

Dans des applications variées dont le spectre dépasse largement celui de la reconnaissance d'images, ces techniques permettent d'atteindre des taux de reconnaissance similaires à ceux de méthodes symboliques plus classiques, telles que les arbres de classification [Safavian 1991, Kothari 2000].

3.2.3 Principales contributions

Nous avons proposé une approche de classification supervisée qui, à la différence de la plupart des méthodes traditionnelles, ne repose pas sur la sélection de concepts ou de règles contextuelles les plus pertinentes dans le treillis pour alimenter un classifieur externe. Plus précisément, **notre approche repose sur une classification par navigation** dans le treillis, de manière similaire aux arbres de décision. En effet, lors de l'étape de classification, les objets à reconnaître naviguent dans le treillis depuis le *bottom* vers le *top* (du haut vers le bas dans la Figure 3.5) par validation (grâce à une distance floue) de leurs attributs discrétisés, jusqu'à atteindre un « concept terminal » qui détermine la classe dans laquelle ils sont catégorisés. Les concepts terminaux sont à rapprocher de la notion de feuilles dans les arbres de décision, et sont représentés par des carrés étiquetés par leur classe dans la Figure 3.5.

Suivant sa définition donnée plus haut, le treillis des concepts est intrinsèquement un outil de représentation des données, plutôt qu'un outil de classification supervisée. Cela explique en partie l'usage qui en est traditionnellement fait pour la classification supervisée, où le treillis sert surtout à sélectionner de l'information concernant les données afin d'alimenter un classifieur externe. Les difficultés liées à son utilisation pour la classification supervisée de données numériques par navigation reposent majoritairement sur deux points :

– Le choix de la méthode de discrétisation : si les caractéristiques d'entrée sont quantitatives (numériques), alors il faut les discrétiser. La difficulté consiste à choisir une méthode qui

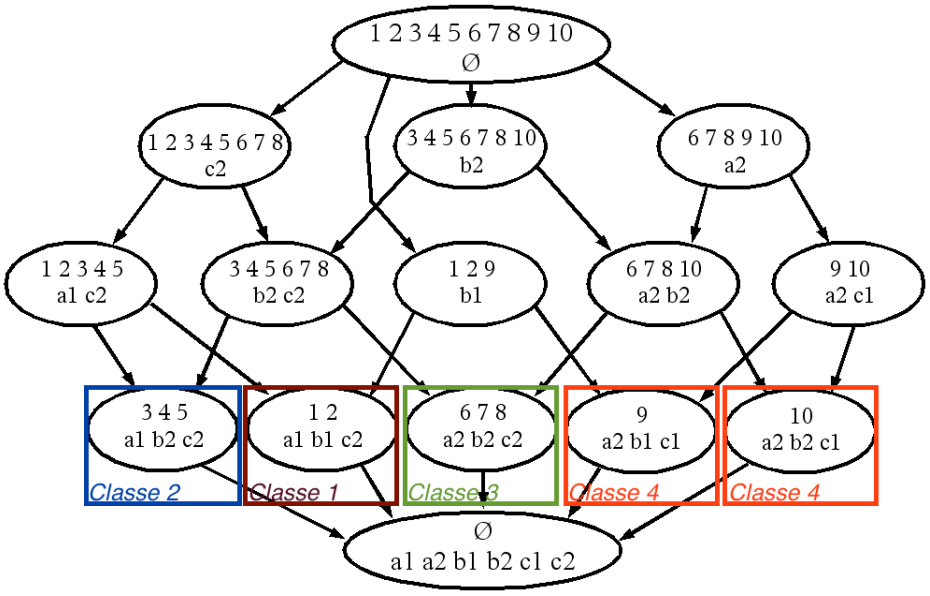


FIGURE 3.5 – Diagramme de Hasse du treillis de la Figure 3.4, étiqueté avec les classes correspondant aux concepts terminaux. *Figure extraite de [Girard 2013].*

soit, d'une part, adaptée à la structure de treillis et qui, d'autre part, favorise le fait que les objets en provenance de classes différentes se retrouvent dans des concepts terminaux différents (critère discriminatif). Le choix de la méthode de discrétisation conditionne également le mode de généralisation vis-à-vis de nouvelles données, notamment au travers du choix d'une mesure de similarité adaptée pour la validation des attributs discrétisés lors de la navigation dans le treillis ;

– Le choix du critère d'arrêt de la discrétisation. Les concepts formels du treillis vérifiant ce critère seront étiquetés avec une classe. Ces concepts sont alors appelés « concepts terminaux » puisque leurs successeurs ne sont pas générés. ;

Nous avons **décliné l'approche générale ci-dessus en deux méthodes** qui diffèrent dans leur apprentissage (notamment dans la technique et le mode de mise en œuvre de la discrétisation), et dans les post-traitements apportés à la structure de données ainsi construite.

Dans la première méthode, à savoir Navigala [Visani 2011a], que nous avons proposée dans le contexte de travaux initiés dans la thèse de Stéphanie Guillas [Guillas 2007], la discrétisation se fait en amont de la génération du treillis. Dans ce cas, le treillis est simplement un outil de représentation et de navigation dans les données qui permet une classification grâce à une analyse grossière des distributions des données dans ses concepts. Cette méthode, qui montre dans la pratique une bonne robustesse vis-à-vis de données dégradées, fait l'objet de la section 3.2.3.1 ci-après.

La famille de treillis des concepts construit par Navigala lorsque toutes les données d'entrée sont quantitatives est une famille de treillis particuliers que nous avons nommés « dichotomiques » et dont nous avons prouvé qu'ils ont des liens structurels avec les arbres de décision [Bertet 2008, Guillas 2009, Bertet 2009]. Plus précisément, il s'agit de liens de fusion et d'inclu-

sion entre treillis dichotomiques et arbres de classification (sous certaines conditions), qui sont définis et prouvés dans la section 3.2.3.2.

Forts de ce résultat théorique, nous avons cherché à créer une structure hybride entre arbres de classification et treillis des concepts qui bénéficie des avantages des deux structures (en particulier la faible complexité des arbres et la robustesse des treillis). C'est le sujet de la thèse de Nathalie Girard, qui a débouché sur la proposition d'une deuxième méthode où la discrétisation se fait cette fois-ci de manière locale, c'est-à-dire localement à l'intérieur de chaque concept, et est guidée par la structure du treillis. Dans ce cas, le treillis devient un outil de classification construit de manière discriminative à l'aide des données d'apprentissage, à la manière de la plupart des arbres de classification. Une phase de post-traitement permet de simplifier la structure afin de garantir une meilleure lisibilité et de diminuer le coût de stockage du treillis. En contrepartie, la structure ainsi construite perd la propriété de treillis et devient donc une structure hybride entre treillis des concepts et arbre de classification. Cette méthode est détaillée en section 3.2.3.3.

Les principales contributions liées à ce travail de recherche sont donc **essentiellement d'ordre méthodologique** et reposent sur trois points, à savoir :

- L'adaptation du treillis des concepts à la problématique de classification par navigation, résultant dans la méthode Navigala ;
- La démonstration de propriétés formelles à propos des liens entre arbres de classification et la famille de treillis produite par Navigala, à savoir les treillis dichotomiques ;
- La définition d'une méthode hybride entre treillis des concepts et arbre de classification pour la classification supervisée.

Les méthodes proposées ont presque totalement été implémentées comme sur-couche de la bibliothèque « *Lattice* » [Bertet 2011], mise à disposition par Karell Bertet sous licence LGPL.

3.2.3.1 Adaptation du treillis des concepts à la classification supervisée par navigation : méthode Navigala

La méthode Navigala, conçue dans le cadre de travaux initiés dans la thèse de Stéphanie Guillas et détaillée dans [Visani 2011a], repose sur une discrétisation préalable (globale) des données numériques. Le critère de discrétisation utilisé est mono-varié, supervisé et discriminatif (au sens où il cherche à séparer les objets appartenant à des classes différentes). Il résulte en des intervalles disjoints étendus en intervalles flous de manière à prendre en compte, lors de la généralisation à de nouvelles données, la distribution des données utilisées pour l'apprentissage. Le critère d'arrêt de la discrétisation est ajusté sur une base de validation.

À l'issue de cette discrétisation, on peut générer le treillis des concepts. Plusieurs algorithmes ont été proposés, dont certains sont incrémentaux, et de complexités calculatoires variables (par exemple une complexité théorique quadratique par élément du treillis dans [Nourine 1999]). La taille du treillis est bornée par $2^{|O+I|}$ dans le pire des cas, et $|O+I|$ dans le meilleur des cas. À noter que cette complexité de stockage peut être contrebalancée par la génération « à la demande » [Bertet 2007] de la portion du treillis nécessaire pour une tâche de classification donnée. La génération à la demande permet *a priori* de réduire la complexité de stockage, mais alourdit la complexité calculatoire de la phase de classification de nouveaux exemples. Cela peut être gênant dans le cas où la reconnaissance de nouveaux exemples est menée en-ligne avec l'utilisateur. Nous ne la considérerons pas ici, puisque dans les contextes applicatifs visés, ce

cas est assez fréquent.

Puis, les concepts terminaux (vérifiant le critère d'arrêt de la discrétisation) sont étiquetés par la classe majoritairement représentée parmi leurs objets. Dans notre cas, le critère d'arrêt de la discrétisation porte sur la pureté des objets du concept en termes de classes.

La classification de nouvelles données se fait par navigation dans le treillis, c'est-à-dire par validations successives, à chaque étape de la navigation, d'un ensemble d'intervalles flous parmi la famille des ensembles d'intervalles flous candidats (correspondants aux concepts successeurs dans le treillis).

La comparaison théorique entre notre approche et les principales approches existantes basées sur des treillis montre que Navigala est plus adaptée à la reconnaissance de classes faiblement représentées, plus robuste au bruit dans les attributs d'entrée, et gère mieux une éventuelle augmentation du nombre de classes que la plupart des approches existantes (basées pour la plupart sur la sélection des concepts ou des règles contextuelles les plus pertinents dans le treillis, comme expliqué plus haut).

Nos expérimentations menées sur les bases de symboles architecturaux GREC 2003 (dont un extrait est montré en Figure 3.3, page 72) et GREC 2005 [7] – composées respectivement de 39 et 150 classes de symboles – montrent que la méthode Navigala est plus adaptée que des méthodes numériques (telles que les SVM) à un usage conjoint avec la signature statistico-structurelle dédiée à la description de symboles détaillée en annexe A et dans [Coustaty 2010]. Même si, avec ce type de signatures, ses résultats restent inférieurs à ceux de l'état de l'art. Lorsqu'elle est utilisée conjointement avec une signature statistique telle que celle de Radon [Tabbone 2006], elle donne des taux de classification similaires à ceux d'un classifieur Bayésien naïf. Elle surpasse nettement l'arbre de décision CART [Breiman 1984] en termes de performances et de robustesse vis-à-vis des différents types de dégradations considérés (bruit additif ou distorsions vectorielles), illustrés en Figure 3.6.

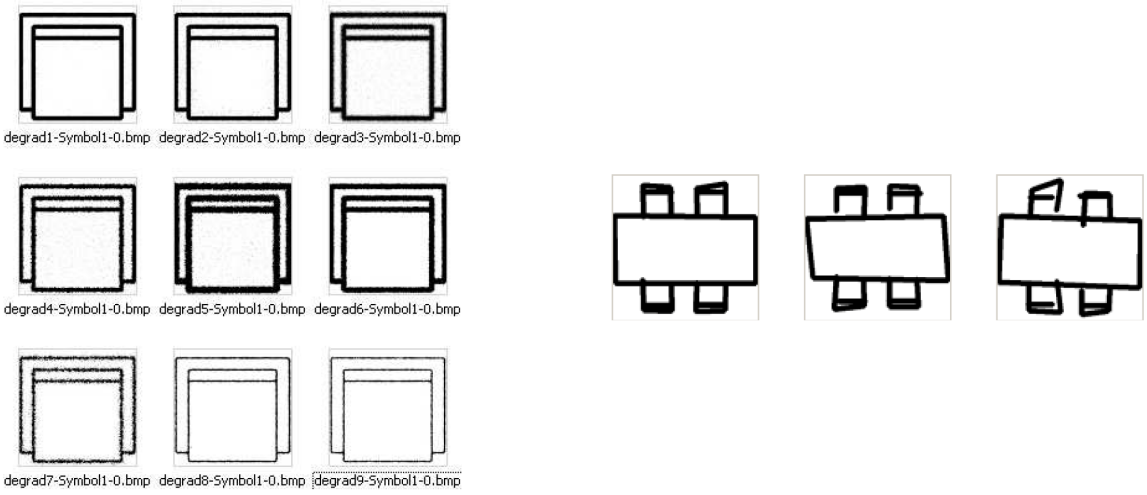


FIGURE 3.6 – Exemples de symboles bruités utilisés pour nos expérimentations. À gauche, du bruit additif est appliqué avec la méthode proposée dans [Kanungo 2000], tandis qu'à droite il s'agit de distorsions vectorielles. *Figure adaptée de [Girard 2013].*

7. http://symbcontestgrec05.loria.fr

Ses performances restent néanmoins nettement inférieures à celles d'un classifieur numérique par SVM (avec noyau RBF) utilisé conjointement avec une signature statistique.

Dans tous les cas, la taille du treillis est très supérieure à celle de l'arbre. Nous avons donc cherché à mieux comprendre les liens formels entre treillis des concepts et arbres de décision, dans l'optique, à terme, de proposer une structure hybride qui tire avantage des deux méthodes (compacité des arbres et meilleure robustesse des treillis).

Pour plus de détails sur l'ensemble de ces comparaisons formelles et expérimentales, le lecteur est invité à se référer à l'annexe E (article [Visani 2011a]) et à la section 4.2. de [Guillas 2007].

3.2.3.2 Liens structurels entre treillis des concepts et arbres de classification

Nous avons étudié les liens formels entre treillis dichotomiques et arbres de décision dans [Bertet 2008, Guillas 2009, Bertet 2009], dans le cas spécifique où l'arbre et le treillis sont générés à partir du même contexte formel (le cas échéant, les attributs d'entrée numériques ayant été préalablement discrétisés). Nous avons démontré les liens d'inclusion et de fusion suivants :

- **Le lien d'inclusion** peut se définir par la proposition suivante : « tout arbre de classification est inclus dans le treillis des concepts, lorsque ces deux structures sont construites à partir du même contexte formel » [8].
- **Le lien de fusion**, lui, est restreint au cas des treillis dits « dichotomiques » que nous avons introduits dans [Bertet 2009]. Un treillis est dichotomique lorsqu'il est construit à partir d'un contexte formel où, à tout attribut du contexte formel, il est possible d'associer un ensemble d'attributs complémentaires (pour une définition plus formelle merci de se référer à la section 4.4.1. de [Girard 2013]). Les treillis générés à partir d'un contexte formel obtenu par discrétisation en intervalles disjoints d'un ensemble d'attributs numériques vérifient naturellement cette propriété de dichotomie ; tout comme (plus largement) l'ensemble des treillis construits à partir d'un ensemble d'attributs qualitatifs dont les modalités sont mutuellement exclusives. Le lien de fusion peut s'énoncer de la manière suivante : « un treillis dichotomique est la fusion (union ensembliste) de tous les arbres de classification, lorsque ces structures sont construites à partir du même contexte formel et que la division de l'arbre de classification se fait jusqu'à ce que les feuilles ne puissent plus être divisées (sans aucune sorte d'élagage) » [9].

8. Cette proposition peut être démontrée à l'aide des trois points suivants, découlant tous de la propriété de fermeture des treillis comme détaillé en section 4.3.1. de [Girard 2013] :
   - Deux nœuds différents d'un arbre de classification sont associés à des concepts différents du treillis (ce que l'on peut aisément démontrer par l'absurde) ;
   - Si deux nœuds sont ancêtres dans l'arbre de classification, alors leurs concepts associés sont en relation selon $\leq$ dans le treillis (démontrable à l'aide de l'isotonie de l'opérateur de fermeture $\beta o \alpha$ sur les attributs) ;
   - Si deux nœuds ne sont pas ancêtres dans l'arbre de classification, alors leurs concepts associés ne sont pas en relation dans le treillis (car les fermetures des ancêtres frères dans l'arbre de ces deux nœuds sont disjointes).

9. Étant donné le lien d'inclusion, pour prouver le lien de fusion, il suffit de montrer que tout concept du treillis appartient à un arbre de classification, ce que nous pouvons faire par construction. En effet, pour tout concept du treillis (O', I'), si l'on construit une structure constituée de (O', I'), de son concept complémentaire (dont l'existence est assurée par le fait que le treillis soit dichotomique), du concept minimal du treillis, et de tous les concepts terminaux successeurs de ce concept et de leurs complémentaires (avec les arcs correspondants), alors cette structure est un arbre de classification. Pour une démonstration plus formelle, merci de se référer à la section 4.4.3. de [Girard 2013].

Ces liens d'inclusion et de fusion peuvent expliquer en partie les meilleures performances en classification du treillis, en comparaison avec un arbre de décision construit seul. Leur découverte nous permet de nous orienter vers la définition d'une structure hybride entre treillis des concepts et arbres de classification, alliant les avantages des deux structures. C'est l'objet de la section ci-après.

3.2.3.3 Conception d'une méthode hybride entre treillis des concepts et arbre de classification

Dans le cadre des travaux réalisés dans le contexte de la **thèse de Nathalie Girard** [Girard 2013], nous avons apporté deux changements majeurs à la méthode Navigala présentée ci-avant afin d'en faire une structure hybride entre treillis des concepts et arbre de classification : nous avons mis en œuvre une discrétisation locale, et simplifié *a posteriori* la structure de treillis pour en réduire la complexité. Ces deux apports sont détaillés ci-après.

a) Discrétisation locale

Nous avons proposé une discrétisation locale pour les treillis des concepts dans [Girard 2009, Girard 2011a, Girard 2011b]. Cette discrétisation est basée sur les propriétés des treillis dichotomiques et s'inspire de la discrétisation locale souvent mise en œuvre pour la construction des arbres de classification. La discrétisation locale, guidée par la structure de treillis, est plus adaptée à celui-ci et l'on peut considérer dans ce cas que le treillis en lui-même fait l'objet d'un apprentissage discriminatif, ce qui n'était pas vraiment le cas dans la méthode Navigala où il servait simplement à représenter et étiqueter les données pour la classification par navigation.

Par contre, la définition d'un algorithme de discrétisation locale adapté aux treillis n'est pas aussi évidente que pour les arbres de décision. En effet, la discrétisation d'un attribut dans un concept engendre une modification de tous les autres concepts contenant cet attribut (du fait de la relation d'ordre entre les concepts), et donc de la structure globale du treillis. Ce n'est pas le cas dans un arbre, où lorsqu'un attribut est discrétisé dans un nœud, cela ne modifie ni ses prédécesseurs ni les autres branches de l'arbre ; l'impact est en effet limité à ses successeurs.

Nous avons donc choisi de baser notre discrétisation locale du treillis sur un **algorithme itératif où chaque étape de discrétisation est menée uniquement sur l'ensemble des coatomes** (concepts en relation directe avec le *top* $\top$ du treillis) qui ne vérifient pas le critère d'arrêt (ici qui ne sont pas suffisamment purs en termes de classes). À chaque étape, seul l'ensemble des coatomes du treillis est généré (et non pas le treillis complet), pour des raisons de complexité. En effet, le nombre de coatomes est $\leq$ au nombre d'objets, alors que le nombre de concepts est dans le pire des cas exponentiel en fonction des données (nombre d'objets + nombre d'attributs). On obtient ainsi le contexte formel. Puis, on peut générer le treillis et étiqueter les concepts terminaux (vérifiant le critère d'arrêt) par la classe majoritairement représentée parmi leurs objets.

Très synthétiquement, chaque étape de discrétisation (mono-variée) comprend les deux phases ci-après. Pour plus de détails, merci de se référer à [Girard 2011a, Girard 2011b].

- pour chaque coatome (O', I'), on sélectionne le meilleur attribut I'^* et son point de coupe optimal $V'^*(I'^*)$ à l'aide d'un critère de division basé sur le χ^2 (choisi pour sa stabilité) ;
- le meilleur couple attribut/point de coupe $(I^*, V^*(I^*))$ pour l'ensemble des coatomes est

ensuite choisi parmi les meilleurs candidats $(I'^*, V'^*(I'^*))$ de chaque coatome. Le critère de sélection utilisé pour cela est basé sur le critère de division ci-dessus, cette fois-ci déployé à un niveau plus global sur l'ensemble des coatomes. Seul cet attribut est discrétisé, le contexte formel est remis à jour et les coatomes re-calculés.

Ce processus est réitéré jusqu'à ce que l'ensemble des coatomes vérifient le critère d'arrêt. On obtient ainsi une discrétisation de chaque attribut quantitatif d'entrée en un ensemble d'intervalles disjoints « consécutifs » couvrant tout $\mathbb{R}$, qui constituent l'ensemble des attributs I.

Lors de la phase de classification de nouveaux exemples, la navigation dans le treillis se fait par validation de ces attributs. Plus précisément, à chaque étape de navigation, un ensemble d'attributs est validé parmi la famille des ensembles d'attributs candidats (des concepts successeurs) selon un critère d'appartenance des caractéristiques des images à classer aux intervalles. La phase de classification de nouveaux exemples est ainsi très peu coûteuse en termes de temps de calcul.

Notons que, tout comme Navigala, cette méthode de classification supervisée est suffisamment générique pour pouvoir être utilisée avec succès sur des données qui ne sont pas nécessairement issues d'images. Nous avons donc mené une **étude expérimentale** sur divers jeux de données. Cette étude expérimentale vise à comparer notre méthode avec discrétisation locale à des SVMs (avec noyaux RBF), à Navigala et à l'arbre de classification ChAID [Kass 1980, Kothari 2000] (également construit avec un critère de division basé sur le χ^2, et avec un pré-élagage basé sur une fusion des nœuds successeurs du nœud courant lorsque leurs distributions sont proches). Les bases utilisées pour cette comparaison expérimentale sont les bases de symboles GREC 2003 et GREC 2005 citées plus haut, mais également trois bases de données extraites du *Machine Learning Repository*[10] et comportant des données autres que des données image (à savoir les bases GLASS, IRIS et Breast Cancer). Ces expérimentations montrent que la discrétisation locale permet de diminuer la taille des treillis (et donc d'améliorer la lisibilité de la structure) tout en conservant leurs bonnes performances (vis-à-vis des arbres de décision et notamment de ChAID). Plus précisément, sur la plupart des bases, lorsqu'on applique une validation croisée, le taux de reconnaissance moyen est meilleur qu'avec ChAID et l'écart-type est plus faible (même si cette différence n'est pas toujours significative statistiquement). On peut donc considérer que le comportement du treillis est plus stable que celui de l'arbre en présence de données de natures variées. Les taux de reconnaissance obtenus avec le treillis discrétisé localement sont similaires à ceux des SVMs avec noyau RBF, sur les bases considérées. Sur les bases d'images de symboles, les taux de classification atteints par le treillis sont largement supérieurs à ceux obtenus par le SVM, lorsque l'on utilise en entrée des deux classifieurs la signature statistico-structurelle présentée en annexe A. Vu que cette signature est intelligible pour un expert, cela permet de fournir à l'humain expert du domaine une vue relativement lisible des règles de classification apprises. En revanche, les signatures statistiques permettent toujours d'obtenir de meilleurs performances que la signature statistico-structurelle. Ce point sera rediscuté en section 3.2.5 « Discussion ».

b) Simplification du treillis

Malgré la diminution de la taille du treillis obtenue grâce à la discrétisation locale,

10. http://archive.ics.uci.edu/ml

cette structure reste largement plus complexe que celle des arbres dans le cas général. Afin de diminuer encore la taille du treillis pour la rapprocher de celle des arbres tout en conservant ses bonnes performances, nous avons étudié diverses possibilités de simplification de la structure de treillis. Nous avons choisi dans un premier temps d'utiliser, *a posteriori* de la construction du treillis, un critère de simplification basé sur l'**indice de stabilité des concepts** [Kuznetsov 2007], propre aux treillis. L'indice de stabilité $\delta(O', I')$ du concept (O', I') est calculé comme suit :

$$\delta(O', I') = \frac{|\ \{O'' \subseteq O'\ tq\ f(O'') = I'\}\ |}{2^{|O'|}} \tag{3.6}$$

Cet indice est un indicateur de la proportion des sous-ensembles d'objets de O' qui ont pour description commune exactement I'. Autrement dit, c'est un indicateur de la dépendance entre les objets de O' et leurs attributs I' selon la relation R. Nous utilisons un algorithme de calcul exact des indices de stabilité de l'ensemble des concepts du treillis basé sur son diagramme de Hasse, en un temps polynomial [Roth 2006]. Nous avons mis au point un algorithme permettant d'ajuster automatiquement le seuil de cet indice de stabilité pour déclencher la simplification.

Logiquement, la complexité calculatoire de la phase d'apprentissage se trouve alourdie par ce traitement supplémentaire de simplification. En contrepartie, on obtient une simplification efficace de la structure, avec une **diminution du nombre de concepts allant jusqu'à 50%**, tout en conservant les performances en classification et la stabilité du treillis. Le coût calculatoire de la phase de classification proprement dite se trouve donc diminué, ce qui est généralement plus intéressant d'un point de vue applicatif que de diminuer la complexité de la phase d'apprentissage.

En revanche, du fait du caractère non-monotone de l'indice de stabilité utilisé, les concepts peuvent être supprimés n'importe où dans la structure. Donc, la structure hiérarchique obtenue ne vérifie pas la propriété de treillis ; c'est pourquoi nous la qualifions d'« hybride ».

3.2.4 Bilan et améliorations possibles

Nous avons proposé une approche de classification supervisée basée sur un treillis des concepts qui, à la différence de la plupart des approches existantes, utilise directement le treillis lors de la phase de classification, en l'occurrence par navigation à la manière d'un arbre de classification. Ainsi, la phase de classification de nouveaux exemples est particulièrement peu coûteuse en temps de calcul. Nous avons décliné cette approche générale en deux méthodes qui diffèrent dans la manière de discrétiser les attributs d'entrée quantitatifs, et dans les post-traitements appliqués le cas échéant sur le treillis.

La première méthode, basée sur une discrétisation globale avec une possibilité de génération à la demande de la portion du treillis concerné lors de la phase de classification, nous a permis de montrer de manière théorique et expérimentale l'intérêt de la classification par navigation dans le treillis.

En nous appuyant sur la définition de liens structurels entre la famille des treillis générés par cette méthode et les arbres de classification, nous avons proposé une deuxième méthode, hybride entre treillis et arbres de classification. Cette méthode est basée sur une discrétisation locale guidée par la structure de treillis, menée uniquement sur les coatomes de la structure pour réduire la complexité, et sur une simplification *a posteriori* de la structure de treillis basée sur un indice de stabilité des concepts dans le treillis.

La structure générée par la deuxième méthode est aussi stable que le treillis des concepts tout en étant moins complexe que celui-ci. L'ordre de grandeur de la taille de la structure obtenue dépend dans la pratique beaucoup de la distribution des données d'entrée, en raison notamment des frontières de séparation « linéaires par morceaux » et parallèles aux axes induites par notre méthode de discrétisation mono-variée. Le nombre de concepts du treillis reste dans tous les cas largement supérieur à celui du nombre de nœuds de l'arbre. Ce point sera discuté dans la section suivante.

L'approche que nous avons proposée est **originale** dans le domaine de l'analyse de données (de par la classification par navigation). Elle l'est également dans le domaine de l'analyse d'images, où les méthodes numériques sont généralement préférées. Ce travail ouvre de nombreuses perspectives d'amélioration [11], qui font l'objet de la suite de cette section.

La première piste d'amélioration envisageable concerne la discrétisation locale. À l'heure actuelle, celle-ci est effectuée de manière mono-variée. Ce qui, dans la pratique, engendre des concepts inutiles au sens de la classification. Le fait de découper en une seule étape de discrétisation locale plusieurs caractéristiques d'entrée (**discrétisation multi-variée**) pourrait donc mener à des treillis bien plus synthétiques, tout en conservant les performances en classification de la structure ainsi créée. Ce type de discrétisation a déjà été étudié pour les arbres de décision : on parle alors d'arbres de décision « multi-variés », qui sont généralement plus synthétiques et plus performants que les arbres de décision monovariés [Brodley 1995]. Cependant, les expressions dérivées à chaque nœud dans le cas multivarié peuvent être plus difficiles à appréhender pour un humain. Nous sommes en train d'initier des travaux en ce sens. Notre objectif est de trouver une solution de discrétisation locale multi-variée adaptée au treillis (où toute décision concernant une division multi-variée prise localement a un impact sur l'ensemble des concepts partageant les attributs concernés par cette décision), et qui offre un bon compromis entre efficacité en temps de calcul, performances et lisibilité de l'information de sortie.

Une deuxième piste d'amélioration possible concerne la simplification de la structure. En effet, la simplification de structure présentée en section précédente, basée sur un indice de stabilité ne dépendant que de la relation entre objets et attributs, n'est pas supervisée et ne vise qu'à réduire la taille de la structure, sans engendrer d'amélioration des performances.

Nous pourrions donc chercher à substituer à cette simplification une méthode visant à **limiter les risques de sur-apprentissage**, à la manière de l'élagage dans les arbres de classification. Cette méthode supervisée pourrait utiliser un critère monotone afin de préserver la structure de treillis. Nous avons déjà mené des études préliminaires en ce sens, mais nous sommes trouvés confrontés à des difficultés liées à la propriété de treillis. En effet, du fait de l'existence d'un concept *meet* pour chaque paire de concepts, si les impacts plus globaux de cette discrétisation locale sont mal maîtrisés, alors on peut obtenir une diminution du nombre d'arcs plutôt que du nombre de concepts, ce qui engendre l'effet contraire à celui recherché (perte en capacité de généralisation). Nous avons une idée de solution qui laisse la possibilité, le cas échéant, de préserver la structure de treillis. Très grossièrement, il s'agit d'un

11. Outre les discussions menées dans le cadre des thèses de Stéphanie Guillas et de Nathalie Girard, co-encadrées avec Karell Bertet, certaines de ces perspectives de recherche ont été discutées avec Rokia Missaoui (PR, Université du Québec en Outaouais), spécialiste à la fois de l'analyse formelle de concepts et de la fouille de données, lors de son séjour de recherche au L3i en juin 2014.

post-élagage permettant de supprimer les concepts offrant le moins bon compromis entre coût de la suppression (en termes de perte de performance) et complexité de la sous-structure ayant pour racine ce concept, un peu à la manière du post-élagage de CART [Breiman 1984] pour les arbres. Sauf que, dans notre cas, l'élagage au niveau d'un concept engendrerait non seulement la suppression de ses successeurs, mais aussi la suppression des prédécesseurs de ses successeurs qui n'ont plus lieu d'être, suivi le cas échéant du rajout d'arcs permettant de conserver la propriété de treillis. Nous sommes en train d'implémenter cette solution.

Une dernière piste d'amélioration possible est d'adapter notre approche afin de permettre au treillis de gérer la **classification multi-étiquettes** (où les étiquettes sont typiquement des mots-clés issus d'annotations des images) [Tsoumakas 2007]. En effet, le treillis semble *a priori* être une structure particulièrement adaptée à ce cas (spécialement lorsque les étiquettes sont organisées selon des structures hiérarchiques). Cela nécessiterait évidemment de modifier le critère de discrétisation, qui ne serait plus basé sur la recherche de frontières perméables entre classes. Cela nécessiterait également de modifier le critère d'étiquetage, qui ne serait plus déclenché uniquement suite à la validation du critère d'arrêt, et nous permettrait donc d'étiqueter des concepts intermédiaires entre le *bottom* et les concepts terminaux.

3.2.5 Discussion

De manière très générale, le pari qui consistait à proposer une approche symbolique issue de l'analyse formelle des concepts était relativement audacieux dans le cadre d'une application à l'analyse d'images, où les méthodes numériques sont généralement préférées. Vu l'intérêt suscité dans les deux communautés scientifiques (celle de l'analyse formelle des concepts et celle de l'analyse d'images) et les bonnes performances obtenues en pratique, on peut considérer que ce choix s'est révélé payant dans une certaine mesure.

L'un des principaux reproches faits par la communauté scientifique de l'analyse d'images vis-à-vis des approches symboliques réside dans la complexité (en stockage et en construction) de la structure.

En ce qui concerne la complexité calculatoire de génération de la structure, celle-ci est effectivement importante (typiquement quadratique par élément du treillis). On peut noter cependant que cela n'est pas forcément gênant dans un contexte de reconnaissance, où elle peut souvent se faire hors-ligne. Une fois la structure complète construite, la phase de classification en elle-même (généralisation à de nouveaux exemples) est très rapide puisqu'il s'agit simplement de faire naviguer l'image à classer dans le treillis, selon des critères d'appartenance des attributs à des ensembles d'intervalles.

Concernant la complexité de stockage, dans notre deuxième méthode, elle est réduite par rapport à un treillis, de par la simplification qui rend la structure hybride entre treillis et arbre. Néanmoins, malgré cette simplification (que nous espérons améliorer très rapidement selon les pistes évoquées ci-dessus), la taille de la structure produite reste largement supérieure à celle d'un arbre de classification. Cependant, vu les liens d'inclusion et de fusion entre arbres et treillis dichotomiques présentés en section précédente, la complexité des treillis serait plutôt à comparer à celles des **forêts aléatoires** [Breiman 2001]. La principale différence entre les deux approches réside dans le fait que les arbres de la forêt sont obtenus par ré-échantillonnage de leurs enregistrements et de leurs attributs (voir section 3.4), tandis que les arbres inclus dans le treillis sont obtenus par l'analyse de toutes les occurrences (objets, attributs) observées

dans le contexte formel. Les forêts aléatoires sont donc *a priori* mieux adaptées à la présence de données aberrantes ou mal préparées (attributs redondants par exemple) et aux espaces presque vides. En contrepartie, le principal avantage de notre approche par rapport aux forêts aléatoires réside dans le fait que nous conservons la lisibilité du graphe, ce qui n'est généralement pas le cas avec les forêts aléatoires. Une comparaison formelle et expérimentale poussée de ces deux types d'approches reste néanmoins à mener.

Notre approche peut être utilisée pour la reconnaissance d'images, mais aussi plus généralement pour la classification supervisée de tous types de données (numériques ou symboliques), ce que nous avons vérifié en pratique sur divers jeux de données de classification supervisée standards. Sa portabilité d'un contexte applicatif à un autre nécessite bien sûr que son apprentissage se fasse avec un jeu de données adapté, mais par contre il ne nécessite l'ajustement d'aucun paramètre dépendant du domaine. Son seul paramètre est en effet le seuil du critère d'arrêt qui peut, soit être ajusté sur une base de validation, soit être traduit sous la forme de l'erreur apparente maximale autorisée et donc, le cas échéant, fixé manuellement par l'utilisateur. Dans tous les cas, son ajustement ne nécessite pas de ré-apprentissage spécifique. En ce sens, **cette méthode est assez générique**.

En revanche, si nous avons montré sa compétitivité avec une méthode numérique telle que les SVM sur divers jeux de données, cette comparaison était basée sur l'usage de SVM avec un noyau fixé *a priori* (ici RBF) quel que soit le jeu de données. Il est cependant possible (voire probable) qu'un SVM utilisant un noyau spécifiquement choisi pour chaque jeu de données donnerait de meilleurs résultats que notre approche. Cela peut porter matière à débat concernant le difficile compromis entre généricité de l'approche et bonnes performances dans un contexte applicatif donné.

Il convient maintenant de se pencher sur les avancées réalisées au travers de ces travaux au regard du fossé existant entre les informations de bas niveau sémantique manipulées par la machine, et les concepts de plus haut niveau attendus par l'utilisateur.

Premièrement, afin de restituer à un expert humain les informations apprises sous une forme qu'il est capable d'appréhender, nous avons proposé une approche de classification supervisée symbolique, qui permet de présenter à l'humain les règles de classification apprises sous une forme graphique. Son avantage principal est donc qu'un expert du domaine d'application peut visuellement retrouver les raisons qui ont poussé la machine à prendre une décision de classification donnée.

Nous avons donc proposé une méthode capable de restituer de manière intelligible à un expert humain les règles de catégorisation apprises... si tant est que l'expert maîtrise suffisamment la sémantique associée aux attributs d'entrée pour comprendre les expressions retournées. C'est le cas dans un grand nombre de domaines applicatifs : par exemple, il sera très utile pour un écologue d'apprendre quelles sont les conditions des facteurs (prévisions météorologiques, de trafic, etc.) qui justifient une prévision de pic de pollution pour le lendemain. Dans le domaine de l'analyse d'images, comme évoqué en annexe A, il est possible de concevoir des signatures qui soient intelligibles pour un expert. Si l'on utilise l'approche symbolique que nous avons proposée conjointement avec la signature statistico-structurelle décrite en annexe A et basée sur la description des relations topologiques entre segments de lignes extraits des symboles, alors il est possible d'obtenir un treillis dont les règles de classification sont relativement lisibles pour l'expert (au prix toutefois d'un certain effort), comme illustré en Figure 3.7. Néanmoins, ce type

de signatures ne sont pas forcément à même de représenter au mieux la diversité d'apparence des images.

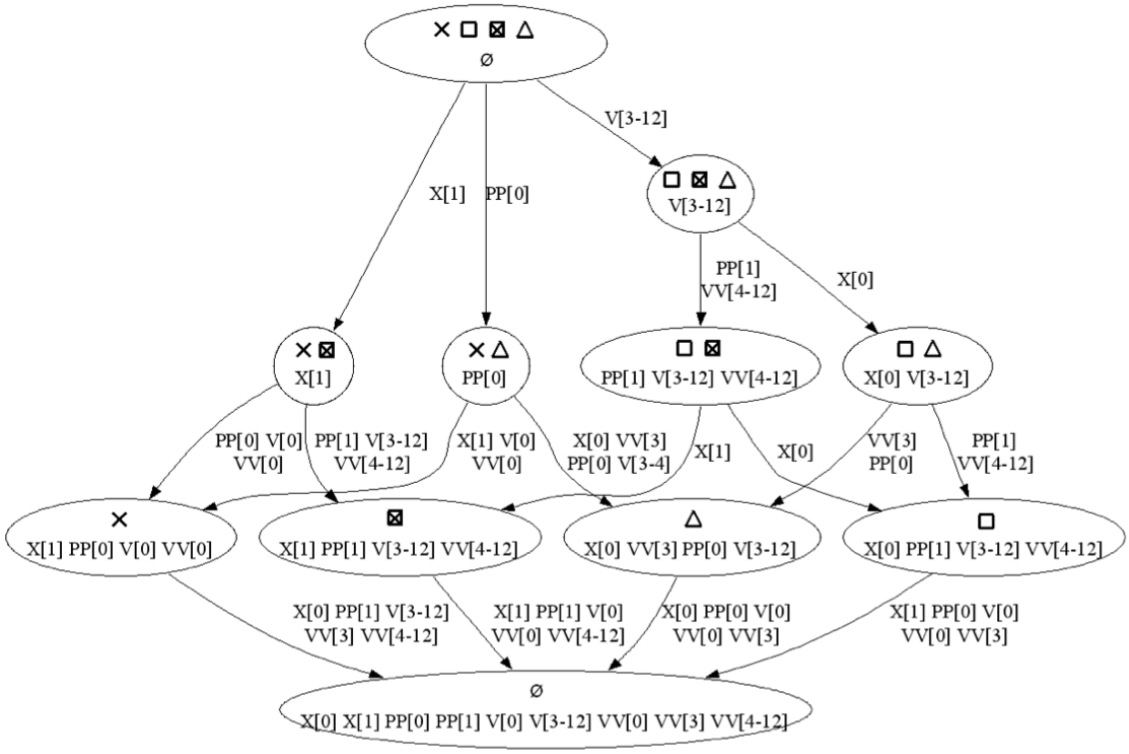


FIGURE 3.7 – Exemple de treillis généré (sur un sous-ensemble de la base GREC 2003) à partir de la signature statistico-structurelle proposée dans [Coustaty 2011] et décrite en section A.3 de l'annexe A, et de la méthode Navigala. Dans chaque concept, nous affichons des vignettes représentatives des objets qui le composent, en plus des attributs discrétisés de la signature. *Figure extraite de [Coustaty 2011], et reprise en Figure A.4 de l'annexe A.*

D'autres signatures, souvent statistiques et moins lisibles pour un humain, permettent généralement d'obtenir de meilleures performances au final, en termes de reconnaissance. Est-on alors contraint de faire un douloureux choix entre lisibilité des informations restituées à l'expert, et performances en pratique ? À vrai dire, tout dépend du sens que l'on accorde à ce terme quelque peu passe-partout d'« informations restituées à l'expert ». Si l'on relaxe un peu notre acception initiale, qui portait sur les « règles de classification apprises », pour l'adapter au domaine de l'image et le remplacer par le « chemin visuel emprunté par la méthode pour classer un exemple donné », **il est possible d'allier lisibilité et performance**. En effet, grâce à la connexion de Galois, chaque concept du treillis comporte à la fois un ensemble d'attributs (caractéristiques visuelles discrétisées) et un ensemble d'objets (images) en relation avec ces attributs. Il est donc possible de représenter le chemin emprunté par une image de manière visuelle, en affichant dans chaque concept de la portion de treillis concerné quelques vignettes des images les plus représentatives (le cas échéant sélectionnées par classe) de ce concept, comme nous l'avons fait dans la Figure 3.7 ci-dessus. Une idée similaire, proposant un outil de navigation dans des bases d'images reposant sur un treillis des concepts, a été exploitée en parallèle de nos travaux dans [Ducrou 2008].

Deuxièmement, intéressons-nous maintenant non plus à la restitution d'informations à l'humain, mais à l'intégration d'informations collectées auprès de l'humain dans le processus de reconnaissance (il s'agit de l'autre angle d'attaque que nous envisageons dans ce manuscrit afin de réduire le fossé sémantique). C'est un point que nous avons négligé jusqu'à présent dans ces travaux. Or, l'analyse formelle des concepts est une théorie de la représentation de connaissances, de la gestion de l'information et de l'analyse de données qui va bien au-delà de l'usage en classification à partir de données de bas niveau sémantique que nous en avons fait ici. **Nous avons jusqu'à présent sous-exploité la puissance de ces outils**, notamment du fait de l'absence de connaissance formalisée de manière explicite dans notre contexte applicatif. Alors même que des travaux récents [Atif 2014] montrent l'intérêt d'utiliser un treillis des concepts pour intégrer des connaissances formalisées explicitement dans un processus d'interprétation d'images. De manière plus générale, l'intégration d'informations collectées auprès de l'humain dans le processus de classification fait partie de nos principales perspectives de recherche, comme détaillé en section 3.6 « Perspectives » clôturant ce chapitre.

À l'issue de cette étude qui concerne essentiellement le bout de la chaîne de traitement traditionnelle de reconnaissance d'images de documents par classification (puisqu'elle porte sur la phase de classification proprement dite et sur la restitution des informations apprises à l'humain), nous allons reprendre notre étude de la chaîne de traitement à rebours. Ainsi, nous allons dans la section suivante nous intéresser particulièrement aux phases de segmentation et de classification (*via* l'extraction de descripteurs), et aux modes de coopération entre ces différentes étapes, au travers d'un cas d'étude applicatif spécifique, à savoir la reconnaissance d'écriture manuscrite. Nous reviendrons sur les travaux présentés ci-dessus dans la conclusion du chapitre, en section 3.5.

3.3 Reconnaissance de mots manuscrits par segmentation semi-explicite et classification à deux niveaux

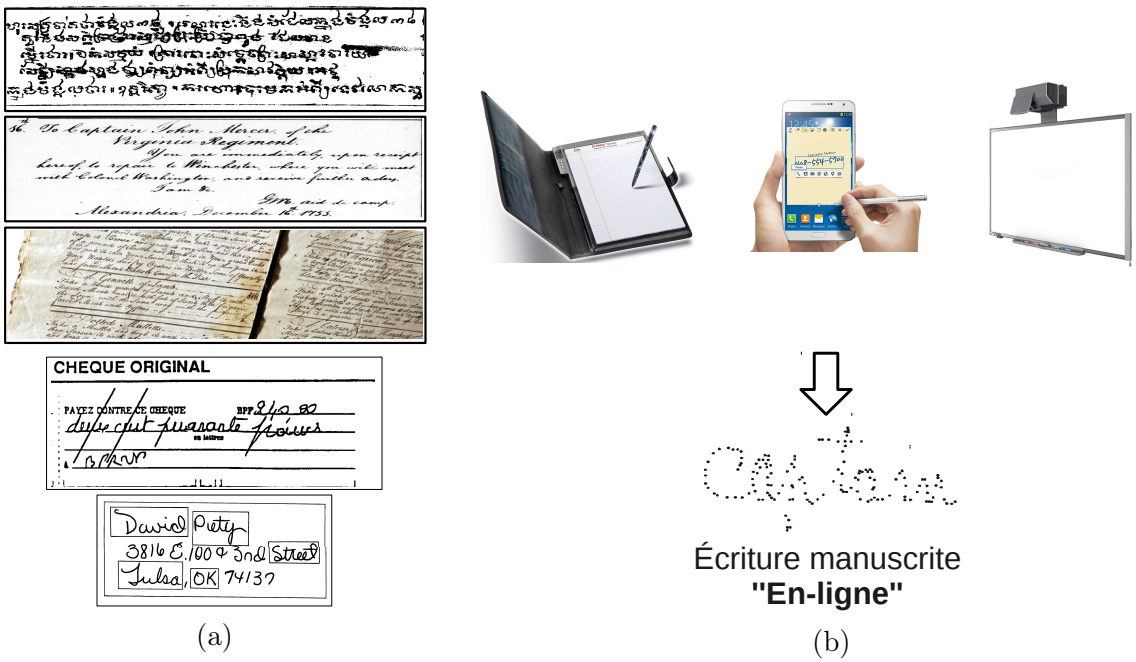
3.3.1 Introduction

3.3.1.1 Contexte de l'étude

Depuis l'apparition du système d'écriture il y a environ 5000 ans, les documents manuscrits constituent un mode de communication majeur dans la société. Après l'émergence dans les années 1960-1970 de systèmes de reconnaissance optique de caractères (communément désignés sous leur acronyme anglais OCR) permettant de reconnaître du texte imprimé ou dactylographié, les chercheurs se sont rapidement intéressés à la reconnaissance d'écriture manuscrite.

Il convient de distinguer deux types de documents manuscrits : les documents hors-ligne et en-ligne (voir Figure 3.8). Tandis que, dans le cas hors-ligne, on s'intéresse à une image de document généralement acquise au travers d'un processus de numérisation, dans le cas en-ligne, on a accès à un signal constitué constitué d'un ensemble d'informations concernant la dynamique du tracé (coordonnées ordonnées dans le temps, poser/lever/inclinaison/pression du stylet, etc.). Les tâches de reconnaissance automatique d'écriture hors-ligne et en-ligne ont donc logiquement chacune leurs spécificités [Plamondon 2000].

L'étude présentée dans cette section a été menée dans le cadre du **projet RecoNomad**. L'objectif principal du projet est la transcription automatique de documents textuels acquis avec une tablette électromagnétique (signal en-ligne).



(a) (b)

FIGURE 3.8 – Exemples (a) d'images de documents manuscrits hors-ligne, et (b) de signal manuscrit (en-ligne). *Figure adaptée de [Prum 2013a].*

3.3.1.2 Objectif visé

Notre objectif est de proposer un système capable de reconnaître du texte manuscrit reposant sur les caractères de l'alphabet latin, minuscules et non accentués, mais à terme notre système doit pouvoir être adaptable à d'autres langages alphabétiques. Le moteur de reconnaissance doit être applicable dans un environnement omni-scripteur (applicable pour n'importe quel scripteur, connu ou inconnu), avec un lexique (ensemble de mots potentiellement présents dans le document) de taille variable et potentiellement importante, sans trop de contraintes concernant le dispositif de capture (résolution, échantillonnage spatial ou temporel, etc.) ; seule une résolution minimale est requise. L'écriture est également complètement libre, à savoir que l'on n'impose pas au scripteur d'écrire les mots avec des caractères isolés, ni de manière cursive ; certains scripteurs ont donc tendance à mélanger les deux, ce qui complexifie la reconnaissance [Tappert 1990]. On ne lui impose pas non plus d'écrire un nombre minimal de mots, ni de structurer les mots en phrases.

Très généralement, la reconnaissance d'écriture manuscrite peut se faire à trois niveaux : au niveau du caractère, du mot ou de la phrase. Du fait de notre contexte applicatif, nous nous intéressons à la reconnaissance de mots, que nous supposons pré-segmentés.

Même si, dans notre cas, nous nous basons sur le signal en-ligne, nous verrons dans la suite que nous allons considérer en entrée de notre système à la fois ce signal en-ligne, mais aussi le signal hors-ligne estimé à partir du signal en-ligne. Par conséquent, nous pouvons considérer ici que nous travaillons à partir d'images structurées dégradées (dans les faits, reconstruites à partir du signal en-ligne) additionnées d'informations concernant la dynamique du tracé. Cette information additionnelle est précieuse car elle peut permettre de lever certaines des ambiguïtés inévitables lors de la reconnaissance de motifs aussi complexes que des mots manuscrits cursifs.

En résumé, **notre objectif est d'identifier des mots manuscrits** à partir de leur **images**, additionnées d'**informations concernant la dynamique de leur tracé** et d'informations de plus haut niveau sémantique concernant le **lexique**. Cela passe par la segmentation de ces mots en unités plus élémentaires et par leur classification, comme nous le verrons dans la suite de cette section. Plus précisément, nous nous intéressons à la portion de la chaîne de traitement traditionnelle de la reconnaissance d'images mise en évidence dans la Figure 3.9 montrée en page suivante. C'est l'objet de la **thèse de Sophea Prum** [12], soutenue en 2013.

Avant de détailler les principales originalités du système proposé vis-à-vis des approches existantes de la littérature, replaçons-les plus précisément à la fois dans le cadre des questions abordées dans ce manuscrit, et vis-à-vis des travaux précédemment évoqués.

3.3.2 Fil conducteur des questions abordées

Tout comme le cas des symboles graphiques évoqué précédemment, les mots sont composés de primitives (ici des caractères) dont l'agencement spatial (structure physique) respecte des règles d'édition spécifiques au langage considéré, et donc les images de mots sont structurées. Néanmoins, la très forte variabilité entre deux occurrences d'un même caractère (essentiellement en fonction du scripteur et du contexte des caractères voisins), et dans l'apparence des liaisons entre ces caractères (du fait de la potentielle cursivité de l'écriture), fait des mots manuscrits

12. Doctorante que j'ai encadrée scientifiquement sous la direction de Jean-Marc Ogier (PR, L3i).

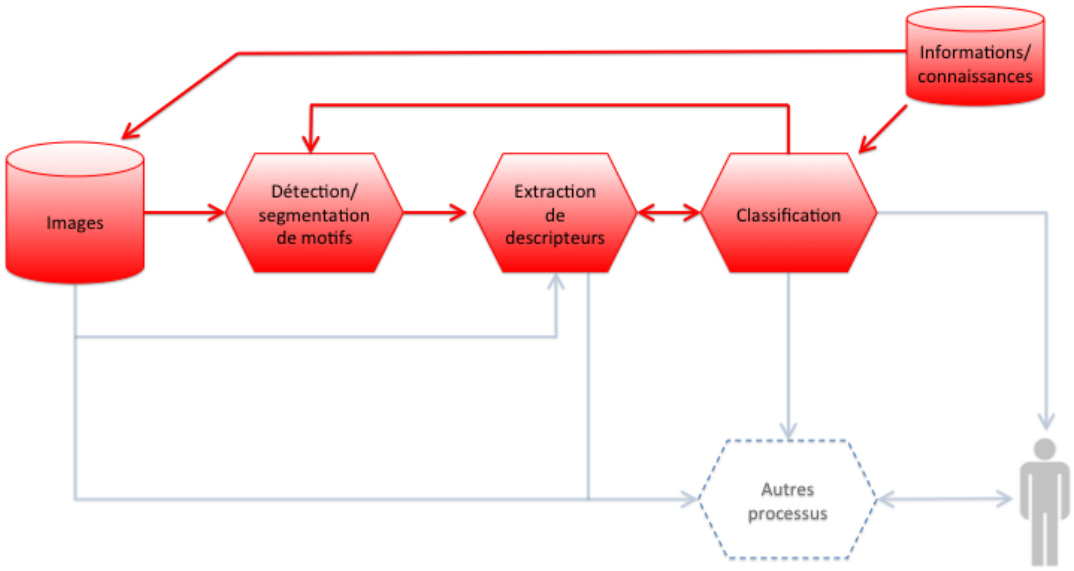


FIGURE 3.9 – Portion de la chaîne de traitement traditionnelle de classification pour la reconnaissance d'images de documents à laquelle on s'intéresse dans cette section.

des motifs plus complexes à reconnaître que les symboles graphiques traités précédemment.

Dans ce contexte, il est difficilement envisageable d'obtenir de bonnes performances avec une méthode de classification supervisée générique telle que la méthode symbolique décrite plus haut. Si l'on souhaite être en mesure d'offrir de bons résultats de reconnaissance en pratique, il convient de proposer un système adapté aux spécificités de la reconnaissance d'écriture manuscrite. Néanmoins, nous sommes quand même en **recherche d'une certaine généricité**, vu que le système proposé devra être omni-scripteur et pouvoir s'adapter à terme à d'autres langages alphabétiques.

À la différence du cas de l'extraction d'invariants évoqué en section 2.3 du chapitre 2 où nous cherchions à découvrir sans aucune information *a priori* les primitives composant le texte, nous disposons d'un jeu d'exemples annotés suffisant pour apprendre, le cas échéant, chacun des caractères de l'alphabet. En contrepartie, on se trouve ici dans un contexte multi-scripteur qui apporte beaucoup de **variabilité dans la collection de documents** et complexifie grandement la tâche par rapport au cas de l'extraction d'invariants, où nous considérions le cas de collections anciennes très homogènes.

L'une des principales spécificités de ce problème par rapport aux problèmes traités précédemment réside dans **la nature très diverse des informations dont on dispose en entrée du processus de catégorisation**. En effet, si, dans l'ensemble des cas d'étude traités précédemment, nous disposions seulement d'informations de bas niveau sémantique concernant l'apparence des images à décrire ou à reconnaître, éventuellement additionnées d'informations de plus haut niveau sémantique concernant les catégories (apportées par la vérité-terrain ou par l'utilisateur), ici nous avons accès à une diversité bien plus importante d'informations.

Plus précisément, nous disposons premièrement d'informations de bas niveau sémantique concernant la dynamique du tracé (signal en-ligne) et l'apparence visuelle de l'image recons-

truite à partir de ce signal en-ligne. Nous avons également accès à des informations de plus haut niveau sémantique sous la forme d'un lexique des mots recherchés. Dès lors que l'on est amené à travailler avec un lexique de potentiellement grande taille (ce qui est notre cas), on ne dispose généralement pas d'un nombre suffisant d'exemples annotés de chaque mot pour apprendre de manière globale, au niveau du mot, la très grande variété des formes possibles pour l'écriture de ce mot. Cela est d'autant plus vérifié dans notre cas applicatif, omni-scripteur et où l'écriture est non contrainte. En conséquence, notre moteur de reconnaissance de mots devra s'appuyer implicitement ou explicitement sur l'information disponible au niveau des caractères (ou du moins au niveau d'un ensemble d'entités élémentaires plus restreint que le mot). On parle d'« approche analytique » [Anquetil 2008] (par opposition aux approches globales, visant à reconnaître directement la forme de chaque mot du lexique). Dans notre contexte applicatif, les catégories à reconnaître (ici les mots du lexique) ne sont pas tous connus directement au travers d'exemples visuels ; nous n'avons accès pour entraîner notre moteur de reconnaissance de mots qu'à des exemples des sous-catégories composant les mots, à savoir les caractères. Qui plus est, ces sous-catégories sont caractérisées par une très forte variabilité intra-classe.

Cette diversité d'information constitue une richesse certaine au regard des cas d'étude auxquels nous nous sommes précédemment confrontés dans ce manuscrit, qu'il nous faudra tenter de mettre à profit au mieux. Nous avons pour cela proposé un système dont les principales originalités sont détaillées dans la section suivante.

3.3.3 Principales originalités du système proposé

Dans un domaine de recherche où la littérature est aussi abondante que celui de la reconnaissance d'écriture manuscrite, il me serait difficile de montrer les principales originalités du système que nous avons proposé sans m'appuyer sur une typologie des approches existantes basée sur un état de l'art fouillé. Afin d'éviter d'alourdir la lecture de ce chapitre néanmoins, j'ai choisi de placer cette typologie détaillée en annexe C, à laquelle le lecteur non spécialiste est encouragé à se référer.

De manière extrêmement résumée (voire grossière), les approches analytiques sont souvent préférées dès lors que le lexique est potentiellement volumineux. Quel que soit le type d'écriture considéré (en-ligne ou hors-ligne), la plupart des approches analytiques sont basées sur trois étapes principales. La première étape consiste en une segmentation du mot en entités plus élémentaires (qui peuvent être des caractères ou des entités encore plus élémentaires, obtenues par exemple par une fenêtre glissante). La séquence de ces entités élémentaires doit alors être reconnue. Pour cela, dans la deuxième phase, on construit un (ou des) modèle(s). Les approches basées sur une « segmentation implicite » (aussi appelées *segmentation-free*) sont basées sur des modèles appris « en aveugle », c'est-à-dire directement à partir de la séquence des entités élémentaires. Dans les approches basées sur une segmentation explicite, en revanche, on passe par une étape de reconnaissance menée au niveau des entités élémentaires (souvent des caractères). Dans tous les cas, les sorties récoltées sont utilisées dans une troisième phase de « décodage » des mots par alignement des sorties du (ou des) modèle(s) obtenus au terme de l'apprentissage avec les mots du lexique.

Les Modèles de Markov Cachés (MMC) étant particulièrement adaptés à la reconnaissance de séquences de taille variable, ils se retrouvent souvent au cœur des systèmes de reconnaissance de mots manuscrits [Hu 2000, Bianne-Bernard 2011]. Et ce, à la fois pour la modélisation des mots et/ou des entités élémentaires le composant, et pour la phase de décodage.

Dans ce cas, l'aspect génératif de l'apprentissage limite les performances de ces approches, surtout étant donnée la faible variabilité entre caractères. Vu que, de plus, la tâche de segmentation explicite d'un mot cursif en caractères est très complexe (que ce soit à partir du signal en-ligne ou de l'image hors-ligne), la communauté porte actuellement un fort intérêt à des approches basées sur une segmentation implicite et un apprentissage discriminatif (souvent mené par des réseaux de neurones) [Graves 2009], le décodage pouvant selon les cas être mis en œuvre directement par le réseau de neurones ou par une autre approche (p. ex. approche Markovienne et/ou basée sur un algorithme de programmation dynamique).

Un désavantage commun à la plupart de ces méthodes est qu'en cas de changement de l'alphabet à reconnaître (voire du style d'écriture du scripteur), il est généralement indispensable de procéder à un ré-apprentissage souvent fastidieux des paramètres du système de reconnaissance (en particulier des paramètres guidant la segmentation implicite ou explicite), ce qui entrave dans une certaine mesure sa généricité.

Sur la foi de cette étude de l'état de l'art, nous avons choisi de nous orienter vers une **approche analytique basée sur un apprentissage discriminatif**. La segmentation devra pouvoir s'adapter à différents styles d'écriture (voire à différents langages) sans nécessiter de ré-apprentissage spécifique. Une vue globale du système proposé est donnée en Figure 3.10.

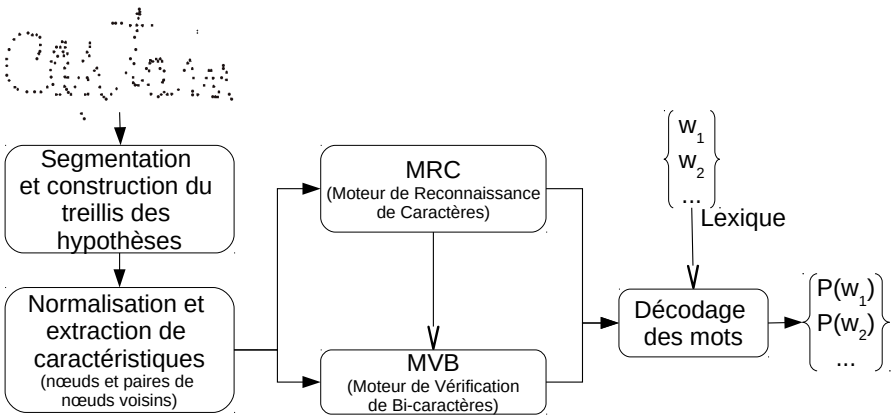


FIGURE 3.10 – Vue globale du système de reconnaissance de mots proposé. *Figure adaptée de [Prum 2013a].*

On commence par sur-segmenter les mots en entités plus élémentaires que les caractères (ici des « graphèmes »). Nous avons choisi de nous appuyer pour cela sur les informations concernant la dynamique du tracé issues du signal en-ligne, mieux à même de faciliter la segmentation que les informations visuelles issues de l'image reconstruite. Les combinaisons de ces graphèmes sont agencées sous la forme d'un treillis.

Afin de s'affranchir du type de dispositif de capture utilisé et notamment du type d'échantillonnage, et de réduire le bruit, les données dans chaque nœud (signal en-ligne et image hors-ligne reconstruite à partir de ce signal) du treillis sont normalisées et pré-traitées. Les caractéristiques [13] utilisées pour composer les signatures décrivant les nœuds du treillis sont

13. Note technique : ces 71 caractéristiques ont été sélectionnées automatiquement par la méthode enveloppante [Jain 1997] Sequential Floating Forward Selection (SFFS) [Pudil 1994] parmi 254 caractéristiques appartenant à 5 familles de caractéristiques en-ligne (locales et globales), et 10 familles de caractéristiques de forme hors-ligne. Parmi ces caractéristiques, 43 sont hors-ligne et 28 sont en-ligne. Pour plus de détails sur les caractéristiques retenues, merci de se référer à [Prum 2013a].

très majoritairement des caractéristiques de forme extraites de l'image reconstruite à partir du signal en-ligne, cette image ayant été normalisée lors d'une phase de pré-traitement ; des caractéristiques extraites localement et globalement du signal en-ligne sont également utilisées dans la signature.

Chaque nœud du treillis est ensuite analysé à deux niveaux : au niveau des caractères par le biais d'un **Moteur de Reconnaissance de Caractères (MRC)** entraîné de manière discriminative, puis au niveau des couples des caractères voisins par le biais du **Moteur de Vérification de Bi-caractères (MVB)**. L'usage du MVB vise, d'une part, à intégrer dans l'analyse le contexte des caractères voisins (qui influe grandement sur l'apparence d'un caractère donné) et, d'autre part, à lever les ambiguïtés dans l'analyse au niveau du caractère dues en particulier à la sur-segmentation systématique des caractères en graphèmes. Plus précisément, le MVB, également entraîné de manière discriminative, vérifie au niveau des couples de caractères voisins les hypothèses émises par le MRC aux niveaux des caractères.

Au final, la reconnaissance au niveau du mot (décodage du mot) est menée par programmation dynamique sur la foi des sorties du MRC et du MVB. Elle prend en compte conjointement les sorties du MRC et du MVB, et est guidée par le lexique. L'algorithme de décodage, bien que conçu spécifiquement pour notre problème, est un algorithme assez classique de programmation dynamique (voir [Prum 2013b] et la section 3.7.3 de [Prum 2013a] pour plus de détails).

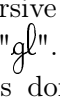
Puisque la phase de segmentation produit un treillis d'hypothèses de segmentation qui seront ensuite validées (ou invalidées) au final lors du décodage du mot (*via* le MRC et le MVB), on peut considérer que les modules de segmentation et de classification coopèrent pour la reconnaissance. De la même manière, puisque les caractéristiques composant la signature des nœuds du treillis sont sélectionnées par une méthode enveloppant le MRC (*cf.* note de bas de page numéro 13 ci-avant), les étapes d'extraction de descripteurs et de classification sont menées de manière collaborative. On couvre donc bien l'ensemble de la portion de la chaîne de traitement montrée en introduction, et plus précisément en Figure 3.9, page 91.

Même si je mets en évidence au fil des sections suivantes des ressemblances avec certaines approches introduites en parallèle de nos travaux dans la littérature, le système que nous proposons est assez original. L'une de ses principales originalités réside dans la **classification à deux niveaux** (caractères et bi-caractères) qui permet de répercuter au niveau plus global du décodage du mot les informations de bas niveau dont on dispose au niveau des graphèmes, en passant par une analyse menée à la fois au niveau des caractères et des couples de caractères voisins.

Une autre de ses originalités repose sur le module de segmentation et ses modes de coopération avec la classification des mots proprement dite. En effet, un peu à la manière d'une approche basée sur une segmentation implicite, notre système repose sur une sur-segmentation des mots en entités plus élémentaires que les caractères. Le fait que nous procédions de manière explicite à la reconnaissance d'entités plus élémentaires que le mot rapproche en revanche notre système des approches basées sur une segmentation explicite. En ce sens, on peut considérer que notre système est basé sur une **segmentation « semi-explicite »**, dont je détaille les principaux avantages dans la section ci-après, tandis que les principales originalités de la classification à deux niveaux seront détaillées dans la section suivante.

3.3.3.1 Segmentation semi-explicite

Afin de pallier les désavantages liés à la fois à l'utilisation des méthodes basées sur une segmentation explicite et implicite, nous nous sommes orientés vers une méthode intermédiaire entre ces deux grands types d'approches, que nous qualifions de semi-explicite. Je présente ci-après notre méthode de sur-segmentation en graphèmes, et la manière dont on construit le treillis d'hypothèses de caractères candidats à partir de ces graphèmes. Autrement dit, je m'intéresse ici à la manière de passer des **formes élémentaires** que l'on retrouve dans le document textuel (ici les graphèmes, à rapprocher des invariants évoqués en section 2.3), à des hypothèses concernant des **classes d'un niveau plus sémantique** (ici les caractères), et pour lesquelles on dispose d'exemples étiquetés.

Lors de la phase de segmentation des mots manuscrits en graphèmes candidats, on sélectionne comme points de segmentation candidats les points de la séquence du signal en-ligne qui sont les maxima et minima locaux sur l'axe des $\vec{y}$ au sein d'un trait d'écriture donné (entre un poser et un lever du stylet), comme illustré dans la partie gauche de la Figure 3.11. Cela engendre dans la très grande majorité des cas une sur-segmentation des caractères en graphèmes. Il peut néanmoins arriver que cette méthode de segmentation conduise à une sous-segmentation du signal, lorsque deux caractères sont enchaînés de manière cursive sans que l'on rencontre ni minimum ni maximum local (p. ex. avec le bi-caractère cursif "  ". Afin d'éviter cela, un post-traitement visant à segmenter arbitrairement les graphèmes dont la projection sur l'axe des $\vec{y}$ est trop grande [14] est appliqué.

Puis, on crée un treillis dont le premier niveau ($L = 1$) est constitué des nœuds $o_{(t,t+1)}$ contenant les graphèmes, et dont les nœuds $o_{(t,t+L)}$ des niveaux $L > 1$ contiennent les combinaisons des graphèmes contenus dans les nœuds $o_{(t,t+1)}$ à $o_{(t+L-1,t+L)}$, comme illustré dans la partie droite de la Figure 3.11.

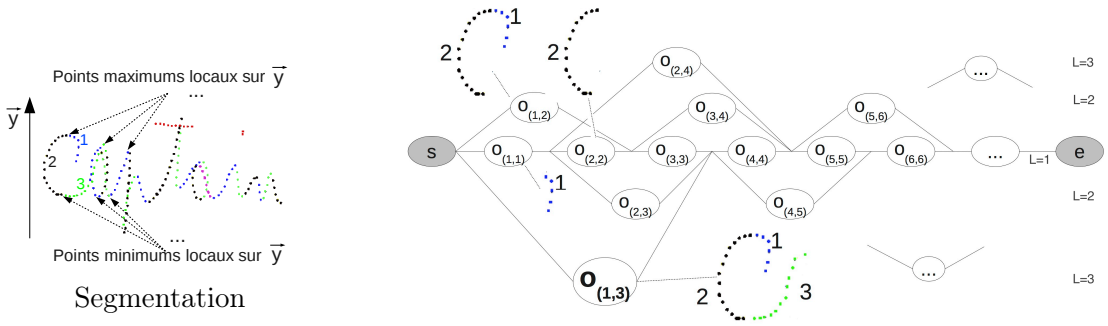


FIGURE 3.11 – Segmentation semi-explicite et création du treillis associé. *Figure adaptée de [Prum 2013a].*

À noter qu'à la différence de beaucoup d'approches basées sur une segmentation obtenue à partir du signal en-ligne, nous gérons les traits d'écriture retardés (barre du 't', point sur le 'i', etc.). En ne nous contentant pas, à l'instar de ces autres approches, de nous fier au lexique et/ou au modèle de langage pour lever les éventuelles ambiguïtés (p. ex. entre un 't' et un

14. Par rapport aux hauteurs du corps du mot et des zones de hampes ou jambages supérieurs et inférieurs, estimées lors de la phase de pré-traitement.

'l'), nous obtenons une amélioration sensible des taux de reconnaissance au niveau du mot (*cf.* [Prum 2013a]).

Nous avons délibérément choisi d'appliquer une méthode de segmentation triviale, dont on sait qu'elle va le plus souvent engendrer une sur-segmentation. En contrepartie, à la différence de la plupart des approches basées sur une segmentation implicite (souvent basées sur une fenêtre glissante dont la taille optimale varie en fonction du style d'écriture), elle ne dépend que du paramètre visant à éviter la sous-segmentation, qui est fixé relativement au signal normalisé en taille. Cela confère à la segmentation semi-explicite une plus grande **robustesse vis-à-vis de variations dans le style d'écriture** (et notamment dans la taille de l'écriture) que la plupart des approches de segmentation implicite (à l'exception de certaines approches basées sur le signal en-ligne qui ne dépendent d'aucun paramètre de segmentation, comme par exemple [Graves 2009]). Cette qualité est particulièrement intéressante dans notre contexte d'étude omni-scripteur. Le principe de notre méthode de segmentation est proche de celui de la méthode qui considère les traits descendants fondamentaux évoquée dans [Anquetil 2008]. La principale différence est que notre approche ne prend pas en compte l'ordre et le sens du tracé, ce qui la rend *a priori* plus facilement adaptable au cas purement hors-ligne.

De plus, comme nous allons le voir, la combinatoire engendrée par cette sur-segmentation est limitée, et cette méthode sera suffisamment facile à appréhender pour permettre à un utilisateur humain de fixer ses paramètres (le cas échéant sous la supervision de la machine), sous réserve qu'il connaisse suffisamment l'alphabet considéré et le fonctionnement de la méthode.

En particulier, il est possible de fixer de cette manière le nombre maximum de graphèmes pour un caractère donné dans l'alphabet. Dans notre système, nous avons ajusté ce paramètre en utilisant un histogramme cumulatif du nombre de graphèmes par caractère, calculé sur une base de validation contenant des caractères. Les valeurs trouvées avec cette méthode (voir section 3.7.3.2 de [Prum 2013a]) rejoignent celles qu'un humain aurait pu inférer ; par exemple, le nombre maximum de graphèmes à considérer pour former le caractère 'c' est de 3 (voir Figure 3.11 pour un exemple de caractère 'ⅽ' contenant 3 graphèmes), contre 6 pour le caractère 'ω', le plus long de l'alphabet latin. Cette information, intégrée au niveau de la reconnaissance de mots, permet de réduire la complexité algorithmique de notre système, mais aussi d'en améliorer les performances en évitant des erreurs de reconnaissance dues à une éventuelle sous-segmentation. On peut déduire de ces valeurs le nombre L_{max} de niveaux du treillis (puisqu'il s'agit du nombre de graphèmes maximal dans un caractère).

Dans le cas des méthodes basées sur une segmentation implicite et un MMC, les paramètres ci-dessus sont à rapprocher des nombres minimums et maximums d'états à prendre en compte pour chaque modèle de caractère. Étant donné que la taille de la fenêtre glissante utilisée pour la segmentation implicite est généralement fixée *a priori*, ces paramètres sont difficilement ajustables par l'utilisateur et requièrent le plus souvent une phase préalable d'apprentissage fastidieuse. Un changement d'alphabet nécessitera généralement une nouvelle phase d'apprentissage. De plus, ces paramètres sont également sensibles au style d'écriture du scripteur, et donc il faudra également les adapter en cas de personnalisation.

En comparaison avec la plupart des méthodes explicites précédemment développées, on évite une phase de segmentation complexe à mettre en œuvre et qui, dans la plupart des cas, nécessite également un ré-apprentissage de ses paramètres en cas de changement d'alphabet ou

de personnalisation (voir annexe C).

Le fait que, dans notre cas, les paramètres de la construction du treillis soient en nombre limité et facilement ajustables en fonction de l'alphabet considéré, rend la segmentation semi-explicite **aisément adaptable à d'autres alphabets** sans nécessiter de ré-apprentissage fastidieux de ses paramètres. Ainsi, dans le contexte du **projet PCSI sur la valorisation du patrimoine khmer** (financé par l'AUF), nous avons adapté avec succès la reconnaissance de caractères manuscrits à l'alphabet khmer : une démonstration est proposée sur le *web*[15]. De plus, s'ils sont correctement fixés *a priori*, ils ne nécessitent pas de ré-ajustement en cas de changements dans le style d'écriture.

En contrepartie de ces avantages, on obtient en sortie une sur-segmentation quasi-systématique (bien que contrôlée) des caractères. En conséquence, seuls peu de nœuds du niveau $L = 1$ du treillis correspondent à un caractère. Le reste du système est adapté à cette sur-segmentation, comme détaillé dans la section suivante.

3.3.3.2 Classification supervisée à deux niveaux avec gestion des nœuds « non caractères »

L'une des principales originalités de notre approche vis-à-vis des approches existantes basées sur une segmentation explicite est d'intercaler entre le MRC et le décodage de mots un niveau de classification intermédiaire qui permet l'intégration du contexte des caractères voisins, et ainsi de corriger les éventuelles erreurs de reconnaissance du MRC liées à la sur-segmentation. À ce niveau intermédiaire, les hypothèses émises par le MRC sont vérifiées au niveau des paires de caractères voisins par un Moteur de Vérification de Bi-caractères (MVB).

Le fait d'étudier l'utilisation d'un Séparateur à Vaste Marge (SVM) pour la reconnaissance des caractères cursifs peut également être vu, dans une certaine mesure, comme une originalité. En effet, même si ce type de classifieurs a prouvé son efficacité au niveau du caractère isolé, certaines difficultés ont freiné jusqu'à récemment son intégration dans des systèmes de reconnaissance de mots cursifs, à l'exception notable de [Bahlmann 2002] et [Ahmad 2009]. Du fait de la sur-segmentation systématique, le SVM devra être capable de gérer efficacement les nœuds qui ne sont pas des caractères (nœuds « non-caractères »).

La suite de cette section est dédiée à la présentation de ces deux originalités de notre système : la conception d'un MRC basé sur des SVMs gérant les nœuds non-caractères et l'utilisation d'un MVB pour intégrer le contexte des caractères voisins dans le décodage des mots.

a) Moteur de Reconnaissance de Caractères (MRC)

Basés sur la théorie statistique de l'apprentissage développée par Vladimir Vapnik dans les années 90 [Vapnik 1998], les SVMs [Schölkopf 1995, Schölkopf 2001] sont une famille de méthodes d'apprentissage supervisé très largement utilisés pour résoudre des problèmes de régression et de discrimination. Le principe de minimisation du risque structurel borne l'erreur de généralisation, ce qui en fait un outil très adapté aux espaces presque vides (ce qui est généralement le cas en présence de grandes dimensions). Outre ces fondements

15. https://www.youtube.com/watch?v=xnXg_CDxEE0

théoriques solides, les raisons principales de cet engouement sont liées à leur faible nombre d'hyper-paramètres et leurs excellents résultats en pratique.

Dans notre cas où nous cherchons un classifieur discriminatif avec un faible nombre de paramètres à ajuster et où chaque nœud du treillis est décrit par une signature composée d'un vecteur de 71 caractéristiques, nous avons choisi d'utiliser un SVM pour construire le MRC et alimenter le processus ultérieur d'exploration du treillis des hypothèses.

Dans la plupart des systèmes de reconnaissance de mots manuscrits, quel que soit l'algorithme utilisé pour le décodage des mots, il prend le plus souvent en entrée des probabilités *a posteriori* (*cf.* annexe C). Or, avec des SVMs, initialement conçus pour des problèmes de classification binaires, il est plus difficile d'obtenir des estimations de probabilités multi-classes *a posteriori* qu'avec des MMC ou même des réseaux de neurones. Cela explique probablement, au moins en partie, le fait que malgré leur popularité dans les domaines de l'apprentissage et de la reconnaissance de formes, **les SVMs ont été peu utilisés pour la reconnaissance d'écriture** manuscrite dans un premier temps.

Néanmoins, durant les dernières décennies, plusieurs solutions ont été proposées pour déduire des estimations de probabilités *a posteriori* à partir des sorties des SVMs. Ces solutions dépendent entre autres de l'extension multi-classes choisie (DAGSVM, un-contre-un, un-contre-tous, etc.). Nous choisissons ici d'utiliser un SVM avec noyau RBF et une stratégie un-contre-un, dont la compétitivité dans le cas général a été démontrée dans [Hsu 2002], et confirmée dans notre contexte expérimental par nos expérimentations. Vu que la variabilité inter-classes est très faible, et ce, particulièrement pour certaines paires de classes (comme par exemple les caractères '𝔬' et '𝔰'), cela peut se comprendre intuitivement. Pour l'estimation des probabilités multi-classes à partir des probabilités par paires, nous utilisons le deuxième algorithme d'optimisation décrit en détails dans [Wu 2004] et implémenté dans LibSVM [Chang 2011] pour sa convergence rapide dans la pratique, et ses bonnes performances.

La plupart des nœuds de notre treillis d'hypothèses contiennent des graphèmes ou des groupes de graphèmes qui ne sont pas des caractères. Le système doit donc être capable de **gérer les nœuds non-caractères**. Cela commence par intégrer la possibilité de reconnaître de tels nœuds au niveau du MRC. Bien que l'on puisse parler de classification « avec rejet » au niveau du MRC, l'objectif n'est pas de rejeter définitivement certains nœuds en sortie du MRC, car cela rendrait notre système trop sensible à d'éventuelles erreurs du MRC. Il s'agit au contraire de transmettre au MVB (respectivement à l'algorithme de décodage des mots) des hypothèses (resp. des valeurs de probabilités *a posteriori*) modifiées en fonction de la vraisemblance qu'un nœud donné soit un non-caractère.

Nous avons donc choisi de nous tourner vers une approche qui consiste à rajouter une classe non-caractères pour l'apprentissage du SVM. Cette classe non-caractères sera entraînée comme toutes les autres en « un-contre-un », et donc son introduction modifie mécaniquement les probabilités *a posteriori* de l'ensemble des caractères candidats, en fonction de la probabilité estimée que le nœud courant soit un non-caractère. Dans la pratique, cette approche montre une bonne capacité à détecter les nœuds non-caractères (bon rappel), mais aura tendance à retourner un nombre élevé de faux positifs (précision très inférieure au rappel).

Ce comportement du classifieur, qui consiste à avoir tendance à assigner préférentiellement un exemple d'entrée à la classe la plus représentée dans la base d'apprentissage (ici la classe non-caractères) est répandu [Sun 2009]. Nos expérimentations préliminaires montrent que, dans notre cas, l'une des stratégies de re-pondération les plus traditionnellement utilisées pour les

SVM un-contre-un pour résoudre des problèmes avec données non équilibrées [Chang 2011], n'est pas suffisante au sens où elle ne fait pas baisser de manière significative le taux de faux positifs. Nous sommes donc en train d'investiguer d'autres techniques permettant de s'accommoder de données non équilibrées dans les SVMs.

Cependant, en pratique, ces faux positifs ne constituent pas nécessairement un problème majeur dans notre cas où aucun rejet n'est définitif. En effet, comme détaillé dans la section 4.2. de [Prum 2013a], malgré ses imperfections, l'intégration du rejet au niveau du MRC nous permet d'améliorer les résultats de manière substantielle en termes de reconnaissance, sans engendrer de surcoût calculatoire notable lors de la phase de reconnaissance de nouveaux mots.

b) Moteur de Vérification de Bi-caractères (MVB)

La forme des caractères voisins influence grandement la forme du caractère courant dans un mot manuscrit, en particulier lorsque celui-ci est écrit de manière cursive, comme illustré en Figure 3.12. En conséquence, nous ne sommes pas les seuls à avoir eu l'idée d'intégrer le contexte des caractères voisins dans la reconnaissance de mots. Parmi les méthodes basées sur une segmentation implicite, on peut citer les travaux présentés en parallèle des nôtres dans [Bianne-Bernard 2011], basés sur la proposition modèles de tri-graphes. Très grossièrement, il s'agit de construire plusieurs modèles par caractère, en fonction des spécificités des caractères voisins (commençant par une minuscule ou majuscule, avec trait ascendant ou descendant, etc.). La construction des modèles de trigraphes s'appuie sur des connaissances formalisées avec l'aide de l'humain, qui devra donc *a priori* être à nouveau sollicité en cas de changement d'alphabet.

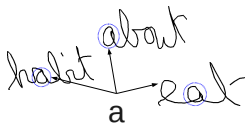


FIGURE 3.12 – Différentes formes que peut prendre le caractère 'a' écrit par un même scripteur, en fonction du contexte des caractères voisins. *Figure extraite de [Prum 2013a].*

L'originalité de l'approche que nous proposons pour intégrer le contexte du caractère courant porte essentiellement sur l'architecture globale du système de reconnaissance où, comme illustré en Figure 3.10 (page 93), on intercale un niveau intermédiaire de vérification des hypothèses émises par le MRC au niveau des paires de caractères voisins avant le décodage des mots. Il s'agit du Moteur de Vérification de Bi-caractères (MVB). Notre MVB est composé d'un classifieur par bi-caractère, entraîné en un-contre-tous. Nous avons choisi une régression logistique [16] pour construire ces classifieurs, car ses résultats de classification sont très proches de ceux d'un SVM linéaire (étant donné que les deux modèles sont basés sur des fonctions de coût asymptotiquement équivalentes [Vapnik 1998], la régression logistique ayant même été à l'origine des travaux de Vapnik sur les SVM). Ses principaux avantages sur le SVM sont que la régression logistique nous fournit des valeurs de probabilités *a posteriori* directement utilisables lors de la phase de décodage, et qu'un SVM [17] était trop coûteux en mémoire, en

16. Avec une régularisation L_2. L'implémentation utilisée est celle de LibSVM.
17. Dans sa formulation classique, sans stratégie de pré-sélection des données comme p. ex. dans [Yu 2003].

raison notamment du nombre de classes de bi-caractères. Pour plus de détails sur le MVB, le lecteur pourra se référer à la section 3.6.2. de [Prum 2013a].

L'évaluation expérimentale détaillée dans la section 4.2.2 de [Prum 2013a] montre que l'intégration du MVB rend le système de reconnaissance de mots plus robuste vis-à-vis des éventuelles erreurs du MRC et améliore les performances du système global en termes de reconnaissance de mots.

3.3.4 Bilan et améliorations possibles

Nous avons proposé un système de reconnaissance de mots manuscrits en-ligne dont l'une des principales originalités repose sur le module de segmentation, et ses modes de coopération avec le module de catégorisation du mot proprement dit. En effet, comme les approches basées sur une segmentation implicite, il est basé sur une sur-segmentation systématique du mot en entités plus élémentaires que des caractères (ici des graphèmes). Le fait que le système procède explicitement à la classification au niveau des caractères le rapproche en revanche des approches basées sur une segmentation explicite. Cette segmentation « semi-explicite », menée à partir du signal en-ligne, est mise en œuvre de manière à s'affranchir des étapes de ré-apprentissage fastidieuses souvent nécessaires, en cas de changement d'alphabet (voire de variations dans les styles d'écriture), avec la plupart des approches basées sur une segmentation implicite ou explicite.

À l'issue de cette segmentation, on construit un treillis dont les nœuds sont composés de groupes de graphèmes qui seront analysés à deux niveaux : au niveau des caractères par un Moteur de Reconnaissance de Caractères (MRC), et au niveau des couples de caractères voisins par un Moteur de Vérification de Bi-caractères (MVB). En vérifiant au niveau des couples de caractères voisins les hypothèses émises par le MRC au niveau des caractères, le MVB permet, d'une part, d'intégrer dans l'analyse le contexte des caractères voisins et, d'autre part, de lever les ambiguïtés dans l'analyse des caractères en se plaçant à un niveau plus global d'analyse. Dans un souci de performance, les deux classifieurs sont entraînés de manière discriminative.

La reconnaissance au niveau du mot (décodage) se fait grâce à un algorithme de programmation dynamique intégrant les sorties du MRC et du MVB, guidé par le lexique.

Il ne s'agissait pas ici de concevoir une méthode (à la différence des travaux présentés en sections 2.2 ou 3.2 par exemple) mais plutôt un nouveau système, basé sur des modules d'analyse sélectionnés avec soin et organisés selon une architecture permettant d'intégrer efficacement les informations hétérogènes dont on dispose (caractéristiques de forme de l'image reconstruite à partir du signal en-ligne, informations sur la dynamique du tracé, exemples de chacun des caractères de l'alphabet et lexique). En ce sens, les contributions liées aux travaux présentés dans cette section sont **plutôt d'ordre applicatif que méthodologique**.

Les expérimentations intensives détaillées dans la thèse de Sophea Prum [Prum 2013a] montrent que le système proposé surpasse une méthode traditionnelle basée sur une segmentation implicite et un MMC utilisé à la fois pour la reconnaissance de caractères et le décodage des mots. Deux démonstrations du fonctionnement de notre système (respectivement pour la reconnaissance de mots manuscrits sur dispositifs tactiles nomades et dans un cadre applicatif lié à l'e-éducation) sont disponibles sur le *web*[18] [19]. Les pistes d'améliorations

18. http://youtu.be/N--a_XXboi0
19. http://youtu.be/ZVQraCE1BLA

possibles sont cependant nombreuses. Les deux pistes qu'il nous semblerait pertinent d'explorer prioritairement sont évoquées ci-après.

La première d'entre elles concerne une meilleure prise en compte conjointe des deux niveaux d'analyse (caractère et bi-caractères), à deux occasions dans le système global.

Premièrement, les **descripteurs utilisés** pour décrire l'ensemble des nœuds du treillis ont été sélectionnés par une méthode « enveloppante » visant à optimiser les résultats du MRC. Donc, ils ne sont pas nécessairement adaptés aux spécificités des bi-caractères, mais sont néanmoins utilisés pour décrire tous les nœuds du treillis (y compris ses niveaux > 1 examinés par le MVB). Puisque chacun des nœuds du treillis est examiné à la fois par le MRC et le MVB, il ne serait pas forcément judicieux de prévoir deux jeux de descripteurs distincts (un pour le MRC, un pour le MVB). En revanche, le problème d'extraire un jeu de caractéristiques qui soit adapté à la fois à la description des caractères et des bi-caractères n'est pas trivial. On pourrait envisager de passer par une sélection conjointe de ces descripteurs avec les deux classifieurs (ce qui n'est *a priori* pas évident d'un point de vue théorique, vu la différence dans le niveau de l'information traitée).

Deuxièmement, l'intégration conjointe des sorties des moteurs de reconnaissance de caractères et de bi-caractères au niveau du **décodage du mot** reste un problème largement ouvert. Pour nos expérimentations, nous avons utilisé un algorithme de programmation dynamique dans lequel nous avons re-pondéré les poids des séquences (ou sous-séquences) comportant deux caractères par les probabilités *a posteriori* du MVB correspondant. Afin d'éviter de favoriser les mots les plus courts, nous avons normalisé les poids ainsi obtenus par la longueur des mots. Cette technique nous a fourni des résultats expérimentaux intéressants, mais mériterait d'être améliorée. Il pourrait s'agir de mettre en œuvre, lors de la phase de décodage, des techniques de fusion d'information tenant mieux compte de la différence de niveau de traitement entre les deux classifieurs.

Une autre piste d'amélioration possible concerne la **personnalisation du système de reconnaissance d'écriture**. En effet, comme nous l'avons vu tout au long de cette section, une part importante des difficultés liées à la reconnaissance d'écriture manuscrite provient de la très grande variabilité entre les écritures de personnes différentes. Afin d'y remédier, de nombreux auteurs cherchent à personnaliser le moteur de reconnaissance vis-à-vis du scripteur [Connel 2002, Schlapbach 2007], qui doit donc être préalablement identifié. De nombreuses méthodes permettant de reconnaître un scripteur ont été proposées dans la littérature [Plamondon 1989, Schlapbach 2008] (le cas échéant en combinant son écriture avec d'autres modalités telles que sa voix [Humm 2009]). Ce type d'approches de personnalisation basées sur une identification préalable du scripteur permettent d'améliorer considérablement les résultats de la reconnaissance de mots, mais elles restreignent la généricité du système, puisqu'elles supposent que chacun des scripteurs possibles est préalablement enregistré. Elles sont donc inapplicables dans notre contexte omni-scripteur. En revanche, grâce en particulier à l'encadrement de trois stagiaires, nous avons étudié la personnalisation vis-à-vis non plus du scripteur, mais du style d'écriture du scripteur. La reconnaissance du style d'écriture est un problème auquel les chercheurs n'ont commencé à s'intéresser que plus récemment [Daher 2012] qu'à celui de la reconnaissance de scripteur. Nous avons proposé dans [Bui 2011, Visani 2012] deux approches de catégorisation du style d'écriture basées sur un apprentissage non supervisé des distributions des signatures des caractères (respectivement des graphèmes). Pour la personnalisation, nous avons utilisé une méthode triviale, basée sur une re-pondération des exemples de la base

d'apprentissage en fonction du style d'écriture du scripteur. Les résultats expérimentaux sont néanmoins encourageants, et nous souhaiterions aller plus loin dans cette voie, en cherchant à concevoir des méthodes de personnalisation plus raffinées, basées sur l'adaptation non plus de la base d'apprentissage, mais du moteur de reconnaissance en lui-même.

Par ailleurs, dans le cadre de nos collaborations avec Andreas Fischer (voir section 1.4), nous sommes en train d'étudier la possibilité de coupler notre module de reconnaissance de style d'écriture (purement statistique) avec une méthode structurelle basée sur les caractéristiques neuro-musculaires du scripteur. Cette méthode structurelle est en cours de conception au sein de Polytechnique Montréal. Nous sommes actuellement en phase de recherche de financements afin de poursuivre cette collaboration dans les meilleures conditions.

3.3.5 Discussion

Les recherches présentées ici constituent les premiers travaux réalisés au sein du laboratoire L3i sur la reconnaissance d'écriture manuscrite. Alors même que la littérature est très abondante sur ce sujet difficile. En raison principalement des contraintes de temps liées au projet RecoNomad et de notre absence de patrimoine logiciel sur ce sujet, nous avons dû concevoir par nous-mêmes la totalité de la chaîne de traitement, et n'avons matériellement pas eu le temps d'en étudier finement ni d'en optimiser chaque brique. En conséquence, comme nous venons de le voir, les améliorations possibles sont nombreuses. À l'issue de la présentation détaillée de cette approche et de ses améliorations possibles, il convient d'apporter des éléments de discussion concernant ses points forts et ses limitations, et les questions posées par cette étude.

Étant donné que l'une des originalités de notre système repose sur une classification à deux niveaux (caractères et bi-caractères), se pose naturellement la question de l'extension de l'approche proposée à des niveaux de reconnaissance supérieurs. On peut ainsi envisager par exemple l'intégration dans le décodage des mots de certains classifieurs de tri-caractères. L'idée serait d'intégrer conjointement le contexte des deux caractères voisins (précédent et suivant le caractère courant), afin de mieux prendre en compte le caractère cursif de l'écriture lors de la reconnaissance de mots. Notre point de vue sur cette question est que l'apport de l'intégration de classifieurs de tri-caractères serait minime. En effet, chaque caractère composant le mot (hormis le premier et le dernier caractère du mot) est actuellement vérifié par deux classifieurs au niveau du bi-caractère (un dont il est le premier caractère, et un dont il est le second), ce qui devrait suffire à intégrer le contexte des deux caractères voisins, si tant est que l'on dispose de suffisamment d'exemples lors des apprentissages du MRC et du MVB pour représenter toute la variété des combinaisons cursives de caractères possibles. En contrepartie de cet apport potentiellement minime, l'intégration de niveaux de reconnaissance supérieurs pose des problèmes techniques liés à la constitution d'une base d'apprentissage de bonne qualité et de taille suffisante au niveau des tri-caractères, et à la complexité calculatoire additionnelle. En outre, elle poserait des problèmes plus scientifiques que nous n'avons jusqu'à présent pas été en mesure de résoudre complètement avec seulement deux niveaux d'analyse. En particulier, le choix de descripteurs adaptés à de multiples niveaux, et l'intégration conjointe des multiples niveaux d'analyse dans le décodage du mot, sont des problèmes non triviaux auxquels nous nous sommes jusqu'à présent attaqué en utilisant des heuristiques, efficaces en pratique, mais ne reposant pas réellement sur une justification théorique solide.

Dès lors, **on peut remettre en cause les choix que nous avons fait** dans la conception de notre système à deux niveaux d'analyse. Il ne s'agit pas ici de remettre en cause le choix de segmentation semi-explicite, qui confère une certaine généricité à l'approche proposée. D'autant plus que les graphèmes obtenus nous permettent également d'identifier le style d'écriture du scripteur courant et de personnaliser le système, améliorant sensiblement les résultats en termes de reconnaissance de mots. Il ne s'agit pas non plus de remettre en cause l'usage de techniques basées sur un apprentissage discriminatif, mieux à même de distinguer des classes parfois proches dans l'espace de représentation de bas niveau.

Par contre, notre architecture mérite d'être discutée. Vu son rôle central dans cette architecture, le MRC est en effet confronté à de nombreux problèmes (énorme variabilité des données d'entrée, gestion des nœuds du treillis non-caractères, variabilité des caractères en fonction du contexte des caractères voisins, etc.), que l'on a cherché à résoudre en particulier en rajoutant un niveau d'analyse.

Néanmoins, l'analyse à deux niveaux des mots (telle que nous la mettons en œuvre) pose un certain nombre de difficultés en pratique. Par exemple, comment notre système résiste-t-il au passage à l'échelle en termes de classes ? Lorsqu'on est en présence de langages qui comportent un nombre important de caractères, comme par exemple le khmer [20] (qui en comporte plus de 80, sans compter les 33 pieds issus des consonnes), le nombre de bi-caractères explose de manière combinatoire. Cela pose un certain nombre de problèmes lors de la vérification qui est faite au niveau du bi-caractère des hypothèses émises au niveau du caractère, et dans l'intégration conjointe des deux niveaux d'analyse lors du décodage du mot. Bien heureusement, la plupart des langages alphabétiques comportent un nombre relativement réduit de lettres, et notre système reste donc applicable pour une grande variété de langages. Et, le cas échéant, on peut toujours envisager une meilleure prise en compte du lexique afin de n'entraîner que les MVB des bi-caractères les plus fréquents (quitte à utiliser une fonction de lissage lors du décodage du mot en l'absence de MVB).

Mais, si l'on rajoute à ces difficultés pratiques les difficultés d'ordre plus théorique décrites plus haut et concernant la représentation et l'analyse à de multiples niveaux de l'information, on peut (avec le recul dont on dispose aujourd'hui et libérés de la pression contractuelle liée à ce travail) se demander s'il ne serait pas plus judicieux de chercher une autre architecture mieux à même de passer de l'information de bas niveau sémantique collectée au niveau des graphèmes à celle, de plus haut niveau sémantique, des mots, tout en tirant parti de l'information dont on dispose au niveau des caractères sous la forme d'exemples annotés. Plutôt que d'adopter une démarche très « reconnaissance de formes » en rajoutant un niveau d'analyse intermédiaire des bi-caractères (entre caractères et mots), on pourrait envisager au contraire d'accorder plus d'importance à l'analyse des graphèmes. L'idée serait de concevoir une architecture mieux à même de gérer la cohérence de la séquence des observations à partir des scores (ou des probabilités estimées) d'appartenance de chaque graphème à chacune des lettres de l'alphabet (et le cas échéant d'une classe « lien inter-caractère »). Le décodage des mots pourrait alors se faire directement à partir de la séquence des graphèmes. Ce changement global d'architecture devrait alors certainement être répercuté sur la manière de sur-segmenter le mot en graphèmes. Le problème de conserver la relative généricité de notre approche de segmentation semi-explicite dans une telle architecture n'est pas trivial.

20. Pour rappel, il s'agit du MRC, et non du moteur de reconnaissance de mots complet, que nous avons adapté au langage khmer.

Nous arrivons maintenant au terme de cette étude qui concerne essentiellement les phases de segmentation et de classification (via l'extraction de descripteurs), et leurs modes de coopération. Nous allons reprendre notre exploration à rebours de la chaîne de traitement de reconnaissance d'images de documents par classification. Dans la section suivante, nous allons nous intéresser au tout début de la chaîne, et en particulier aux stratégies possibles pour pallier un éventuel manque d'images annotées en entrée de cette chaîne. Nous reviendrons sur les travaux présentés ci-dessus dans la conclusion du chapitre, en section 3.5.

3.4 Génération d'images semi-synthétiques de documents

3.4.1 Introduction

Depuis le début de ce chapitre, nous nous sommes essentiellement intéressés aux phases de segmentation des motifs dans l'image, de classification (et le cas échéant à leurs modes de coopération), ainsi qu'à la restitution des informations apprises à un utilisateur humain. Dans les **travaux en cours** présentés dans cette section, nous allons nous pencher sur les manières de contourner les problèmes qui peuvent survenir dans cette chaîne de traitement lorsque les informations d'entrée sont très partielles.

Dans certains cadres applicatifs, il est particulièrement difficile d'accéder au nombre d'images requises pour mettre en œuvre les traitements nécessaires à leur reconnaissance. C'est entre autres souvent le cas lorsque l'on s'intéresse aux **documents anciens**, surtout lorsque ceux-ci sont précieux ou endommagés, et doivent donc être manipulés avec précaution. C'est un problème auquel nous sommes confrontés dans le **projet ANR DIGIDOC**, qui se situe dans le contexte général de la numérisation de documents, et plus précisément celle des documents précieux et anciens.

Outre cette difficulté d'accès aux images proprement dites, un obstacle supplémentaire réside dans l'acquisition des annotations liées aux images. Si l'on prend l'exemple du texte, des systèmes capables de transcrire de manière complètement automatique un document ancien (depuis l'analyse de la mise en page jusqu'à la reconnaissance de mots et leur alignement avec l'image) ont été proposés [Fischer 2014] très récemment. Cependant, vu la complexité de la tâche et en raison en particulier du fait que les documents anciens ne suivent pas des règles orthographiques strictes, leurs performances restent à l'heure actuelle en-deçà de celles attendues pour alimenter un système de reconnaissance. L'annotation des documents anciens reste donc à l'heure actuelle largement manuelle, ou du moins assistée par un opérateur humain. Cela pose à la fois des problèmes de coût et de qualité variable des annotations (notamment en fonction de l'opérateur).

Afin de tenter de pallier ces problèmes, plusieurs initiatives de *crowdsourcing* ont vu le jour ces dernières années, certaines sous la forme de jeux pour les rendre plus attractives. On peut par exemple citer la plateforme *web* du projet tranScriptorium [21], qui permet à des internautes volontaires de transcrire manuellement les manuscrits du philosophe anglais Jeremy Bentham (1748–1832). Si ce type d'initiatives permet effectivement de réduire le coût et de vérifier (statistiquement) la qualité des annotations, elles doivent faire face à des difficultés techniques difficiles à résoudre automatiquement avec la précision requise, comme par exemple l'alignement de la transcription et de l'image [Gatos 2014]. De plus, elles concernent essentiellement des documents écrits dans un langage que le grand public est à même d'appréhender, et excluent de fait certains documents parmi les plus anciens.

Les effets cumulés de la rareté des images de documents anciens et des difficultés liées à leur annotation font que l'on dispose bien souvent de très peu d'information en entrée de la chaîne de reconnaissance.

21. http://blogs.ucl.ac.uk/transcribe-bentham/

3.4.1.1 Difficultés liées à la rareté des données annotées et stratégies de remédiation possibles

La rareté des images annotées disponibles en entrée de la chaîne de reconnaissance pose un certain nombre de problèmes, à la fois pour l'entraînement des approches de reconnaissance, et pour la caractérisation de leurs performances.

En ce qui concerne l'**entraînement des approches de reconnaissance**, la principale difficulté est que, vu la grande taille des signatures typiquement utilisées pour décrire les images, un jeu de données très fourni est généralement requis. En effet, en l'absence d'un nombre suffisant d'exemples pour l'apprentissage, le problème à résoudre est mal conditionné, ce qui peut entraîner une instabilité du classifieur ou un phénomène de sur-apprentissage. C'est la malédiction de la dimensionnalité (voir section 1.3.2). Ce problème est particulièrement prégnant avec certaines techniques modernes basées sur un apprentissage profond, qui nécessitent généralement un très grand nombre d'exemples, mais permettent en contrepartie d'obtenir une excellente précision sur des problèmes complexes.

Intéressons-nous maintenant à la **caractérisation des performances** des approches de reconnaissance. Premièrement, si les bases utilisées pour cette caractérisation contiennent trop peu d'images, se pose le problème de l'aspect significatif (statistiquement parlant) des conclusions que l'on peut en tirer. Deuxièmement, si les images utilisées sont trop peu représentatives des différentes variations/dégradations que l'on peut retrouver dans la pratique, il est très difficile d'évaluer la robustesse des algorithmes de reconnaissance vis-à-vis de ces variations. D'autant plus que la plupart des bases d'images disponibles ne contiennent que très peu (voire pas) d'information sur les dégradations présentes dans les images.

Afin de pallier ces difficultés, plusieurs types de stratégies sont communément utilisés, parmi lesquels nous pouvons citer les deux types ci-après.

Le premier type de stratégies repose sur des **techniques de ré-échantillonnage** statistique, qui consistent à apporter, généralement aléatoirement, des petits changements dans la base de données annotées. Ces techniques sont utilisées bien au-delà de la reconnaissance d'images, pour diverses applications d'analyse de données. Elles peuvent être mises en œuvre lors de la phase d'entraînement, ou pour la caractérisation de performances. En ce qui concerne l'entraînement, l'objectif est généralement de construire des modèles/classifieurs plus stables par agrégation de modèles/classifieurs appris sur des versions rééchantillonnées de la base de données annotées disponible en entrée. En ce qui concerne la caractérisation de performances, l'objectif est généralement de diminuer les biais liés au manque de données. Il peut s'agir par exemple d'évaluer les performances moyennes d'un algorithme de reconnaissance à partir de plusieurs versions ré-échantillonnées des bases d'apprentissage, de validation et éventuellement de test (p. ex. validation croisée voire technique du *leave-one-out*).

Quelle que soit la manière de mettre en œuvre le ré-échantillonnage (que ce soit lors de la phase d'apprentissage, de validation ou de test), son efficacité est naturellement limitée par le nombre et la variabilité des enregistrements annotés disponibles en entrée.

Plusieurs chercheurs se sont donc tournés vers un second type de stratégies, qui consiste à **générer des images semi-synthétiques** et la vérité-terrain associée. Ces images sont le plus souvent obtenues à partir d'images réelles, en y apportant des modifications synthétiques. Le cas échéant, certaines approches permettent de composer une nouvelle image à partir d'éléments

extraits de diverses images réelles. Les images semi-synthétiques obtenues peuvent être utilisées à la fois pour enrichir l'apprentissage (« ré-apprentissage ») et pour évaluer la robustesse des approches de reconnaissance d'images sur des données dont la vérité-terrain est maîtrisée.

Il existe plusieurs travaux traitant du **ré-apprentissage** de systèmes de reconnaissance à partir de données semi-synthétiques. On peut citer par exemple les travaux récents présentés dans [Opitz 2014], qui concernent la reconnaissance de texte dans des scènes naturelles. L'approche proposée repose sur un apprentissage profond de réseaux de neurones convolutionnels. L'entraînement se fait sur 60000 images d'apprentissage, parmi lesquelles 55000 sont semi-synthétiques et obtenues grâce à des petites distorsions du texte présent dans la base d'apprentissage. Il est frappant de noter que l'on atteint presque les mêmes taux de reconnaissance avec une base d'apprentissage composée uniquement d'images semi-synthétiques qu'avec toutes les images, tandis que les taux de reconnaissance chutent drastiquement si l'on n'utilise que les 5000 images réelles. Le ré-apprentissage d'approches de reconnaissance de mots manuscrits a également été étudié, en particulier dans [Varga 2003a, Varga 2003b] pour le cas des documents modernes. Des améliorations prometteuses des performances ont été rapportées grâce à l'introduction de données semi-synthétiques dans la base d'apprentissage, bien que ces données semi-synthétiques soient obtenues par des modèles de déformation assez basiques, et soient peu réalistes visuellement.

Les applications les plus abondantes concernent l'**évaluation de la robustesse**. Dans le domaine de l'analyse d'images de documents, il n'est pas rare que des compétitions soient organisées sur des documents modifiés de manière synthétique. Les bases générées pour cette occasion sont alors mises à la disposition de la communauté des chercheurs, qui peuvent les réutiliser comme *benchmark* afin de confronter leurs approches à celles de la littérature. Ces bases sont parfois organisées en fonction des variations/dégradations présentes dans les images. C'est par exemple le cas des bases de symboles architecturaux que nous avons utilisées pour nos expérimentations présentées en section 3.2.

Les initiatives listées ci-dessus concernent surtout des documents modernes. Cela peut être expliqué, en partie, par le manque d'approches de génération d'images semi-synthétiques permettant d'obtenir des images visuellement semblables à des images de documents anciens (le réalisme visuel étant une qualité appréciée par les chercheurs, notamment pour l'évaluation de performances).

3.4.1.2 Objectif visé

Notre objectif est de proposer des solutions permettant de pallier les problèmes liés à la rareté ou au manque d'exhaustivité des bases d'images annotées de documents anciens actuellement disponibles. Pour cela, notre proposition est de concevoir un système capable, à partir d'un faible nombre d'images d'entrée correspondant à des documents modernes ou anciens, de générer des **images semi-synthétiques de documents anciens**.

Il s'agit donc de rajouter une étape supplémentaire de génération de documents semi-synthétiques dans la chaîne de traitement traditionnelle de la reconnaissance d'images. Cette étape supplémentaire est mise en évidence dans la Figure 3.13, en page suivante. C'est le sujet de la **thèse de Kieu Van Cuong**[22], qui s'inscrit dans le contexte du projet ANR DIGIDOC.

22. Doctorant financé par le projet ANR DIGIDOC et en co-supervision entre le L3i et le LABRI (Bordeaux).

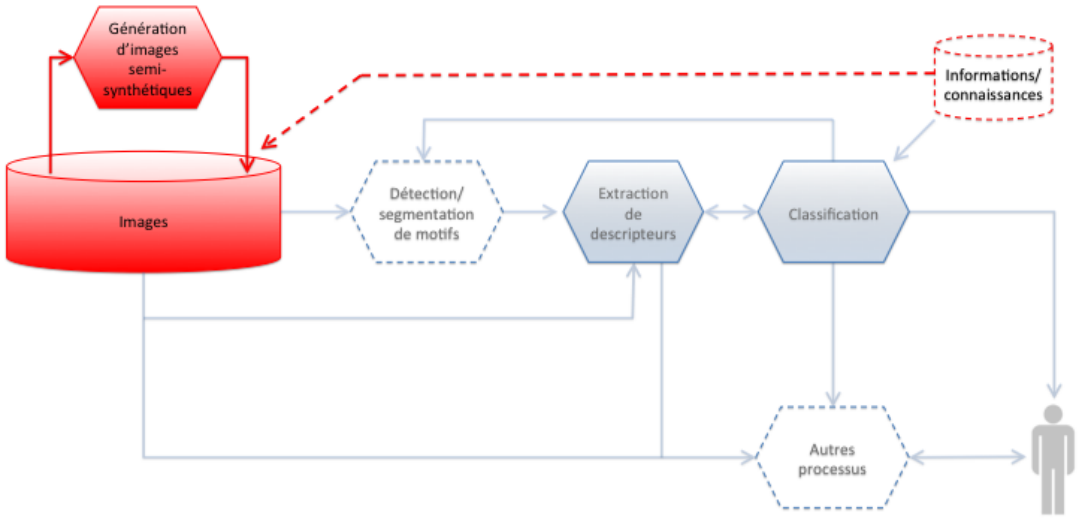


FIGURE 3.13 – Étape supplémentaire dans la chaîne de traitement traditionnelle de classification pour la reconnaissance d'images de documents faisant l'objet de cette section.

Afin de générer des images semi-synthétiques de documents anciens, nous nous appuyons sur la proposition de **méthodes/modèles de dégradation** permettant de modifier les images d'entrée de manière à y rajouter des défauts fréquemment rencontrés dans des documents anciens. Le générateur de documents synthétiques que nous utilisons a été développé dans le cadre du projet DIGIDOC et est décrit dans [Journet 2010]. La Figure 3.14 présente une vue simplifiée de cet outil de génération de documents semi-synthétiques, focalisée sur l'étape qui nous intéresse, à savoir celle de dégradation des images.

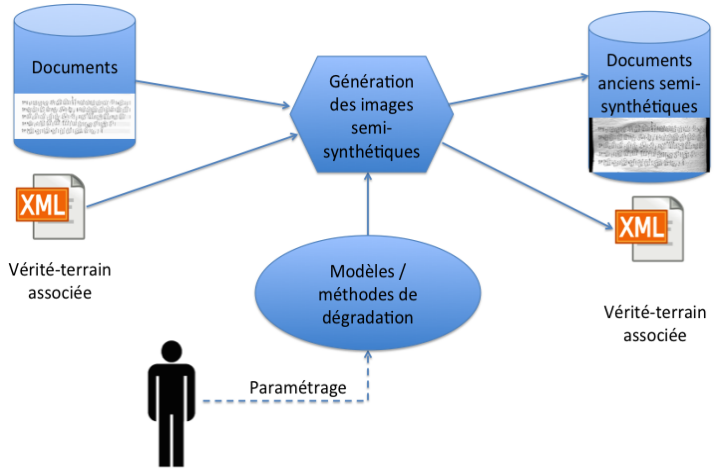


FIGURE 3.14 – Vue simplifiée de l'outil de génération de documents semi-synthétiques anciens utilisé. Les pointillés correspondent à des étapes optionnelles.

Si les images d'entrée sont annotées par leur vérité-terrain, notre système permet de fournir une vérité-terrain associée aux images semi-synthétiques générées à partir de ces images d'entrée (en tenant compte, le cas échéant, des dégradations apportées dans l'image). Dans la mesure du possible, nous cherchons à ce que le type (voire le niveau) des dégradations apportées à l'image puissent être choisies par l'humain. L'idée est de lui permettre de générer

des bases d'images semi-synthétiques rangées par nature/niveau de dégradation.

Maintenant que le contexte applicatif et scientifique de cette étude, ainsi que nos objectifs, sont définis, nous pouvons passer à une description plus détaillée des principales originalités du système que nous utilisons.

3.4.2 Principales originalités du système

Les principales originalités du système que nous utilisons reposent sur les méthodes/modèles de dégradation sur lesquels il s'appuie.

Plusieurs modèles de dégradation ont été proposés dans la littérature [Baird 1993, Kanungo 1993, Kanungo 2000, Zhai 2003, Liang 2008, Smith 2008, Moghaddam 2009]. Les dégradations modélisées sont typiquement liées aux processus d'impression et de digitalisation. Il peut s'agir par exemple de modéliser les bruits ou distorsions qui peuvent survenir en cas de mauvais paramétrage de l'imprimante ou du scanner, l'aspect bombé d'un livre près de la reliure (lorsqu'il n'est pas numérisé à plat), ou encore les phénomènes de « transvision » (apparition du verso sur le recto) qui peuvent notamment apparaître lorsque l'opérateur humain chargé de la numérisation omet de placer un calque sombre derrière le verso. Certains travaux visent également à modéliser des défauts physiques intrinsèques aux documents, comme par exemple la diffusion ou l'affadissement de l'encre. Plus généralement, ces travaux visent le plus souvent à modéliser les dégradations présentes dans les documents dans un but final de restauration. Seuls peu de ces modèles sont applicables en niveaux de gris et/ou adaptés aux dégradations intrinsèques des documents anciens.

Nous avons proposé une méthode et un modèle de dégradation, dont la principale originalité réside dans le fait qu'ils visent à reproduire, en niveaux de gris, certains des dommages subis par les documents anciens. À la différence de nos travaux présentés précédemment, il s'agit de méthodes de traitement – et non d'analyse – d'images, puisqu'on récupère en sortie une nouvelle image, et non une information supplémentaire concernant l'image d'entrée. Étant donné qu'ils sortent légèrement du cadre des questions abordées dans ce manuscrit, dédié à l'analyse d'images, je ne présenterai ci-après que brièvement ces travaux, avant de me concentrer sur leurs applications en analyse d'images (dans la section suivante).

La première méthode de dégradation est une **méthode de distorsion 3D**, que nous avons présentée dans [Kieu 2013a], et qui vise à imiter les froissures, pliures, etc. du papier. Ces distorsions surviennent avec l'âge des documents et/ou à cause des procédés utilisés pour leur numérisation. En effet, on ne peut souvent pas mettre un document ancien à plat, pour des raisons de conservation. Cette méthode prend en entrée une image de document réelle ou synthétique (2D) et un maillage 3D, obtenu suite à la digitalisation d'un document ancien (distordu du fait de son âge) à partir d'un scanner 3D. La méthode plaque la texture du document d'entrée sur le maillage 3D, et renvoie une image semi-synthétique de même contenu que l'image d'entrée, mais distordue selon le maillage 3D (voir Figure 3.15). L'application du modèle d'illumination de Phong [Phong 1975] permet d'obtenir des variations de niveaux de gris visuellement réalistes dans les zones de crêtes ou de vallées. À partir d'un unique maillage ou de ses sous-parties, une multitude d'images peut être générée à partir de différentes images de documents réels ou synthétiques, modernes ou anciens. À la différence de la plupart des modèles de distorsion géométrique de documents précédemment proposés [Kanungo 1994, Liang 2008], cette

méthode permet de générer des distorsions en 3D (et non en 2D), et ainsi d'obtenir des images semi-synthétiques d'aspect plus réaliste. Pour plus de détails, merci de se référer à [Kieu 2013a].

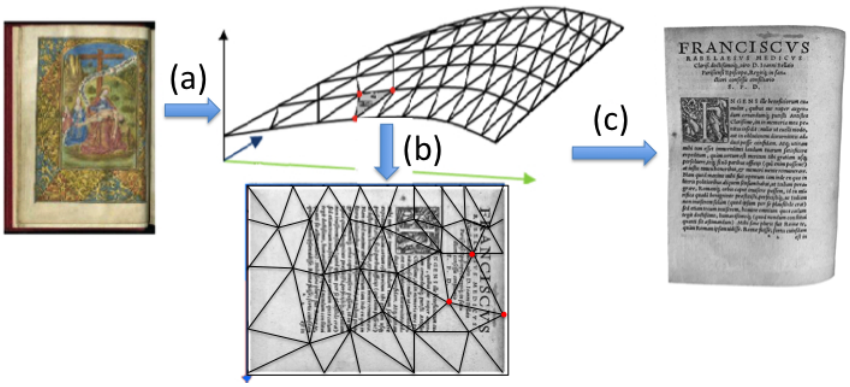


Figure 3.15 – Principe de la méthode de distorsion 3D. (a) Obtention d'un maillage 3D par digitalisation 3D d'une page de document ancien naturellement déformée. (b) Plaquage de la texture d'un document d'entrée quelconque (non déformé) sur le maillage 3D. (c) Obtention d'une image (2D) semi-synthétique simulant une distorsion 3D du document d'entrée et application du modèle de Phong afin de rendre le résultat visuellement plus réaliste.

La deuxième approche de dégradation que nous avons proposée est un **modèle de bruit local** détaillé dans [Kieu 2012b], et dont une version facilement paramétrable a été introduite dans [Kieu 2013b]. Ce modèle de bruit local vise à imiter les petites taches d'encre apparues lors de l'écriture, du dessin ou de l'impression de documents anciens, et les effacements partiels de l'encre dûs à l'âge du document. Le principe de ce modèle de dégradation est de générer des taches en des points sélectionnés aléatoirement, préférentiellement à proximité des bords des pixels d'avant-plan (déduits à partir de l'image binarisée), selon une procédure inspirée de [Kanungo 1994] et modifiée de manière à contrôler le nombre de points sélectionnés.

À la différence du modèle introduit dans [Kanungo 1994] puis réutilisé dans [Delalandre 2010, Visani 2011b] et qui vise plutôt à imiter les artefacts liés à la digitalisation/binarisation de documents, ce modèle permet de générer des taches en niveaux de gris et non binaires, de formes, d'orientations et de tailles variables, et d'aspect réaliste pour des documents anciens. Comme nous le verrons dans la section suivante qui traite des applications, ce modèle peut être appliqué sur du texte ou sur du graphique, mais par simplicité nous nous référerons ici à l'exemple du texte. Selon la configuration du système (manuelle ou automatique), l'utilisateur peut choisir :

- Le niveau de bruit (faible, moyen, fort), qui permet de fixer le nombre N_{sp} de points sources de dégradation (points de dégradation) en fonction du nombre de composantes connexes du document d'entrée ;
- Le cas échéant, les pourcentages des différents types de dégradations : taches isolées (p_1), taches connectées aux caractères (p_2), ou taches engendrant une perte de connectivité du caractère (p_3).

Puis, le système génère automatiquement des zones de bruit d'apparence variée et réaliste, vérifiant le cas échéant les contraintes données par l'utilisateur, selon une procédure illustrée en Figure 3.16. Ce modèle est donc facilement paramétrable par un expert souhaitant contrôler le niveau de bruit des images qu'il génère (en sachant, par exemple, que les tâches engendrant une

perte de connectivité du caractère seront plus susceptibles de mettre à mal des techniques de reconnaissance du texte). Pour plus de détails sur ce modèle de dégradation, merci de se référer à [Kieu 2012b, Kieu 2013b].

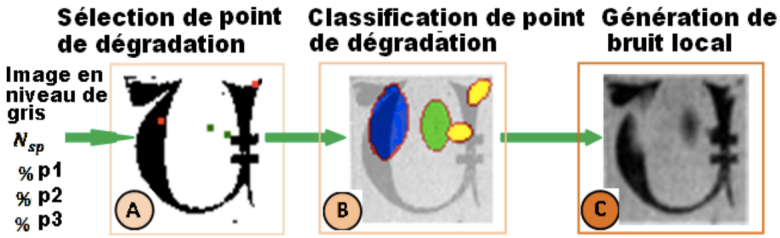


FIGURE 3.16 – Principe du modèle de bruit local appliqué sur des caractères. (A) Sélection aléatoire des points de dégradation (à proximité des caractères) dans l'image binarisée. (B) Selon la localisation du point de dégradation, on l'affecte à l'un des trois types p_1, p_2 ou p_3 (de manière à éviter les trop grosses taches qui rendent le résultat peu réaliste) (C) Génération d'une tache (zone de bruit de forme prédéterminée) en niveaux de gris. Ici, les taches sont des ellipses dont les orientations, tailles et valeurs de pixels sont choisies selon des distributions visant à rendre le résultat visuellement réaliste. *Figure adaptée de [Kieu 2013b].*

Grâce à ces deux méthodes/modèles, on peut générer de grands volumes d'images semi-synthétiques associées à leur vérité-terrain, tout en maîtrisant le type et (dans le cas du modèle de bruit local) le niveau des dégradations présentes dans le document. Ces travaux ont suscité un grand intérêt de la part de la communauté scientifique nationale et internationale en analyse d'images de documents, donnant lieu à plusieurs applications à l'évaluation de performances et au ré-apprentissage de méthodes de traitement ou de reconnaissance d'images. Les principales applications à la reconnaissance d'images de documents, menées en collaboration avec différents partenaires, sont détaillées ci-après.

3.4.3 Applications

3.4.3.1 Application à l'évaluation de la robustesse

En ce qui concerne l'évaluation de la robustesse de systèmes de reconnaissance d'images de documents, nous avons organisé en collaboration avec le **CVC (Barcelone)** une **compétition internationale pour la détection des lignes de portée** dans des documents musicaux manuscrits [Visani 2013]. Cette compétition s'est tenue à l'occasion d'ICDAR 2013 (conférence internationale de référence dans le domaine de la reconnaissance et de l'analyse d'images de documents) et de son *workshop* satellite GREC, spécialisé dans la reconnaissance de documents graphiques. La détection des lignes de portée est un premier pas nécessaire et crucial vers des tâches telles que la reconnaissance des notes/symboles musicaux, ou du scripteur.

Plus précisément, nous avons évalué la robustesse de huit méthodes proposées par des participants du monde entier sur des images que nous avons générées semi-synthétiquement. Pour cela nous avons, à partir d'une base d'images réelles de documents musicaux modernes manuscrits (base MUSCIMA [23] [Fornés 2012]), généré un ensemble d'images semi-synthétiques plus

23. Disponible sur http://www.cvc.uab.es/cvcmuscima/index_database.html.

ou moins dégradées. Ces images constituent la base ICDAR 2013 [24], qui est composée de 6000 images de documents musicaux entièrement semi-synthétiques. Elle comprend une base d'apprentissage (4000 images) et une base de test (2000 images), les images de test ayant bien sûr été générées à partir d'images réelles (et, le cas échéant, de maillages 3D) non utilisés pour la génération de la base d'apprentissage.

Les images d'apprentissage comme de test sont réparties en 13 sous-ensembles, générés soit en appliquant une seule source de bruit à la fois, soit en combinant les deux. Ces sous-ensembles correspondent à des niveaux de dégradation variables, obtenus en changeant les maillages et en faisant varier le niveau du bruit local (nombre et taille des régions de bruit), comme illustré en Figure 3.17.

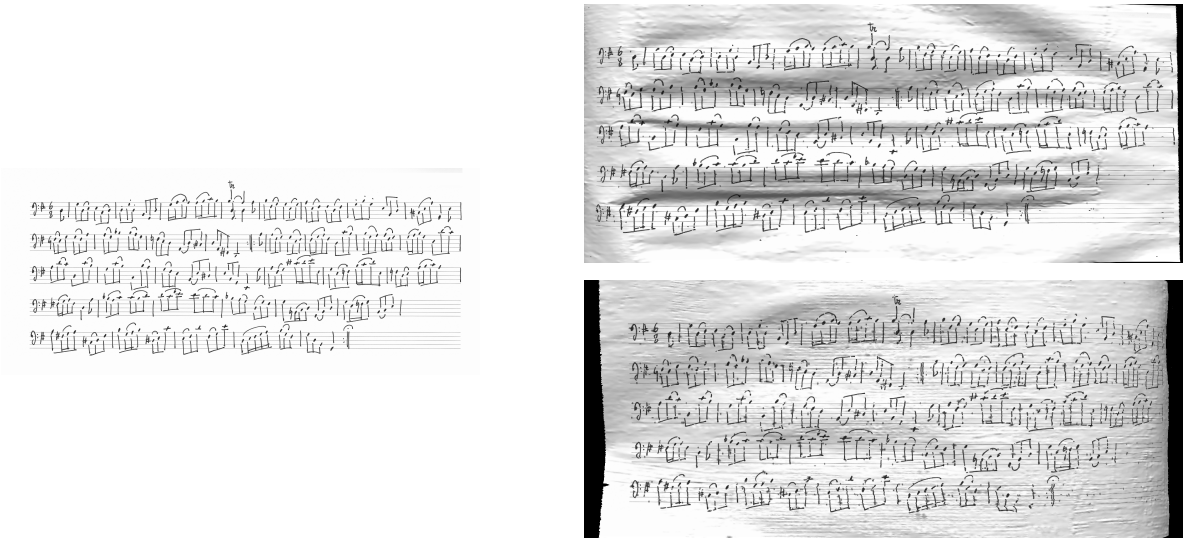


FIGURE 3.17 – De gauche à droite : une image réelle (extraite de la base CVC-MUSCIMA) et deux images semi-synthétiques obtenues en utilisant un niveau croissant de bruit local, et deux maillages 3D différents.

La vérité-terrain (localisation de la portée musicale) associée à chaque image semi-synthétique est facilement déduite à partir de la vérité-terrain des images réelles et (le cas échéant) de la méthode de distorsion 3D, et est disponible pour chaque image de la base.

L'analyse des résultats de la compétition nous a permis d'établir un classement des performances des huit méthodes évaluées vis-à-vis de chacun des deux types de dégradation et de leurs combinaisons. Cette analyse nous a également permis de caractériser leur robustesse vis-à-vis de la nature de la dégradation (bruit local, distorsions 3D locales ou plus globales), et le cas échéant de son niveau (pour le bruit local). Nous avons cherché à expliquer ces différences de comportement, en fonction des caractéristiques des algorithmes évalués. Pour plus de précisions concernant l'analyse de ces résultats, merci de se référer à l'article étendu post-conférence en chapitre de livre [Fornés 2014].

La communauté scientifique est d'ores et déjà en train de s'approprier la base d'images ICDAR 2013, que nous avons rendue publique à l'issue de la compétition. Par exemple, à la confé-

24. Disponible sur http://www.cvc.uab.es/cvcmuscima/competition2013.

rence ICPR qui s'est tenue un an plus tard, deux équipes de chercheurs n'ayant pas participé à la compétition (localisées en Corée [Bui 2014] et au Brésil [Dos Santos Montagner 2014]) ont présenté des articles où ils utilisaient cette base pour comparer les performances de leurs approches de détection de lignes de portée à d'autres approches de l'état de l'art.

3.4.3.2 Applications au ré-apprentissage

En ce qui concerne l'amélioration de l'apprentissage par l'utilisation d'images semi-synthétiques, nous avons mené à bien, en collaboration avec d'autres chercheurs, deux applications touchant au domaine de la reconnaissance d'images de documents anciens.

La première application a été menée en collaboration avec le **LaBRI** dans le cadre du projet ANR DIGIDOC. Malgré le fait que de nombreux algorithmes de reconnaissance de documents reposent encore sur une binarisation préalable, le problème du choix *a priori* de l'algorithme de binarisation le plus adapté à un document donné n'est pas résolu. Dans ce contexte, nous avons cherché à améliorer l'apprentissage d'une **méthode permettant de prédire l'erreur de différents algorithmes de binarisation**. Cette méthode de prédiction, développée au LaBRI et présentée dans [Rabeux 2013], permet de choisir, parmi 11 algorithmes de binarisation de référence, l'algorithme le plus adapté à une image de document donnée. L'une des principales limitations de cette méthode de prédiction est qu'elle repose sur un apprentissage supervisé, dans le cadre applicatif de la binarisation où la vérité-terrain est fastidieuse et coûteuse à produire. À tel point que la base DIBCO 2011 [Pratikakis 2011], très largement utilisée pour l'évaluation des algorithmes de binarisation, comporte seulement 50 pages annotées.

Les résultats expérimentaux, obtenus sur DIBCO 2011 et détaillés dans [Kieu 2014], montrent que l'on peut atteindre 15% de diminution des taux d'erreur de l'algorithme (sur 36 pages) en injectant dans la base d'apprentissage des images semi-synthétiques générées à partir d'images réelles de la base d'apprentissage (initialement composée de 14 pages). Les images semi-synthétiques sont obtenues avec le modèle de bruit local que nous avons introduit dans [Kieu 2012b] et présenté dans la section 3.4.2 ci-avant. **La réduction du taux d'erreur de prédiction est significative statistiquement** [25], alors même que l'enrichissement de la base d'apprentissage se fait sans addition d'information réelle (c'est-à-dire sans utilisation d'autres données réelles que celles de la base d'apprentissage). Cela prouve la pertinence de notre approche et du modèle de bruit local que nous avons proposé. Plus précisément, on constate que l'erreur de prédiction converge à partir du doublement de la taille de la base d'apprentissage.

La deuxième application a été menée en collaboration avec le laboratoire **CENPARMI** de l'université de Concordia pour la **reconnaissance d'écriture manuscrite** dans des documents anciens. Dans notre protocole expérimental détaillé dans [Fischer 2013], différents modèles de dégradation ont été appliqués aux images d'apprentissage afin d'enrichir la base d'apprentissage avec des données semi-synthétiques. Comme précédemment, cet ajout de données d'apprentissage se fait sans addition d'information réelle (c'est-à-dire sans utilisation d'autres données réelles que celles de la base d'apprentissage). Parmi les modèles de dégradation utilisés, on retrouve notre modèle de bruit local, le modèle de bruit local présenté dans [Kanungo 1994] et le

25. On obtient une p-valeur inférieure à 1% avec un test de Student mené de la manière suivante : la méthode de prédiction est d'abord entraînée avec une base de 14 images réelles et testée sur une base réelle de 36 images. En parallèle, la méthode de prédiction est entraînée avec une base de 14 images réelles et 12 images semi-synthétiques générées à partir de ces images réelles, puis testée avec une base réelle de 36 images.

modèle de distorsion géométrique introduit dans [Liang 2008]. Ces sources de dégradation sont appliquées sur des images binaires de lignes de texte afin de générer des exemples d'apprentissage supplémentaires pour entraîner le moteur de reconnaissance d'écriture manuscrite décrit dans [Fischer 2012, Marti 2001]. Ce dernier est basé sur un Modèle de Markov Caché (MMC) avec segmentation implicite (voir annexe C), appliqué au niveau de la ligne de texte.

Sur les bases d'images médiévales de Saint Gall [26] et de Parzival [27], notre analyse montre que les performances en termes de reconnaissance d'écriture peuvent être significativement améliorées par l'ajout d'images semi-synthétiques dans la base d'apprentissage (initialement composée de respectivement 468 et 2237 lignes de texte). Plus précisément, la réduction du taux d'erreur peut atteindre plus de 16% sur la base de Saint Gall, et plus de 20% sur la base de Parzival, **cette diminution étant significative statistiquement** dans les deux cas [28]. Cette réduction est optimale lorsque le bruit local que nous avons proposé est combiné à une autre source de dégradation (plus précisément au modèle de déformation géométrique de [Liang 2008] pour Saint Gall, et au modèle de dégradation local de [Kanungo 1993] pour Parzival), à des niveaux faibles de dégradation. Il semblerait donc que l'injection d'exemples d'apprentissage semi-synthétiques ne peut être bénéfique que si ces exemples ne sont pas trop dégradés. Pour plus de détails, le lecteur est invité à se référer à [Fischer 2013]. Les meilleurs résultats ont été rapportés avec un triplement de la taille de la base d'apprentissage (pour deux sources de dégradation). Bien qu'une analyse plus fine de l'effet du pourcentage d'exemples semi-synthétiques injectés dans la base d'apprentissage reste à mener, en accord avec l'expérimentation précédente sur la binarisation, il semblerait que le ratio optimal se situe autour d'une image semi-synthétique par image réelle et par source de dégradation.

Que ce soit au travers d'applications à la caractérisation de performances ou au ré-apprentissage, nous avons montré dans cette section l'intérêt pratique de notre approche de génération d'images semi-synthétiques. Dans la section suivante, nous allons nous intéresser au problème de l'évaluation des résultats du système de génération.

3.4.4 Évaluation des résultats du système de génération d'images

La plupart des chercheurs cherchent à valider leurs méthodes/modèles de dégradation en se basant sur une **évaluation statistique** de la pertinence de leurs résultats.

Une méthode d'évaluation statistique a été proposée dans [Kanungo 2000] pour la validation d'un modèle de dégradation visant à imiter les bruits binaires locaux (à proximité des caractères) liés au processus de numérisation dans des documents modernes binaires. Elle est basée sur un test d'hypothèse statistique de permutation permettant de comparer de manière non paramétrique les distributions des valeurs de pixels des caractères extraits, d'une part, d'une base dégradée artificiellement et, d'autre part, d'une base réelle. Dans leur cas, ce protocole est relativement aisé à mettre en œuvre. En effet, il s'agit de comparer des populations dont chaque observation (composée d'un caractère réel ou synthétique) est décrite par un vecteur composé uniquement de '0' ou de '1' et peut être considérée comme indépendante statistiquement. Dans notre cas où les images (réelles comme dégradées) sont en niveaux de gris, où les dégradations peuvent avoir un impact plus global sur le document (cas des distorsions 3D), et où de plus nous manquons à la fois de données réelles et de métadonnées concernant leur niveau

26. www.iam.unibe.ch/fki/databases/iam-historical-document-database/saint-gall-database
27. www.iam.unibe.ch/fki/databases/iam-historical-document-database/parzival-database
28. Selon un test de Student avec une erreur de première espèce de 5%.

de dégradation naturelle, ce type de protocoles est largement moins aisé à appliquer. Une autre approche d'évaluation statistique, proposée dans [Li 1996], consiste simplement à montrer que la distribution des taux d'erreur d'un OCR appliqué sur des documents réels est similaire à celle d'un OCR appliqué sur des documents semi-synthétiques avec un niveau de dégradation équivalent.

Dans notre contexte, en raison des difficultés à constituer des bases réelles de niveau de dégradation similaire aux bases que nous générons (du fait du manque de données réelles et de métadonnées concernant leur niveau de dégradation), nous n'avons pas pu appliquer strictement cette méthodologie. Nos études préliminaires montrent cependant que l'évolution des taux d'erreur retournés par un OCR sur des données semi-synthétiques générées avec notre modèle de bruit local est cohérente avec ce que l'on aurait pu en attendre (les taches engendrant une déconnexion des caractères ont un plus fort impact sur les taux d'erreur que les taches connectées aux caractères, tandis que les taches isolées n'ont pratiquement aucun effet). Cependant, ce protocole manque de généricité, en ce sens que ce n'est pas parce qu'un OCR a le même comportement vis-à-vis d'une dégradation réelle et d'une dégradation synthétique qu'il en sera de même d'autres tâches d'analyse d'images. De plus, ce protocole ne permet en aucun cas de valider l'aspect réaliste des dégradations générées, puisqu'on pourrait envisager qu'un défaut qui ne serait absolument pas réaliste visuellement aurait le même impact sur les OCR qu'un défaut qui serait réaliste. Or, en particulier dès lors qu'il s'agit d'évaluation de performances, la communauté scientifique exige un fort niveau de réalisme visuel des bases utilisées.

Nous avons donc préféré nous focaliser prioritairement sur une **évaluation qualitative** (visuelle) de l'aspect réaliste des images semi-synthétiques. Pour cela, nous avons développé un jeu : « réelle ou synthétique ? », où les participants doivent choisir, pour chaque série de 4 images, laquelle de ces images est semi-synthétique. C'est ce que Li *et al.* appelaient dans [Li 1996] le « test de Turing ». Nous avons effectué des tests préliminaires sur tablette lors de la conférence ICDAR 2013, qui ont donné des résultats très encourageants puisque la plupart des participants commettaient de nombreuses erreurs, malgré le fait qu'ils étaient tous des experts de traitement ou d'analyse d'images de documents. Suite à ce premier retour d'expérience, nous sommes en train de finaliser le développement d'une application *web* afin de pouvoir collecter plus de retours et de procéder à une évaluation statistique. Une version β est disponible en ligne [29].

Nous sommes donc en bon chemin pour évaluer nos méthodes/modèles de dégradation de manière qualitative, du point de vue de l'humain (visuellement parlant). Mais, qu'en est-il du point de vue de la machine ? Nous avons d'ores et déjà montré l'efficacité du modèle de bruit local pour diverses applications de ré-apprentissage, ce qui valide dans une certaine mesure son intérêt du point de vue de la machine. En effet, si l'injection de tels exemples a permis d'améliorer l'apprentissage, alors cela signifie que nos données semi-synthétiques sont suffisamment proches des exemples réels (du moins selon ce qu'en perçoit la machine), tout en apportant une certaine variété dans les bases d'apprentissage. Mais, du fait de la perception tronquée qu'a la machine des images (au travers des descripteurs) rien ne nous dit que leur aspect visuellement réaliste ait une quelconque importance pour la machine.

Nos études montrent cependant que l'amélioration des performances après ré-apprentissage est significativement supérieure à celle obtenue avec du bruit « poivre et sel » aléatoirement

29. www.doconcloud.org:8080/DoQuBookWeb/groundtruth/findthefake_game/game_presentation.jsf

rajouté dans les images [Kieu 2014]. Nos expérimentations préliminaires montrent qu'il en est de même lorsque l'on compare nos résultats après ré-apprentissage à ceux obtenus avec du bruit blanc appliqué directement sur les signatures (et non dans les images). Ces travaux sont en train d'être poursuivis et étendus au cas de la méthode de distorsion 3D que nous avons proposée, au travers notamment de la coopération avec le **laboratoire DIVA de l'université de Fribourg** pour la détection de zones/lignes de texte.

3.4.5 Bilan et améliorations possibles

Nous avons proposé une méthode et un modèle de dégradation, dont la principale originalité réside dans le fait qu'ils visent à reproduire, en niveaux de gris, certains des dommages subis par les documents anciens. Grâce à ces méthodes/modèles de dégradation, nous pouvons générer, à partir d'images de documents réels (modernes ou anciens), une grande variété d'images semi-synthétiques. Si les images d'entrée sont accompagnées d'une vérité-terrain, alors nous pouvons retourner en sortie la vérité-terrain associée aux images semi-synthétiques, le cas échéant en prenant en compte les dégradations appliquées. Les images ainsi générées présentent quelques-uns des défauts fréquemment rencontrés dans des documents anciens. Leur aspect visuellement réaliste, et le fait que nous puissions contrôler au moins partiellement la nature (et le cas échéant le niveau) des dégradations présentes dans les images a suscité un grand intérêt de la part de la communauté scientifique, qui s'est associée à nous pour mener conjointement des campagnes d'évaluation de performances et des applications de ré-apprentissage.

En ce sens, ces travaux ont donné lieu à des **contributions plutôt applicatives** dans le contexte de la reconnaissance d'images. Suite à ces travaux, nous avons pu identifier trois pistes d'amélioration principales, détaillées ci-après.

La première piste d'amélioration possible concerne la manière dont un **utilisateur humain peut contrôler la nature et le niveau de distorsion 3D** qu'il souhaite rajouter dans une image donnée. En effet, si le modèle de bruit local est paramétrable en nature et en quantité par un humain, il n'en est pas de même pour notre méthode de distorsion 3D. Dans ce dernier cas, la dégradation dépend de maillages issus de la numérisation 3D de documents réels. Cela rend le résultat visuellement réaliste, et l'humain peut toujours choisir le maillage ou la portion de maillage à appliquer en fonction de la nature de la distorsion désirée, mais il est impossible à l'heure actuelle de contrôler finement le niveau de dégradation. Or, un tel contrôle pourrait être une plus-value non négligeable, à la fois dans des applications de caractérisation de performances et de ré-apprentissage.

Notre première idée a été de déformer les maillages dont on dispose de manière à accentuer (ou à diminuer) les déformations existantes. Nos travaux préliminaires ont montré qu'il est très difficile de le faire automatiquement et de manière réaliste sans un modèle des distorsions 3D que l'on peut retrouver dans un document. Or, un tel modèle n'est pas disponible à l'heure actuelle, et nous pouvons penser qu'il ne pourra le cas échéant être appris qu'au travers d'un grand nombre d'exemples. Mais, il nous est actuellement difficile d'acquérir des grands volumes d'images, vu que nous n'avons qu'un accès limité au scanner 3D. Nous avons donc préféré (au moins dans un premier temps) adopter une démarche plus pragmatique en commençant par proposer des mesures permettant de qualifier plus finement le type, la localisation et le niveau de distorsion présents dans les maillages 3D dont on dispose (évaluation qualitative).

L'idée, à terme, est de pouvoir s'aider de ces travaux pour constituer des bases d'images semi-synthétiques de niveaux de distorsion semblables. Cela nous permettra entre autres de

nous attaquer à des applications de caractérisation systématique de la robustesse d'algorithmes d'analyse d'images vis-à-vis d'un niveau croissant de distorsion 3D, comme nous pouvons d'ores et déjà le faire avec le modèle de bruit local.

Plus généralement, encouragés par le très fort intérêt que la communauté a porté à ces bases, nous souhaitons aller plus loin en mettant à la disposition de la communauté une **plateforme de génération** de documents semi-synthétique complète. L'idée est qu'un utilisateur expert puisse lui-même générer ses propres images semi-synthétiques de documents anciens dont il pourrait contrôler le contenu (texte, fond, fonte, etc.), la structure (physique et logique), et le cas échéant la nature et le niveau des dégradations qu'il souhaite y ajouter. Des modèles de contenu pourraient être appris automatiquement (le cas échéant en interaction avec l'humain) à partir d'exemples de documents réels. L'utilisateur typiquement visé est soit un chercheur en analyse d'images de documents souhaitant caractériser les performances des approches qu'il propose, soit un utilisateur final manipulant couramment des outils d'analyse d'images, comme par exemple des OCR dont il souhaiterait comparer les performances.

En 2013, nous avons déposé auprès de l'ANR un projet « Jeunes Chercheurs » sur ce sujet, avec pour chercheurs impliqués Nicholas Journet (MCF, LaBRI, porteur du projet) et Jean-Philippe Domenger (PR, LaBRI). Ce projet, dans un contexte de très forte sélectivité, n'a pas été retenu suite à ce premier dépôt (bien qu'il ait passé la première phase d'évaluation). Mais, vu les besoins de la communauté et l'appui d'industriels et d'utilisateurs finaux (tels que la BnF par exemple), nous en avons déposé en 2014 une version retravaillée auprès de l'ANR.

La troisième piste d'amélioration concerne l'**utilisation des images semi-synthétiques dans la phase de reconnaissance** proprement dite. En effet, en vue d'améliorer les performances en reconnaissance, nous avons jusqu'à présent privilégié des applications où seule la base d'apprentissage était enrichie avec des exemples semi-synthétiques. Notre but était essentiellement d'évaluer l'intérêt de nos dégradations du point de vue de la machine. Or, il serait certainement possible de tirer un plus large bénéfice des méthodes/modèles de dégradation que nous avons proposé, en utilisant également des versions dégradées des images à reconnaître. En effet, très récemment, certains travaux [Krizhevsky 2012, Paulin 2014] ont montré la pertinence de telles stratégies pour la classification supervisée d'images. Plus précisément, dans [Paulin 2014] par exemple, des descripteurs issus de plusieurs versions transformées des images originales (d'apprentissage comme de test) sont agrégés pour constituer la signature de ces images. Ce sont ces signatures qui sont utilisées lors de la phase de reconnaissance. Une amélioration significative des taux de reconnaissance est rapportée par rapport à un ré-apprentissage seul. Nous pourrions facilement appliquer ce type de stratégies avec notre modèle de dégradation local, puisqu'il est très peu complexe d'un point de vue calculatoire et qu'il peut être paramétré automatiquement. Concernant notre méthode de distorsion 3D, bien plus coûteuse à mettre en œuvre dans son implémentation actuelle, cela nécessiterait un travail considérable d'optimisation. Nous pourrions alternativement nous tourner vers d'autres sources de dégradation, comme par exemple le modèle de transparence de [Moghaddam 2009], ou encore le modèle de distorsion 2D de [Liang 2008], qui ont été ré-implémentés dans le contexte du projet DIGIDOC.

3.4.6 Discussion

L'un des principaux problèmes auxquels nous nous sommes intéressés dans cette section est la caractérisation des performances d'approches de reconnaissance d'images de documents

anciens. Grâce à la proposition de méthodes/modèles de dégradation visuellement réalistes, nous avons généré des bases d'images de niveaux de dégradation variables dont on peut (au moins partiellement) contrôler la nature et/ou le niveau. Cela nous a permis de mettre à la disposition de la communauté des chercheurs des bases qui leur permettent de comparer de manière systématique la robustesse des approches de reconnaissance d'images qu'ils proposent à d'autres approches de la littérature.

Néanmoins, malgré l'intérêt certain des chercheurs pour de telles bases (en particulier dès lors que leur contenu est visuellement réaliste), on peut s'interroger sur le **biais introduit dans l'évaluation** de leurs approches par l'usage de données modifiées synthétiquement. Comment s'assurer que les algorithmes testés se comporteraient de la même manière sur des données réelles que sur ces données synthétiques ? La réponse à cette question ne pourra être apportée sans une évaluation poussée (au moins statistique) de l'adéquation entre les distributions des images générées et celles des images réelles. Comme nous l'avons vu plus haut, il s'agit d'un problème qui n'est pas trivial dans le contexte d'images de documents anciens, dès lors que l'on cherche à s'y attaquer de manière indépendante d'une application spécifique (OCR, binarisation, etc.).

Un autre des principaux problèmes auxquels nous avons cherché à nous attaquer au travers de cette section est la manière de faire bénéficier au mieux le processus de catégorisation des informations contenues dans la base d'apprentissage, lorsque celles-ci sont en nombre limité. Notre proposition est de déduire, à partir de ces informations partielles, de nouvelles informations pour alimenter l'apprentissage. Pour cela, nous avons proposé de générer, à partir des images réelles de la base d'apprentissage, des images semi-synthétiques obtenues par dégradation des images originales, et de les rajouter dans la base d'apprentissage. Nous avons montré au travers de diverses applications que ce « ré-apprentissage » améliore significativement les performances des processus de reconnaissance.

Ces résultats montrent l'efficacité dans la pratique de nos approches de dégradation et, dans une certaine mesure, leur pertinence du point de vue de la machine. Se pose néanmoins la question très délicate d'une **évaluation des méthodes/modèles de dégradation** proposés qui soit non plus expérimentale, mais formelle. Cette question reste à ce jour largement ouverte au sein de la communauté scientifique.

3.5 Conclusion du chapitre

Au travers des travaux présentés dans ce chapitre, nous nous sommes intéressés à la reconnaissance d'images structurées (et plus précisément d'images de documents symboliques) par classification.

Nous avons commencé par nous attaquer au problème de la conception d'une méthode de classification supervisée permettant de restituer les informations de catégorisation apprises par la machine à l'humain. Pour cela, nous avons proposé une approche hybride entre treillis des concepts et arbre de décision. Le choix d'une méthode symbolique est en soi assez original dans le domaine de l'analyse d'images de documents, où les approches numériques sont généralement préférées. Grâce à l'utilisation de diverses stratégies qui ont permis de pallier les principaux désavantages typiques des méthodes symboliques, l'approche proposée permet d'atteindre des performances comparables aux approches numériques.

Cette approche repose sur très peu de paramètres, dont aucun ne dépend directement du domaine d'application. Elle peut donc être transposée avec succès dans une grande variété d'applications d'analyse de données, bien au-delà de la reconnaissance d'images de documents, ce que nous avons vérifié dans la pratique sur divers jeux de données.

Néanmoins, de par l'usage de bas niveau sémantique que nous faisons des treillis, on peut considérer que l'on sous-exploite dans ces travaux la puissance des outils relevant de l'analyse formelle des concepts, notamment en termes de représentation des connaissances, et que ces aspects mériteraient d'être approfondis.

Puis, nous nous sommes intéressés à la manière de faire coopérer classification et segmentation, en présence de données hétérogènes. Plus précisément, on s'est intéressé au cadre applicatif de la reconnaissance d'écriture manuscrite en-ligne, où les informations d'entrée sont de nature et de niveaux sémantiques divers (signal en-ligne, descripteurs de l'image hors-ligne reconstruite à partir de ce signal en-ligne, lexique, etc.). Notre approche repose sur une segmentation semi-explicite des mots en caractères et sur une classification à deux niveaux (caractère et bi-caractère) permettant d'intégrer le contexte des caractères voisins, et de lever les ambigüités liées à la phase de segmentation, lors de la reconnaissance du mot. Cette dernière se fait grâce au décodage de la séquence des caractères/bi-caractères candidats, basé sur un algorithme de programmation dynamique et assisté par un lexique. Au travers d'une discussion fournie, nous avons montré les avantages et les limitations de cette approche originale dans le domaine d'application très spécifique et riche de la reconnaissance d'écriture manuscrite.

Enfin, nous nous sommes attaqués aux difficultés rencontrées dans la pratique lorsque les données annotées sont en nombre insuffisant pour mener à bien l'apprentissage et/ou la caractérisation des performances. Au-delà du cadre applicatif des documents anciens ou rares traités dans ce manuscrit, c'est un cas qui se présente fréquemment dans le domaine de la reconnaissance d'images de documents, où l'annotation est souvent associée à des problèmes de coût et/ou de fiabilité. Nous avons choisi, à partir des images annotées disponibles, d'en générer des versions semi-synthétiques, mais néanmoins d'apparence réaliste, permettant d'enrichir des bases d'apprentissage, de validation et/ou de test. Si nous avons pu montrer l'intérêt de ce choix dans la pratique, une validation plus formelle des méthodes/modèles utilisés reste un problème largement ouvert. Vu en particulier l'intérêt suscité par ces travaux au sein de la communauté scientifique, on peut considérer qu'ils mériteraient d'être poursuivis et étendus.

3.6 Perspectives

Dans ce chapitre, nous nous sommes focalisés sur la reconnaissance d'images de documents symboliques (graphiques et/ou textuels) par classification supervisée. La distinction historiquement faite par les chercheurs entre images de documents montrant des scènes naturelles (images naturelles) et images de documents symboliques (images de documents) leur sert essentiellement à des fins pratiques, notamment pour travailler dans un cadre applicatif fixé, où les objectifs attendus sont plus simples à formuler, et où des modèles de connaissance spécifiques peuvent permettre, le cas échéant, d'orienter l'analyse. Cette distinction historique tend néanmoins à s'éroder [30].

Dans mes travaux futurs, je souhaiterais généraliser mes recherches concernant la reconnaissance d'images au cas d'images naturelles. Ce qui ne signifie pas abandonner mes travaux concernant l'analyse d'images de documents. Au contraire, il s'agit plutôt de trouver des **thèmes de recherche fédérateurs** qui pourraient être mis à profit à la fois pour des images de documents, et pour des images naturelles.

Les avancées récentes dans la reconnaissance d'images de documents ont été évoquées au fil de ce chapitre, notamment au travers des différents cas d'étude que nous avons exploré. Il convient maintenant de faire un point rapide sur l'état des recherches actuelles concernant la reconnaissance d'images naturelles.

Les progrès récents dans le domaine de l'apprentissage artificiel, favorisés par la mise à disposition de bases d'images annotées gigantesques permettant de mener à bien l'entraînement des méthodes ou des modèles, ont engendré une véritable rupture dans les performances atteignables pour la reconnaissance d'objets dans des images. On peut citer à titre d'exemple la base ImageNet [Deng 2009] composée de 14 millions d'images en 2011, et organisée selon la structure sémantique de WordNet [Fellbaum 1998]. Si l'on prend l'exemple du concours de classification d'images ImageNet organisé à partir de cette base, le taux d'erreur (avec 5 prédictions par image) est passé de 15% en 2012 à 11% en 2013, pour atteindre moins de 7% en 2014. Ainsi, si le problème de la reconnaissance d'objets « à plat » dans des images n'est pas encore complètement résolu, la communauté scientifique est aujourd'hui dotée de moyens de calculs de plus en plus performants (avec notamment l'avènement des processeurs graphiques) et d'un ensemble de solutions efficaces (par exemple basées sur l'apprentissage profond) pour le mener à bien de manière relativement satisfaisante, généralement sous réserve que le nombre de données annotées disponibles pour l'apprentissage soit très important.

Une large part de l'effort de recherche dans le domaine de la reconnaissance d'images naturelles est donc en train de se tourner vers (liste non exhaustive) :

– La reconnaissance d'objets « de grain fin » (*fine-grained object recognition*), où il ne s'agit plus simplement de distinguer un chien d'un oiseau ou une voiture d'un vélo, mais de reconnaître les sous-catégories de ces objets, comme par exemple l'espèce d'oiseau, la race de chien ou encore le modèle de voiture, en se basant le cas échéant sur la détection de parties de l'image permettant de distinguer ces sous-catégories [Yao 2012] ;

30. Par exemple, puisque l'information que l'on peut tirer du texte présent dans une scène naturelle donnée est souvent à forte valeur ajoutée pour la compréhension de la scène, des chercheurs spécialisés dans le traitement et l'analyse d'images de scènes naturelles se sont intéressés à y détecter/reconnaître du texte [Wolf 2004, Minetto 2014]. Inversement, les chercheurs spécialisés en analyse de documents symboliques s'inspirent largement des techniques initialement proposées pour l'analyse d'images naturelles.

- La reconnaissance de classes inconnues jusque-là (*zero-shot classification*), où il s'agit de transposer les méthodes/modèles/informations appris sur un ensemble de classes connues, à de nouvelles classes. On parle parfois de « transfert des connaissances » apprises sur un ensemble de classes données vers un nouvel ensemble de classes [Rohrbach 2011] ;
- La classification multi-étiquettes, où les classes ne sont pas mutuellement exclusives (en d'autres termes, une même image peut appartenir à plusieurs classes) [Boutell 2004]. Dans les images naturelles, une application-phare est celle de l'annotation d'images avec des attributs (aussi appelés *tags*). La principale motivation de ces travaux est de chercher à réduire le fossé sémantique entre le contenu de bas niveau de l'image et la compréhension que peut en avoir un humain, au travers de l'annotation des images par un ensemble d'attributs (souvent sous la forme de mots-clés tels que : « scène d'extérieur », « mer », « lac », « personne », etc.). Les usages ultérieurs des attributs sont nombreux. D'une part, la connaissance des attributs d'images peut aider d'autres processus de reconnaissance d'images, voire servir à une interprétation de plus haut niveau sémantique de la scène. D'autre part, les bases d'images ainsi annotées pourront par la suite être requêtées à l'aide de ces mots-clés, ce qui est souvent plus facile et intuitif pour un humain qu'une recherche par l'exemple visuel traditionnellement envisagée en CBIR.

Comme nous venons de l'illustrer au travers de nos travaux sur la génération d'images semi-synthétiques, les besoins en termes d'outils d'annotation sont tout aussi nombreux dans le domaine de l'analyse d'images de documents que pour les images naturelles. C'est donc sur ce thème de la **classification multi-étiquette** pour l'annotation d'images que je souhaiterais poursuive mes travaux concernant la reconnaissance d'images.

L'objectif de l'annotation d'images est de prédire les attributs pertinents, étant donnée une image et un dictionnaire des attributs d'images plus ou moins flexible. La plupart des travaux existants se focalisent sur des tâches d'annotation menées, soit de manière manuelle (où la flexibilité du dictionnaire et la qualité des annotations peuvent être variables[31]), soit de manière complètement automatique [Zhang 2012]. Les approches les plus traditionnelles d'annotation automatique se font grâce à un apprentissage supervisé mené indépendamment pour chacun des attributs [Grangier 2008, Guillaumin 2009]. D'autres approches se basent sur l'étude des corrélations entre attributs à l'intérieur des images, éventuellement analysées au travers des multiples instances (p. ex. des régions) qui les composent [Zhou 2012].

Dans le suite de mes travaux, je souhaite plus précisément m'intéresser à la question moins explorée de l'**annotation semi-automatique** des images, par interaction avec l'utilisateur. Cette interaction se ferait durant la phase d'annotation d'une nouvelle image, et non lors de la phase d'apprentissage comme c'était le cas dans les travaux décrits au chapitre 2. Les applications finales relèvent de l'assistance à l'annotation manuelle, à la différence de la plupart des approches qui visent à obtenir une annotation complètement automatique. Les utilisateurs visés sont typiquement des archivistes.

À noter que les attributs obtenus par annotation pourraient, à terme, être réutilisés comme descripteurs en entrée d'une catégorisation d'images, ou pour l'interprétation d'images.

31. On peut citer comme illustration la banque d'images GettyImages `http://www.gettyimages.com/` dont le dictionnaire et la qualité des annotations sont contrôlées, et le site communautaire de partage de photographies Flickr `https://www.flickr.com/` où la qualité et l'aspect informatif des annotations sont parfois limités.

Durant son doctorat, Thomas Mensik, qui a reçu le prix de thèse AFRIF en 2012, a montré que le fait de mettre en œuvre l'annotation d'images naturelles de manière interactive permettait d'en améliorer sensiblement les performances [Mensink 2012]. Plus précisément, l'approche proposée dans cette thèse repose sur l'apprentissage de modèles structurés qui permettent de prédire les attributs des images en fonction de leur contenu [32], en prenant en compte explicitement les dépendances observées entre les attributs dans les images de la collection. Cet apprentissage est effectué hors-ligne, à partir d'une base d'images préalablement annotées. Dans la phase d'annotation proprement dite (en-ligne avec l'utilisateur), ces modèles permettent d'assigner à chaque image un ensemble d'attributs, avec des scores de confiance associés. Le retour de l'utilisateur sur la pertinence (ou non) de certains de ces attributs peut être propagé grâce à la structure, afin de modifier les scores d'autres attributs. La structure permet donc de mieux tirer parti des retours de l'utilisateur que dans un scénario interactif où l'apprentissage aurait été mis en œuvre indépendamment pour chacun des attributs. L'amélioration de la précision est substantielle par rapport aux approches automatiques, même avec peu de retours de l'utilisateur. Les attributs à présenter à l'utilisateur sont sélectionnés grâce au modèle, de manière à ce que leur annotation minimise l'incertitude associée aux autres étiquettes.

Outre l'intérêt de l'interactivité pour mettre en œuvre l'annotation de manière semi-automatique, ces travaux ont montré la pertinence de s'appuyer sur des structures modélisant de manière explicite la dépendance entre attributs. Ils ont montré de plus l'**importance de la topologie de telles structures**. Les auteurs ont dans un premier temps choisi d'utiliser des structures arborescentes dont chaque nœud représente un attribut, et les arcs représentent la dépendance entre attributs, sous la forme de modèles CRF (*Conditional Random Fields*). Étant donné que, dans la pratique, chaque attribut dépend de tous les autres, il est quasiment impossible de trouver la structure arborescente optimale. Ils ont expérimenté deux méthodes permettant de l'approximer, avec des résultats mitigés. Après avoir envisagé différentes solutions, c'est finalement un modèle basé sur un mélange d'arbres qui leur a permis de représenter de la manière la plus exhaustive possible les dépendances entre attributs, et donc d'obtenir les meilleurs résultats. Chaque arbre dans le mélange est appris indépendamment, et leurs sorties sont moyennées pour obtenir les annotations finales. Un peu à la manière des forêts aléatoires, l'aspect explicite de la méthode est donc perdu au profit de ses performances. Un désavantage de l'approche est que le fait que chaque arbre soit appris indépendamment peut *a priori* engendrer un biais dans les dépendances entre attributs prises en compte dans le mélange final.

Notre proposition est de chercher un modèle basé sur une structure de treillis (le cas échéant simplifiée) qui nous permettrait de prendre en compte de manière explicite un plus grand nombre de dépendances entre attributs qu'une structure d'arbre. Cette prise en compte explicite permettrait de réduire les possibles biais liés à l'utilisation d'un mélange d'arbres, et d'obtenir une représentation explicite des dépendances entre attributs.

Il y a essentiellement deux manières de s'attaquer à ce problème. La première consiste à chercher un modèle graphique adapté à l'inférence d'une telle structure, construite à partir des descripteurs d'images (ou d'instances des images), et de leurs annotations. Vu la complexité que l'on peut attendre d'une telle structure, la recherche d'une méthode d'inférence exacte ne nous semble pas réalisable. Si nous choisissons cette première solution, alors il nous faudra donc trouver une solution approximative qui soit suffisamment exhaustive pour décrire au mieux les

32. Le contenu des images étant décrit au travers de vecteurs de Fisher [Perronnin 2010].

dépendances entre attributs, ce qui est loin d'être aisé. Une deuxième manière de s'attaquer à ce problème est de chercher à définir, à partir d'un ensemble d'arbres approximés de l'inférence exacte de CRFs, un treillis (ou du moins une version simplifiée d'un treillis) permettant de « fusionner [33] » ces arbres. Quelle que soit l'approche adoptée, on est confronté à un problème scientifique difficile à résoudre.

Un autre problème auquel nous souhaiterions nous attaquer dans ce contexte concerne le cas où, en plus des données annotées, on dispose pour alimenter l'apprentissage d'images non annotées (ou partiellement annotées). Le problème est alors de tirer parti de la distribution de ces dernières (et/ou de leurs instances) dans l'espace de représentation de bas niveau sémantique pour enrichir la modélisation des dépendances entre attributs, et au final les résultats de l'annotation. Il s'agit d'un problème d'apprentissage semi-supervisé.

À ces problèmes scientifiques s'ajoutent des **questions de recherche** non moins complexes, parmi lesquelles on peut énoncer les suivantes :

- Est-il possible, théoriquement et d'un point de vue pratique, de faire évoluer un tel modèle de prédiction des attributs en fonction des retours de l'utilisateur (cette évolution se faisant hors-ligne), tout en maintenant l'interaction lors de la phase d'annotation (en-ligne) ?
- Jusqu'à quel point est-il possible de réduire la part des images annotées parmi les images utilisées dans l'apprentissage, tout en conservant des performances satisfaisantes ?
- Est-il possible de sélectionner les attributs à soumettre à l'utilisateur à l'aide d'un apprentissage actif qui viserait, non pas à minimiser l'incertitude associée aux autres attributs dans la collection, mais directement à maximiser les performances du modèle en termes d'annotation ?
- Est-il envisageable d'adapter une telle approche à un dictionnaire flexible (ce qui pourrait permettre à l'utilisateur de suggérer ses propres mots-clés hors du dictionnaire par exemple) ?
- Le cas échéant, comment gérer la variabilité dans la qualité des annotations (dans les exemples utilisés pour l'apprentissage et parmi les retours de l'utilisateur) ?

S'agissant de perspectives de recherche abordant des problèmes complexes liés en particulier à la théorie des probabilités, à la modélisation, à l'apprentissage et aux interactions homme-machine, elles ne pourront être menées qu'à moyen, voire à long terme.

Certains de ces problèmes pourront être attaqués d'un point de vue applicatif au travers du projet ARCHIVES, qui va débuter en 2015 pour une durée de 42 mois, et grâce auquel nous allons pouvoir numériser des centaines de documents retraçant l'histoire du Fleuve Rouge au travers d'une très grande diversité de photographies, de cartes et de documents textuels.

33. Nous avons déjà travaillé sur les liens structurels entre arbres et treillis (voir section 3.2), mais ces études portaient spécifiquement sur les treillis des concepts, où les arcs ne sont pas valués, alors qu'ici le problème est de fusionner des arbres valués. En dehors de ces considérations, le lien de fusion que nous avons démontré ne concerne que des jeux d'attributs mutuellement exclusifs (produisant des treillis « dichotomiques »), comme expliqué en page 80. Or, ici, chaque attribut a une dépendance avec l'autre. Nos travaux précédents ne peuvent donc pas s'appliquer ici.

Les historiens de l'École Française d'Extrême-Orient, qui souhaitent exploiter le contenu de ces documents pour leurs recherches, sont en effet très intéressés par la possibilité d'obtenir une assistance dans la lourde tâche d'annoter ces images. En outre, nous sommes en train d'initier des discussions avec la Bibliothèque nationale de France, par ailleurs notre partenaire dans le cadre du projet DIGIDOC, qui pourrait également être intéressée, en tant qu'utilisateur final, par les approches développées.

3.7 Faits marquants liés à ces contributions

3.7.1 Synthèse des faits marquants

Section	Type de contribution majoritaire	Encadrement	Projets	Collaborations (co-publication ou co-encadrement)	Publications
Section 3.2	Méthodologique	1 thèse [TH-Gira13] 1 stage [ST-Bacc10]	– –	– –	1 thèse [Girard 2013] 1 revue internationale [Visani 2011a] 2 revues fr. [Coustaty 2010, Bertet 2009] 3 conf. internationales [Guillas 2009] [Girard 2011a, Girard 2011b] 2 conf. francophones [Bertet 2008, Girard 2009]
Section 3.3	Méthodologique et applicatif	1 thèse [TH-Prum13] 3 stages [ST-Ayadi12] [ST-Dao11] [ST-Bui10]	Projet Eurêka RECONOMAD	CENPARMI, Montréal	1 thèse [Prum 2013a] 5 conf. internationales [Prum 2010a, Prum 2010b] [Bui 2011, Visani 2012] [Prum 2013b]
Section 3.4	Méthodologique et applicatif	1 thèse [TH-KieuXX] 2 Stages [ST-Tran14] [ST-Bui11]	Projet ANR DIGIDOC	LaBRI, Bordeaux CENPARMI, Montréal CVC, Barcelone	1 chapitre de livre [Fornés 2014] 5 conf. internationales [Kieu 2012b, Visani 2013, Kieu 2013a] [Kieu 2013b, Fischer 2013] 2 conf. fr. [Kieu 2012a, Kieu 2014]

TABLE 3.2 – Faits marquants liés aux contributions présentées dans le chapitre 3.

3.7.2 Encadrements en lien avec ces contributions [34]

[TH-Gira13] N. Girard *Vers une approche hybride mêlant arbre de classification et treillis des concepts pour l'indexation d'images*. Thèse de doctorat co-supervisée par moi-même et Karell Bertet, directrice de thèse, L3i/université de La Rochelle. Thèse soutenue à l'université de La Rochelle le 5 Juillet 2013. Manuscrit en français [Girard 2013].

[TH-Prum13] S. Prum *Vers une approche discriminante pour la reconnaissance de mots manuscrits en-ligne utilisant des modèles de bi-caractères*. Thèse de doctorat supervisée par moi-même avec Jean-Marc Ogier comme directeur de thèse, L3i/université de La Rochelle. Thèse soutenue à l'université de La Rochelle le 8 Novembre 2013. Manuscrit en anglais [Prum 2013a].

[TH-KieuXX] Kieu V.C. *Génération d'images semi-synthétiques de documents anciens pour la caractérisation de performances et le ré-apprentissage de méthodes de traitement et d'analyse*. Thèse de doctorat co-supervisée par moi-même, Nicholas Journet et Jean-Philippe Domenger (LaBRI, Bordeaux). Directeur de thèse : Rémy Mullot, L3i/université de La Rochelle. Thèse commencée en Novembre 2011,en fin de rédaction.

[ST-Tran14] Tran V.D. Stage de Master 2. *Distortion level evaluation for semi-synthetic ancient images*. 2014.

[ST-Ayadi12] M. Ayadi. Stage de Master 1. *Techniques d'adaptation de moteurs de reconnaissance d'écriture vis-à-vis du scripteur*. 2012.

[ST-Bui11] Bui T.M.A. Stage de Master 2. *Modèle de dégradation d'images de documents anciens pour la génération de données synthétiques et analyse de complexité des bases d'images*. 2011.

[ST-Dao11] Dao N.B. Stage de Master 1. *Création automatique de profils d'écriture*. 2011.

[ST-Bui10] Bui Q.A. Stage de Master 2. *Identification du scripteur pour la reconnaissance d'écriture manuscrite*. 2010.

[ST-Bacc10] G. Le Baccon. Stage de Master 1. *Étude et implémentation de méthodes de construction d'arbres de classification*. 2010.

34. Pour plus de détails sur ces encadrements, merci de se référer à la section 2.3 du Tome I.

Chapitre 4

Conclusion et perspectives

Vers l'exploitation interactive de collections d'images

Ce manuscrit est organisé autour des deux grands objectifs auxquels nous nous intéressons : la description et la reconnaissance d'images. Dans les deux cas, les approches que nous avons déployées reposent sur une catégorisation des images d'entrée, ou de motifs de ces images. Cette distinction entre description et reconnaissance fait sens notamment du point de vue de l'usage ultérieur qui sera fait des catégories apprises, et de l'information dont on dispose en entrée du processus.

Du point de vue de l'usage ultérieur de l'information apprise *via* la catégorisation, dans le cas de la description, la catégorisation est un moyen utile pour synthétiser, résumer et/ou organiser les images ou les collections d'images en des groupes d'images similaires selon certains critères (en général d'assez bas niveau sémantique). Tandis que, dans le cas de la reconnaissance, on recherche des catégories qui sont souvent pré-établies et de plus haut niveau sémantique, pouvant *in fine* servir plus directement à une compréhension du contenu des images ou des collections d'images.

Dans la pratique, cela se concrétise par des différences dans la nature et la quantité des informations dont on dispose en entrée du processus de catégorisation. Dans les deux cas, on bénéficie d'une information de bas niveau sémantique sous la forme de descripteurs visuels automatiquement extraits des images, éventuellement additionnée de quelques informations du domaine.

Dans le cas de la description, la catégorisation doit généralement être menée sans connaître *a priori* les catégories perçues par un humain dans la collection. Nous avons donc choisi, après avoir confié à la machine la tâche de découvrir des groupes d'images similaires selon leur description de bas niveau sémantique, de faire interagir l'utilisateur avec ces groupes d'images de manière à guider un rapprochement de ces groupes vers les concepts de plus haut niveau qu'il perçoit dans la collection.

Dans le cas de la reconnaissance au contraire, on dispose d'emblée d'informations plus complètes concernant les catégories perçues par un humain dans la collection, puisque certaines des images la composant ont été préalablement annotées par (ou sous la supervision de) l'humain. Dans ce contexte, une part importante de notre effort a porté sur la recherche de manières adéquates d'intégrer dans le processus de reconnaissance l'ensemble des informations disponibles, ces dernières étant parfois de natures hétérogènes et, dans certains cas, très partielles.

Tout au long de ces travaux de recherche, nous avons conservé une même ligne conductrice. Nous nous sommes attelés à réduire la complexité pour la machine d'extraire et de restituer,

essentiellement à partir de représentations de bas niveau sémantique des données et en l'absence de connaissance du domaine formalisée explicitement, des informations qui font sens pour l'humain. D'un point de vue scientifique, nous avons choisi pour cela de privilégier les approches d'apprentissage, en nous appuyant sur des notions issues notamment des domaines des statistiques, de la reconnaissance de formes, du traitement du signal et de l'intelligence artificielle.

Ces approches ont abouti à des résultats qui ont été valorisés aussi bien d'un point de vue scientifique que d'un point de vue applicatif, au travers de cas d'étude concrets souvent en lien avec une activité contractuelle.

Si on les replace dans le cadre plus large de l'analyse d'image, on peut considérer que l'ensemble des travaux décrits dans ce manuscrit visent à indexer (au sens le plus large du terme) les images selon des critères qui font sens pour l'humain. Cette indexation a pour but final de faciliter l'exploitation du contenu des images par l'humain (que ce soit pour des opérations de visualisation/navigation dans la base d'images ou pour rechercher une information donnée dans une image de document par exemple). L'ensemble des types d'images plus ou moins structurées considérés au fil de ce manuscrit (scènes naturelles ou documents symboliques) se conforment à la définition d'images de documents, le terme « document » étant entendu sous son acception la plus large : « tout objet servant à conserver ou à transmettre de l'information tangible pour l'être humain » [Ingold 2002]. D'un point de vue très général, on peut donc considérer que l'ensemble des travaux présentés dans ce manuscrit visent à concevoir des outils pouvant servir à indexer des images de documents de niveaux de structuration variables, en vue d'une exploitation ultérieure de ces documents par l'humain.

Et les besoins sont nombreux. Dans un contexte de transition numérique et dans une recherche d'efficacité, de nombreuses entreprises et administrations se tournent vers la dématérialisation de leurs documents. Selon les prévisions de Xerfi [1], la croissance du marché de la dématérialisation en France devrait se situer autour des 8% en valeur par an jusqu'en 2017. Dans un souci de préservation, de sauvegarde et/ou de diffusion du patrimoine documentaire, de nombreux centres d'archives ou bibliothèques se lancent dans des campagnes de numérisation massives. On peut citer l'exemple du projet Gallica mené par la Bibliothèque nationale de France (BnF), qui a débouché sur la mise en ligne de plus de deux millions de documents.

Derrière ces grandes tendances se cachent une multitude d'initiatives plus ciblées sur des besoins spécifiques. Citons l'exemple du projet ARCHIVES qui va débuter l'année prochaine. Des centaines de documents, gardés aux archives nationales d'Hanoï, retracent l'histoire du Fleuve Rouge au travers de photographies, de cartes et de documents textuels. À partir de ce corpus de documents de natures très diverses, les historiens de l'École Française d'Extrême Orient cherchent à retracer une chronologie des événements survenus sur le Fleuve Rouge. La question est alors : une fois dématérialisés, que faire de ces documents pour assister l'historien dans sa tâche ? Vu leur diversité et leur niveau de dégradation, une indexation basée sur une reconnaissance systématique de leur contenu ne semble pas envisageable à l'aide des outils actuels. Elle ne correspond d'ailleurs pas forcément à un besoin des utilisateurs finaux (ici les historiens). Ces derniers souhaitent en effet surtout avoir accès à des outils qui leur permettent de naviguer dans la collection de documents selon divers critères (temporels ou géographiques par exemple), afin de retrouver par eux-mêmes l'information qu'ils recherchent. En d'autres termes, ils ont **besoin d'un outil permettant de les assister dans l'exploitation** qu'ils

1. Acteur majeur dans le domaine des études économiques sectorielles : http://www.xerfi.com.

feront de la collection de documents. Au travers de cet exemple applicatif concret, c'est toute une variété de chercheurs dans des disciplines telles que l'histoire, la géographie ou les lettres qui partagent finalement des besoins similaires.

Portée par ces besoins, la direction de recherche que je souhaite privilégier dans les années à venir consiste à me pencher sur les différents moyens pour amener la machine à assister un utilisateur final dans l'exploitation de corpus d'images de documents de contenus variés, en accordant un rôle-clé à l'utilisateur final du système.

En lignes conductrices de ces travaux futurs, on retrouve bien évidemment les deux principales **perspectives qui se sont dégagées à l'issue des deux chapitres précédents**, et certains des aspects évoqués dans les discussions agrémentant ces chapitres. Il convient ici de les hiérarchiser et de les replacer dans cette optique d'exploitation de corpus d'images variés.

Les deux perspectives de recherche que je souhaite privilégier dans les années à venir concernent :

- **L'apprentissage de descripteurs structurés**[2] sémantiques, c'est-à-dire plus conformes aux concepts perçus par l'utilisateur dans les images que la représentation de bas niveau sémantique que peut en faire automatiquement la machine à partir des données pixellaires. Il s'agit d'apprendre un modèle capable de hiérarchiser, à partir d'informations partielles données interactivement par l'utilisateur et de critères spatiaux dans l'image, un ensemble de descripteurs visuels locaux extraits automatiquement des images. Nous nous appuierons pour cela sur notre expérience acquise au travers de nos travaux sur la catégorisation interactive semi-supervisée, tout en cherchant cette fois-ci à répercuter les interventions de l'utilisateur à un plus bas niveau que celui des catégories, à savoir celui des descripteurs. La hiérarchie de descripteurs ainsi apprise pourrait être réutilisée pour des images de contenus variés dans de multiples applications à l'exploitation d'images : navigation dans la collection de documents, recherche d'images, annotation automatique ou semi-automatique, etc.

- **La classification multi-étiquettes**[3] (et le cas échéant multi-instances) d'images. Les groupes d'images perçus par l'humain ne sont pas forcément de frontières perméables et donc il convient d'être capable d'associer une même image à de multiples attributs (mots-clés), plus conformes à la diversité de ce que perçoit l'utilisateur dans l'image. L'apprentissage se ferait cette fois-ci sans intervention en-ligne de l'utilisateur, à partir d'une collection d'images (au moins partiellement) préalablement annotées par leurs attributs. Nous proposons pour cela d'adopter une approche basée sur un modèle structuré permettant de prédire, à partir de descripteurs de l'image et de la distribution des attributs dans les images de la collection, les attributs à associer à une nouvelle image. Nous pourrons pour cela nous appuyer, d'une part, sur l'expérience acquise au travers de nos travaux sur les structures hiérarchiques dérivées de treillis (pour chercher à les étendre notamment à la modélisation de la dépendance entre attributs) et, d'autre part, sur notre connaissance de l'apprentissage supervisé et semi-supervisé. L'application visée est l'annotation semi-automatique du contenu des images avec des mots-clés. L'utilisateur serait sollicité lors de la phase d'annotation afin de valider (ou d'infirmer)

2. Voir section 2.5 : « Perspectives ».
3. Voir section 3.6 : « Perspectives ».

certains des attributs inférés pour une image donnée par le système, ces retours ciblés permettant d'améliorer la précision de l'annotation en contrepartie d'un faible effort de l'utilisateur.

L'exploitation de ces pistes de recherche nécessitera notamment d'étendre et de mieux formaliser nos travaux visant à déduire, à partir d'une information partielle obtenue grâce à un humain (par le biais d'une vérité-terrain ou par interaction), une information supplémentaire pouvant améliorer les performances du système au final. Cela passe par une **validation plus formelle de la qualité de l'information déduite**.

Au travers de ces perspectives, se posent notamment les **questions de recherche** suivantes :

- Comment faire en sorte de guider la machine vers les résultats attendus par l'utilisateur, en s'appuyant efficacement, non seulement sur les informations en provenance de l'humain (contenues dans la vérité-terrain et/ou obtenues par interaction avec l'utilisateur), mais aussi sur les descripteurs des images (parfois nombreuses) pour lesquelles on ne dispose pas de ces informations ? Comment évaluer les résultats obtenus en tenant compte conjointement de ces deux niveaux d'analyse ?

- Dans quelle mesure est-il possible de s'aider de ces informations en provenance de l'humain pour construire/alimenter des modèles de connaissance permettant à leur tour d'améliorer l'apprentissage ?

- Jusqu'à quel point peut-on efficacement guider la machine vers des résultats plus en adéquation avec les concepts perçus dans les images par l'utilisateur, tout en offrant à ce dernier une certaine flexibilité dans l'usage du système ? Le cas échéant, comment gérer la variabilité dans la qualité des informations en provenance de l'humain ?

- Dans quelle mesure est-il possible/souhaitable de chercher à « faire bénéficier » le système conjointement des interactions de multiples utilisateurs (historiens et géographes par exemple) percevant parfois des concepts différents dans les images ?

L'ensemble de ces perspectives, et les questions qui en découlent, constitueront le fil rouge de mes activités de recherche dans les années à venir.

Références bibliographiques

[Agrawal 2007] R. Agrawal, W. Changhua, W.I. Grosky et F. Fotouhi. *Image Clustering Using Visual and Text Keywords*. In proceedings of the International Symposium on Computational Intelligence in Robotics and Automation (CIRA), pages 49–54, 2007. (Cité en page 33.)

[Ahmad 2009] A.R. Ahmad, C. Viard-Gaudin et M. Khalid. *Lexicon-Based Word Recognition Using Support Vector Machine and Hidden Markov Model*. In proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 161–165, 2009. (Cité en pages 97 et 170.)

[Ai 2013] L.F. Ai, J.Q. Yu, Y.F. He et T. Guan. *High-dimensional indexing technologies for large scale content-based image retrieval : a review*. Journal of Zhejiang University Science C, vol. 14, no. 7, pages 505–520, 2013. (Cité en page 30.)

[Alpaydin 2004] E. Alpaydin. Introduction to machine learning (adaptive computation and machine learning). MIT Press, 539 pages, 2004. (Cité en page 2.)

[André 1990] J. André et V. Quint. Le document électronique, chapitre Structures et modèles de documents, pages 3–60. INRIA (ed.), 1990. (Cité en pages 11 et 68.)

[Anquetil 2008] E. Anquetil. *Reconnaissance d'écriture manuscrite et interaction homme-document*. Habilitation à diriger des recherches, Université de Rennes I, 125 pages, 2008. (Cité en pages 92, 96, 169 et 170.)

[Arrivault 2005] D. Arrivault, N. Richard, C. Fernandez-Maloigne et P. Bouyer. *Collaboration Between Statistical and Structural Approaches for Old Handwritten Characters Recognition*. In proceedings of the IAPR international workshop on Graph based Representation in pattern recognition (GbR), pages 291–300, 2005. (Cité en page 153.)

[Atif 2014] J. Atif, C. Hudelot et I. Bloch. *Explanatory reasoning for image understanding using formal concept analysis and description logics*. IEEE Transactions on Systems, Man and Cybernetics : Systems, vol. 44, no. 5, pages 552–570, 2014. (Cité en pages 72 et 88.)

[Avila 2013] S. Avila, N. Thome, M. Cord, E. Valle et A.A. Araújo. *Pooling in image representation : The visual codeword point of view*. Computer Vision and Image Understanding, vol. 117, no. 5, pages 453–465, 2013. (Cité en pages 27 et 61.)

[Bahlmann 2002] C. Bahlmann, B. Haasdonk et H. Burkhardt. *Online handwriting recognition with support vector machines - a kernel approach*. In proceedings of the International Conference on Frontiers in Handwriting Recognition (ICFHR), pages 49–54, 2002. (Cité en page 97.)

[Baird 1993] H.S Baird. *Document image defect models and their uses*. In proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 62–67, 1993. (Cité en page 109.)

[Basu 2004] S. Basu, M. Bilenko et R.J. Mooney. *A probabilistic framework for semi-supervised clustering*. In proceedings of the ACM SIGKDD international conference on Knowledge Discovery and Data mining, pages 59–68, 2004. (Cité en pages 33, 40, 41 et 167.)

[Basu 2008] S. Basu, I. Davidson et K. Wagstaff. Constrained clustering : Advances in algorithms, theory, and applications. Chapman & Hall/CRC, 472 pages, 2008. (Cité en page 34.)

[Belhumeur 1997] P.N. Belhumeur, J.P. Hespanha et D.J. Kriegman. *Eigenfaces vs Fisherfaces : Recognition Using Class Specific Linear Projection*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pages 711–720, 1997. (Cité en page 155.)

[Bellman 1961] R. Bellman. Adaptative Control Processes : A Guided Tour. Princeton University Press, 255 pages, 1961. (Cité en page 13.)

[Bengio 1995] Y. Bengio, Y. Le Cun, C. Nohl et C. Burges. *LeRec : A NN/HMM Hybrid for On-Line Handwriting Recognition*. Neural computation, vol. 7, no. 6, pages 1289–1303, 1995. (Cité en page 171.)

[Bertet 2007] K. Bertet, S. Guillas et J.M. Ogier. *Extensions of Bordat's algorithm for attributes*. In proceedings of the international conference on Concept Lattices and their Applications (CLA), pages 38–49, 2007. (Cité en page 78.)

[Bertet 2008] K. Bertet, S. Guillas, M. Visani et J.M. Ogier. *A propos des liens entre arbre de décision et treillis dichotomique*. In actes du Colloque International Francophone sur l'Écrit et le Document (CIFED), pages 25–30, 2008. (Cité en pages 77, 80 et 125.)

[Bertet 2009] K. Bertet, M. Visani et N. Girard. *Treillis dichotomiques et arbres de décision*. Traitement du Signal, vol. 26, no. 5, pages 409–418, 2009. (Cité en pages 77, 80 et 125.)

[Bertet 2011] K. Bertet. *Structure de treillis : Contributions structurelles et algorithmiques ; quelques usages pour des données image*. Habilitation à diriger des recherches, Université de La Rochelle, 192 pages, 2011. (Cité en pages 17, 18 et 78.)

[Bianne-Bernard 2011] A.L. Bianne-Bernard, F. Menasri, R.A.H. Mohamad, C. Mokbel, C. Kermorvant et L. Likforman-Sulem. *Dynamic and Contextual Information in HMM Modeling for Handwritten Word Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 10, pages 2066–2080, 2011. (Cité en pages 92, 99, 169 et 170.)

[Bilenko 2004] M. Bilenko, S. Basu et R. J. Mooney. *Integrating constraints and metric learning in semi-supervised clustering*. In proceedings of the International Conference on Machine Learning (ICML), pages 81–88, 2004. (Cité en page 42.)

[Birkhoff 1967] G. Birkhoff. Lattice theory, volume 25. American Mathematical Society, 3ème édition, 418 pages, 1967. (Cité en page 73.)

[Bishop 2006] C.M. Bishop. Pattern recognition and machine learning. Information Sciences and Statistics. Springer, 743 pages, 2006. (Cité en page 2.)

[Biswas 2012] A. Biswas et D.W. Jacobs. *Active image clustering : Seeking constraints from humans to complement algorithms*. In proceedings of the IEEE international conference on Computer Vision and Pattern Recognition (CVPR), pages 2152–2159, 2012. (Cité en pages 34 et 43.)

[Bloch 2014] I. Bloch, R. Clouard, M. Revenu et O. Sigaud. Panorama de l'intelligence artificielle - ses bases méthodologiques ses développements - volume 3, chapitre 7 : Intelligence artificielle et reconnaissance des formes, vision, apprentissage, pages 1165–1195. Cépaduès, France, 2014. (Cité en page 2.)

[Borth 2008] D. Borth, C. Schulze, A. Ulges et T.M. Breuel. *Navidgator - Similarity Based Browsing for Image and Video Databases*. In proceedings of the German conference on artificial intelligence (KI), pages 22–29, 2008. (Cité en pages 8, 25 et 33.)

[Bouchard 2005] G. Bouchard. *Les modèles génératifs en classification supervisée et applications à la catégorisation d'images et à la fiabilité industrielle*. Thèse de doctorat, Université Joseph Fourier, Grenoble 1, 230 pages, 2005. (Cité en pages 10, 173 et 174.)

[Boureau 2012] Y.L. Boureau. *Learning Hierarchical Feature Extractors For Image Recognition*. PhD thesis, Department of Computer Science, New York University, 167 pages, 2012. (Cité en page 60.)

[Boutell 2004] M.R Boutell, J. Luo, X. Shen et C.M. Brown. *Learning multi-label scene classification*. Pattern recognition, vol. 37, no. 9, pages 1757–1771, 2004. (Cité en page 121.)

[Breiman 1984] L. Breiman, J.H. Friedman, R.A. Olshen et C.J. Stone. Classification And Regression Trees. Wadsworth and Brooks, Monterey, CA., 368 pages, 1984. (Cité en pages 79 et 85.)

[Breiman 2001] L. Breiman. *Random forests*. Machine learning, vol. 45, no. 1, pages 5–32, 2001. (Cité en page 85.)

[Brodley 1995] C.E. Brodley et P.E. Utgoff. *Multivariate decision trees*. Machine Learning, vol. 19, no. 1, pages 45–77, 1995. (Cité en page 84.)

[Brut 2011] M. Brut, J.L. Soubie et F. Sèdes. *Integrated Cooperative Framework for Project Resources Allocation*. In Integrated Computing Technology, chapitre 4, pages 40–49. Volume 165 de Communication in Compter and Information Science, Springer, 2011. (Cité en page 56.)

[Bui 2011] Q.A. Bui, M. Visani, S. Prum et J.M. Ogier. *Writer Identification using TF-IDF for Cursive Handwritten Word Recognition*. In proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 844–848, 2011. (Cité en pages 101 et 125.)

[Bui 2012] Q.A. Bui, M. Visani et R. Mullot. *Système générique et omni-langage de navigation dans des bases de documents anciens basé sur de la recherche de mots par composition interactive de requêtes*. In actes du Colloque International Francophone sur l'Écrit et le Document (CIFED), pages 413–418, 2012. (Cité en pages 45 et 63.)

[Bui 2013] Q.A. Bui, M. Visani et R. Mullot. *Invariants extraction method applied in an omni-language old document navigating system*. In proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 1357–1361, 2013. (Cité en pages 47, 49 et 63.)

[Bui 2014] H.N. Bui, I.S. Na et S.H. Kim. *Staff Line Removal Using Line Adjacency Graph and Staff Line Skeleton for Camera-Based Printed Music Scores*. In proceedings of the International Conference on Pattern Recognition (ICPR), pages 2787–2789, 2014. (Cité en page 113.)

[Bunke 2011] H. Bunke et K. Riesen. *Recent advances in graph-based pattern recognition with applications in document analysis*. Pattern Recognition, vol. 44, no. 5, pages 1057 – 1067, 2011. (Cité en pages 25, 73, 153 et 158.)

[Burghouts 2009] G.J Burghouts et J.M. Geusebroek. *Performance evaluation of local colour invariants*. Computer Vision and Image Understanding, vol. 113, no. 1, pages 48–62, 2009. (Cité en page 61.)

[Cachier 2001] P. Cachier, J.F. Mangin, X. Pennec, D. Rivière, D. Papadopoulos-Orfanos, J. Régis et N. Ayache. *Multisubject non-rigid registration of brain MRI using intensity and geometric features*. In proceedings of the international conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pages 734–742, 2001. (Cité en page 25.)

[Caillault 2005] E. Caillault, C. Viard-Gaudin et A.R. Ahmad. *MS-TDNN with Global Discriminant Trainings*. In proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 856–861, 2005. (Cité en page 171.)

[Carpineto 1993] C. Carpineto et G. Romano. *Galois : An order-theoretic approach to conceptual clustering*. In proceedings of the International Conference on Machine Learning (ICML), pages 33–40, 1993. (Cité en page 76.)

[Chang 2011] C.C. Chang et C.J. Lin. *LIBSVM : A Library for Support Vector Machines*. ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, pages 1–27, 2011. (Cité en pages 98 et 99.)

[Chatfield 2011] K. Chatfield, V. Lempitsky, A. Vedaldi et A. Zisserman. *The devil is in the details : an evaluation of recent feature encoding methods*. In proceedings of the British Machine Vision Conference (BMVC), pages 1–12, 2011. (Cité en page 61.)

[Chen 1995] M.Y. Chen, A. Kundu et S.N. Srihari. *Variable duration hidden Markov model and morphological segmentation for handwritten word recognition*. IEEE Transactions on Image Processing, vol. 4, no. 12, pages 1675–1688, 1995. (Cité en page 170.)

[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, 2000. (Cité en page 33.)

[Chen 2005] Y. 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, 2005. (Cité en pages 30 et 33.)

[Chen 2007] C.H. Chen, W. Härdle et A. Unwin. Handbook of data visualization. Springer, 950 pages, 2007. (Cité en page 2.)

[Chen 2009] C.H. Chen. Handbook of pattern recognition and computer vision. World Scientific, 984 pages, 2009. (Cité en page 2.)

[Chen 2012] N. Chen et V.K. Prasanna. *Semantic Image Clustering Using Object Relation Network*. In proceedings of the international conference on Computational Visual Media (CVM), pages 59–66, 2012. (Cité en page 33.)

[Chuan 2011] X. Chuan, Z. Yang et Y. Dan. *Ontology based Image Semantics Recognition using Description Logics*. International Journal of Advancements in Computing Technology, vol. 3, no. 10, pages 1–8, 2011. (Cité en page 28.)

[Cohn 1996] D. Cohn, Z. Ghahramani et M.I. Jordan. *Active learning with statistical models*. Journal of Artificial Intelligence Research, vol. 4, pages 129–145, 1996. (Cité en pages 39 et 43.)

[Connel 2002] S. Connel et A.K. Jain. *Writer adaptation of online handwriting models*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, pages 434–437, 2002. (Cité en page 101.)

[Cord 2008] M. Cord et C. Padraig. Machine learning techniques for multimedia. Springer, 290 pages, 2008. (Cité en pages 2, 39 et 43.)

[Coustaty 2008] M. Coustaty, S. Guillas, M. Visani, K. Bertet et J.M. Ogier. *On the joint use of a Structural Signature and a Galois Lattice Classifier for Symbol Recognition*. In Graphics Recognition : Recent Advances and New Opportunities, Selected Papers from GREC 2007, chapitre 7, pages 61–70. Volume 5046 de Lecture Notes in Computer Science (LNCS), Springer, 2008. (Cité en pages 63 et 157.)

[Coustaty 2010] M. Coustaty, S. Guillas, M. Visani, K. Bertet et J.M. Ogier. *Reconnaissance de symboles à partir d'une signature structurelle flexible et d'un classifieur de type treillis de Galois*. Revue de Technique et Science Informatiques, vol. 29, no. 6, pages 665–690, 2010. (Cité en pages 79 et 125.)

[Coustaty 2011] M. Coustaty, K. Bertet, M. Visani et J.M. Ogier. *A New Adaptive Structural Signature for Symbol Recognition By Using a Galois Lattice as a Classifier*. IEEE Transactions on Systems, Man, and Cybernetics. Part B : Cybernetics, vol. 41, no. 4, pages 1136–1148, 2011. (Cité en pages 63, 73, 87, 158 et 159.)

[Cover 2012] T. Cover et J. Thomas. Elements of information theory. Series in Telecommunications. John Wiley & Sons, 776 pages, 2012. (Cité en page 174.)

[Cutting 1992] D.R. Cutting, J.O. Pedersen, D. Karger et J.W. Tukey. *Scatter/Gather : a cluster-based approach to browsing large document collections*. In proceedings of the international ACM SIGIR conference on research and development in Information Retrieval, pages 318–329, 1992. (Cité en page 34.)

[Daher 2012] H. Daher, D. Gaceb, V. Eglin, S. Bres et N. Vincent. *Unsupervised categorization method of graphemes on handwritten manuscripts : application to style recognition*. In proceedings of the SPIE conference on Document Recognition and Retrieval (DRR), 2012. (Cité en page 101.)

[Dalal 2005] N. Dalal et B. Triggs. *Histograms of oriented gradients for human detection*. Computer Vision and Pattern Recognition, vol. 1, pages 886–893, 2005. (Cité en page 26.)

[Datta 2008] R. Datta, D. Joshi, J. Li et J.Z. Wang. *Image Retrieval : Ideas, Influences, and Trends of the New Age*. ACM Computing Surveys, vol. 40, no. 2, pages 1–60, 2008. (Cité en pages 8 et 25.)

[Davey 1991] B.A. Davey et H.A. Priestley. Introduction to lattices and orders. Cambridge University Press, 2ème édition, 300 pages, 1991. (Cité en page 73.)

[Davidson 2005] I. Davidson et S.S. Ravi. *Agglomerative Hierarchical Clustering with Constraints : Theoretical and Empirical Results*. In proceedings of the international conference on Knowledge Discovery in Databases (KDD), pages 59–70, 2005. (Cité en page 33.)

[Delalandre 2010] M. Delalandre, E. Valveny, T. Pridmore et D. Karatzas. *Generation of synthetic documents for performance evaluation of symbol recognition and spotting systems*. International Journal on Document Analysis and Recognition, vol. 13, pages 187–207, 2010. (Cité en page 110.)

[Delaye 2008] A. Delaye, S. Macé et E. Anquetil. *Hybrid statistical-structural on-line Chinese character recognition with fuzzy inference system*. In proceedings of the International Conference on Pattern Recognition (ICPR), pages 1–4, 2008. (Cité en page 153.)

[Deng 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li et L. Fei-Fei. *ImageNet : A Large-Scale Hierarchical Image Database*. In proceedings of the IEEE international conference on Computer Vision and Pattern Recognition (CVPR), pages 1–8, 2009. (Cité en page 120.)

[Depeursinge 2014] A. Depeursinge, C. Kurtz, S. Napel, C. Beaulieu et D. Rubin. *Predicting Visual Semantic Descriptive Terms from Radiological Image Data : Preliminary Results with Liver Lesions in CT*. IEEE Transactions on Medical Imaging, vol. 99, pages 1–8, 2014. (Cité en page 28.)

[Deselaers 2008] T. Deselaers, D. Keysers et H. Ney. *Features for image retrieval : an experimental comparison*. Information Retrieval, vol. 11, no. 2, pages 77–107, 2008. (Cité en page 27.)

[Donahue 2014] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng et T. Darrell. *Decaf : A deep convolutional activation feature for generic visual recognition*. In proceedings of the International Conference on Machine Learning (ICML), pages 1–10, 2014. (Cité en page 26.)

[Dos Santos Montagner 2014] I. Dos Santos Montagner, R. Hirata et N.S.T. Hirata. *A Machine Learning based method for Staff Removal*. In proceedings of the International Conference on Pattern Recognition (ICPR), pages 3162–3167, 2014. (Cité en page 113.)

[Dosch 2000] P. Dosch, K. Tombre, C. Ah-Soon et G. Masini. *A complete system for the analysis of architectural drawings*. International Journal on Document Analysis and Recognition, vol. 3, no. 2, pages 102–116, 2000. (Cité en page 73.)

[Du 2009] L. Du, L. Ren, L. Carin et Dunson D.B. *A Bayesian Model for Simultaneous Image Clustering, Annotation and Object Segmentation*. In proceedings of the International Conference on Advances in Neural Information Processing Systems, pages 486–494, 2009. (Cité en page 33.)

[Dubey 2010] A. Dubey, I. Bhattacharya et S. Godbole. *A Cluster-Level Semi-supervision Model for Interactive Clustering*. In proceedings of the joint conference on Machine Learning and Knowledge Discovery in Databases, pages 409–424, 2010. (Cité en page 33.)

[Ducrou 2008] J. Ducrou et P. Eklund. *An intelligent user interface for browsing and searching MPEG-7 images using concept lattices*. International journal of foundations of computer science, vol. 19, no. 02, pages 359–381, 2008. (Cité en page 87.)

[Duda 2012] R.O. Duda, P.E. Hart et D.G. Stork. Pattern classification (2nd edition). John Wiley & Sons, New York, 680 pages, 2012. (Cité en pages 9 et 173.)

[Duffner 2005] S. Duffner et C. Garcia. *A Connexionist Approach for Robust and Precise Facial Feature Detection in Complex Scenes*. In proceedings of the IEEE symposium on Image and Signal Processing and Analysis (ISPA), pages 1–6, 2005. (Cité en pages 155 et 156.)

[Eakins 1999] J. Eakins, M. Graham et T. Franklin. *Content-based Image Retrieval*. Library and Information Briefings, vol. 85, pages 1–15, 1999. (Cité en page 30.)

[El-Yacoubi 1999] M.A. El-Yacoubi, M. Gilloux, R. Sabourin et C.Y. Suen. *An HMM-Based Approach for Off-Line Unconstrained Handwritten Word Modeling and Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 8, pages 752–760, 1999. (Cité en page 170.)

[España-Boquera 2011] S. España-Boquera, Castro-Bleda M.J., Gorbe-Moya J. et Zamora-Martinez F. *Improving Offline Handwritten Text Recognition with Hybrid HMM/ANN Models*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 4, pages 767–779, 2011. (Cité en page 171.)

[Fellbaum 1998] C. Fellbaum. Wordnet : an electronical lexical database. Cambridge, MA : MIT Press, 501 pages, 1998. (Cité en page 120.)

[Felzenszwalb 2010] P.F. Felzenszwalb, R.B. Girshick, D. McAllester et D. Ramanan. *Object Detection with Discriminatively Trained Part-Based Models*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pages 1627–1645, 2010. (Cité en page 33.)

[Fischer 2009] A. Fischer, M. Wuthrich, M. Liwicki, V. Frinken, H. Bunke, G. Viehhauser et M. Stolz. *Automatic transcription of handwritten medieval documents*. In proceedings of the IEEE International Conference on Virtual Systems and Multimedia, pages 137–142, 2009. (Cité en page 171.)

[Fischer 2012] A. Fischer, A. Keller, V. Frinken et H. Bunke. *Lexicon-free handwritten word spotting using character HMMs*. Pattern Recognition Letters, vol. 33, pages 934–942, 2012. (Cité en page 114.)

[Fischer 2013] A. Fischer, M. Visani, V.C. Kieu et C.Y. Suen. *Generation of Learning Samples for Historical Handwriting Recognition Using Image Degradation*. In proceedings of the international workshop on Historical document Imaging and Processing (HIP), pages 73–79, 2013. (Cité en pages 113, 114 et 125.)

[Fischer 2014] A. Fischer, M. Baechler, A. Garz, M. Liwicki et R. Ingold. *A Combined System for Text Line Extraction and Handwriting Recognition in Historical Documents*. In proceedings of the IAPR international workshop on Document Analysis Systems (DAS), pages 71–75, 2014. (Cité en page 105.)

[Fornés 2012] A. Fornés, A. Dutta, A. Gordo et J. Lladós. *CVC-MUSCIMA : a Ground Truth of Handwritten Music Score Images for Writer Identification and Staff Removal*. International Journal on Document Analysis and Recognition (IJDAR), vol. 15, no. 3, pages 243–251, 2012. (Cité en page 111.)

[Fornés 2014] A. Fornés, V.C. Kieu, M. Visani, N. Journet et A. Dutta. *The ICDAR/GREC 2013 Music Scores Competition : Staff Removal*. In Graphics Recognition. New Trends and Challenges, Revised Selected Papers from GREC 2013, chapitre 16. Volume 8746 de Lecture Notes in Computer Science (LNCS), Springer, 15 pages, 2014. (Cité en pages 112 et 125.)

[Fouquier 2012] G. Fouquier, J. Atif et I. Bloch. *Sequential model-based segmentation and recognition of image structures driven by visual features and spatial relations*. Computer Vision and Image Understanding, vol. 116, no. 1, pages 146–165, 2012. (Cité en page 11.)

[Fournier 2001] J. Fournier, M. Cord et S. Philipp-Foliguet. *RETIN : a content-based image indexing and retrieval system*. Pattern Analysis and Applications, special issue on image indexation, vol. 4, pages 153–173, 2001. (Cité en pages 8, 25 et 31.)

[Frinken 2012] V. Frinken, A. Fischer, R. Manmatha et H. Bunke. *A novel word spotting method based on recurrent neural networks*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 2, pages 211–224, 2012. (Cité en pages 54 et 55.)

[Ganter 1999] B. Ganter et R. Wille. Formal concept analysis, mathematical foundations. Springer Verlag, Berlin, 1999. (Cité en page 73.)

[Gao 2005] B. Gao, T-Y Liu, T. Qin, X Zheng, Q.S. Cheng et W.Y. Ma. *Web image clustering by consistent utilization of visual features and surrounding texts*. In proceedings of the ACM International Conference on Multimedia, pages 112–121, 2005. (Cité en page 33.)

[Garcia 2004] C. Garcia et M. Delakis. *Convolutional Face Finder : A Neural Architecture for Fast and Robust Face Detection*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 11, pages 1408–1423, 2004. (Cité en page 155.)

[Gatos 2014] B. Gatos, G. Louloudis, T. Causer, K. Grint, V. Romero, J.A. Sánchez, A.H. Toselli et E. Vidal. *Ground-Truth Production in the tranScriptorium Project*. In proceedings of the IAPR international workshop on Document Analysis Systems (DAS), pages 237–241, 2014. (Cité en page 105.)

[Ghosh 2011] J. Ghosh et A. Acharya. *Cluster ensembles.* Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery, vol. 1, no. 4, pages 305–315, 2011. (Cité en page 10.)

[Girard 2009] N. Girard, K. Bertet et M. Visani. *Vers une discrétisation locale pour les treillis dichotomiques.* In actes des rencontres de la Société Francophone de Classification (SFC), pages 113–116, 2009. (Cité en pages 81 et 125.)

[Girard 2011a] N. Girard, K. Bertet et M. Visani. *A local discretization of continuous data for lattices : Technical aspects.* In proceedings of the international conference on Concept Lattices and their Applications (CLA), pages 409–412, 2011. (Cité en pages 81 et 125.)

[Girard 2011b] N. Girard, K. Bertet et M. Visani. *Local discretization of numerical data for Galois Lattices.* In proceedings of the IEEE International Conference on Tools with Artificial Intelligence (ICTAI), pages 902–903, 2011. (Cité en pages 81 et 125.)

[Girard 2013] N. Girard. *Vers une approche hybride mêlant arbre de classification et treillis de Galois pour l'indexation d'images.* Thèse de doctorat, Université de La Rochelle, France, 285 pages, 2013. (Cité en pages 18, 71, 75, 76, 77, 79, 80, 81, 125 et 126.)

[Goh 2013] H. Goh. *Learning deep visual representations.* Thèse de doctorat, Université Pierre et Marie Curie-Paris 6, 167 pages, 2013. (Cité en pages 26 et 60.)

[Grangier 2008] D. Grangier et S. Bengio. *A discriminative kernel-based approach to rank images from text queries.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 8, pages 1371–1384, 2008. (Cité en page 121.)

[Graves 2009] A. Graves, M. Liwicki, Fernandez S., R. Bertolami, H. Bunke et J. Schmidhuber. *A Novel Connectionist System for Unconstrained Handwriting Recognition.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, pages 855–868, 2009. (Cité en pages 93, 96 et 171.)

[Grewe 1995] L. Grewe et A.C. Kak. *Interactive learning of a multiple-attribute hash table classifier for fast object recognition.* Computer Vision and Image Understanding, vol. 61, no. 3, pages 387–416, 1995. (Cité en page 32.)

[Grira 2005] N. Grira, M. Crucianu et N. Boujemaa. *Semi-supervised image database categorization using pairwise constraints.* In proceedings of the IEEE International Conference on Image Processing (ICIP), volume 3, pages 1228–1231, 2005. (Cité en page 33.)

[Grira 2008] N. Grira, M. Crucianu et N. Boujemaa. *Active semi-supervised fuzzy clustering.* Pattern Recognition, vol. 41, no. 5, pages 1834–1844, 2008. (Cité en pages 34 et 43.)

[Guillas 2007] S. Guillas. *Reconnaissance d'objets graphiques détériorés : approche fondée sur un treillis de Galois.* Thèse de doctorat, Université de La Rochelle, 209 pages, 2007. (Cité en pages 17, 71, 77 et 80.)

[Guillas 2009] S. Guillas, K. Bertet, M. Visani, J.M. Ogier et N. Girard. *Some links between decision tree and dichotomic lattice.* In proceedings of the international conference on Concept Lattices and their Applications (CLA), pages 193–205, 2009. (Cité en pages 77, 80 et 125.)

[Guillaumin 2009] M. Guillaumin, T. Mensink, J. Verbeek et C. Schmid. *Tagprop : Discriminative metric learning in nearest neighbor models for image auto-annotation.* In proceedings of the International Conference on Computer Vision (ICCV), pages 309–316, 2009. (Cité en page 121.)

[Haralick 1979] R. M. Haralick. *Statistical and structural approaches to texture*. Proceedings of the IEEE, vol. 67, no. 5, pages 786–804, 1979. (Cité en page 153.)

[He 2004] J. He, A.H. Tan, C.L. Tan et S.Y. Sung. *On Quantitative Evaluation of Clustering Systems*. In Clustering and Information Retrieval, volume 11 of *Network Theory and Applications*, pages 105–133. Springer US, 2004. (Cité en page 39.)

[Hsu 2002] C.W. Hsu et C.J. Lin. *A comparison of methods for multiclass support vector machines*. Neural Networks, IEEE Transactions on, vol. 13, no. 2, pages 415–425, 2002. (Cité en page 98.)

[Hu 2000] J. Hu, S.G. Lim et M.K. Brown. *Writer Independent On-line Handwriting Recognition using an HMM Approach*. Pattern Recognition, vol. 33, no. 1, pages 133–147, 2000. (Cité en pages 92, 169 et 170.)

[Hudelot 2008] C. Hudelot, J. Atif et I. Bloch. *Fuzzy Spatial Relation Ontology for Image Interpretation*. Fuzzy Sets and Systems, vol. 159, pages 1929–1951, 2008. (Cité en page 160.)

[Humm 2009] A. Humm, J. Hennebert et R. Ingold. *Combined handwriting and speech modalities for user authentication*. IEEE Transactions on Systems, Man and Cybernetics, Part A : Systems and Humans, vol. 39, no. 1, pages 25–35, 2009. (Cité en page 101.)

[Hwang 2004] B.W. Hwang, M.C. Roh et S.W. Lee. *Performance Evaluation of Face Recognition Algorithms on Asian Face Database*. In proceedings of the IEEE international conference on automatic Face and Gesture Recognition (FGR), pages 278–283, 2004. (Cité en page 155.)

[Iandola 2014] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell et K. Keutzer. *DenseNet : Implementing Efficient ConvNet Descriptor Pyramids*. Technical report, University of California, Berkeley, 11 pages, 2014. (Cité en page 26.)

[Ingold 2002] R. Ingold. *Analyse et reconnaissance d'images de documents*. Techniques de l'ingénieur. Documents numériques : technologies d'acquisition et de restitution, pages 1–12, 2002. (Cité en pages 67, 68 et 128.)

[Jain 1997] A. Jain et D. Zongker. *Feature selection : Evaluation, application, and small sample performance*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 2, pages 153–158, 1997. (Cité en pages 27, 60, 76 et 93.)

[Jain 2010] A.K. Jain. *Data clustering : 50 years beyond K-means*. Pattern Recognition Letters, vol. 31, no. 8, pages 651–666, 2010. (Cité en pages 9, 32 et 39.)

[Javidi 2002] B. Javidi. Image recognition and classification : algorithms, systems, and applications. CRC Press, 506 pages, 2002. (Cité en page 68.)

[Jing 2012] Y. Jing, H.A. Rowley, J. Wang, D. Tsai, C. Rosenberg et M. Covell. *Google Image Swirl : A Large-Scale Content-Based Image Visualization System*. In proceedings of the international World Wide Web Conference (WWW), pages 539–540, 2012. (Cité en pages 8, 25 et 32.)

[Jolion 2001] J.M. Jolion. *Feature similarity*. In principles of visual information retrieval, chapitre 5, pages 121–143. Springer, 2001. (Cité en pages 24 et 154.)

[Journet 2010] N Journet, A. Vialard et J.P. Domenger. *Analyse de fontes anciennes : de la génération de données synthétiques à la reconnaissance*. In actes du Colloque International Francophone sur l'Écrit et le Document (CIFED), pages 1–16, 2010. (Cité en page 108.)

[Kanungo 1993] T. Kanungo, R. M Haralick et I. Phillips. *Global and local document degradation models*. In proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 730–734, 1993. (Cité en pages 109 et 114.)

[Kanungo 1994] T. Kanungo, R.M. Haralick et I. Phillips. *Nonlinear global and local document degradation models*. International Journal of Imaging Systems and Technology, vol. 5, pages 220–230, 1994. (Cité en pages 109, 110 et 113.)

[Kanungo 2000] T. Kanungo, R.M. Haralick, H.S. Baird, W. Stuezle et D. Madigan. *A statistical, nonparametric methodology for document degradation model validation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 11, pages 1209–1223, 2000. (Cité en pages 79, 109 et 114.)

[Kass 1980] G.V. Kass. *An Exploratory Technique for Investigating Large Quantities of Categorical Data*. Journal of the Royal Statistical Society, Series C (Applied Statistics), vol. 29, no. 2, pages 119–127, 1980. (Cité en page 82.)

[Kieu 2012a] V.C. Kieu, N. Journet, M. Visani, J.P. Domenger et R. Mullot. *Génération d'images semi-synthétiques de documents anciens*. In actes du Colloque International Francophone sur l'Écrit et le Document (CIFED), pages 415–421, 2012. (Cité en page 125.)

[Kieu 2012b] V.C. Kieu, M. Visani, N. Journet, J.P. Domenger et R. Mullot. *A character degradation model for grayscale ancient document images*. In proceedings of the International Conference on Pattern Recognition (ICPR), pages 685–688, 2012. (Cité en pages 110, 111, 113 et 125.)

[Kieu 2013a] V.C. Kieu, N. Journet, M. Visani, J.P. Domenger et R. Mullot. *Semi-synthetic Document Image Generation Using Texture Mapping on Scanned 3D Document Shapes*. In proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 489–493, 2013. (Cité en pages 109, 110 et 125.)

[Kieu 2013b] V.C. Kieu, M. Visani, N. Journet, R. Mullot et J.P. Domenger. *An Efficient Parametrization of Character Degradation Model for Semi-synthetic Image Generation*. In proceedings of the international workshop on Historical document Imaging and Processing (HIP), pages 29–35, 2013. (Cité en pages 110, 111 et 125.)

[Kieu 2014] V.C. Kieu, M. Mehri, V. Rabeux, N. Journet et M. Visani. *Génération d'images semi-synthétiques de documents anciens à des fins d'évaluation de performances et d'apprentissage*. In actes de la Semaine du Document Numérique et de la Recherche d'Information (SDNRI - CIFED/CORIA), pages 1–16, Tours, France, 2014. (Cité en pages 113, 116 et 125.)

[Kinoshita 2005] Y. Kinoshita, N. Nitta et N. Babaguchi. *Interactive Clustering of Video Segments for Media Structuring*. In IEEE International Conference on Multimedia and Expo (ICME), pages 630–633, 2005. (Cité en page 33.)

[Kittler 1998] J. Kittler, M. Hatef, R.P.W. Duin et J. Matas. *On combining classifiers*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 3, pages 226–239, 1998. (Cité en page 10.)

[Klein 2002] D. Klein, S.D. Kamvar et C.D. Manning. *From Instance-level Constraints to Space-Level Constraints : Making the Most of Prior Knowledge in Data Clustering*. In proceedings of the International Conference on Machine Learning (ICML), pages 307–314, 2002. (Cité en page 33.)

[Konidaris 2007] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis et S.J. Perantonis. *Keyword-guided word spotting in historical printed documents using synthetic data and user feedback*. International Journal on Document Analysis and Recognition, vol. 9, pages 167–177, 2007. (Cité en page 46.)

[Kothari 2000] R. Kothari et M. Dong. Pattern recognition : From classical to modern approaches, chapitre Decision trees for classification : A review and some new results, pages 169–184. S.R. Pal, A. Pal (eds), World Scientific, 2000. (Cité en pages 76 et 82.)

[Krause 2014] J. Krause, T. Gebru, J. Deng, L.-J. Li et L. Fei-Fei. *Learning Features and Parts for Fine-Grained Recognition*. In proceedings of the International Conference on Pattern Recognition (ICPR), invited paper, 8 pages, 2014. (Cité en page 26.)

[Krishnamachari 1999] S. Krishnamachari et M. Abdel-Mottaleb. *Image browsing using hierarchical clustering*. In proceedings of the IEEE International Symposium on Computers and Communications, pages 301–307, 1999. (Cité en page 33.)

[Krizhevsky 2012] A. Krizhevsky, I. Sutskever et G.E. Hinton. *ImageNet Classification with Deep Convolutional Neural Networks*. In proceedings of the international conference on advances in Neural Information Processing Systems (NIPS), pages 1097–1105, 2012. (Cité en page 117.)

[Kulis 2009a] B. Kulis, S. Basu, I. Dhillon et R.J. Mooney. *Semi-supervised Graph Clustering : A Kernel Approach*. Machine Learning, vol. 74, no. 1, pages 1–22, 2009. (Cité en page 33.)

[Kulis 2009b] B. Kulis et K. Grauman. *Kernelized Locality-Sensitive Hashing for Scalable Image Search*. In proceedings of the IEEE International Conference on Computer Vision, pages 2130–2137, 2009. (Cité en page 32.)

[Kullback 1959] S. Kullback. Information theory and statistics. John Wiley & Sons, 418 pages, 1959. (Cité en page 174.)

[Kuznetsov 2007] S. Kuznetsov. *On stability of a formal concept*. Annals of Mathematics and Artificial Intelligence, vol. 49, pages 101–115, 2007. (Cité en page 83.)

[Kvalseth 1987] T.O. Kvalseth. *Entropy and Correlation : Some Comments*. IEEE Transactions on Systems, Man and Cybernetics, vol. 17, pages 517–519, 1987. (Cité en page 49.)

[Käster 2003] T. Käster, V. Wendt et G. Sagerer. *Comparing Clustering Methods for Database Categorization in Image Retrieval*. In Pattern Recognition, volume 2781 of *Lecture Notes in Computer Science*, pages 228–235. Springer, 2003. (Cité en pages 30, 33 et 37.)

[Lai 2012a] H.P. Lai, M. Visani, A. Boucher et J.M. Ogier. *An experimental comparison of clustering methods for content-based indexing of large image databases*. Pattern Analysis and Applications, vol. 15, no. 4, pages 345–366, 2012. (Cité en pages 36 et 63.)

[Lai 2012b] H.P. Lai, M. Visani, A. Boucher et J.M. Ogier. *Unsupervised and semi-supervised clustering for large image database indexing and retrieval*. In proceedings of the IEEE-RIVF international conference on computing and communication technologies (RIVF), pages 276–281, 2012. (Cité en pages 38 et 63.)

[Lai 2013a] H.P. Lai. *Towards an interactive index structuring system for content-based image retrieval in large image databases*. Thèse de doctorat, Université de La Rochelle, France, 179 pages, 2013. (Cité en pages 16, 32, 35, 39, 41, 63, 64, 161, 164, 165 et 167.)

[Lai 2013b] H.P. Lai, M. Visani, A. Boucher et J.M. Ogier. *A new Interactive Semi-Supervised Clustering model for large image database indexing*. Pattern Recognition Letters, Special

Issue on Partially Supervised Learning for Pattern Recognition, pages 1–48, 2013. (Cité en pages 40, 41, 63 et 175.)

[Lai 2014a] H.P. Lai, M. Visani, A. Boucher et J.M. Ogier. *Towards an interactive index structuring system for content-based image retrieval in large image databases.* Electronic Letters on Computer Vision and Image Analysis (ELCVIA), Special Issue on Recent PhD Thesis Dissemination, vol. 13, no. 2, pages 45–46, 2014. (Cité en pages 32 et 63.)

[Lai 2014b] H.P. Lai, M. Visani, A. Boucher et J.M. Ogier. *Unsupervised and interactive semi-supervised clustering for large image database indexing and retrieval.* Fundamenta Informaticae, vol. 130, pages 1–18, 2014. (Cité en pages 38, 63, 163 et 167.)

[Lallican 2000] P. M. Lallican, C. Viard-Gaudin et S. Knerr. *From off-line to on-line handwriting recognition.* In proceedings of the International Conference on Frontiers in Handwriting Recognition (ICFHR), pages 303–312, 2000. (Cité en page 47.)

[Lam 2000] L. Lam. *Classifier combinations : implementations and theoretical issues.* In Multiple classifier systems, pages 77–86. Springer, 2000. (Cité en page 10.)

[Lange 2004] T. Lange, V. Roth, M.L. Braun et J. M. Buhmann. *Stability-based validation of clustering solutions.* Neural computation, vol. 16, no. 6, pages 1299–1323, 2004. (Cité en pages 49 et 168.)

[Laurent 2003] C. Laurent, N. Laurent et M. Visani. *Color Image Retrieval Based on Wavelet Salient Features Detection.* In proceedings of the international workshop on Content-Based Multimedia Indexing (CBMI 03), pages 327–334, 2003. (Cité en pages 26 et 63.)

[Lazebnik 2006] S. Lazebnik, C. Schmid et J. Ponce. *Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories.* In proceedings of the IEEE international conference on Computer Vision and Pattern Recognition (CVPR), volume 2, pages 2169–2178, 2006. (Cité en page 60.)

[Le 2012] V.P. Le, M. Visani, C.D. Tran et J.M. Ogier. *Logo Spotting For Document Categorization.* In proceedings of the International Conference on Pattern Recognition (ICPR), pages 3484–3487, 2012. (Cité en pages 55 et 63.)

[Le 2013] V.P. Le, M. Visani, C.D. Tran et J.M. Ogier. *Improving Logo Spotting and Matching for Document Categorization by a Post-Filter based on Homography.* In proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 270–274, 2013. (Cité en pages 55 et 63.)

[Le 2014] V.P. Le, N. Nayef, M. Visani, J.M. Ogier et C.D. Tran. *Logo Spotting and Recognition for Document Retrieval.* In proceedings of the International Conference on Pattern Recognition (ICPR), pages 3056–3061, 2014. (Cité en pages 55 et 63.)

[Lelis 2009] L. Lelis et J. Sander. *Semi-supervised density-based clustering.* In proceedings of the International Conference on Data Mining (ICDM), pages 842–847, 2009. (Cité en page 33.)

[Lew 2006] M.S. Lew, N. Sebe, C. Djeraba et R. Jain. *Content-based Multimedia Information Retrieval : State of the Art and Challenges.* ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 2, no. 1, pages 1–19, 2006. (Cité en page 2.)

[Leydier 2009] Y. Leydier, A. Ouji, F. Lebourgeois et H. Emptoz. *Towards an omnilingual word retrieval system for ancient manuscripts.* Pattern Recognition, vol. 42, no. 9, pages 2089–2105, 2009. (Cité en page 46.)

[Li 1996] Y. Li, D. Lopresti, G. Nagy et A. Tomkins. *Validation of image defect models for optical character recognition.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 2, pages 99–107, 1996. (Cité en page 115.)

[Li 2008] J. Li et N.M. Allinson. *A comprehensive review of current local features for computer vision.* Neurocomputing, vol. 71, no. 10-12, pages 1771–1787, 2008. (Cité en page 26.)

[Li 2013] K. Li, X. Lu, H. Ling, L. Liu, T. Feng et Z. Tang. *Detection of Overlapped Quadrangles in Plane Geometric Figures.* In proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 260–264, 2013. (Cité en pages 73 et 159.)

[Liang 2008] J. Liang, D. De Menthon et D.S. Doermann. *Geometric Rectification of Camera-Captured Document Images.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, pages 591–605, 2008. (Cité en pages 109, 114 et 117.)

[Likas 2003] A. Likas, N. Vlassis et J. Verbeek. *The global k-means clustering algorithm.* Pattern Recognition, vol. 36, no. 2, pages 451–461, 2003. (Cité en pages 36 et 51.)

[Liu 2014] L. Liu et X. Lu. *Plane Geometry Figure Retrieval with Bag of Shapes.* In proceedings of the IAPR international workshop on Document Analysis Systems (DAS), pages 1–5, 2014. (Cité en page 160.)

[Liwicki 2007] M. Liwicki, A. Graves, H. Bunke et J. Schmidhuber. *A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks.* In proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 367–371, 2007. (Cité en page 170.)

[Llados 2003] J. Llados et G. Sanchez. *Symbol recognition using graphs.* In proceedings of the International Conference on Image Processing (ICIP), pages 49–52, 2003. (Cité en pages 73 et 158.)

[Lladós 2012] J. Lladós, M. Rusiñol, A. Fornés, D. Fernández et A. Dutta. *On the influence of word representations for handwritten word spotting in historical documents.* International Journal of Pattern Recognition and Artificial Intelligence, vol. 26, no. 5, pages 1–25, 2012. (Cité en page 54.)

[Lowe 2004] D.G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints.* International Journal on Computer Vision, vol. 60, no. 2, pages 91–110, 2004. (Cité en pages 26 et 37.)

[Lu 2003] G. Lu, D. Zhang et K. Wang. *Palmprint recognition using eigenpalms features.* Pattern Recognition Letters, vol. 24, no. 9, pages 1463–1467, 2003. (Cité en page 157.)

[Ma 1999] W.Y. Ma et B.S. Manjunath. *NETRA : A toolbox for navigating large image databases.* ACM Multimedia Systems, vol. 7, no. 3, pages 184–198, 1999. (Cité en pages 8, 25 et 33.)

[Maillot 2008] N. Maillot et M. Thonnat. *Ontology based complex object recognition.* Image and Vision Computing, vol. 26, no. 1, pages 102–113, 2008. (Cité en page 28.)

[Marinai 2006] S. Marinai, E. Marino et G. Soda. *Font Adaptive Word Indexing of Modern Printed Documents.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 8, pages 1187–1199, 2006. (Cité en page 46.)

[Marti 2001] U.V. Marti et H. Bunke. *Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system.* International Journal of Pattern Recognition and Artificial Intelligence, vol. 15, pages 65–90, 2001. (Cité en page 114.)

[Mensink 2012] T. Mensink. *Learning Image Classification and Retrieval Models.* PhD thesis, Université de Grenoble, INRIA-Grenoble, and Xerox Research Centre Europe, 176 pages, 2012. (Cité en page 122.)

[Mephu-Nguifo 1993] E. Mephu-Nguifo. *Une nouvelle approche basée sur le treillis de Galois pour l'apprentissage de concepts.* Mathématiques et Sciences Humaines, vol. 124, pages 19–38, 1993. (Cité en page 76.)

[Minetto 2014] R. Minetto, N. Thome, M. Cord, N.J. Leite et J. Stolfi. *SnooperText : A text detection system for automatic indexing of urban scenes.* Computer Vision and Image Understanding, vol. 122, pages 92–104, 2014. (Cité en page 120.)

[Moghaddam 2009] R.F. Moghaddam et M. Cheriet. *Low quality document image modeling and enhancement.* International Journal on Document Analysis and Recognition, vol. 11, no. 4, pages 183–201, 2009. (Cité en pages 109 et 117.)

[Nagy 2000] G. Nagy. *Twenty years of document image analysis in PAMI.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pages 38–62, 2000. (Cité en page 67.)

[Nguyen 2010] X.V. Nguyen, J. Epps et J. Bailey. *Information Theoretic Measures for Clusterings Comparison : Variants, Properties, Normalization and Correction for Chance.* Journal of Machine Learning Research, vol. 11, pages 2837–2854, 2010. (Cité en page 49.)

[Njiwoua 1999] P. Njiwoua et E. Mephu-Nguifo. *Améliorer l'apprentissage à partir d'instances grâce à l'induction de concepts : le système CIBLe.* Revue d'intelligence Artificielle, vol. 13, no. 2, pages 413–440, 1999. (Cité en page 76.)

[Norouzi 2011] M. Norouzi et D.J. Fleet. *Minimal Loss Hashing for Compact Binary Codes.* In proceedings of the International Conference on Machine Learning, pages 353–360, 2011. (Cité en page 32.)

[Nourine 1999] L. Nourine et O. Raynaud. *A fast algorithm for building lattices.* Information processing letters, vol. 71, no. 5, pages 199–204, 1999. (Cité en page 78.)

[Oosthuizen 1988] G. Oosthuizen. *The use of a Lattice in Knowledge Processing.* PhD thesis, University of Strathclyde, Glasgow, 1988. (Cité en page 76.)

[Opitz 2014] M. Opitz, M. Diem, S. Fiel, F. Kleber et R. Sablatnig. *End-to-End Text Recognition using Local Ternary Patterns, MSER and Deep Convolutional Nets.* In proceedings of the IAPR international workshop on Document Analysis Systems (DAS), pages 187–190, 2014. (Cité en page 107.)

[Paulin 2014] M. Paulin, J. Revaud, Z. Harchaoui, F. Perronnin et C. Schmid. *Transformation Pursuit for Image Classification.* In proceedings of the IEEE international conference on Computer Vision and Pattern Recognition (CVPR), pages 1–8, 2014. (Cité en pages 26 et 117.)

[Pentland 1994] A.D. Pentland, B. Moghaddam et T. Starner. *View-Based and Modular Eigenspaces for Face Recognition.* In proceedings of the IEEE Computer Society Conference on Pattern Recognition, pages 84–91, 1994. (Cité en page 155.)

[Perronnin 2010] F. Perronnin, J. Sánchez et T. Mensink. *Improving the fisher kernel for large-scale image classification.* In proceedings of the European Conference on Computer Vision (ECCV), pages 143–156, 2010. (Cité en pages 27 et 122.)

[Philbin 2008] J. Philbin, O. Chum, M. Isard, J. Sivic et A. Zisserman. *Lost in quantization : Improving particular object retrieval in large scale image databases.* In proceedings of the

IEEE international conference on Computer Vision and Pattern Recognition (CVPR), pages 1–8, 2008. (Cité en page 61.)

[Phong 1975] Bui Tuong Phong. *Illumination for computer generated pictures.* Communications of the ACM, vol. 18, no. 6, pages 311–317, 1975. (Cité en page 109.)

[Plamondon 1989] R. Plamondon et G. Lorette. *Automatic signature verification and writer identification – the state of the art.* Pattern Recognition, vol. 22, no. 2, pages 107–131, 1989. (Cité en page 101.)

[Plamondon 2000] R. Plamondon et S.N. Srihari. *On-Line and Off-Line Handwriting Recognition : A Comprehensive Survey.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pages 63–84, 2000. (Cité en page 89.)

[Pratikakis 2011] I. Pratikakis, B. Gatos et K. Ntirogiannis. *ICDAR 2011 Document Image Binarization Contest (DIBCO2011).* In proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 1506–1510, 2011. (Cité en page 113.)

[Prum 2010a] S. Prum, M. Visani et J.M. Ogier. *Cursive on-line Handwriting word recognition using a bi-character model for large lexicon applications.* In proceedings of the International Conference on Frontiers in Handwriting Recognition (ICFHR), pages 194–199, 2010. (Cité en page 125.)

[Prum 2010b] S. Prum, M. Visani et J.M. Ogier. *On-line Handwriting word recognition using a bi-character model.* In proceedings of the International Conference on Pattern Recognition (ICPR), pages 2700–2703, 2010. (Cité en page 125.)

[Prum 2013a] S. Prum. *On the use of a discriminant approach for handwritten word recognition based on bi-character models.* Thèse de doctorat, Université de La Rochelle, France, 203 pages, 2013. (Cité en pages 18, 89, 93, 94, 95, 96, 99, 100, 125 et 126.)

[Prum 2013b] S. Prum, M. Visani, A. Fischer et J.M. Ogier. *A Discriminative Approach to On-Line Handwriting Recognition Using Bi-Character models.* In proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 364–368, 2013. (Cité en pages 94 et 125.)

[Pudil 1994] P. Pudil, J. Novovicová et J. Kittler. *Floating search methods in feature selection.* Pattern Recognition Letters, vol. 15, no. 11, pages 1119–1125, 1994. (Cité en page 93.)

[Rabeux 2013] V. Rabeux, N. Journet, A. Vialard et J.P. Domenger. *Quality Evaluation of Degraded Document Images for Binarization Result Prediction.* IJDAR, vol. 1, pages 1–13, 2013. (Cité en page 113.)

[Ranzato 2007] M. Ranzato, F. J. Huang, Y. L. Boureau et Y. Le Cun. *Unsupervised learning of invariant feature hierarchies with applications to object recognition.* In proceedings of the IEEE international conference on Computer Vision and Pattern Recognition (CVPR), pages 1–8, 2007. (Cité en pages 26 et 60.)

[Rohrbach 2011] M. Rohrbach, M. Stark et B. Schiele. *Evaluating knowledge transfer and zero-shot learning in a large-scale setting.* In proceedings of the IEEE international conference on Computer Vision and Pattern Recognition (CVPR), pages 1641–1648, 2011. (Cité en page 121.)

[Ros 2009] J. Ros, C. Laurent et J.M. Jolion. *A bag of strings representation for image categorization.* Journal of Mathematical Imaging and Vision, vol. 35, no. 1, pages 51–67, 2009. (Cité en pages 27 et 61.)

[Roth 2006] C. Roth, S.A. Obiedkov et D.G. Kourie. *Towards Concise Representation for Taxonomies of Epistemic Communities*. In proceedings of the international conference on Concept Lattices and their Applications (CLA), pages 240–255, 2006. (Cité en page 83.)

[Rousseeuw 1987] P.J. Rousseeuw. *Silhouettes : a graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics, vol. 20, no. 1, pages 53–65, 1987. (Cité en page 40.)

[Rubner 2000] Y. Rubner, C. Tomasi et L.J. Guibas. *The earth mover's distance as a metric for image retrieval*. International Journal of Computer Vision, vol. 40, no. 2, pages 99–121, 2000. (Cité en pages 30 et 33.)

[Russ 2010] J.C Russ. The image processing handbook. 822 pages, CRC press, 2010. (Cité en page 2.)

[Safavian 1991] S. R. Safavian et D. Landgrebe. *A survey of decision tree classifier methodology*. IEEE transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pages 660–674, 1991. (Cité en page 76.)

[Sahami 1995] M. Sahami. *Learning classification rules using lattices*. In proceedings of the European Conference on Machine Learning (ECML), pages 343–346, 1995. (Cité en page 76.)

[Schaefer 2010] G. Schaefer. *A next generation browsing environment for large image repositories*. Multimedia Tools and Applications, vol. 47, no. 1, pages 105–120, 2010. (Cité en page 32.)

[Schenk 2006] J. Schenk et G. Rigoll. *Novel Hybrid NN/HMM Modelling Techniques for Online Handwriting Recognition*. In Guy Lorette, editeur, proceedings of the International Conference on Frontiers in Handwriting Recognition (ICFHR), pages 1–5, 2006. (Cité en page 171.)

[Schlapbach 2007] A. Schlapbach et H. Bunke. *A writer identification and verification system using HMM based recognizers*. Pattern Analysis and Applications, vol. 10, no. 1, pages 33–43, 2007. (Cité en page 101.)

[Schlapbach 2008] A. Schlapbach. Writer identification and verification, volume 311 of *Dissertations in Artifical Intelligence*. IOS Press, 148 pages, 2008. (Cité en page 101.)

[Schroder 2000] M. Schroder, H. Rehrauer, K. Seidel et M. Datcu. *Interactive learning and probabilistic retrieval in remote sensing image archives*. IEEE Transactions on Geoscience and Remote Sensing, vol. 38, no. 5, pages 2288–2298, 2000. (Cité en page 32.)

[Schölkopf 1995] B. Schölkopf, C. Burges et V. Vapnik. *Extracting Support Data for a Given Task*. In proceedings of the international conference on Knowledge Discovery and Data Mining, pages 252–257, 1995. (Cité en page 97.)

[Schölkopf 2001] B. Schölkopf et A.J. Smola. Learning with Kernels. MIT Press, Cambridge, MA., 648 pages, 2001. (Cité en page 97.)

[Sèdes 2002] F. Sèdes et H. Martin. *Recherche d'information multimédia*. In Actes des deuxièmes assises nationales du GdR I3, pages 197–215, 2002. (Cité en page 2.)

[Sivic 2003] J. Sivic et A. Zisserman. *Video Google : a text retrieval approach to object matching in videos*. In Proceeding of the IEEE International Conference on Computer Vision, pages 1470–1477, 2003. (Cité en pages 27, 37 et 61.)

[Smeulders 2000] A. W. Smeulders, M. Worring, S. Santini, A. Gupta et R. Jain. *Content-based image retrieval at the end of the early years*. IEEE Transactions on Pattern Analysis

and Machine Intelligence, vol. 22, no. 12, pages 1349–1380, 2000. (Cité en pages 8, 12 et 25.)

[Smith 2008] E.H.B. Smith. *Modeling image degradations for improving OCR.* In proceedings of the EUropean SIgnal Processing COnference (EUSIPCO), pages 1–5, 2008. (Cité en page 109.)

[Strecha 2012] C. Strecha, A.M. Bronstein, M.M. Bronstein et P. Fua. *LDAHash : improved matching with smaller descriptors.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 1, pages 66–78, 2012. (Cité en page 32.)

[Strehl 2003] A. Strehl et J. Ghosh. *Cluster ensembles - a knowledge reuse framework for combining multiple partitions.* The Journal of Machine Learning Research, vol. 3, pages 583–617, 2003. (Cité en pages 10 et 49.)

[Su 2009] Z. Su, Z. Cao et Y. Wang. *Stroke extraction based on ambiguous zone detection : a preprocessing step to recover dynamic information from handwritten Chinese characters.* International Journal on Document Analysis and Recognition, vol. 12, no. 2, pages 109–121, 2009. (Cité en page 48.)

[Sun 2009] Y. Sun, A.K. Wong et M.S. Kamel. *Classification of Imbalanced Data : a Review.* International Journal of Pattern Recognition and Artificial Intelligence, vol. 23, no. 4, pages 687–719, 2009. (Cité en page 98.)

[Sun 2012] Y. Sun. *Symmetry and Feature Selection in Computer Vision.* PhD thesis, University of California Riverside, 167 pages, 2012. (Cité en page 60.)

[Tabbone 2006] S. Tabbone, L. Wendling et J.P. Salmon. *A new shape descriptor defined on the Radon transform.* Computer Vision and Image Understanding, vol. 102, no. 1, pages 42–51, 2006. (Cité en pages 73, 79 et 158.)

[Tappert 1990] C.C. Tappert, C.Y. Suen et T. Wakahara. *State of the Art in On-Line Handwriting Recognition.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 8, pages 787–808, 1990. (Cité en page 90.)

[Tay 2001] Y.H. Tay, M. Khalid, P.M. Lallican, S. Knerr et C. Viard-Gaudin. *An Analytical Handwritten Word Recognition System with Word-level Discriminant Training.* In proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 726–730, 2001. (Cité en page 170.)

[Tefas 2001] A. Tefas, C. Kotropoulos et I. Pitas. *Using Support Vector Machines to Enhance the Performance of Elastic Graph Matching for Frontal Face Authentication.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, pages 735–746, 2001. (Cité en page 155.)

[Tombre 2006] K. Tombre, S. Tabbone et P. Dosch. *Musings on symbol recognition.* In Graphics Recognition. Ten Years Review and Future Perspectives, pages 23–34. Lecture Notes in Computer Science (LNCS), 2006. (Cité en page 72.)

[Trupin 2005] É. Trupin. *La reconnaissance d'images de documents : Un panorama.* Traitement du signal, vol. 22, no. 3, pages 159–189, 2005. (Cité en pages 11, 67 et 68.)

[Tsoumakas 2007] G. Tsoumakas et I. Katakis. *Multi-label classification : An overview.* International Journal of Data Warehousing and Mining (IJDWM), vol. 3, no. 3, pages 1–13, 2007. (Cité en page 85.)

[Tsoumakas 2010] G. Tsoumakas, I. Katakis et I. Vlahavas. Data mining and knowledge discovery handbook, chapitre Mining multi-label data, pages 677–685. Springer, 2010. (Cité en page 10.)

[Turk 1991] M.A. Turk et A.D. Pentland. *Eigenfaces for Recognition.* Journal of Cognitive Neuroscience, vol. 3, pages 71–86, 1991. (Cité en pages 25 et 154.)

[Van de Sande 2010] K.E.A. Van de Sande, T. Gevers et C.G.M. Snoek. *Evaluating Color Descriptors for Object and Scene Recognition.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pages 1582–1596, 2010. (Cité en page 37.)

[Van der Maaten 2009] L.J.P Van der Maaten, E.O Postma et H.J. Van den Herik. *Dimensionality reduction : A comparative review.* Journal of Machine Learning Research, vol. 10, no. 1-41, pages 66–71, 2009. (Cité en pages 27 et 76.)

[Vapnik 1998] V. Vapnik. Statistical Learning Theory. John Wiley & Sons, Inc., New York, 768 pages, 1998. (Cité en pages 2, 97, 99 et 174.)

[Varga 2003a] T. Varga et H. Bunke. *Effects of Training Set Expansion in Handwriting Recognition Using Synthetic Data.* In proceedings of the Conference of the International Graphonomics Society, pages 200–203, 2003. (Cité en page 107.)

[Varga 2003b] T. Varga et H. Bunke. *Generation of Synthetic Training Data for an HMM-based Handwriting Recognition System.* In proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 618–622, 2003. (Cité en page 107.)

[Viola 2001] P. Viola et M. Jones. *Rapid Object Detection using a Boosted Cascade of Simple Features.* In proceedings of the IEEE international conference on Computer Vision and Pattern Recognition (CVPR), volume 1, pages 511–518, 2001. (Cité en page 155.)

[Visani 2004a] M. Visani, C. Garcia et J.M. Jolion. *Two-Dimensional-Oriented Linear Discriminant Analysis for Face Recognition.* In proceedings of the International Conference on Computer Vision and Graphics (ICCVG), pages 1008–1017, 2004. (Cité en pages 63 et 155.)

[Visani 2004b] M. Visani, C. Garcia et C. Laurent. *Comparing Robustness of Two-Dimensional PCA and Eigenfaces for Face Recognition.* In proceedings of the International Conference on Image Analysis and Recognition (ICIAR), volume 2, pages 717–724, 2004. (Cité en page 63.)

[Visani 2005a] M. Visani. *Vers de nouvelles approches discriminantes pour la reconnaissance automatique de visages.* Thèse de doctorat, Institut National des Sciences Appliquées de Lyon, France, 229 pages, 2005. (Cité en pages 15, 25, 155, 156 et 173.)

[Visani 2005b] M. Visani, C. Garcia et J.M. Jolion. *Bilinear Discriminant Analysis for Face Recognition.* In proceedings of the International Conference on Advances in Pattern Recognition (ICAPR), volume 2, pages 247–256, 2005. (Cité en pages 63 et 156.)

[Visani 2005c] M. Visani, C. Garcia et J.M. Jolion. *Face Recognition using Modular Bilinear Discriminant Analysis.* In proceedings of the international conference on Visual Information Systems (VIS), pages 24–34, 2005. (Cité en pages 63 et 156.)

[Visani 2005d] M. Visani, C. Garcia et J.M. Jolion. *Normalized Radial Basis Function Networks and Bilinear Discriminant Analysis for Face Recognition.* In proceedings of the IEEE international conference on Advanced Video and Signal based Surveillance (AVSS), pages 342–347, 2005. (Cité en pages 63 et 156.)

[Visani 2005e] M. Visani, C. Garcia et J.M. Jolion. *Une Nouvelle Méthode de Représentation des Visages pour leur Reconnaissance : l'Analyse Discriminante Bilinéaire.* In actes de la conférence COmpression et REprésentation des Signaux Audiovisuels (CORESA), pages 103–108, 2005. (Cité en pages 63 et 156.)

[Visani 2006] M. Visani, C. Garcia et C. Laurent. *Method for Recognising Faces by means of a two-dimensional Linear Discriminant Analysis.* Brevet international (PCT) numéro PCT/FR2004/001395, 2006. (Cité en pages 63 et 155.)

[Visani 2011a] M. Visani, K. Bertet et J.M. Ogier. *Navigala : an Original Symbol Classifier Based on Navigation through a Galois Lattice.* International Journal of Pattern Recognition and Artificial Intelligence, vol. 25, no. 4, pages 449–473, 2011. (Cité en pages 77, 78, 80, 125 et 175.)

[Visani 2011b] M. Visani, O. Ramos et S. Tabbone. *A Protocol to Characterize the Descriptive Power and the Complementarity of Shape Descriptors.* International Journal on Document Analysis and Recognition, vol. 14, no. 11, pages 87–100, 2011. (Cité en pages 27, 63, 110 et 175.)

[Visani 2012] M. Visani, Q.A. Bui et S. Prum. *On-line cursive handwriting characterization using TF-IDF scores of graphemes.* In short papers proceedings of the IAPR international workshop on Document Analysis Systems (DAS), pages 20–21, 2012. (Cité en pages 101 et 125.)

[Visani 2013] M. Visani, V.C. Kieu, A. Fornés et N. Journet. *The ICDAR 2013 Music Scores Competition : Staff Removal.* In proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pages 1439–1443, 2013. (Cité en pages 111 et 125.)

[Vishwanathan 2010] N.N. Vishwanathan, R.K. Schraudolph et Karsten M.B. *Graph Kernels.* Journal of Machine Learning Research, pages 1201–1242, 2010. (Cité en page 153.)

[von Luxburg 2012] U. von Luxburg, B. Williamson, I. Guyon*et al. Clustering : Science or art ?* Journal of Machine Learning Research, vol. 27, pages 65–80, 2012. (Cité en page 58.)

[Vondrick 2013] C. Vondrick, A. Khosla, T. Malisiewicz et A. Torralba. *HOGgles : Visualizing Object Detection Features.* In proceedings of the International Conference on Computer Vision (ICCV), pages 1–8, 2013. (Cité en page 26.)

[Wagstaff 2001] K. Wagstaff, C. Cardie, S. Rogers et S. Schrödl. *Constrained K-means Clustering with Background Knowledge.* In proceedings of the International Conference on Machine Learning (ICML), pages 577–584, 2001. (Cité en page 33.)

[Wang 2012] J. Wang, S. Kumar et S.F. Chang. *Semi-supervised hashing for large scale search.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 12, pages 2393–2406, 2012. (Cité en page 32.)

[Witten 2011] I.H. Witten et E. Frank. Data mining : Practical machine learning tools and techniques. Morgan Kaufmann, 664 pages, 2011. (Cité en page 2.)

[Wolf 2004] C. Wolf et J.M. Jolion. *Extraction and recognition of artificial text in multimedia documents.* Formal Pattern Analysis & Applications, vol. 6, no. 4, pages 309–326, 2004. (Cité en page 120.)

[Wu 2004] T.F. Wu, C.J. Lin et R.C. Weng. *Probability Estimates for Multi-class Classification by Pairwise Coupling.* Journal of Machine Learning Research, vol. 5, pages 975–1005, 2004. (Cité en page 98.)

[Wulsin 2012] D. Wulsin, S. Jensen et B. Litt. *A hierarchical Dirichlet process model with multiple levels of clustering for human EEG seizure modeling.* In proceedings of the International Conference on Machine Learning (ICML), pages 1–8, 2012. (Cité en page 61.)

[Xie 2002] Z. Xie, W. Hsu, Z. Liu et M.L. Lee. *Concept lattice based composite classifiers for high predictability*. Journal of Experimental & Theoretical Artificial Intelligence, vol. 14, no. 2-3, pages 143–156, 2002. (Cité en page 76.)

[Xie 2013] L. Xie, Z. Deng et S. Cox. *Special issue on multimodal joint information processing in human machine interaction : recent advances*. Multimedia Tools and Applications, 2013. (Cité en page 2.)

[Yang 2006] L. Yang et R. Jin. *Distance metric learning : A comprehensive survey*. Technical report, 51 pages, Michigan State Universiy, 2006. (Cité en page 41.)

[Yao 2012] B. Yao, G. Bradski et L. Fei-Fei. *A codebook-free and annotation-free approach for fine-grained image categorization*. In proceedings of the IEEE international conference on Computer Vision and Pattern Recognition (CVPR), pages 3466–3473, 2012. (Cité en page 120.)

[Yoo 2004] T.S. Yoo. Insight into images : principles and practice for segmentation, registration, and image analysis, volume 203. Wesllesley Massachusetts, AK Peters, 393 pages, 2004. (Cité en page 2.)

[Yu 2003] H. Yu, J. Yang et J. Han. *Classifying large data sets using SVMs with hierarchical clusters*. In proceedings of the ACM SIGKDD international conference on Knowledge Discovery and Data mining, pages 306–315, 2003. (Cité en page 99.)

[Yuan 2007] J. Yuan, Y. Wu et M. Yang. *Discovery of collocation patterns : from visual words to visual phrases*. In proceedings of the IEEE international conference on Computer Vision and Pattern Recognition (CVPR), pages 1–8, 2007. (Cité en page 61.)

[Zenou 2004] E. Zenou et M. Samuelides. *Utilisation des treillis de Galois pour la caractérisation d'ensembles d'images*. In actes de la conférence Reconnaissance de Formes et Intelligence Artificielle (RFIA), pages 395–404, 2004. (Cité en page 76.)

[Zhai 2003] J. Zhai, L. Wenyin, D. Dori et Q. Li. *A line drawings degradation model for performance characterization*. In proceedings of the International Conference on Document Analysis and Recognition (ICDAR), volume 2, pages 1020–1020, 2003. (Cité en page 109.)

[Zhang 1996] T. Zhang, R. Ramakrishnan et M. Livny. *BIRCH : an efficient data clustering method for very large databases*. ACM SIGMOD Record, vol. 25, no. 2, pages 103–114, 1996. (Cité en page 37.)

[Zhang 2011] Y. Zhang, Z. Jia et T. Chen. *Image retrieval with geometry-preserving visual phrases*. In proceedings of the IEEE international conference on Computer Vision and Pattern Recognition (CVPR), pages 809–816, 2011. (Cité en page 61.)

[Zhang 2012] D. Zhang, M.M. Islam et G. Lu. *A review on automatic image annotation techniques*. Pattern Recognition, vol. 45, no. 1, pages 346 – 362, 2012. (Cité en pages 10, 31, 33, 56 et 121.)

[Zhou 2003] X.S. Zhou et T.S. Huang. *Relevance feedback in image retrieval : A comprehensive review*. Multimedia Systems, vol. 8, no. 6, pages 536–544, 2003. (Cité en page 31.)

[Zhou 2012] Z.H. Zhou, M.L. Zhang, S.J. Huang et Y.F. Li. *Multi-instance multi-label learning*. Artificial Intelligence, vol. 176, no. 1, pages 2291–2320, 2012. (Cité en page 121.)

Annexes

Annexe A

Conception de descripteurs visuels dédiés à un certain type d'images

A.1 Introduction

De nombreux chercheurs font encore la distinction entre deux grands types d'approches en reconnaissance de forme : les approches statistiques et les approches structurelles, comme en témoigne la pérennisation des congrès internationaux SPR (Statistical techniques on Pattern Recognition) et SSPR (Structural and Syntactic Pattern Recognition), qui se tiennent de manière conjointe tous les deux ans en marge d'ICPR. Tandis que les approches statistiques se basent généralement sur une représentation des formes de l'image par le biais d'un vecteur, les approches structurelles, elles, représentent les formes de manière séquentielle ou structurée (chaînes, graphes, etc.) [Haralick 1979]. Ces dernières sont donc réputées plus adaptées aux images structurées. Leurs sorties sont de manière générale plus faciles à interpréter par un humain. Cependant, les approches purement structurelles souffrent le plus souvent d'une complexité calculatoire importante et de performances réduites en comparaison avec les approches statistiques. Ces deux types d'approches étant complémentaires, la tendance actuelle consiste à les faire collaborer [Arrivault 2005], voire à les combiner [Delaye 2008] au sein d'une même signature afin de tirer avantage de leurs qualités respectives. Parmi les travaux les plus emblématiques de ce rapprochement, on peut citer les approches structurelles exploitant des noyaux - empruntés aux méthodes issues de l'apprentissage statistique - sur des graphes [Vishwanathan 2010, Bunke 2011]. Une distinction ferme entre approches statistiques et structurelles semble donc à l'heure actuelle un peu dépassée dès lors que l'on s'intéresse à un système d'analyse d'images global, même si cette distinction garde une certaine pertinence lorsque l'on restreint le propos à la description des images proprement dite, et en particulier à l'extraction de descripteurs d'images.

On peut donc distinguer les descripteurs d'images statistiques des descripteurs structurels. Tandis que les premiers se présentent sous la forme d'un vecteur obtenu grâce à une étude des distributions radiométriques et spatiales des pixels des images, les seconds permettent de représenter les formes de l'image de manière séquentielle ou structurée, en se basant sur une analyse de l'organisation spatiale d'éléments d'intérêt (« primitives ») dont sont composées les images.

Comme expliqué en section 2.1.1, la littérature regorge de descripteurs (essentiellement statistiques) visant à décrire les images tout-venant selon leur apparence visuelle. La procédure d'extraction d'un descripteur visuel à partir d'une image engendre par définition des pertes d'information par rapport au signal original. Quelle que soit la tâche visée au final, les signatures composées à partir de ces descripteurs doivent en contrepartie vérifier un certain nombre de « bonnes propriétés ». En particulier, elles doivent être le plus **compactes** possible (notamment afin de restreindre le cas échéant les difficultés liées à la malédiction de la dimensionnalité

lors de l'usage ultérieur qui en sera fait), et **invariantes** vis-à-vis d'un certain nombre de transformations ou de bruits présents dans l'image. La liste des invariances souhaitées dépend de l'application visée et conditionne en grande partie le choix des descripteurs à extraire de l'image. Outre ces propriétés de compacité et d'invariance, on cherche souvent à ce que les signatures de deux images de contenus différents (au sens de ce que les descripteurs cherchent à décrire, p. ex. forme, texture, couleur) soient différentes. Cette propriété (parfois appelée « **capacité de discrimination** ») est particulièrement recherchée lorsque la tâche visée au final est basée sur la comparaison des signatures de différentes images. Une bonne capacité de discrimination sera donc recherchée, à des degrés divers, pour la détection de copies, le *clustering*, la navigation dans des bases d'images, la reconnaissance ou la recherche d'images. Afin de permettre une comparaison efficace des signatures de différentes images, celles-ci sont en général introduites conjointement avec une mesure de similarité adaptée à cette signature [Jolion 2001] (mesure spécialement conçue ou, du moins, en adéquation avec ses propriétés).

Parfois, notamment dans le cas d'applications dédiées à un type d'images en particulier, nous disposons explicitement ou implicitement d'informations du domaine, et en particulier d'informations *a priori* sur la nature et/ou l'agencement spatial des éléments d'intérêt composant l'image ou le motif. Il peut donc être utile de prendre en compte cette information pour définir des descripteurs d'images adaptés spécifiquement à ce type d'images. Comme évoqué en section 1.3.1.2 (« Images considérées »), c'est le cas des deux types d'images que nous considérons dans la suite de cette annexe : les images de visages et les images de documents techniques.

Considérons d'abord le cas des visages. Malgré leurs très grandes variations d'apparence, ils sont tous composés de caractéristiques faciales (coin des yeux, iris, yeux, nez, bouche, etc.) agencées physiquement et logiquement de manière similaire. En ce sens, nous aborderons en section A.2 le problème de leur description sous l'angle de l'extraction de descripteurs par projection statistique à partir d'images de « structure figée ».

Considérons maintenant le cas de symboles architecturaux issus d'images de documents binaires ou binarisés et composés de primitives tels que des segments ou des arcs de cercle. Dans le cas qui nous intéresse, la liste des primitives possibles est connue, et leur apparence complètement déterminée par quelques paramètres. À la différence des visages cependant, l'agencement spatial des primitives varie selon le symbole (c'est d'ailleurs ce qui nous permet de distinguer les différents symboles). Nous pouvons donc considérer que les images de symboles sont très structurées (mais de structure non figée), et je présenterai en section A.3 une signature statistico-structurelle que nous avons conçue pour ce cas spécifique.

A.2 Conception de descripteurs statistiques pour des images de structure figée

Nous considérons que les images de visages sont de structure figée, dans le sens où les visages sont tous composés de caractéristiques faciales (yeux, nez, bouche, etc.) agencés de manière similaire quel que soit l'individu. Bien sûr, l'apparence de ces caractéristiques faciales (couleur, forme, texture, etc.) varie en fonction de l'individu. Cependant, lorsque l'on cherche à décrire ou à reconnaître des images de visages, la prise en compte explicite ou implicite de la présence de cette structure dans les images de visages peut être utile. Par exemple, la technique très réputée des *eigenfaces*, proposée par Turk et Pentland en 1991 dans [Turk 1991] et qui consiste à appliquer une ACP (Analyse en Composantes Principales) sur une basse

d'entraînement en niveaux de gris et à décrire chaque visage par les coefficients de sa projection sur le sous-espace principal, ne fonctionne que parce que tous les visages ont à peu près la même structure (logiquement et physiquement), comme on peut le voir facilement dans la Figure A.1. Il en est de même pour la technique des *fisherfaces* [Belhumeur 1997], qui consiste à appliquer séquentiellement une ACP puis une ADL (Analyse Discriminante Linéaire). L'ADL permet d'obtenir des signatures plus compactes et plus discriminantes (puisqu'elle est basée sur un apprentissage supervisé), tout en augmentant leur invariance vis-à-vis des facteurs extérieurs à l'apparence des visages en eux-mêmes (changements d'illumination par exemple) qui affectent beaucoup les *eigenfaces*. Pour autant, ces techniques de projection statistique ne vont pas jusqu'à la prise en compte explicite de la structure des visages, défaut partiellement pallié par la technique des sous-espaces modulaires [Pentland 1994], qui est basée sur de multiples ACP locales réalisées autour de chacune des caractéristiques faciales. Il existe par contre des techniques structurelles prenant en compte de manière explicite la structure des visages et qui ont fait leurs preuves pour la description des visages, comme par exemple la méthode introduite dans [Tefas 2001].

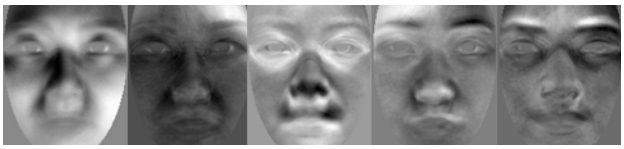


FIGURE A.1 – Les 5 premières *eigenfaces* (associées aux plus grandes valeurs propres) obtenues sur une sous-base de l'Asian Face Database PF01 [Hwang 2004], comptant 107 personnes et 4 vues par personne. Chaque *eigenface* correspond à un axe principal obtenu par application de l'ACP sur les vecteurs obtenus par concaténation des pixels de l'image, et remis sous la forme d'une image pour leur visualisation. Chaque visage de la collection est ensuite décrit par ses coefficients de projection dans la base composée des *eigenfaces*. Autrement dit, chaque visage peut être considéré comme une combinaison linéaire des *eigenfaces* extraites de la base, et le vecteur composé de ces coefficients constitue son descripteur. *Figure extraite de [Visani 2005a].*

Dans le cadre de ma thèse [Visani 2005a], j'ai proposé des descripteurs par projection statistique spécifiques pour la description de visages (en niveaux de gris), préalablement détectés [Viola 2001, Garcia 2004] et normalisés en utilisant une localisation des caractéristiques faciales [Duffner 2005]. Étant donnés ces pré-traitements, les signatures qui en dérivent ne doivent être que partiellement invariantes vis-à-vis des translations, rotations dans le plan de l'image et changements d'échelle. En revanche, ces signatures doivent être dotées d'une certaine robustesse vis-à-vis de changements de pose de la tête (hors du plan de l'image), de variations dans l'expression faciale et d'occultations (port de lunettes, moustache, cache-nez, etc.).

Tout comme la technique des *fisherfaces*, les descripteurs que j'ai proposés sont basées sur une ADL, ce qui leur confère par nature un caractère discriminant. À la différence de la technique des *fisherfaces* qui considère l'image de visage en entrée comme un immense vecteur composé de la concaténation des lignes de pixels, les descripteurs que j'ai proposés prennent mieux en compte la structure bidimensionnelle des images de visages.

Plus précisément, j'ai proposé un descripteur appelé « Analyse Discriminante Linéaire 2D-orientée » (ADL2Do) qui revient (en résumé) à appliquer une ADL sur les lignes (respectivement les colonnes) des images de visages [Visani 2004a, Visani 2006]. Ce descripteur présente plusieurs

avantages par rapport au descripteur issu des *fisherfaces*. Tout d'abord, en appliquant l'analyse multimensionnelle des données sur les lignes (ou les colonnes) et non sur les très grands vecteurs image issus d'une concaténation des lignes de pixels, on contourne le problème de la singularité de la matrice de covariance intra-classes (problème lié à la malédiction de la dimensionnalité), et ainsi il n'est pas nécessaire d'appliquer préalablement à l'ADL une technique de réduction de dimension (comme par exemple l'ACP dans le cas des *fisherfaces*). Le coût et l'instabilité lors de l'extraction des descripteurs est également réduite par rapport à la plupart des descripteurs par projection statistique basées sur une ADL de l'état de l'art [Visani 2005a]. Quand on l'applique à la reconnaissance de visages avec une simple stratégie au plus proche voisin et une distance Euclidienne, la signature issue de l'ADL2Do donne de meilleurs résultats que les *fisherfaces*, et ce, qu'elle soit appliquée en lignes ou en colonnes. Par contre, la taille du descripteur est plus importante qu'avec la plupart de ces techniques.

Une analyse poussée des matrices de confusion obtenues durant l'analyse des performances en reconnaissance a montré que les deux versions de l'ADL2Do (en ligne et en colonne) ont un comportement complémentaire, à savoir qu'il n'est pas rare qu'un visage soit correctement reconnu avec une signature composée de l'un des deux descripteurs, mais pas avec l'autre.

J'ai donc proposé un autre descripteur, que j'ai nommé « Analyse Discriminante Bilinéaire » (ADB), et qui peut être considéré comme réellement bidimensionnel, puisqu'il combine efficacement les deux versions (en ligne et en colonne) de l'ADL2Do. Deux algorithmes itératifs basés sur une application alternative des deux versions de l'ADL2Do ont été détaillés dans [Visani 2005b, Visani 2005e], dont un qui permet de déterminer automatiquement le nombre de vecteurs propres. Si elle n'est pas prouvée formellement, la convergence est en pratique atteinte au bout d'un nombre limité d'itérations, comme nous l'avons montré dans [Visani 2005a] à l'issue d'une campagne d'évaluation intensive. La signature issue de ce descripteur est de taille très réduite en comparaison avec celle obtenue par ADL2Do ; sa taille est comparable à celle de la plupart des signatures par projection statistique de l'état de l'art basées sur une ADL. L'ADB est également plus stable selon ses paramètres, et robuste vis-à-vis de changements de pose, de variations dans l'expression faciale et d'occultations que l'ADB et la plupart des signatures existantes basées sur une ADL (pour plus de détails merci de se référer à la section 5.2. de [Visani 2005a]).

Du fait de la nature même de cette signature, il n'est pas surprenant que la distance Euclidienne soit particulièrement adaptée pour la comparaison de signatures ADB, ce que j'ai vérifié expérimentalement dans la section 5.2.3. de ma thèse. Lorsqu'elle est combinée avec un réseau de fonctions à base radiale normalisée pour une application finale de reconnaissance, cette signature permet d'obtenir des taux d'identification en monde ouvert (avec rejet) compétitifs avec les autres approches basées sur une ADL et proposées jusque-là [Visani 2005d].

Puisque l'on dispose de la localisation des caractéristiques faciales [Duffner 2005] et afin de prendre en compte de manière plus explicite la structure des visages, j'ai également proposé dans [Visani 2005c] une version modulaire de l'ADB, à savoir l'ADBM (Analyse Discriminante Bilinéaire Modulaire). Plusieurs ADB sont extraites depuis différentes régions faciales (et depuis la totalité du visage), puis combinées dans un but de reconnaissance. Les résultats montrent une robustesse accrue, en particulier vis-à-vis de changements d'expression faciales et d'occultations, en compensation d'une complexité calculatoire plus importante lors de la phase d'extraction de la signature.

Les descripteurs que nous avons proposés dans le cadre de ma thèse et brièvement décrites

ci-dessus permettent de prendre en compte la structure bidimensionnelle des images de visages, à la différence de la plupart des descripteurs basés sur une projection statistique et proposés jusque-là. Outre les visages, ces descripteurs pourraient être utilisés pour tout autre type d'objets (2D ou 3D) dont la structure est figée, comme par exemple la technique des *eigenfaces* qui a été réutilisée pour la reconnaissance de paumes de la main dans [Lu 2003].

Cependant, de par la nature statistique de ces descripteurs, le niveau de prise en compte de la structure des images n'atteint pas celui des descripteurs structurels. La section suivante présente brièvement une signature statistico-structurelle dédié à la description d'images de symboles techniques.

A.3 Conception d'une signature statistico-structurelle pour des images très structurées

Tout comme les visages, les symboles extraits de documents techniques sont composés d'éléments d'intérêt. On peut cependant constater deux différences majeures avec le cas des visages. Premièrement, dans le cas où l'alphabet des symboles est fixé et où ces symboles sont imprimés et monochromes, ces éléments d'intérêt appartiennent à une liste fermée (lignes, arcs de cercle, etc.), comme on peut le voir en Figure A.2 avec un extrait de la base GREC 2003 [1] que nous utilisons comme cas d'usage. De plus, l'apparence de ces éléments d'intérêt est beaucoup moins complexe que celles des caractéristiques faciales, et est complètement déterminée par quelques paramètres (largeur, point de départ, point d'arrivée et orientation dans le cas d'un segment de ligne par exemple). On parle alors de « primitives » de l'image. Deuxièmement, les relations topologiques entre primitives, voire leur agencement spatial, sont variables et permettent généralement de distinguer facilement des symboles différents.

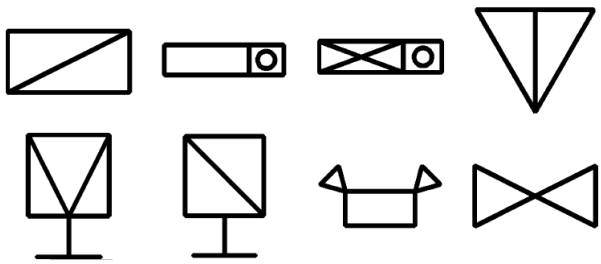


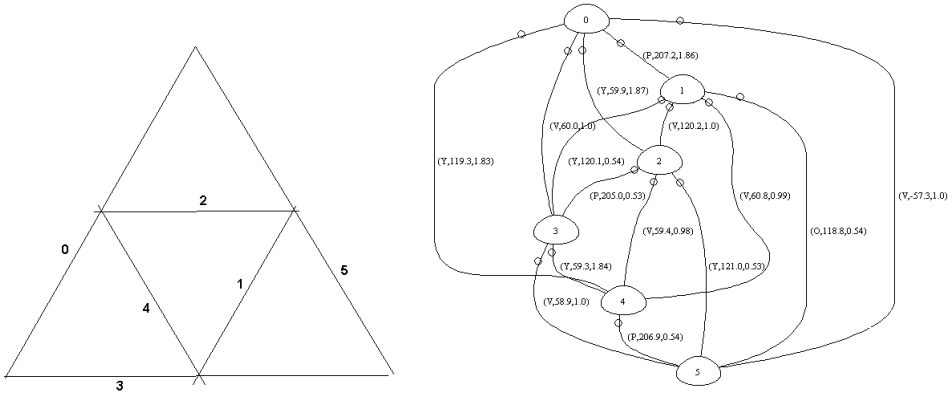
FIGURE A.2 – Symboles extraits de la base GREC 2003, utilisée pour notre étude.

Dans le cadre du stage de Master 2 de Mickaël Coustaty, nous nous sommes intéressés à la description et à la reconnaissance de symboles techniques. Étant donné que ces symboles peuvent se retrouver à différentes échelles et sous différents angles dans les documents techniques, les signatures visant à les décrire doivent être invariantes à la rotation et à l'échelle. Nous avons proposé un descripteur structurel qui n'est pas basé sur un apprentissage, mais plutôt sur le calcul d'un graphe permettant de caractériser les relations topologiques entre segments de ligne (préalablement détectés en utilisant une transformée de Hough spécialement adaptée au cas des segments) [Coustaty 2008]. Chaque nœud du graphe correspond à un segment, chaque arc à la description du type de relation entre deux segments. Les arcs sont valués par :

1. www.cvc.uab.es/grec2003/SymRecContest/index.htm

– le type de relation (en X, en Y, en V, « parallèle » (P) ou « autres » (O)) ;
– une valeur caractéristique de la relation (distance entre segments pour les relations « parallèle », angle entre segments pour tous les autres types de relations) ;
– le ratio des longueurs des deux segments.

Un exemple de l'ensemble des segments extraits d'un symbole et du graphe topologique associé est donné en Figure A.3.



(a) (b)

FIGURE A.3 – Exemple des segments extraits d'un symbole graphique (a), et du graphe topologique associé (b). On remarque le bruit vectoriel sur les segments extraits des symboles, bruit lié au processus de détection. *Figure adaptée de [Coustaty 2011].*

Ce graphe topologique permet de décrire fidèlement les symboles de manière invariante aux changements d'échelle et aux rotations, mais d'une manière peu compacte, comme c'est souvent le cas des descripteurs structurels. Afin de pouvoir facilement et plus rapidement comparer deux symboles différents dans un but de reconnaissance, nous en avons dérivé une signature statistico-structurelle basée sur la description des différents chemins que l'on peut extraire de ce graphe, en considérant la matrice d'adjacence M [2] du graphe et ses puissances M^n, où $n \in \mathbb{N}^*$. On obtient ainsi des « sacs de chemins » qui peuvent être appréhendés par un expert humain (car de niveau sémantique intermédiaire), ou comparés automatiquement en utilisant une simple mesure de similarité ou un classifieur plus raffiné. Dans [Coustaty 2011], nous avons montré la relative efficacité de la signature issue de ce descripteur quand elle est utilisée conjointement avec la méthode de classification symbolique Navigala basée sur un treillis des concepts [3], malgré le fait que les performances enregistrées restent très en-deçà de celles atteintes avec une signature statistique telle que celle de Radon [Tabbone 2006] par exemple. La Figure A.4 montre une vue d'un treillis construit à partir de cette signature et d'un sous-ensemble de symboles de la base GREC 2003, et illustrent la lisibilité de la combinaison de cette signature structurelle avec un treillis, laissant entrevoir l'intérêt d'approches de navigation dans la base de symboles plutôt que de reconnaissance.

La signature issue de ce descripteur présente deux originalités majeures par rapport aux autres signatures structurelles ou statistico-structurelles précédemment proposées pour la reconnaissance de symboles [Llados 2003, Bunke 2011], et même par rapport à des signatures

2. Cette matrice d'adjacence est calculée après discrétisation des deux attributs numérique servant à valuer les arcs.
3. Plus de détails sur ce classifieur sont donnés en section 3.2.

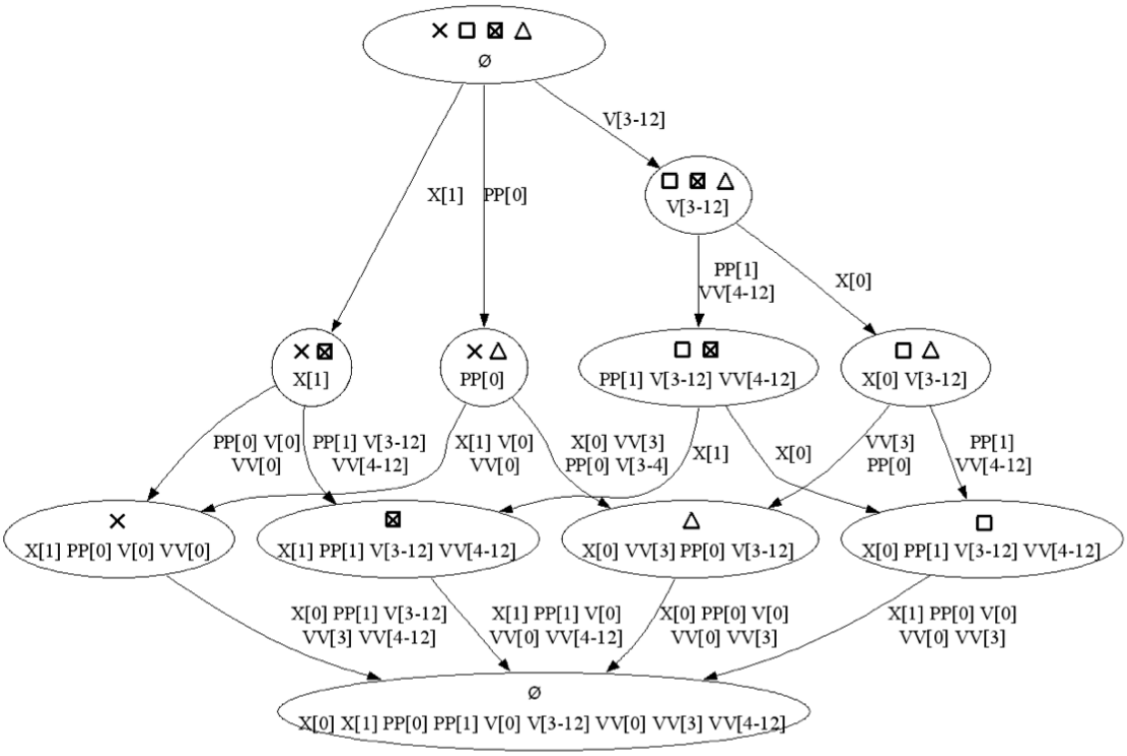


FIGURE A.4 – Exemple de treillis généré à partir de la signature proposée et de la méthode Navigala (sur un sous-ensemble de la base GREC 2003). *Figure extraite de [Coustaty 2011].*

plus récemment proposées [Li 2013]. Premièrement, plutôt que de rechercher des formes prédéfinies (triangles, rectangles, etc.) dans les images de symboles, la signature se base sur une description dynamique de ses primitives plus élémentaires (ici des segments de lignes). Cela confère à la signature une meilleure capacité d'adaptation vis-à-vis de nouveaux symboles, et une certaine robustesse, en particulier vis-à-vis de bruits vectoriels. Deuxièmement, le fait de nous ramener à une représentation numérique par le biais de « sacs de chemins » nous permet de contourner les difficultés liées à la comparaison de signatures structurelles, tout en conservant un niveau intermédiaire de sémantique intelligible pour un humain expert du domaine.

En revanche, afin de décrire pleinement des symboles techniques plus complexes, il conviendrait bien évidemment de prendre en considération d'autres types de primitives et leurs relations comme par exemple les arcs de cercle. Cela demanderait plusieurs changements des graphes topologiques utilisés. En particulier, les nœuds deviendraient valués par le type de primitive, et les attributs des relations entre, d'une part, ces nouvelles primitives et, d'autre part, les anciennes primitives (segments) devraient être définis. Cela est possible, mais pas évident, et les possibilités sont nombreuses. C'est pourquoi nous avons choisi pour l'instant de considérer ce descripteur comme un simple descripteur de l'agencement entre segments de droite, quitte à la combiner avec d'autres descripteurs statistiques et/ou structurels permettant de décrire les autres propriétés du symbole au sein d'une signature décrivant de manière plus exhaustive l'apparence visuelle du symbole graphique.

La robustesse de notre approche vis-à-vis du bruit (assez limitée du fait notamment de notre méthode de détection des segments de ligne) pourrait être améliorée en caractérisant de manière floue les relations spatiales entre primitives, en particulier dans le cas de la relation « autres » entre segments. Cela pourrait se faire par exemple au travers de l'utilisation d'ontologies de relations spatiales floues [Hudelot 2008].

De plus, de nombreuses possibilités s'ouvrent à nous quant à la manière de définir les « sacs de chemins ». Depuis la publication de nos travaux, d'autres auteurs se sont penchés sur des problématiques assez similaires [Liu 2014] et les pistes en ce sens sont nombreuses.

A.4 Discussion

Ces travaux sur l'extraction de descripteurs dédiés à un certain type d'images, qui ne sont pas basés sur une catégorisation d'images, sortent un peu du cadre de ce manuscrit. Néanmoins, ils sont en lien avec l'un des verrous que nous avons cherché à attaquer au cours de nos recherches, à savoir la manière d'intégrer efficacement dans le système des informations de nature potentiellement variable. Ici on dispose, d'une part, du contenu pixellaire des images de la collection (qui est homogène puisque composée d'un unique type d'images) et, d'autre part, d'une information concernant la structuration de ces images. Bien que cette information du domaine soit non formalisée explicitement, nous avons cherché à la prendre en compte lors de la conception des signatures. On obtient ainsi des « descripteurs de grain fin », dédiés au type d'images de la collection.

En revanche, ces descripteurs sont par essence peu génériques. Si ceux que nous avons proposés dans le cas des images de visages peuvent dans une certaine mesure être réutilisés sur d'autres collections homogènes représentant des images de structure figée, le descripteur proposé pour les symboles graphiques ne peut dans son état actuel prétendre décrire seul et de manière efficace d'autres types de motifs.

C'est l'une des raisons principales qui nous ont poussé à suspendre en 2009 nos travaux sur la conception de descripteurs visuels d'images dédiés à un type d'images en particulier, pour nous tourner la conception de descripteurs capables de caractériser des contenus plus variés (voir chapitre 2).

Il n'empêche que ces descripteurs dédiés se révèlent très utiles dans des applications de reconnaissance, dès lors que l'on est en mesure de détecter ce type d'objets dans les images, comme illustré au chapitre 3.

Annexe B

Présentation de la méthode de *clustering* semi-supervisé interactif proposée

B.1 Introduction

En raison de ses bonnes propriétés et de ses excellentes performances en pratique (voir section 2.2.3.1), nous avons sélectionné l'algorithme BIRCH pour l'étendre au contexte semi-supervisé et interactif de la section 2.2 du présent manuscrit. Dans cette annexe, je commence par présenter brièvement cet algorithme, de manière à pouvoir par la suite détailler puis discuter brièvement des améliorations possibles de la méthode de *clustering* semi-supervisé interactif que nous avons conçue.

B.2 Algorithme BIRCH

L'algorithme BIRCH est un algorithme de *clustering* adapté aux grandes bases de données de grandes dimensions. Cette méthode permet de construire une structure de données hiérarchique sous la forme d'un *Clustering Feature Tree* (arbre CF) illustré en Figure B.1. Dans cet arbre, chaque nœud contenant M enregistrements décrits au travers de leurs signatures $\{x_i | i = 1, ..., M\}$ est lui-même décrit par le « vecteur CF [1] » suivant :

$$CF = (M, LS, SS) \tag{B.1}$$

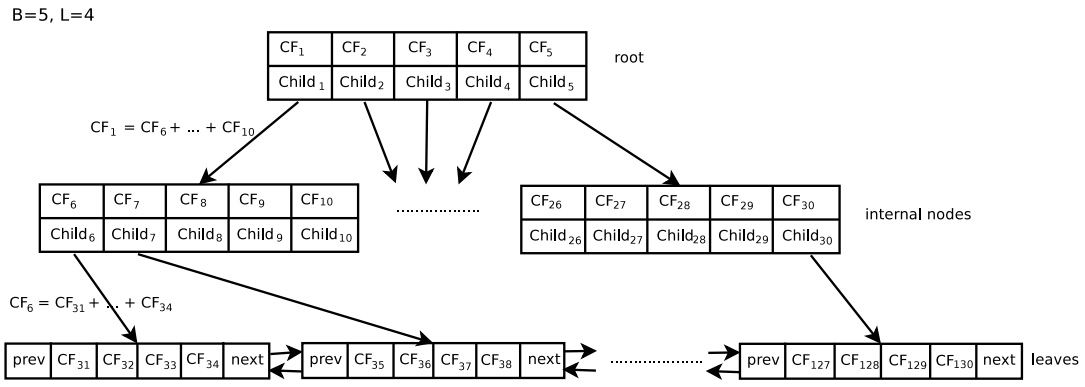


Figure B.1 – Exemple d'un arbre CF construit avec l'algorithme BIRCH et les valeurs de paramètres $B = 5$ et $L = 4$. *Figure extraite de [Lai 2013a].*

1. Cette représentation permet d'obtenir facilement certaines statistiques basiques telles que la moyenne, le rayon et le diamètre moyen de chaque nœud ainsi que différentes distances Euclidiennes entre deux nœuds. De plus, le vecteur CF d'un nœud interne peut être très rapidement calculé à l'aide des vecteur CF de ses descendants.

où LS et SS sont respectivement la somme et la somme au carré des signatures des M enregistrements de ce nœud ($LS = \sum_{i=1}^{M} x_i$; $SS = \sum_{i=1}^{M} x_i.x_i$).

La construction de l'arbre CF est dirigée par trois paramètres B, L et T déterminant la taille de l'arbre. En effet, un arbre CF doit respecter les contraintes suivantes :

- chaque nœud interne contient au plus B éléments $[CF_i, child_i]$, où CF_i est le vecteur CF du $i^{\text{ème}}$ nœud-fils, référencé par le pointeur $child_i$;
- chaque feuille contient au plus L éléments $[CF_i]$; elle contient également deux pointeurs *prev* et *next* permettant de relier les feuilles entre elles. Par ailleurs, toutes les feuilles sont situées au même niveau de l'arbre ;
- chaque vecteur CF d'une feuille doit avoir un diamètre inférieur à un seuil T.

L'arbre CF est créé par insertion incrémentale des données dans l'arbre. Un nouvel enregistrement est rajouté à la racine et descend dans l'arbre en sélectionnant à chaque étape le vecteur CF le plus proche parmi les nœuds fils du nœud courant, et ce, jusqu'à atteindre une feuille. L'enregistrement est alors rajouté à cette feuille, sous réserve que l'on n'enfreigne aucune des contraintes liées aux trois paramètres de l'algorithme. En cas d'infraction de l'une de ces contraintes, une nouvelle feuille est créée. Si l'infraction porte sur le paramètre L (la feuille contient déjà L éléments), alors la nouvelle feuille sera issue de la division de la feuille initiale en deux feuilles contenant moins de L enregistrements chacune. Les changements liés à cette division sont alors rétro-propagés vers la racine de l'arbre. Si par contre l'infraction porte sur le paramètre T, alors la nouvelle feuille ne contiendra que ce nouvel enregistrement, ce qui permet d'isoler les valeurs aberrantes afin de leur appliquer, le cas échéant, un traitement particulier.

Une fois l'arbre CF construit selon une heuristique itérative permettant d'ajuster automatiquement le paramètre T, nous obtenons en général majoritairement des feuilles relativement équilibrées en termes de nombre d'enregistrements, mais qui ne sont pas nécessairement représentatives des groupes d'images similaires (qui peuvent être déséquilibrés, comme nous l'avons évoqué plus haut). Nous pouvons utiliser n'importe quelle méthode pour procéder au *clustering* de ces feuilles, chaque feuille étant considérée comme un unique enregistrement décrit par son vecteur CF. La technique la plus couramment utilisée est celle des k-moyennes.

L'algorithme de *clustering* BIRCH dépend donc de deux paramètres B et L (auxquels il faut rajouter le paramètre T qui peut être déterminé automatiquement et, le cas échéant, le nombre k de *clusters* pour le *clustering* des feuilles), et est sensible à l'ordre de présentation des données. En revanche, il produit une structure hiérarchique des données et permet d'identifier facilement les valeurs aberrantes qui sont situées dans des feuilles presque vides et isolées. La complexité calculatoire très faible de sa construction ou de sa mise à jour, ainsi que son caractère incrémental, en font un outil adapté à notre contexte applicatif où les volumes de données à catégoriser peuvent être importants, mais où le traitement doit être rapide du fait de l'interactivité.

En pratique, c'est lui qui donne les meilleurs résultats en termes de *clustering* d'images au sens des mesure d'évaluation externes [2], quand il est utilisé conjointement avec une signature extraite à l'aide de l'algorithme rgSIFT et d'une discrétisation par sacs de mots (avec 200 mots).

2. Voir section 2.2.4.

B.3 Présentation détaillée de la méthode proposée

Nous avons introduit dans [Lai 2014b] une méthode basée sur une **extension semi-supervisée et interactive de l'algorithme BIRCH** et l'avons intégrée au système de *clustering* semi-supervisé interactif dont un aperçu a été donné en section 2.2.2. En re-définissant et en ré-arrangeant les feuilles de l'arbre CF en fonction des retours de l'utilisateur, **cette approche ré-organise localement la collection de données selon les souhaits de l'humain**.

Les contraintes entre paires d'images peuvent être de deux types : *must-link* ou *cannot-link*. Par exemple, dans la version la plus basique de notre algorithme, lorsque l'utilisateur déplace une image X d'un *cluster* C_1 vers un cluster C_2, l'image X devient liée par des contraintes *must-link* à l'image prototype de C_2 (sauf en cas de retour négatif sur cette image) et à toutes les images que l'utilisateur a étiquetées comme appartenant à C_2 lors de l'itération interactive courante (et le cas échéant à l'image prototype de C_2). De la même manière, cette image X devient liée par des contraintes *cannot-link* à toutes les images étiquetées par l'utilisateur comme C_1 (et le cas échéant à l'image prototype de C_1) lors de l'itération interactive courante.

L'approche proposée repose sur la notion de voisinage N_p, défini comme un regroupement d'images de taille intermédiaire qui nous permet de passer des contraintes ML et CL entre paires d'images (déduites des interactions de l'utilisateur, réalisées au niveau des images) à des contraintes ML_{CF} et CL_{CF} entre feuilles CF qui pourront directement être intégrées dans le processus de *clustering* semi-supervisé interactif. Elle est basée sur les étapes suivantes (détaillées dans les paragraphes suivants) :

1. Np, $CannotNp \leftarrow \{\Phi\}$;

2. Procéder au *clustering* initial (non supervisé) des images de la collection avec BIRCH ;

3. Tant que l'utilisateur n'est pas satisfait des *clusters* obtenus :

- **(a)** ML_{CF}, CL_{CF}, PF, NF, ML, $CL \leftarrow \{\Phi\}$;
- **(b)** Présenter c *clusters* et p images par *cluster* à l'utilisateur à l'aide de l'interface interactive présentée en Figure 2.3, et le laisser interagir librement jusqu'à ce qu'il décide d'arrêter l'itération interactive courante ;
- **(c)** Recevoir la liste PF d'images positives et NF d'images négatives données par l'utilisateur pour chacun des *clusters* avec lesquels il vient d'interagir ;
- **(d)** Mettre à jour la liste Np de voisinages et de leurs contraintes *cannot-link* par paires $CannotNp$, en fonction de PF et NF ;
- **(e)** En déduire les listes des contraintes par paires entre images ML (contraintes *must-link*) et CL (contraintes *cannot-link*) ;
- **(f)** En fonction de la liste Np de voisinages et de leurs contraintes *cannot-link* par paires $CannotNp$, diviser certaines feuilles de l'arbre CF, et mettre à jour l'arbre CF en fonction de ces divisions ;
- **(g)** Déduire des contraintes ML et CL entre images les listes ML_{CF} et CL_{CF} de contraintes (respectivement *must-link* et *cannot-link*) entre les paires de nouvelles feuilles de l'arbre CF ;
- **(h)** Appliquer l'algorithme de *clustering* semi-supervisé proposé, en intégrant comme information supervisée les contraintes par paires ML_{CF} et CL_{CF} entre feuilles CF, et en considérant chaque feuille CF comme un enregistrement d'entrée du *clustering*.

Le *clustering* initial **(étape 2)**, complètement non-supervisé, est mené sur les sacs de mots extraits des images en utilisant un descripteur rgSIFT (et 200 mots visuels). La structure hiérarchique des données (ici l'arbre CF) est construit en utilisant l'algorithme BIRCH dont un aperçu est donné ci-avant, et les *clusters* sont obtenus à l'aide d'un regroupement des feuilles CF de l'arbre selon l'algorithme des k-moyennes (chaque feuille CF_i étant considérée comme un unique enregistrement décrit par sa moyenne LS_i).

À chaque itération interactive **(étape 3)**, plutôt que d'utiliser directement les contraintes entre paires d'images ML et CL déduites des retours de l'utilisateur collectés dans l'étape 3(c), notre méthode de *clustering* semi-supervisé cherche à les intégrer sous la forme de contraintes ML_{CF} et CL_{CF} entre paires de feuilles de l'arbre CF. Plus précisément, ces contraintes portent sur les nouvelles feuilles de l'arbre (dont certaines ont pu être divisées en fonction des retours de l'utilisateur lors de l'étape 3(f)), et sont déduites des retours de l'utilisateur pendant l'étape 3(g). Cela nous permet finalement (étape 3(h)) de procéder au *clustering* semi-supervisé au niveau des feuilles CF et non des images, ce qui est original par rapport à la plupart des méthodes de la littérature. Cette procédure engendre une réduction du nombre de contraintes à intégrer dans l'algorithme, tout en maximisant la quantité d'information supervisée prise en compte pour un nombre fixé de clics de l'utilisateur. L'objectif est de réduire le temps de calcul tout en conservant les performances du *clustering* (voire en les améliorant) par rapport à un *clustering* intégrant l'information supervisée directement sous la forme de contraintes entre paires d'images. Pour déduire depuis les interactions de l'utilisateur les nouvelles feuilles CF et les contraintes qui les lient deux à deux, on passe par la formalisation des contraintes au niveau intermédiaire des voisinages (étape 3(d)), que l'on généralise sous la forme de contraintes entre paires d'images (étape 3(e)) .

Les étapes 3(c) à 3(h) sont détaillées dans les paragraphes suivants.

L'information concernant les *clusters* qui peut être directement déduite des retours de l'utilisateur dans l'**étape 3(c)** consiste en une liste PF d'images positives, et une liste NF d'images négatives, pour chaque *cluster* C_h.

L'**étape 3(d)** consiste à mettre à jour la liste des voisinages Np, et la liste $CannotNp$ des contraintes *cannot-link* entre ces voisinages, à partir des listes PF et NF. Considérons qu'il existe des contraintes *must-link* entre toutes les images positives d'un *cluster* donné, et des contraintes *cannot-link* entre chaque image positive et chaque image négative d'un même *cluster*. Un voisinage Np est défini comme un ensemble d'images liées par des contraintes *must-links*. Deux voisinages Np_i et Np_m sont considérées comme liés par une contrainte *cannot-link* s'il existe une contrainte *cannot-link* entre une image du voisinage Np_i et une image du voisinage Np_m. Le processus de mise à jour de ces voisinages est détaillé dans [Lai 2013a] et résumé en Figure B.2. Supposons que nous nous plaçons à la première itération interactive, c'est-à-dire juste après le *clustering* totalement non supervisé initial, et que la partie gauche de la Figure B.2 représente le résultat de ce *clustering* initial. Le *cluster* 1 contient les feuilles CF_1 et CF_2, le *cluster* 2 contient les feuilles CF_3 et CF_4, tandis que le *cluster* 3 contient les trois feuilles CF_5, CF_6 et CF_7. Quand il interagit avec le système, l'utilisateur donne ses retours au niveau des images, et non des feuilles de l'arbre CF. Si l'on prend l'exemple des images du *cluster* 3, les signes « + » signifient que l'utilisateur étiquette les images x_{15}, x_{16} et x_{17} comme positives pour ce *cluster*, et la flèche de gauche signifie qu'il change l'étiquette de l'image x_5 afin de l'assigner au *cluster* 1. La partie droite de la Figure B.2 illustre le résultat de l'étape 3(d). L'image x_5

est ajoutée au voisinage Np_1 contenant toutes les images positives du *cluster* 1, tandis que l'on rajoute une contrainte *cannot-link* entre les voisinages Np_3 et Np_1 (contenant respectivement toutes les images positives du *cluster* 3 et du *cluster* 1). Les images x_6 et x_7, qui reçoivent un retour négatif de l'utilisateur respectivement pour les *clusters* 2 et 1 (signes « - »), sont à l'origine de la création des nouveaux voisinages Np_5 et Np_4.

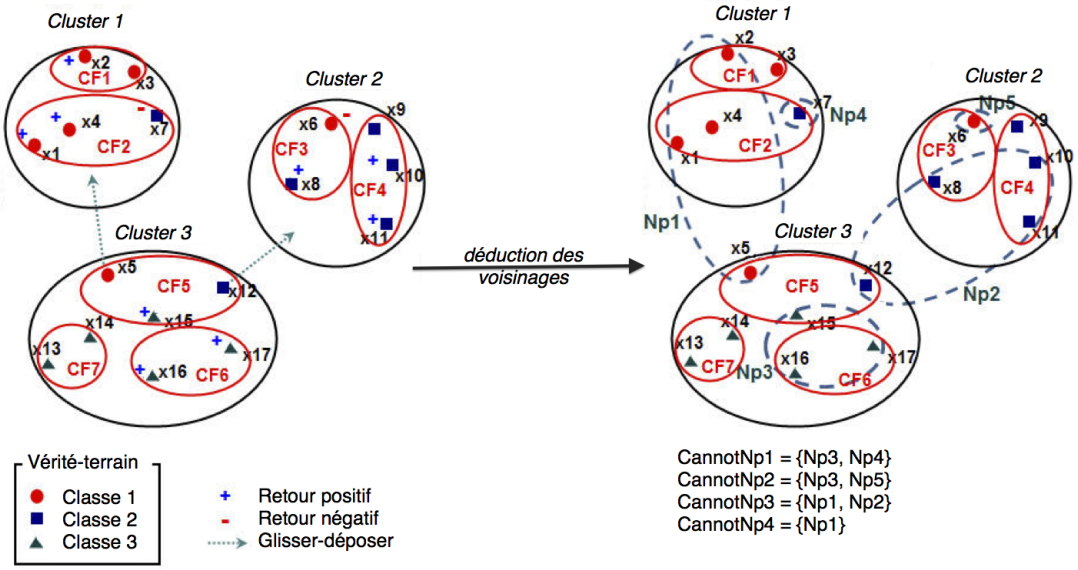


FIGURE B.2 – Exemple de déduction des voisinages basée sur les retours de l'utilisateur (étape 3(d)). À gauche on voit les signatures d'images x_i, les feuilles CF (ellipses rouges) et les *clusters* initiaux (ellipses noires), ainsi que les retours de l'utilisateur (signes « + » et « - »). Les voisinages Np_i déduits (ellipses bleues en pointillés), ainsi que la liste des contraintes cannot-link $CannotNp_i$ entre le voisinage Np_i et les autres voisinages, sont donnés dans la partie droite de la figure. Ici la vérité-terrain est donnée à titre indicatif (légende en bas à gauche), afin de mieux comprendre les interactions de l'utilisateur. *Figure adaptée de [Lai 2013a].*

À partir des listes PF et NF ainsi que des informations de voisinage Np et $CannotNp$, nous pouvons définir lors de l'**étape 3(e)** une liste CL de contraintes *cannot-link* et une liste ML de contraintes *must-link* entre paires d'images x_i. Plusieurs stratégies sont possibles pour déduire les contraintes entre paires d'images à partir des retours de l'utilisateur (dont la méthode exhaustive utilisée dans la version la plus basique de notre algorithme et évoquée plus haut). Le choix de cette stratégie aura un impact très important sur la complexité calculatoire et les performances de la méthode proposée. Nous avons procédé au total à la comparaison de six stratégies de déduction des contraintes dans [Lai 2013a]. Certaines sont basées sur la notion de voisinage pour des raisons d'efficacité. Ce point est expliqué plus en détails en section 2.2.3.2.

Avant de procéder au re-*clustering* des feuilles, il convient de diviser les feuilles de l'arbre CF qui ne sont pas suffisamment « pures ». C'est le cas des feuilles qui, suite aux derniers retours de l'utilisateur, contiennent au moins deux images liées par une contrainte *cannot-link* (exemple de la feuille CF_5), ou encore des feuilles qui ont à la fois un lien *cannot-link* et un lien *must-link* avec les images d'une autre feuille CF. C'est l'objet de l'**étape 3(f)**. Pour éviter une étude exhaustive des contraintes entre paires d'images, nous nous appuyons sur une notion intermédiaire de « noyau » d'une feuille CF_i, défini comme étant l'ensemble d'images de CF_i qui

sont incluses dans un même voisinage. Par exemple, dans la Figure B.2, la feuille CF_2 comprend deux noyaux, à savoir $\{x_1, x_4\}$ pour le voisinage Np_1 et $\{x_7\}$ pour le voisinage Np_4. Ses deux noyaux étant liés par un *cannot-link* (en raison de l'étiquetage négatif de x_7, et de l'étiquetage positif de x_1 et x_4, dans le *cluster* 1), la feuille CF_2 doit être découpée en deux. Si les deux noyaux n'étaient liés par aucune contrainte *cannot-link*, alors on les présenterait à l'utilisateur et, en fonction de ses retours, on déterminerait s'il convient de découper la feuille contenant ces deux noyaux, ou non.

Une fois certaines feuilles divisées, l'arbre doit être modifié localement en fonction de ces divisions. Plus précisément, les divisions de feuilles sont propagées vers les ancêtres de ces feuilles, ce qui se concrétise vers l'ajout de nouveaux pointeurs dans les nœuds intermédiaires, voire de nouveaux nœuds internes ou de nouveaux niveaux dans l'arbre (si la contrainte liée au paramètre B est violée).

Une fois les feuilles de l'arbre CF redéfinies de manière être suffisamment pures, nous devons lors de l'**étape 3(g)** inférer les contraintes ML_{CF} et CL_{CF} entre paires de feuilles CF à partir des contraintes entre paires d'images ML et CL. La stratégie utilisée ici est très simple, puisqu'on est maintenant sûrs qu'aucune feuille ne contient deux images liées par une contrainte *cannot-link*, et qu'aucune n'a à la fois un lien *cannot-link* et un lien *must-link* avec les images d'une autre feuille CF. On considère donc simplement que, si une feuille CF_i contient au moins une image x_l liée par une contrainte *cannot-link* (respectivement *must-link*) à une image $x_m \in CF_j$, alors les feuilles CF_i et CF_j sont liées par une contrainte *cannot-link* : $(CF_i, CF_j) \in CL_{CF}$ (respectivement *must-link* : $(CF_i, CF_j) \in ML_{CF}$).

La dernière **étape 3(h)** consiste à grouper les nouvelles feuilles de l'arbre CF en tenant compte des retours de l'utilisateur. Pour cela, nous avons proposé une méthode de *clustering* semi-supervisé à appliquer sur l'ensemble de toutes les feuilles $S_{CF} = (CF_1, ..., CF_m)$ de l'arbre CF, sous les contraintes entre paires de feuilles ML_{CF} et CL_{CF} déduites lors de l'étape 3(g). Nous cherchons à optimiser la fonction objective suivante, inspirée de celle d'un algorithme des k-moyennes qui serait mené de manière semi-supervisée en considérant chaque feuille CF comme un enregistrement d'entrée :

$$\begin{aligned} J_{obj} &= \sum_{CF_i \in S_{CF}} D(CF_i, \mu(CF_i)) \\ &+ \sum_{\substack{(CF_i, CF_j) \in ML_{CF} \\ K(CF_i) \neq K(CF_j)}} w N_{CF_i} N_{CF_j} D(CF_i, CF_j) \\ &+ \sum_{\substack{(CF_i, CF_j) \in CL_{CF} \\ K(CF_i) = K(CF_j)}} \overline{w} N_{CF_i} N_{CF_j} (D_{max} - D(CF_i, CF_j)) \end{aligned} \tag{B.2}$$

où :

- Le premier terme de l'équation (B.2) mesure la somme des distances entre chaque feuille CF_i et la moyenne $\mu(CF_i)$ du *cluster* correspondant $K(CF_i)$;
- Les deuxième et troisième termes de l'équation (B.2) représentent respectivement le coût de violation des contraintes *must-link* et *cannot-link* entre feuilles CF, où :
  - w et $\overline{w}$ sont des constantes spécifiant respectivement le coût de violation d'une contrainte *must-link* et d'une contrainte *cannot-link* entre deux images ;
  - N_{CF_i} est le nombre d'images contenues dans la feuille CF_i. Par conséquent, une

contrainte entre deux feuilles CF_i et CF_j correspond en fait à $N_{CF_i} \times N_{CF_j}$ contraintes entre paires d'images. C'est pourquoi ce terme multiplicatif est utilisé dans les deuxième et troisième membres de l'équation (B.2) ;
- Le coût de la violation d'une contrainte entre deux feuilles CF_i et CF_j est défini comme une fonction linéaire de la distance $D(CF_i, CF_j)$ qui les sépare, de manière à ce que ce coût soit d'autant plus élevé que les deux feuilles sont éloignées (dans le cas d'une contrainte *must-link*) ou proches (dans le cas d'une contrainte *cannot-link*). Ici, nous avons choisi de considérer pour $D(CF_i, CF_j)$ la distance Euclidienne entre leurs moyennes LS_{CF_i} et LS_{CF_j} (voir équation (B.1), page 161). Pour une discussion sur la distance utilisée, merci de se référer à la section 2.2.5 ;
- D_{max} est la distance maximum entre deux feuilles CF parmi l'ensemble S_{CF}.

L'algorithme que nous proposons afin de minimiser la fonction objective (B.2) à la fin de chaque itération interactive est très similaire à l'algorithme inspiré de la méthode des k-moyennes proposé dans le cas du HMRF-kmeans [Basu 2004]. Il repose sur deux étapes, qui seront réitérées séquentiellement à la fin de chaque itération interactive jusqu'à convergence : une étape de ré-étiquetage des feuilles et une étape de ré-estimation des centres des *clusters*. Cet algorithme est détaillé dans [Lai 2014b] et dans [Lai 2013a].

Puisque la fonction objective de l'équation (B.2) ne peut que diminuer à chaque itération de cet algorithme, **on converge vers un minimum local dans chaque itération interactive** [Lai 2013a]. Par contre, la fonction objectif est modifiée entre deux itérations interactives du fait de l'ajout de nouvelles contraintes. Il est par exemple possible que l'utilisateur rajoute à l'itération interactive $t+1$ des contraintes en contradiction avec les contraintes qu'il avait données aux itérations interactives précédentes $1 \dots t$. Donc, on ne peut garantir le fait que la fonction objective diminue entre deux itérations interactives successives, et il est impossible de garantir théoriquement la convergence globale de l'algorithme. Cependant, il est possible de la vérifier en pratique lorsque les souhaits de l'utilisateur ne varient pas trop avec le temps. Dans nos expérimentations où la vérité-terrain est utilisée par l'utilisateur pour ses retours (comme expliqué en section suivante), l'algorithme converge en un nombre d'itérations qui dépend évidemment de la taille de la base mais reste raisonnable.

Pour une analyse plus poussée de la complexité calculatoire de la méthode proposée dans son ensemble, merci de se référer à la section 5.3. de [Lai 2013a].

B.4 Améliorations possibles

On peut remarquer que l'utilisation du terme D_{max} dans l'équation (B.2) peut théoriquement rendre le coût des contraintes *cannot-link* sensibles aux valeurs aberrantes et mener à une sous-estimation uniforme de l'ensemble des coûts liés à des contraintes *cannot-link* en comparaison avec les violations *must-link*. Plusieurs stratégies sont possibles pour remédier à ce problème, que nous n'avons pas observé dans la pratique malgré nos nombreuses expérimentations sur quatre bases d'images de taille petite à modérée (comptant entre 1000 et 30000 images). La plus immédiate d'entre elles consisterait à filtrer les valeurs aberrantes pour le calcul de D_{max}, ce que la structure produite par BIRCH permet (voir plus haut).

Une caractéristique de l'algorithme utilisé pour le *clustering* semi-supervisé à la fin de chaque itération interactive (dans l'étape 3(h) de notre méthode) est qu'il est inspiré de l'algorithme des k-moyennes et ne permet donc pas de déterminer automatiquement le nombre de *clusters*

k à constituer. En fonction de l'application visée, cela pourrait constituer un désavantage de cette approche. Plusieurs pistes sont possibles pour pallier ce problème. L'une des plus évidentes consiste à construire plusieurs modèles de *clustering* avec différentes valeurs de k, puis à sélectionner le meilleur modèle, par exemple à l'aide d'une mesure de stabilité [Lange 2004] qui serait modifiée de manière à prendre en compte les retours de l'utilisateur. Mais cette étape additionnelle de sélection de modèle engendrerait inévitablement un surcoût en termes de temps de calcul, ce qui est préjudiciable dans le cas interactif où l'utilisateur est dans l'attente de la solution de *clustering* fournie par le système. Une autre piste est de partir d'un nombre de *clusters* fixé, puis de permettre à l'utilisateur de le faire évoluer au cours de chaque itération interactive en lui proposant explicitement des fonctionnalités de division ou de fusion de *clusters* (c'est d'ailleurs un mode d'interaction que nous avons implémenté dans le cadre de nos travaux en cours sur l'extraction d'invariants depuis des documents textuels, voir section 2.3). L'apprentissage actif pourrait nous être très utile pour choisir les *clusters* à proposer à l'utilisateur dans cette optique de fusion/division (voir section 2.2.5).

Annexe C

Typologie des approches existantes pour la reconnaissance d'écriture manuscrite

C.1 Introduction

La diffusion scientifique est plus abondante concernant la reconnaissance d'écriture manuscrite hors-ligne que concernant l'écriture en-ligne, cette dernière menant souvent à la conception de systèmes propriétaires de reconnaissance embarqués dans le dispositif de capture [Anquetil 2008]. Nous nous focaliserons dans la suite de ce mémoire sur les approches analytiques, les approches globales ayant été progressivement abandonnées au fil des dernières décennies, avec notamment l'augmentation de la taille des lexiques considérés et l'impossibilité de disposer de suffisamment d'exemples pour apprendre correctement chacun des mots.

Quel que soit le type d'écriture considéré (en-ligne ou hors-ligne), la plupart des approches analytiques sont basées sur trois étapes principales. La première étape consiste en une segmentation du mot en entités plus élémentaires (qui peuvent être des caractères ou des entités encore plus élémentaires, obtenues par exemple par une fenêtre glissante). La séquence de ces entités élémentaires doit alors être reconnue. Pour cela, dans la deuxième phase, on construit un (ou des) modèle(s). Les approches basées sur une segmentation implicite (aussi appelées *segmentation-free*) sont basées sur des modèles appris « en aveugle », c'est-à-dire directement à partir de la séquence des entités élémentaires. Dans les approches basées sur une segmentation explicite, en revanche, on passe par une étape de reconnaissance menée au niveau des entités élémentaires (souvent des caractères). Dans tous les cas, les sorties récoltées sont utilisées dans une troisième phase de « décodage » des mots par alignement des sorties du (ou des) modèle(s) obtenus avec les mots du lexique. Les Modèles de Markov Cachés (MMC) étant particulièrement bien adaptés à la reconnaissance de séquences de taille variable, ils se retrouvent souvent au cœur des systèmes de reconnaissance de mots manuscrits [Hu 2000, Bianne-Bernard 2011]. Et ce, à la fois pour la modélisation des mots et/ou des entités élémentaires le composant, et pour la phase de décodage.

Ces deux types d'approches (analytiques avec segmentation explicite et analytique avec segmentation implicite) sont détaillées dans les deux sections ci-après.

C.2 Approches analytiques avec segmentation explicite

Les approches basées sur une segmentation explicite visent à segmenter le mot en caractères, avant de reconnaître chaque caractère dans la séquence composant le mot. Que ce soit à partir d'une image hors-ligne ou d'un signal en-ligne, la segmentation explicite est une tâche difficile, surtout en présence d'une écriture cursive ou non-contrainte. Les techniques de segmentation

mises en œuvre sont donc en général relativement complexes et s'appuient souvent sur une analyse morphologique du tracé (p. ex. une analyse des contours [El-Yacoubi 1999] ou des outils de morphologie mathématique [Chen 1995]). La sortie du module de segmentation est un ensemble d'hypothèses de segmentation (graphèmes susceptibles de constituer des caractères), souvent organisées en un graphe d'hypothèses.

La seconde étape de décodage des mots est généralement basée sur une exploration de ce graphe utilisant les sorties d'un Moteur de Reconnaissance de Caractères (MRC). Le MRC prend comme entrée les nœuds du graphe de segmentation explicite - ou plus exactement des descripteurs statistiques extraits de ces nœuds - et renvoie des probabilités ou des vraisemblances associées aux caractères des mots du lexique. L'exploration du graphe pour le décodage des mots est souvent formalisée à l'aide d'un MMC et permet de sélectionner les meilleurs points de segmentation en se basant sur ces probabilités ou vraisemblances. On peut alors parler d'approches *collaboratives*, au sens où la segmentation et la reconnaissance sont menées conjointement et de manière à s'enrichir mutuellement.

Certains systèmes sont basés sur des MMCs à la fois pour l'apprentissage du MRC et pour l'exploration du graphe [El-Yacoubi 1999]. Dans ce cas, le MRC est basé sur des modèles de caractères (construits par MMC) qui sont *génératifs*, c'est-à-dire appris indépendamment les uns des autres, et donc qui ne cherchent pas spécialement à séparer les classes (voir annexe D). Vu la faible variabilité inter-classes, cela peut causer des ambiguïtés entre caractères ou entre caractères et entités plus élémentaires, surtout lorsque la segmentation est imparfaite.

Afin de s'affranchir de ces désavantages liés à l'utilisation de modèles génératifs, d'autres approches hybrides utilisent plutôt des méthodes discriminatives pour construire le MRC, et un MMC pour guider l'exploration du graphe. Parmi les méthodes discriminatives utilisées, on peut citer les réseaux de neurones [Tay 2001] ou plus récemment les Séparateurs à Vaste Marge (SVM) [Ahmad 2009]. Dans [Anquetil 2008], Éric Anquetil décrit le système ResifMot, basé sur un système de reconnaissance en-ligne de caractères s'appuyant sur la théorie des sous-ensembles flous et des systèmes d'inférence floue, et une exploration du graphe des hypothèses par fusion des connaissances basées sur la logique floue.

Si cette dernière méthode intègre des informations de cohérence spatiale inter-caractères, la majorité des approches basées sur une segmentation explicite ne prennent pas en compte le contexte des caractères voisins. Un désavantage supplémentaire de ce type d'approches est que l'étape de segmentation est cruciale, puisque les erreurs de segmentation peuvent être source d'erreurs lors de la phase de reconnaissance. Or, cette tâche est très difficile et requiert généralement la constitution d'une base d'apprentissage composée de caractères segmentés à partir de mots cursifs. Malgré le nombre important de travaux de recherche entrepris, il n'existe pas à l'heure actuelle de méthode de segmentation qui fasse l'unanimité parmi les chercheurs du domaine. Ce type d'approches s'est donc vu progressivement supplanté par les approches basées sur une segmentation implicite, et décrites ci-après.

C.3 Approches analytiques avec segmentation implicite

Inspirées des travaux effectués dans le domaine de la reconnaissance de la parole, la plupart des méthodes récemment proposées sont basées sur une segmentation implicite, que ce soit pour la reconnaissance d'écriture hors-ligne [Bianne-Bernard 2011] ou en-ligne [Hu 2000, Liwicki 2007]. C'est-à-dire qu'à la différence des méthodes basées sur une segmenta-

tion explicite, on ne cherche pas à segmenter le signal d'entrée en caractères de manière explicite pour la reconnaissance. Alternativement, lors de la segmentation implicite, le plus souvent on sur-segmente systématiquement les caractères composant le mot en entités élémentaires de manière triviale, par exemple en faisant passer une fenêtre glissante sur l'image (hors-ligne) ou la séquence de points (en-ligne). Puis, un (ou des) modèle(s) préalablement appris permet(tent) de reconnaître la séquence de caractères lors de la phase de décodage des mots. Cette dernière est souvent basée sur un alignement de modèles de caractères (d'où la notion de segmentation implicite), et éventuellement de modèles de liaisons inter-caractères permettant d'intégrer le contexte des caractères voisins.

Tout comme dans le cas des approches avec segmentation explicite, on retrouve des approches utilisant des MMC à la fois pour construire des modèles de caractères (et le cas échéant des modèles de liaisons inter-caractères), et pour la phase de décodage des mots. Mais, dans le cas implicite, l'apprentissage des modèles de caractères et inter-caractères MMC se fait généralement de manière *aveugle*, c'est-à-dire que les modèles de caractères/inter-caractères sont appris depuis des mots entiers, sans segmentation explicite préalable en caractères. C'est en ce sens que les méthodes basées sur une segmentation implicite sont parfois qualifiées de « sans segmentation » (*segmentation-free* en anglais). L'entraînement de ces modèles requiert donc un nombre d'exemples (mots) important et bien contrôlé. En outre, ces modèles sont *génératifs* et donc généralement moins performants que des méthodes discriminatives. De plus, les paramètres des modèles MMC (taille de la fenêtre glissante, nombre d'états par modèle de caractère/inter-caractère, nombre de Gaussiennes dans les mélanges de Gaussiennes, etc.) sont le plus souvent difficiles à ajuster, même si certaines stratégies visent à apprendre automatiquement ou semi-automatiquement ces paramètres [Fischer 2009].

Ces désavantages expliquent en partie l'engouement pour des approches hybrides neuro-Markoviennes [España-Boquera 2011], combinant réseaux de neurones et MMC. Dans ce cas, le réseau de neurones sert à construire un (ou des) modèle(s) permettant de reconnaître la séquence de caractères de manière discriminative à partir d'une séquence d'observations d'entités élémentaires « en aveugle », c'est-à-dire sans segmentation préalable en caractères. Ce type d'approches hybrides bénéficient du caractère discriminant et donc des meilleurs performances des réseaux de neurones, tout en conservant la capacité des MMCs à mener à bien le décodage en intégrant les informations de contexte au niveau du mot. Différents types de réseaux de neurones peuvent être utilisés : perceptron multi-couches ou réseaux de neurones récurrents dans [Schenk 2006], réseaux de neurones convolutionnels dans [Bengio 1995] ou *time-delay neural networks* dans [Caillault 2005] et dont le principe de convolution permet de gérer la cohérence de la séquence des observations.

Plus récemment, Graves *et al.* ont introduit dans [Graves 2009] une approche purement neuronale où un réseau de neurones récurrent est directement entraîné pour la reconnaissance de la séquence de caractères composant le mot. Le réseau de neurones récurrent est configuré avec une couche de sortie CTC (*Connectionist Temporal Classification*), qui utilise le réseau pour passer directement de la séquence des observations d'entrée à la séquence des classes de sortie, et une architecture BLSTM (Bidirectional Long Short-Term Memory) afin d'intégrer efficacement le contexte des caractères voisins. Les résultats rapportés en termes de reconnaissance de mots, quand le réseau de neurones est utilisé conjointement avec un dictionnaire et un modèle de langage au niveau de la ligne de texte, sont excellents.

Cependant, pour les approches hybrides comme pour les approches purement neuronales, il

reste très difficile de fixer les paramètres du réseau de neurones (nombre de couches cachées et/ou de neurones dans les couches cachées, etc.), et la plupart du temps des approches heuristiques sont employées. De plus, l'entraînement de ces méthodes et leur optimisation restent comme les méthodes génératives très gourmandes en données d'apprentissage. Cela rend la conception de systèmes personnalisés difficile.

Annexe D

Approches génératives et discriminatives pour la classification supervisée

Dans sa formulation la plus simple, la classification supervisée de données consiste à construire, à partir variables d'entrée X (souvent appelées variables prédictives, covariables ou régresseurs en statistiques) [1], un « classifieur » permettant de prédire un ensemble de variables qualitatives de sortie Y (classes). Le classifieur est appris à partir de l'observation des valeurs conjointes de X et Y sur un ensemble d'exemples souvent appelé base d'apprentissage ou d'entraînement.

Les nombreuses méthodes de classification supervisée basées sur un apprentissage et proposées dans la littérature [Duda 2012] peuvent être grossièrement rangées selon la fonction de coût que l'on minimise lors de l'apprentissage. On distingue alors les méthodes génératives et les méthodes discriminatives [Bouchard 2005].

La différence entre les méthodes génératives et discriminatives, détaillée ci-après, peut être résumée de la manière suivante : tandis que les approches génératives cherchent avant tout à modéliser les classes, les approches discriminatives, elles, cherchent à modéliser les frontières entre classes afin de les séparer au mieux.

Plus précisément, les **approches génératives** cherchent à modéliser la distribution conjointe de toutes les variables, à savoir les entrées X, les sorties Y, et le cas échéant les variables cachées ou latentes Z, et des probabilités *a priori* correspondantes et notamment $\mathbb{P}[X/Y]$. Cette modélisation se fait généralement sur la foi d'hypothèses probabilistes concernant les données (ce sont donc des méthodes désignées par le terme « paramétriques » dans le domaine de l'analyse de données). Puis, le plus souvent, les paramètres du classifieur sont appris par maximisation de la vraisemblance $Pr[X/Y]\mathbb{P}[Y]$. On compte parmi les approches génératives les méthodes Bayésiennes (modèles graphiques aussi appelés réseaux Bayésiens, classifieur naïf de Bayes, etc.), mais aussi les Modèles de Markov Cachés (MMC, largement évoqués en annexe C, et les méthodes d'analyse discriminante (linéaire, quadratique, à noyau, flexible, mixte, etc.), que j'ai utilisées dans ma thèse [Visani 2005a] et à l'intitulé trompeur.

Les **approches discriminatives** (également appelées approches conditionnelles), quant à elles, cherchent à estimer directement les probabilités *a posteriori* $P(Y/X)$ d'appartenance de chaque enregistrement de X à reconnaître dans chacune des classes de Y, ou bien elles cherchent à apprendre une fonction de décision permettant de passer directement des observations X aux classes Y. Elles sont entraînées par sélection du bon compromis entre complexité du modèle et minimisation des coûts de classification sur une base d'apprentissage et/ou une base de validation étiquetée(s) (c'est-à-dire que l'on connaît à la fois X et Y) ; l'utilisation d'une base de validation sert généralement à éviter tout sur-apprentissage des données. Parmi les

1. Dans le contexte de ce manuscrit il s'agit des signatures des images, voir chapitre 2.

approches discriminatives, on retrouve la régression logistique (linéaire, quadratique, etc.), les k plus proches voisins, les noyaux de Parzen, les Séparateurs à Vaste Marge (SVM), les réseaux de neurones, etc. Les approches discriminatives peuvent être paramétriques (comme par exemple la régression logistique) ou non-paramétriques (comme par exemple les SVM).

Les performances de toute méthode d'apprentissage sont limitées par la complexité de la règle de classification recherchée en regard de la quantité d'information dont on dispose (ici déterminée par le nombre et la qualité des données d'apprentissage). On parle de borne de Shannon en théorie de l'information [Cover 2012], de borne de Cramer-Rao en statistiques [Kullback 1959] et de borne de Vapnik en théorie de l'apprentissage [Vapnik 1998]. De très nombreux chercheurs se sont intéressés à la comparaison des approches génératives et discriminatives. Le principal avantage des méthodes génératives est d'ajouter des informations liées au contexte pour enrichir l'apprentissage, grâce à la modélisation de la distribution jointe de toutes les variables (y compris des variables d'entrée et des variables latentes ou cachées). Cela revient à faire des hypothèses supplémentaires (par rapport à une approche discriminative), et ainsi de restreindre le domaine de recherche des solutions, ce qui peut en théorie engendrer une diminution des taux d'erreur. Mais, dans la pratique, les attributs d'entrée X et les variables latentes ou cachées ont le plus souvent une distribution inconnue, ce qui rend leur modélisation très difficile.

Les approches discriminatives sont donc en général préférées dès lors que toute modélisation probabiliste des données est vouée à l'échec, et que les données d'apprentissage sont en nombre suffisant pour éviter le sur-apprentissage. D'autant plus que, en se basant sur la minimisation d'une fonction de coût empirique (avec éventuellement une pénalisation de la complexité de la règle de décision), qui correspond à l'objectif concret recherché, elles donnent de meilleurs taux de classification que les approches génératives dans la plupart des problèmes réels de classification supervisée [Bouchard 2005]. En effet, comme Vapnik le constatait dans [Vapnik 1998] : « *one should solve the [classification] problem directly and never solve a more general problem as an intermediate step.* ».

Il est important de noter que cette distinction historique tend à perdre de son sens, puisque durant la dernière décennie de nombreuses approches combinant des méthodologies discriminatives et génératives ont vu le jour, et ce, dans des domaines variés incluant la vision par ordinateur. En particulier, les approches neuronales basées sur un apprentissage profond, en plein essor, reposent largement sur un apprentissage hybride génératif/discriminatif. En effet, ces approches sont typiquement composées d'une première phase de « pré-apprentissage » génératif, suivie d'une phase discriminative d'affinement (« fine-tuning »).

Annexe E

Recueil d'articles publiés dans des revues internationales

Cette annexe est composée d'un recueil de 3 articles parus récemment dans des revues d'audience internationale, à savoir :

[Lai 2013b] H.P. Lai, M. Visani, A. Boucher et J.M. Ogier. *A new Interactive Semi-Supervised Clustering model for large image database indexing*. Pattern Recognition Letters, Special Issue on Partially Supervised Learning for Pattern Recognition, pages 1–48, DOI : 10.1016/j.patrec.2013.06.014, 2013.

[Visani 2011a] M. Visani, K. Bertet et J.M. Ogier. *Navigala : an Original Symbol Classifier Based on Navigation through a Galois Lattice*. International Journal of Pattern Recognition and Artificial Intelligence, vol. 25, no. 4, pages 449–473, DOI : 10.1142/S0218001411008634, 2011.

[Visani 2011b] M. Visani, O. Ramos et S. Tabbone. *A Protocol to Characterize the Descriptive Power and the Complementarity of Shape Descriptors*. International Journal on Document Analysis and Recognition, vol. 14, no. 11, pages 87–100, DOI : 10.1109/TSMCB.2011.2108646, 2011.

Pattern Recognition Letters 37 (2014) 94–106

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A new interactive semi-supervised clustering model for large image database indexing

Hien Phuong Lai [a,b,c,*], Muriel Visani [a], Alain Boucher [a,b,c], Jean-Marc Ogier [a]

[a] L3I, Université de La Rochelle, Avenue M. Crépeau, 17042 La Rochelle cedex 1, France
[b] IFI, Equipe MSI; IRD, UMI 209 UMMISCO, Institut de la Francophonie pour l'Informatique, 42 Ta Quang Buu, Hanoi, Vietnam
[c] Vietnam National University, Hanoi, Vietnam

ARTICLE INFO

Article history:
Available online 27 June 2013

Keywords:
Semi-supervised clustering
Interactive learning
Image indexing

ABSTRACT

Indexing methods play a very important role in finding information in large image databases. They organize indexed images in order to facilitate, accelerate and improve the results for later retrieval. Alternatively, clustering may be used for structuring the feature space so as to organize the dataset into groups of similar objects without prior knowledge (unsupervised clustering) or with a limited amount of prior knowledge (semi-supervised clustering).

In this paper, we introduce a new interactive semi-supervised clustering model where prior information is integrated via pairwise constraints between images. The proposed method allows users to provide feedback in order to improve the clustering results according to their wishes. Different strategies for deducing pairwise constraints from user feedback were investigated. Our experiments on different image databases (Wang, PascalVoc2006, Caltech101) show that the proposed method outperforms semi-supervised HMRF-kmeans (Basu et al., 2004).

1. Introduction

Content-Based Image Retrieval (CBIR) refers to the process which uses visual information (usually encoded using color, shape, texture feature vectors, etc.) to search for images in the database that correspond to the user's queries. Traditional CBIR systems generally rely on two phases. The first phase is to extract the feature vectors from all the images in the database and to organize them into an efficient index data structure. The second phase is to efficiently search in the indexed feature space to find the most similar images to the query image.

With the development of many large image databases, an exhaustive search is generally intractable. Feature space structuring methods (normally called indexing methods) are therefore necessary for facilitating and accelerating further retrieval. They can be classified into space partitioning methods and data partitioning methods.

Space partitioning methods (KD-tree (Bentley and Sep., 1975), KDB-tree (Robinson, 1981), LSD-tree (Henrich et al., 1989), GridFile (Nievergelt et al., 1988) etc.) generally divide the feature space into cells (sometimes referred to as "buckets") of fairly similar cardinality (in terms of number of images per cell), without taking into account the distribution of the images in the feature space. Therefore, dissimilar points may be included in a same cell while similar points may end up in different cells. The resulting index is therefore not optimal for retrieval, as the user generally wants to retrieve similar images to the query image. Moreover, these methods are not designed to handle high dimensional data, while image feature vectors commonly count hundreds of elements.

Data partitioning methods (B-tree (Bayer and McCreight, 1972), R-trees (Guttman, 1984; Sellis et al., 1987; Beckmann et al., 1990), SS-tree (White and Jain, 1996), SR-tree (Katayama and Satoh, 1997), X-tree (Berchtold et al., 1996) etc.) also integrate information about image distribution in the feature space. However, the limitations on the cardinality of the space cells remain, causing the resulting index to be non-optimal for retrieval, especially in the case where groups of similar objects are unbalanced, i.e. composed of different numbers of images.

Our claim is that using clustering instead of traditional indexing to organize feature vectors, results in indexes better adapted to high dimensional and unbalanced data. Indeed, clustering aims to split a collection of data into groups (clusters) so that similar objects belong to the same group and dissimilar objects are in different groups, with no constraints on the cluster size. This makes the resulting index better optimized for retrieval. In fact, while in traditional indexing methods it might be difficult to fix the number of objects in each bucket (especially in the case of unbalanced data),

* Corresponding author at: L3I, Université de La Rochelle, Avenue M. Crépeau, 17042 La Rochelle cedex 1, France. Tel.: +33 6 46 51 12 32; fax: +33 5 46 45 82 42.
E-mail addresses: hien_phuong.lai@univ-lr.fr (H.P. Lai), muriel.visani@univ-lr.fr (M. Visani), alainboucher12@gmail.com (A. Boucher), jean-marc.ogier@univ-lr.fr (J.-M. Ogier).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.06.014

clustering methods have no limitation on the cardinality of the clusters, objects can be grouped into clusters of very different sizes. Moreover, using clustering might simplify the relevance feedback task, as the user might interact with a small number of cluster prototypes rather than numerous single images.

Because feature vectors only capture low level information such as color, shape or texture, there is a semantic gap between high-level semantic concepts expressed by the user and these low-level features. The clustering results are therefore generally different from the intent of the user. Our work aims to involve users in the clustering phase so that they can interact with the system in order to improve the clustering results. The clustering methods should therefore produce a hierarchical cluster structure where the initial clusters may be easily merged or split. We are also interested in clustering methods which can be incrementally built in order to facilitate the insertion or deletion of new images by the user. It can be noted that incrementality is also very important in the context of huge image databases, when the whole dataset cannot be stored in the main memory. Another very important point is the computational complexity of the clustering algorithm, especially in an interactive online context where the user is involved.

In the case of large image database indexing, we may be interested in traditional clustering (unsupervised) (Jain et al., 1999; Xu et al., 2005) or semi-supervised clustering (Basu et al., 2002, 2004; Dubey et al., 2010; Wagstaff et al., 2001). While no information about ground truth is provided in the case of unsupervised clustering, a limited amount of knowledge is available in the case of semi-supervised clustering. The provided knowledge may consist of class labels (for some objects) or pairwise constraints (must-link or cannot-link) between objects.

In Lai et al. (2012a), we proposed a survey of unsupervised clustering techniques and analyzed the advantages and disadvantages of different methods in a context of huge masses of data where incrementality and hierarchical structuring are needed. We also experimentally compared five methods (global k-means (Likas et al., 2003), AHC (Lance and Williams, 1967), R-tree (Guttman, 1984), SR-tree (Katayama and Satoh, 1997) and BIRCH (Zhang et al., 1996)) with different real image databases of increasing sizes (Wang, PascalVoc2006, Caltech101, Corel30k) (the number of images ranges from 1000 to 30,000) to study the scalability of different approaches relative to the size of the database. In Lai et al. (2012b), we presented an overview of semi-supervised clustering methods and proposed a preliminary experiment of an interactive semi-supervised clustering model using the HMRF-kmeans (Hidden Markov Random Fields kmeans) clustering (Basu et al., 2004) on the Wang image database in order to analyze the improvement in the clustering process when user feedback is provided.

There are three main parts to this paper. Firstly, we propose a new interactive semi-supervised clustering model using pairwise constraints. Secondly, we investigate different methods for deducing pairwise constraints from user feedback. Thirdly, we experimentally compare our proposed semi-supervised method with the widely known semi-supervised HMRF-kmeans method.

This paper is structured as follows. A short review of semi-supervised clustering methods is presented in Section 2. Our interactive semi-supervised clustering model is proposed in Section 3. Some experiments are presented in Section 4. Some conclusions and further works are provided in Section 5.

2. A short review of semi-supervised clustering methods

For unsupervised clustering only similarity information is used to organize objects; in the case of semi-supervised clustering a small amount of prior knowledge is available. Prior knowledge is either in the form of class labels (for some objects) or pairwise constraints between objects. Pairwise constraints specify whether two objects should be in the same cluster (must-link) or in different clusters (cannot-link). As the clusters produced by unsupervised clustering may not be the ones required by the user, this prior knowledge is needed to guide the clustering process for resulting clusters which are closer to the user's wishes. For instance, for clustering a database with thousands of animal images, an user may want to cluster by animal species or by background landscape types. An unsupervised clustering method may give, as a result, a cluster containing images of elephants with a grass background together with images of horses with a grass background and another cluster containing images of elephants with a sand background. These results are ideal when the user wants to cluster by background landscape types. But they are poor when the user wants to cluster by animal species. In this case, must-link constraints between images of elephants with a grass background and images of elephants with a sand background and cannot-link constraints between images of elephants with a grass background and images of horses with a grass background are needed to guide the clustering process. The objective of our work is to make the user interact with the system so as to define easily these constraints with only a few clicks. Note that the available knowledge is too poor to be used with supervised learning, as only a very limited ratio of the available images are considered by the user at each step. In general, semi-supervised clustering methods are used to maximize intra-cluster similarity, to minimize inter-cluster similarity and to keep a high consistency between partitioning and domain knowledge.

Semi-supervised clustering has been developed in the last decade and some methods have been published to date. They can be divided into semi-supervised clustering with labels, where partial information about object labels is given, and semi-supervised clustering with constraints, where a small amount of pairwise constraints between objects is given.

Some semi-supervised clustering methods using labeled objects have been put forward: seeded-kmeans (Basu et al., 2002), constrained-kmeans (Basu et al., 2002), etc. Seeded-kmeans and constrained-kmeans are based on the k-means algorithm. Prior knowledge for these two methods is a small subset of the input database, called seed set, containing user-specified labeled objects of k different clusters. Unlike k-means algorithm which randomly selects the initial cluster prototypes, these two methods use the labeled objects to initialize the cluster prototypes. Following this we repeat, until convergence, the re-assignment of each object in the dataset to the nearest prototype and the re-computation of the prototypes with the assigned objects. The seeded-kmeans assigns objects to the nearest prototype without considering the prior labels of the objects in the seed set. In contrast, the constrained-kmeans maintains the labeled examples in their initial clusters and assigns the other objects to the nearest prototype. An interactive cluster-level semi-supervised clustering was proposed in Dubey et al. (2010) for document analysis. In this model, knowledge is progressively provided as assignment feedback and cluster description feedback after each interactive iteration. Using assignment feedback, the user moves an object from one cluster to another cluster. Using cluster description feedback, the user modifies the feature vector of any current cluster (e.g. increase the weighting of some important words). The algorithm learns from all the feedback to re-cluster the dataset in order to minimize average distance between points and their cluster centers while minimizing the violation of constraints corresponding to feedback.

Among the semi-supervised clustering methods using pairwise constraints between objects, we can cite COP-kmeans (constrained-kmeans) (Wagstaff et al., 2001), HMRF-kmeans (Hidden Markov Random Fields Kmeans) (Basu et al., 2004), semi-supervised kernel-kmeans (Kulis et al., 2005), etc. The input data of these

methods is data set X, a set of must-link constraints M and a set of cannot-link constraints C. In COP-kmeans, points are assigned to clusters without violating any constraint. A point x_i is assigned to its closest cluster μ_j unless a constraint is violated. If x_i cannot be placed in μ_j, we continue attempting to assign x_i to the next cluster in the sorted list of clusters by ascending order of distances with x_i until a suitable cluster is found. The clustering fails if no solution respecting the constraints is found. While the constraint violation is strictly prohibited in COP-kmeans, it is allowed with a violation cost (penalty) in HMRF-kmeans and in semi-supervised kernel-kmeans. The objective function to be minimized in the semi-supervised HMRF-kmeans is as follows:

$$J_{HMRF\_Kmeans} = \sum_{x_i \in X} D(x_i, \mu_{l_i}) + \sum_{(x_i,x_j)\in M, l_i \neq l_j} w_{ij} + \sum_{(x_i,x_j)\in C, l_i = l_j} \overline{w_{ij}} \quad (1)$$

where w_{ij} ($\overline{w_{ij}}$) is the penalty cost for violating a must-link (cannot-link) constraint between x_i and x_j, l_i refers to the cluster label of x_i, and $D(x_i, \mu_{l_i})$ measures the distance between x_i and its corresponding cluster center μ_{l_i}. The violation cost of a pairwise constraint may be either a constant or a function of the distance between the two points specified in the pairwise constraint as follows:

$$w_{ij} = wD(x_i, x_j) \quad (2)$$

$$\overline{w_{ij}} = \overline{w}(D_{max} - D(x_i, x_j)) \quad (3)$$

where w and $\overline{w}$ are constants specifying the cost for violating a must-link or a cannot-link constraint. D_{max} is the maximal distance between two points in the data set. We can see that, to ensure the most difficult constraints are respected, higher penalties are assigned to violations of must-link constraints between points which are distant and to violations of cannot-link constraints between points which are close. The term D_{max} in Eq. (3) can make the cannot-link penalty term sensitive to extreme outliers, but all cannot-link constraints are treated in the same way, so even in the presence of extreme outliers, there would be no cannot-link constraint favored compared to the others. The objective function in Eq. (1) is also sensitive to outliers. We can reduce this sensitivity by using an outlier filtering technique or by replacing the term D_{max} by the maximum distance between two clusters. HMRF-kmeans first initializes the k cluster centers based on user-specified constraints, as described in Basu et al. (2004). After the initialization step, an iterative relocation approach similar to k-means is applied to minimize the objective function. The iterative algorithm represents the repetition of the assignment phase of each point to the cluster which minimizes its contribution to the objective function and the re-estimation phase of the cluster centers minimizing the objective function. The semi-supervised kernel-kmeans (Kulis et al., 2005) is similar to the HMRF-kmeans, but calculates the objective function in a transformed space instead of the original space using a kernel function mapping as follows:

$$J_{SS\_Kernel\_Kmeans} = \sum_{x_i \in X} \|\phi(x_i) - \overline{\phi_{l_i}}\|^2 - \sum_{(x_i,x_j)\in M, l_i = l_j} w_{ij} + \sum_{(x_i,x_j)\in C, l_i = l_j} \overline{w_{ij}} \quad (4)$$

where $\phi(x_i)$ is the kernel function mapping, $\overline{\phi_{l_i}}$ is the centroid of the cluster containing x_i and w_{ij} ($\overline{w_{ij}}$) is the penalty cost for violating a must-link (cannot-link) constraint between x_i and x_j. In the second term of Eq. (4), instead of adding a penalty cost for a must-link violation if the two points are in different clusters Kulis et al. (2005) give a reward for must-link constraint satisfaction if the two points are in the same cluster, by subtracting the corresponding penalty term from the objective function.

3. Proposed interactive semi-supervised clustering model

In this section, we present our proposed interactive semi-supervised clustering model. In our model, the initial clustering is carried out without any prior knowledge, using an unsupervised clustering method. In Lai et al. (2012a) we discussed the adequation between different unsupervised clustering methods and our applied context (involving user interactivity) as well as experimentally compared different unsupervised clustering methods (global k-means (Likas et al., 2003), AHC (Lance and Williams, 1967), R-tree (Guttman, 1984), SR-tree (Katayama and Satoh, 1997), BIRCH (Zhang et al., 1996)). Our conclusion was that BIRCH is the most suitable to our context. BIRCH is less sensitive to variations in its parameters. Moreover, it is incremental, it provides a hierarchical structure of clusters and it outperforms other methods in the context of a large database (best results and best computational time in our tests). Therefore, BIRCH is chosen for the initial unsupervised clustering in our model. After the initial clustering, the user views the clustering results and provides feedback to the system. The pairwise constraints (must-link, cannot-link) are deduced, based on user feedback; the system then re-organizes the clusters by considering the constraints. The re-clustering process is done using the proposed semi-supervised clustering described in Section 3.2. The interactive process (user provides feedback and system reorganizes the clusters) is repeated until the clustering result satisfies the user. The interactive semi-supervised clustering model contains the following steps:

1. Initial clustering using BIRCH unsupervised clustering.
2. *Repeat*:
   (a) Receive feedback from the user and deduce pairwise constraints.
   (b) Re-organize the clusters using the proposed semi-supervised clustering method.

until the clustering result satisfies the user.

3.1. BIRCH unsupervised clustering

Let us briefly describe the BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) unsupervised clustering method (Zhang et al., 1996). The idea of BIRCH is to build a Clustering Feature Tree (CF-tree).

We define a CF-vector, summarizing information of a cluster including N vectors $(\vec{x_1}, \ldots, \vec{x_N})$, as a triplet $CF = (N, \overrightarrow{LS}, SS)$, where $\overrightarrow{LS}$ and SS are respectively the linear sum and the square sum of vectors $\left(\overrightarrow{LS} = \sum_{i=1}^{N} \vec{x_i};\ SS = \sum_{i=1}^{N} \vec{x_i}^2\right)$. From the CF-vectors, we can simply compute the centroid, the radius (average distance from points to the centroid) of a cluster and also the distance between two clusters (e.g. the Euclidean distance between their centroids). A CF-tree is a balanced tree having three parameters B, L and T:

- Each internal node contains, at most, B elements of the form $[CF_i, child_i]$ where $child_i$ is a pointer to its ith child node and CF_i is the CF-vector of this child.
- Each leaf node contains, at most, L entries of the form $[CF_i]$, it also contains two pointers, *prev* and *next*, to link leaf nodes.
- Each entry CF_i represents the information of a group of points which are close together. Each entry CF_i of a leaf node must have a radius lower than a threshold T (threshold condition).

The CF-tree is created by successively inserting points into the tree. A new point is preferably inserted in the closest CF_i of the closest leaf, if the threshold condition is not violated. If it is impos-

sible, a new CF_j is created for the new point. The corresponding internal and leaf nodes must be split if necessary. After creating the CF-tree, we can use any clustering method (AHC, k-means, etc.) to cluster all leaf entries CF_i. In our work, we use k-means for clustering the leaf entries, as it is suitable to be used with our proposed semi-supervised clustering in the interactive phase.

3.2. Proposed semi-supervised clustering method

At each interactive iteration, our semi-supervised clustering method is applied after receiving feedback from the users for re-organizing the clusters according to their wishes. Our semi-supervised clustering method considers the set of all leaf entries $S_{CF} = (CF_1, \ldots, CF_m)$ of the CF-tree. Supervised information is provided as two sets of pairwise constraints between CF entries deduced from user feedback: must-links $M_{CF} = \{(CF_i, CF_j)\}$ and cannot-links $C_{CF} = \{(CF_i, CF_j)\}$. $(CF_i, CF_j) \in M_{CF}$ implies that CF_i, CF_j and therefore all points which are included in these two entries should belong to the same cluster, while $(CF_i, CF_j) \in C_{CF}$ implies that CF_i and CF_j should belong to different clusters. The objective function to be minimized is as follows:

$$J_{obj} = \sum_{CF_i \in S_{CF}} D(CF_i, \mu_{l_i}) + \sum_{(CF_i, CF_j) \in M_{CF}, l_i \neq l_j} w N_{CF_i} N_{CF_j} D(CF_i, CF_j) + \sum_{(CF_i, CF_j) \in C_{CF}, l_i = l_j} \overline{w} N_{CF_i} N_{CF_j} (D_{max} - D(CF_i, CF_j)) \quad (5)$$

where:

- The first term measures the distortion between each leaf entry CF_i and the corresponding cluster center μ_{l_i}, l_i refers to the cluster label of CF_i.
- The second and the third terms represent the penalty costs for respectively violating the must-link and cannot-link constraints between CF entries. w and $\overline{w}$ are constants specifying the violation cost of a must-link and a cannot-link between two points. As an entry CF_i represents the information of a group of N_{CF_i} points, a pairwise constraint between two entries CF_i and CF_j corresponds to $N_{CF_i} \times N_{CF_j}$ constraints between points of these two entries. The violation cost of a pairwise constraint between two entries CF_i, CF_j is thus a function of their distance $D(CF_i, CF_j)$ and of the number of points included in these two entries. D_{max} is the maximum distance between two CF entries in the data set. Therefore, higher penalties are assigned to violations of must-link between entries that are distant and of cannot-link between entries which are close. As in HMRF-kmeans, the term D_{max} can make the cannot-link penalty term sensitive to extreme outliers, and could be replaced by the maximum distance between two clusters if the database contains extreme outliers.

In our case, we use the most frequently used squared Euclidean distance as distortion measure. The distance between two entries $CF_i = (N_{CF_i}, \overrightarrow{LS}_{CF_i}, SS_{CF_i})$, $CF_j = (N_{CF_j}, \overrightarrow{LS}_{CF_j}, SS_{CF_j})$ is calculated as the distance between their means as follows:

$$D(CF_i, CF_j) = \sum_{p=1}^{d} \left(\frac{LS_{CF_i}(p)}{N_{CF_i}} - \frac{LS_{CF_j}(p)}{N_{CF_j}} \right)^2 \quad (6)$$

where d is the number of dimensions of the feature space.

The proposed semi-supervised clustering is as follows:

Input: Set of leaf entries $S_{CF} = \{CF_i\}_{i=1}^{m}$ which are clustered into K clusters with the corresponding centroids $\{\mu_h\}_{h=1}^{K}$,
set of must-link constraints $M_{CF} = \{(CF_i, CF_j)\}$
set of cannot-link constraints $C_{CF} = \{(CF_i, CF_j)\}$.

Output: New disjoint K clusters of S_{CF} such that the objective function in Eq. (5) is locally minimized.

Method:

1. Set $t \leftarrow 0$
2. Repeat until convergence
   (a) Re-assignment step: Given $\{\mu_h^{(t)}\}_{h=1}^{K}$, re-assign cluster labels $\{l_i^{(t+1)}\}_{i=1}^{m}$ of entries $\{CF_i\}_{i=1}^{m}$ to minimize the objective function.
   (b) Re-estimation step: Given cluster labels $\{l_i^{(t+1)}\}_{i=1}^{m}$, re-calculate the cluster centroids $\{\mu_h^{(t+1)}\}_{h=1}^{K}$ to minimize the objective function.
   (c) $t \leftarrow t + 1$.

In the re-assignment step, given the current cluster centers, each entry CF_i is re-assigned to the cluster μ_h which minimizes its contribution to the objective function as follows:

$$J_{obj}(CF_i, \mu_h) = D(CF_i, \mu_h) + \sum_{(CF_i, CF_j) \in M_{CF}, h \neq l_j} w N_{CF_i} N_{CF_j} D(CF_i, CF_j) + \sum_{(CF_i, CF_j) \in C_{CF}, h = l_j} \overline{w} N_{CF_i} N_{CF_j} (D_{max} - D(CF_i, CF_j)) \quad (7)$$

We can see that the optimal assignment of each CF entry also depends on the current assignment of the other CF entries due to the violation cost of pairwise constraints in the second and third terms of Eq. 7. Therefore, after all entries are re-assigned, they are randomly re-ordered, and the re-assignment process is repeated until no CF entry changes its cluster label between two successive iterations.

In the re-estimation step, given the cluster labels $\{l_i^{(t+1)}\}_{i=1}^{m}$ of all CF entries, the cluster centers $\{\mu_h\}_{h=1}^{K}$ are re-calculated in order to minimize the objective function of the current assignment. For simple calculation, each cluster center is also represented in the form of a CF-vector. By using the squared Euclidean measure, the CF-vector of each cluster prototype μ_h is calculated based on CF entries which are assigned to this cluster as follows:

$$N_{\mu_h} = \sum_{l_i = h} N_{CF_i} \quad (8)$$

$$\overrightarrow{LS}_{\mu_h} = \sum_{l_i = h} \overrightarrow{LS}_{CF_i} \quad (9)$$

$$SS_{\mu_h} = \sum_{l_i = h} SS_{CF_i} \quad (10)$$

We can see that in each re-assignment step, each entry CF_i moves to a new cluster μ_h if its contribution to the objective function is decreased with this re-assignment. Therefore, the objective function J_{obj} is decreased or unchanged after the re-assignment step. And in each re-estimation step, the mean of the CF-vector of each cluster μ_h corresponds to the mean of the CF entries (and therefore the points) in this cluster, that minimizes the contribution of μ_h to the component $\sum_{CF_i \in S_{CF}} D(CF_i, \mu_{l_i})$ of J_{obj}. The penalty terms of J_{obj} are not functions of the centroid, thus they do not take part in cluster center re-estimation. Therefore, the objective function J_{obj} will decrease or remain the same in the re-estimation step. Since J_{obj} is bounded below and decreases after each re-assignment and re-estimation steps, the proposed semi-supervised clustering will converge to a (at least local) minimum in each interactive iteration.

After each interactive iteration, new constraints are given to the system. These new constraints might be in contradiction with some of the ones previously deduced by the system from the earlier user interactive iterations. For this reason and also for computational time matters, our system omits at each step some of the

constraints deduced at earlier steps. Therefore, the objective function J_{obj} may be different between different interactive iterations. And the convergence of the interactive semi-supervised model is thus not guaranteed. But we can verify the convergence of the model, practically, by determining, at the end of all interactive iterations, the global objective function which considers all feedback given by the user in all interactive iterations and then by verifying if this global objective function has improved or not after different interactive steps. This is a part of our current work.

3.3. Interactive interface

In order to allow the user to view the clustering results and to provide feedback to the systems, we implement an interactive interface as shown in Fig. 1.

The rectangle at the bottom right corner of Fig. 1 is the principal plane representing all presented clusters by their prototype images. In our system, the maximum number of cluster prototypes presented to the user on the principal plane is fixed at 30. The prototype image of each cluster is the most representative image of that cluster chosen as follows. In our model, we use the internal measure Silhouette-Width (SW) (Rousseeuw and Nov., 1987) to estimate the quality of each image in a cluster. The higher the SW value of an image in a cluster, the more representative this image is for the cluster. The prototype image of a cluster is thus the image with the highest SW value in the cluster. Any other internal measure could be used instead. The position of the prototype image of each cluster in the principal plane represents the position of the corresponding cluster center. It means that, if two cluster centers are close (or distant) in the n-dimensional feature space, their prototype images are close (or distant) in the 2D principal plane. For representing the cluster centers which are n-dimensional vectors in 2D plane, we use Principal Component Analysis (PCA) (Pearson, 1901); the principal plane consists of the two principal axes associated with the highest eigenvalues. The importance of an axis is represented by its inertia (the sum of the squared elements of this axis (Abdi and Williams, 2010)) or by the percentage of its inertia in the total inertia of all axes. In general, if the two principal axes explain (cumulatively) greater or equal to 80% of the total inertia, the PCA approach could lead to a nice 2D-representation of the prototype images. In our case, the accumulated inertia explained by the two first principal axes is about 65% for the Wang and PascalVoc2006 databases and about 20% for the Caltech101 and Corel30k image databases. As only a maximum of 30 clusters (and therefore 30 prototype images) can be shown to the user in an interactive iteration, a not very nice 2D-representation of prototype images does not influence on the results as long as the user can distinguish between the prototype images and have a rough idea of the distances between the clusters. When there are some prototype images which overlap each other, a slight modification of the PCA components can help to separate these images.

By clicking on a prototype image in the principal plane, the user can view the corresponding cluster. In Fig. 1, each cluster selected by the user is represented by a circle:

- The prototype image of this cluster is located at the center of the circle.
- The 10 most representative images (images with the highest SW values), which have not received feedback from the user in the previous iterations, are located in the first circle of images around the prototype image, near the center.

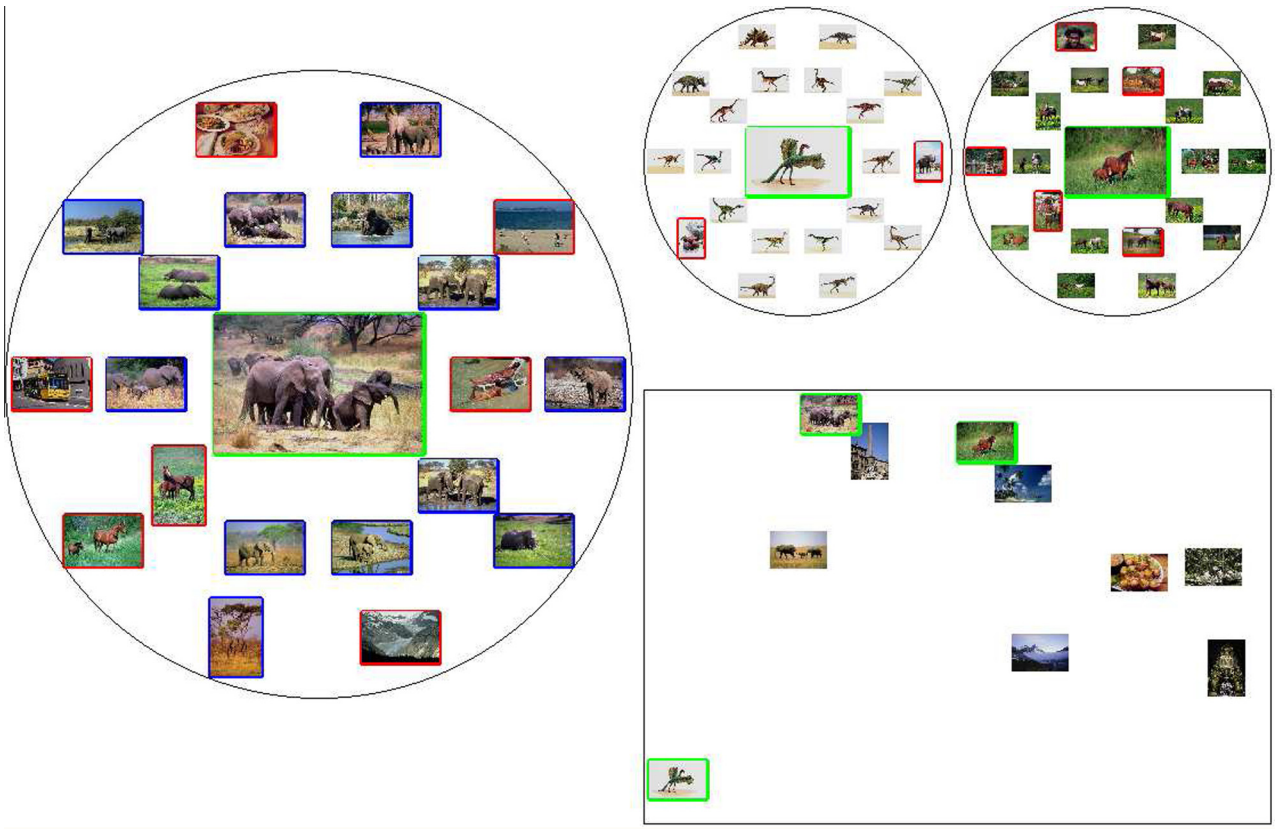


Fig. 1. 2D interactive interface. The rectangle at the bottom right corner represents the principal plane consisting of the two first principal axes (obtained by PCA) of the prototype images of all clusters. Each circle represents the details of a particular cluster selected by the user.

- The 10 least representative images (images with the smallest SW values), which have not received feedback from the user in the previous iterations, are located in the second circle of images around the prototype image, close to the cluster border.

By showing, for each iteration, the images which have not received user feedback in previous iterations, we wish to obtain feedback for different images.

The user can specify positive feedback and negative feedback (images in Fig. 1 with blue and red borders respectively[1]) for each cluster. The user can also change the cluster assignment of a given image by dragging and dropping the image from the original cluster to the new cluster. When an image is changed from cluster A to cluster B, it is considered as negative feedback for cluster A and positive feedback for cluster B. Therefore, after each interactive iteration, the process returns a positive image list and a negative image list for each cluster with which the user has interacted.

3.4. Pairwise constraint deduction

In each interactive iteration, user feedback is in the form of positive and negative images, while the supervised input information of the proposed semi-supervised clustering method are pairwise constraints between CF entries. Therefore, we have to deduce the pairwise constraints between CF entries from the user feedback.

At each interactive iteration and for each interacted cluster, all positive images should be in this cluster while negative images should move to another cluster. We consider that each image in the positive set is linked to each image in the negative set by a cannot-link, while all images in the positive set are linked by must-links. If we assume that all feedback is coherent between different interactive iterations, we try to group images, which should be in the same cluster according to the user feedback of all interactive iterations, in a group called *neighborhood*. We define:

- $Np = \{Np_i\}$ is the neighborhood list, each neighborhood $Np_i = \{x_j\}$ including a list of images which should be in a same cluster.
- $CannotNp = \{cannotNp_i\}$, each element $cannotNp_i = \{n_j\}$ including labels of the neighborhoods which should not be in the same cluster as Np_i. Two neighborhoods Np_i and Np_j are called cannot-link neighborhoods if there is at least one cannot-link between a point of Np_i and a point of Np_j.

After receiving the list of feedback in the current iteration, the lists *Np* and *CannotNp* are updated as follows:

1. Update based on positive feedback: For each cluster μ_h which receives interaction from the user:
   (a) Initialize $n_h \leftarrow -1$, n_h indicates the neighborhood including positive images of the cluster μ_h.
   (b) If all positive images of μ_h are not included in any neighborhood → create a new neighborhood for these positive images and assign n_h as the index of this neighborhood.
   (c) If some positive images of μ_h are already included in one or multiple neighborhoods → merge these neighborhoods (in the case of multiple neighborhoods) into one single neighborhood, insert the other positive images which are not included in any neighborhood to this neighborhood and update n_h as the index of this neighborhood. Also update the set *CannotNp* to signify that neighborhoods that had cannot-link with one of the neighborhoods which has merged, now have cannot-link with the new neighborhood.
2. Update based on negative feedback: For each negative image x_j of each cluster μ_h which receives interaction from the user:
   (a) If x_j is not included in any neighborhood → create a new neighborhood for x_j.
   (b) If x_j is already included in the neighborhood Np_{n_j}, and Np_{n_h} is the neighborhood corresponding to the positive images of the cluster μ_h, update the corresponding $cannotNp_{n_j}$ and $cannotNp_{n_h}$ to signify that Np_{n_j} and Np_{n_h} have cannot-link.

As we assume that the user feedback is coherent among different interactive iterations, all images in a same neighborhood should be in a same cluster and images of cannot-link neighborhoods should be in different clusters. There may be cannot-link images belonging to the same CF_i. There may also be simultaneous must-link and cannot-link between images of CF_i and images of CF_j. In such cases, these CF entries should be split into purer CF entries. To do so, we define a *seed* of an entry CF_i as a subset of images of CF_i so that the images of this *seed* are included in a same neighborhood. Therefore, an entry CF_i may contain some *seeds* corresponding to different neighborhoods and other images which are not included in any other neighborhood. Cannot-link may or may not exist between seeds of a CF entry. With each CF entry that should be split, we present the user with each pair of seeds, which do not have cannot-link between them, to demand more information (for each seed, the image which is closest to the center of the seed is presented):

- If the user indicates that there is must-link between these two seeds, these seeds and also their corresponding neighborhoods are merged.
- If the user indicates that there is cannot-link between these two seeds, update the corresponding *cannotCF* lists specifying that their two corresponding neighborhoods have cannot-link between them.

An entry CF_i is split as follows: if CF_i has p seeds, it should be split into p different CF entries; each new CF entry contains all points of a seed; every other point of CF_i which is not included in any seed is assigned to the CF entry corresponding to the closest seed. By splitting the necessary CF entries into purer CF entries, we can eliminate the case where cannot-link exists between images of a same CF or where must-link and cannot-link exist simultaneously between images of two different CF entries. Subsequently, pairwise constraints between CF entries can be deduced based on pairwise constraints between images as follows: if there is must-link (or respectively cannot-link) between two images of two CF entries, a must-link (or respectively cannot-link) is created between these two CF entries.

Concerning pairwise constraints between images, a simple and complete way to deduce them is to create must-link between each pair of images of a same neighborhood, and to create, for each pair of cannot-link neighborhoods (Np_i, Np_j), cannot-link between each image of Np_i and each image of Np_j. By deducing pairwise constraints between images in this way, the number of constraints between images can be very high, and therefore the number of constraints between CF entries could also be very high. The processing time of the semi-supervised clustering in the next phase could thus be very high due to the high number of constraints. There are different strategies for deducing pairwise constraints between images that could reduce the number of constraints and also the processing time. One of them is presented in Fig. 2 and others are described and tested in Section 4. In Fig. 2, must-links are created between positive images of each cluster while cannot-link are created between positive and negative images of each cluster (note

[1] For interpretation of color in Fig. 1, the reader is referred to the web version of this article.

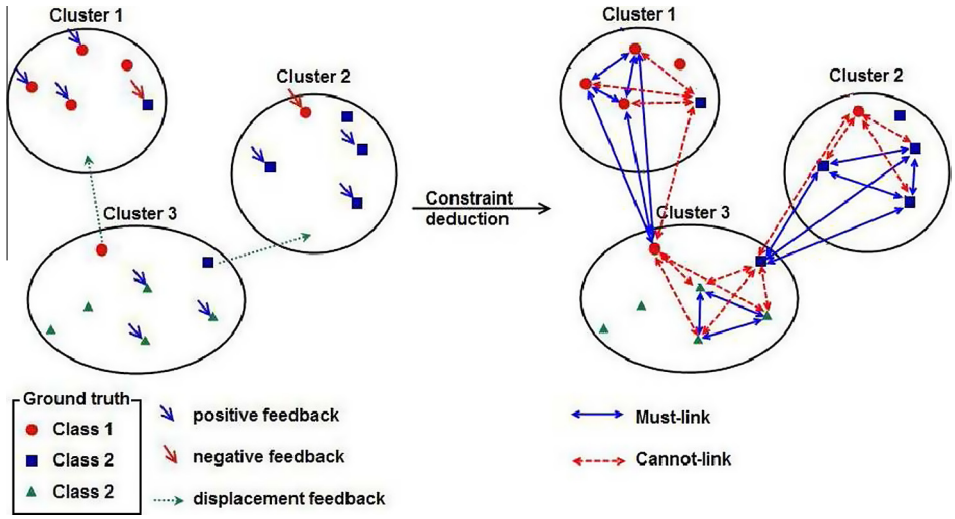


Fig. 2. Example of pairwise constraint deduction between images from the user feedback.

the displacement feedback corresponding to a negative image of the source cluster and a positive image of the destination cluster).

4. Experiments

In this section, we present some experimental results of our interactive semi-supervised clustering model. We also, experimentally, compare our semi-supervised clustering model with the semi-supervised HMRF-kmeans. When using the semi-supervised HMRF-kmeans in the re-clustering phase, the initial unsupervised clustering is k-means.

4.1. Experimental protocol

In order to analyze the performance of our interactive semi-supervised clustering model, we use different image databases (Wang[2] (1000 images divided into 10 classes), PascalVoc2006[3] (5304 images divided into 10 classes), Caltech101[4] (9143 images divided into 101 classes)). Note that in our experiments we use the same number of clusters as the number of classes in the ground truth. As presented in Section 3.3, the cluster prototype images are shown to the user on the principal plane; users can choose to view and interact with any cluster in which they are interested. For databases which have a small number of classes, such as Wang and PascalVoc2006, all prototype images can be shown on the principal plane. For databases which have a large number of classes, such as Caltech101, only a part of the prototype images can be shown for visualization. In our system, the maximum number of cluster prototypes shown to the user in each iteration is fixed at 30. We use two simple strategies for choosing clusters to be shown for each iteration: 30 clusters chosen randomly or iteratively chosen pairs of closest clusters until there are 30 clusters.

The external measures compare the clustering results with the ground truth, thus they are compatible for estimating the quality of the interactive clustering involving user interaction. As different external measures analyze the clustering results in a similar way (see Lai et al. (2012a)), we use, in this paper, the external measure V-measure (Rosenberg and Hirschberg, 2007). The greater the V-measure values are, the better the results (compared to the ground-truth).

Concerning feature descriptors, we implement the local descriptor rgSIFT (Van de Sande et al., 2008), an extension for color image of the SIFT descriptor (Lowe, 2004), that today is widely used for its high performance. The SIFT descriptor detects interest points from an image and describes the local neighborhood around each interest point by a 128-dimensional histogram of local gradient directions of image intensities. The rgSIFT descriptor of each interest point is computed as the concatenation of the SIFT descriptors calculated for the r and g components of the normalized RGB color space (Van de Sande et al., 2008) and the SIFT descriptor in the intensity channel, resulting in a 3×128-dimensional vector. The "Bag of words" (Sivic and Zisserman, 2003) approach is chosen to group local features of each image into a single vector. It consists in two steps. Firstly, K-means clustering is used to group local features of all images in the database according to a number *dictSize* of clusters. We then generate a dictionary containing *dictSize* visual words which are the centroids of these clusters. The feature vector of each image is a *dictSize* dimension histogram representing the frequency of occurrence of the visual words in the dictionary, by replacing each local descriptor of the image by the nearest visual word. Our experiments in Lai et al. (2012a) show that local descriptors are better than global descriptors regarding the external measures and the value $dictSize = 200$ is a good trade-off between the size of the feature vector and the performance. Therefore, in our experiments, we use the rgSIFT descriptor together with a visual word dictionary of size 200.

In order to undertake the interactive tests automatically, we implement a software agent, later referred to as "user agent" that simulates the behavior of the human user when interacting with the system (assuming that the agent knows all the ground truth containing the class label for each image). At each interactive iteration, clustering results are returned to the user agent by the system; the agent simulates the behavior of the user giving feedback to the system. For simulating the user behavior, we suggest some rules:

- At each interactive iteration, the user agent interacts with a fixed number of c clusters.
- The user agent uses two strategies for choosing clusters: randomly chosen c clusters, or iteratively chosen pairs of closest clusters until there are c clusters.

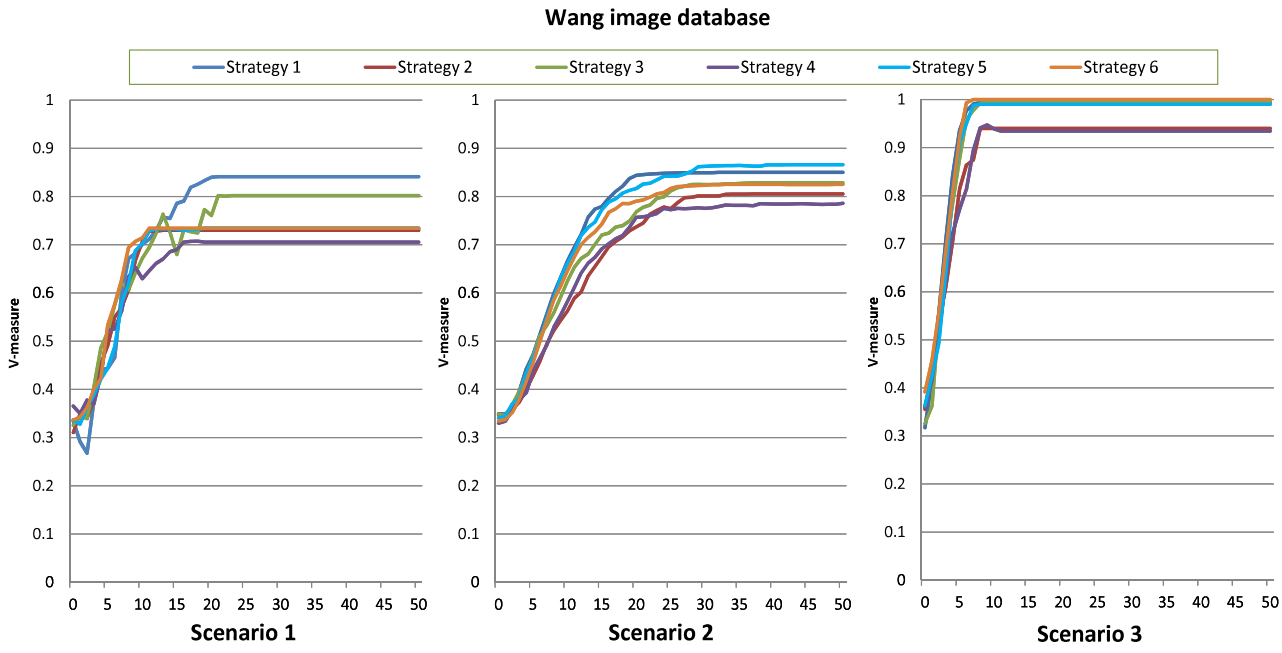
[2] http://wang.ist.psu.edu/docs/related/.
[3] http://pascallin.ecs.soton.ac.uk/challenges/VOC/.
[4] http://www.vision.caltech.edu/Image Datasets/Caltech101/.

Table 1
D75ifferent strategies for deducing pairwise constraints between images based on user feedback and on neighborhood information.

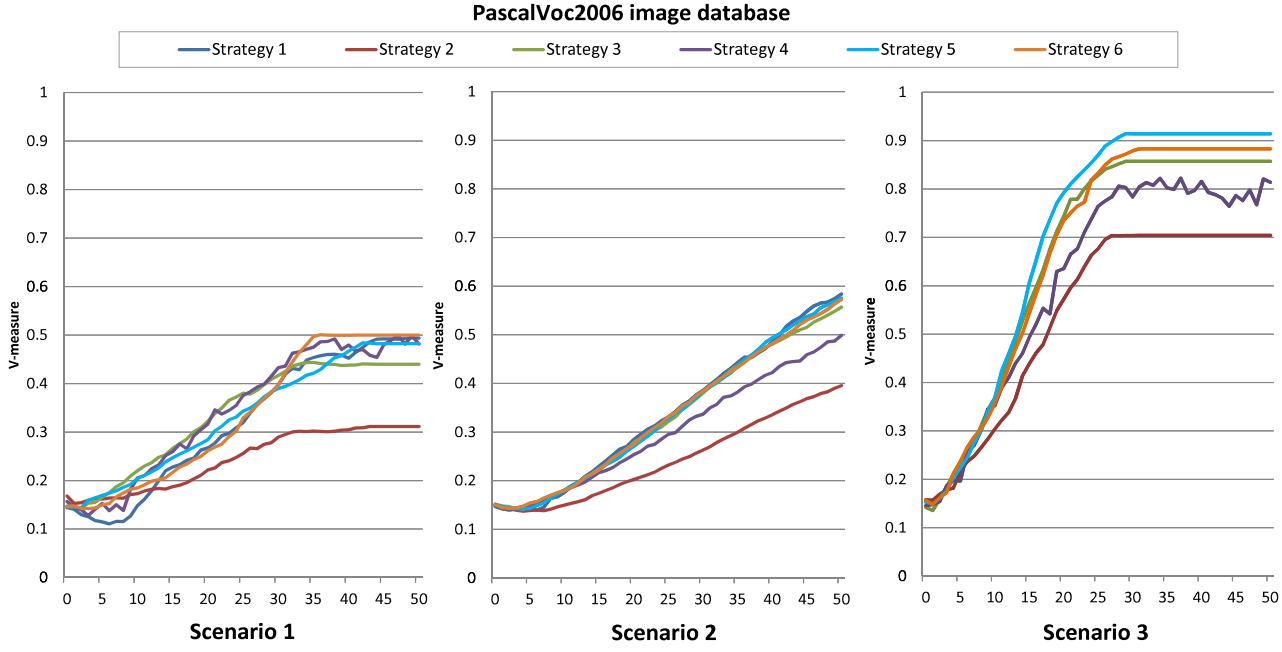
No.	Take into account	Details
1	• *All* user constraints of *all* interactive iterations • *All* deduced constraints of *all* interactive iterations	All constraints are created based on the neighborhood information: • Must-link between each pair of images of each neighborhood • Cannot-link between each image of each neighborhood $Np_i \in Np$ and each image of each neighborhood having cannot-link with Np_i (listed in $cannotNp_i$)
2	• *All* user constraints of *all* interactive iterations • *None* of deduced constraints	In each iteration, all possible user constraints are created: • Must-link between each pair of positive images of each cluster • Cannot-link between each pair of a positive image and a negative image of a same cluster
3	• *All* user constraints of *all* interactive iterations • *All* deduced constraints in the *current* iteration (deduced constraints in the previous iterations are eliminated)	• In each iteration, all possible user constraints are created as in Strategy 2 • Deduced constraints in the current iteration are created while updating the neighborhoods as follows: – If there is a must-link (or cannot-link) $(x_i, x_j), x_j \in Np_m$, deduced must-links (or cannot-links) $(x_i, x_l), \forall x_l \in Np_m$ are created – If there is a must-link (or cannot-link) $(x_i, x_j),\ x_i \in Np_m,\ x_j \in Np_n$, deduced must-links (or cannot-links) $(x_k, x_l),\ \forall x_k \in Np_m,\ \forall x_l \in Np_n$ are created
4	• User constraints between images and cluster centers of *all* interactive iterations • Deduced constraints between images and cluster centers in the *current* iteration (deduced constraints in the previous iterations are eliminated)	In each iteration, the positive image having the best internal measure (SW) value among all positive images of each cluster is the center of this cluster • Must-link/cannot-link user constraints are created in each iteration between each positive/negative image and the corresponding cluster center • Deduced constraints in the current iteration are created while updating the neighborhoods as follows: – If x_i and x_j must be in the same (or different) clusters (based on user feedback), $x_j \in Np_m$, deduced must-links (or cannot-links) are created between x_i and each center image of Np_m – If x_i and x_j must be in the same (or different) clusters (based on user feedback), $x_i \in Np_m, x_j \in Np_n$, deduced must-links (or cannot-links) are created between x_i and each center image of Np_n and between x_j and each center image of Np_m
5	• User constraints (must-links between the most distant images and cannot-links between the closest images) of *all* iterations • Deduced constraints (must-links between the most distant images and cannot-links between the closest images) of *all* iterations	• User constraints are created for each cluster in each iteration as follows: must-links are successively created between two positive images (at least one of them is not selected by any must-link) that have the longest distance until all positive images of the cluster are connected by these must-links; cannot-links are created between each negative image and the nearest positive image of the cluster • Deduced constraints are created in each iteration as follows: must-links for each neighborhood are successively created between two images that have the longest distance until all images of this neighborhood are connected by these must-links; cannot-links are deduced, for each pair of cannot-link neighborhoods (Np_i, Np_j), between each image of Np_i and the nearest image of Np_j and between each image of Np_j and the nearest image of Np_i
6	Same idea as in strategy 5, but the size of the neighborhoods is considered while creating deduced cannot-links	User constraints and deduced must-link constraints are created as in Strategy 5. For each pair of cannot-link neighborhoods, deduced cannot-links are only created between each image of the neighborhood that has the least number of images and the nearest image of the neighborhood that has the most images

- The user agent determines the image class (in the ground truth) corresponding to each cluster by the most represented class among the 21 presented images of the cluster. The number of images of this class in the cluster must be greater than a threshold *MinImages*. If this is not the case, this cluster can be considered as a noise cluster. In our experiments, $MinImages = 5$ for databases having a small number of classes (Wang, PascalVoc2006), and $MinImages = 2$ for databases having a large number of classes (Caltech101).
- When several clusters (among chosen clusters) correspond to a same class, the cluster in which the images of this class are the most numerous (among the 21 shown images of the cluster) is chosen as the principal cluster of this class. The classes of the other clusters are redefined as usual, but neutralize the images from this class.
- In each chosen cluster, all images, where the result of the algorithm corresponds to the ground truth, are labeled as positive samples of this cluster, while the others are negative samples of this cluster. All negative samples are moved to the cluster (among chosen clusters) corresponding to their class in the ground truth.

As presented in Section 3.4, we have to deduce pairwise constraints between images based on user feedback in each iteration and also on the neighborhood information. User feedback is in the form of positive and negative images of each cluster (the image which is displaced from one cluster to another cluster is considered as a negative image of the source cluster and a positive image of the destination cluster). The neighborhood information is in the form of the lists $Np = \{Np_i\}$ and $CannotNp = \{cannotNp_i\}$, where each neighborhood Np_i contains images which should be in a same cluster and $cannotNp_i$ identifies the list of neighborhoods having cannot-link with Np_i. Neighborhood information is deduced from user feedback during all interactive iterations, as presented in Section 3.4. Pairwise constraints between images will be used directly for the semi-supervised HMRF-Kmeans, while they have to be deduced into pairwise constraints between CF entries (see Section 3.4) to be used by our proposed semi-supervised clustering. We divide pairwise constraints between images into two kinds: *user constraints* and *deduced constraints*. *User constraints* are created directly, based on user feedback in each iteration, while *deduced constraints* are created by deduction rules. For instance, in the first iteration, the user marks x_1, x_2 as positive images and x_3 as a negative image of cluster μ_i; while in the second iteration, he marks x_1 and x_4 as positive images of cluster μ_j. The created user constraints are: must-link between positive images in the first iteration (x_1, x_2), must-link between positive images in the second iteration (x_1, x_4), and cannot-links between positive and negative images in the first iteration $(x_1, x_3),\ (x_2, x_3)$. As there are must-links $(x_1, x_2),\ (x_1, x_4)$, there is also a deduced must-link (x_2, x_4). In addition deduced



(a) Results on the Wang image database



(b) Results on the PascalVoc2006 image database

Fig. 3. Results of our proposed interactive semi-supervised clustering model during 50 interactive iterations on the Wang and PascalVoc2006 image databases, using 6 strategies for deducing pairwise constraints. The horizontal axis specifies the number of iterations.

cannot-link (x_3, x_4) is created, based on the must-link (x_1, x_4) and the cannot-link (x_1, x_3). We can see that deduced constraints can be created based on neighborhood information. In our experiments, we use different strategies for deducing pairwise constraints between images. These strategies are detailed in Table 1.

4.2. Experimental results

4.2.1. Analysis of different strategies for deducing pairwise constraints between images

The first set of experiments aims at evaluating the performance of our interactive semi-supervised clustering model using different strategies for deducing pairwise constraints between images. Note that constraints between CF entries should be deduced from constraints between images, before being used in the re-clustering phase. We use the Wang and the PascalVoc2006 image databases for these experiments. For these two databases, we propose three test scenarios (note that c specifies the number of clusters which are chosen for interacting in each iteration):

- Scenario 1: $c = 5$ closest clusters are chosen.
- Scenario 2: $c = 5$ clusters are randomly chosen.
- Scenario 3: $c = 10$, all cluster are chosen (Wang and PascalVoc2006 both have 10 clusters).

Table 2
Average standard deviation of 10 executions of the scenario 2 after 50 interactive iterations corresponding to the experiments of our proposed interactive semi-supervised clustering model shown in Fig. 3(a) and (b).

	Average standard deviation	
	Wang database	PascalVoc2006 database
Strategy 1	0.033	0.022
Strategy 2	0.044	0.017
Strategy 3	0.045	0.025
Strategy 4	0.047	0.022
Strategy 5	0.036	0.024
Strategy 6	0.044	0.026

Table 3
Processing time after 50 interactive iterations of the experiments of our proposed interactive semi-supervised clustering model shown in Fig. 3(a) and (b).

	Scenario 1	Scenario 2	Scenario 3
Wang database			
Strategy 1	1 h 58′	2 h 24′	1 h 41′
Strategy 2	9′	12′	10′
Strategy 3	31′	19′	47′
Strategy 4	8′	9′	8′
Strategy 5	8′	9′	9′
Strategy 6	6′	8′	8′
PascalVoc2006 database			
Strategy 1	16 d 12 h	14 d 11 h	
Strategy 2	2 h 55′	4 h 02′	5 h 6′
Strategy 3	3 h 23′	6 h 39′	6 h 22′
Strategy 4	1 h 9′	1 h 33′	2 h 17′
Strategy 5	3 h 33′	4 h 42′	3 h 10′
Strategy 6	1 h 3′	1 h 21′	2 h

Note that our experiments are carried out automatically, i.e. the feedback is given by a software agent simulating the behaviors of the human user when interacting with the system. In fact, the human user can give feedback by clicking for specifying the positive and/or negative images of each cluster or by dragging and dropping the image from a cluster to another cluster. For each cluster selected by the user, only 21 images of this cluster are displayed (see Fig. 1). Therefore, for interacting with 5 clusters (scenarios 1, 2) or 10 clusters (scenario 3), the user has to realize respectively a maximum of 105 or 210 mouse clicks in each interactive iteration. These upper bounds do not depend on neither the size of the database nor the pairwise constraint deduction strategy, and in practice the number of clicks that the user has to provide is far lower. However, the number of deduced constraints may be much greater than the user's clicks (and this number depends on the database size and on the pairwise constraint deduction strategy). When applying the interactive semi-supervised clustering model in the indexing phase, the user is generally required to provide as much feedback as possible for having a good indexing structure which could lead to better results in the further retrieval phase. Therefore, in the case of the indexing phase, the proposed number of clicks seems tractable.

Fig. 3(a) and (b) show, respectively, the results during 50 interactive iterations of our proposed interactive semi-supervised clustering model on the Wang and PascalVoc2006 image databases, with the three proposed scenarios. The results are shown according to 6 strategies for deducing pairwise constraints presented in Table 1. The vertical axis specifies the V-measure values, while the horizontal axis specifies the number of iterations. Note that with each selected cluster, the user agent gives all possible feedback. Therefore, for each scenario, the numbers of user feedback are equivalent between different iterations and between different strategies. As in scenario 2, clusters are randomly chosen, we realize this scenario 10 times for each database. The curves of the scenario 2 shown in Fig. 3(a) and (b) represent the mean values of the V-measure over these 10 executions at each iteration. The average standard deviation of each strategy after 50 iterations is presented in Table 2. The corresponding execution time for these experiments is presented in Table 3 (note that for the scenario 2, the average execution times of 10 executions are shown). The experiments are executed using a normal PC with 2 GB of RAM.

We can see that the clustering results progress, in general, after each interactive iteration, in which the system re-clusters the dataset by considering the constraints deduced from accumulated user feedback. In most cases, the clustering results converge after only a few iterations. This may be due to the fact that no new knowledge is provided. Moreover, we can easily see that the clustering results are better and converge more quickly when the number of chosen clusters (and therefore the number of constraints) in each interactive iteration is higher (scenario 3 gives better results and converges more quickly than scenarios 1 and 2). In addition, for both image databases, scenario 2, in which clusters are randomly chosen for interacting, gives better results than scenario 1, in which the closest clusters are chosen. When selecting the closest clusters there may be only several clusters that always receive user feedback; thus the constraint information is less than when all the clusters could receive user feedback when we randomly select the clusters.

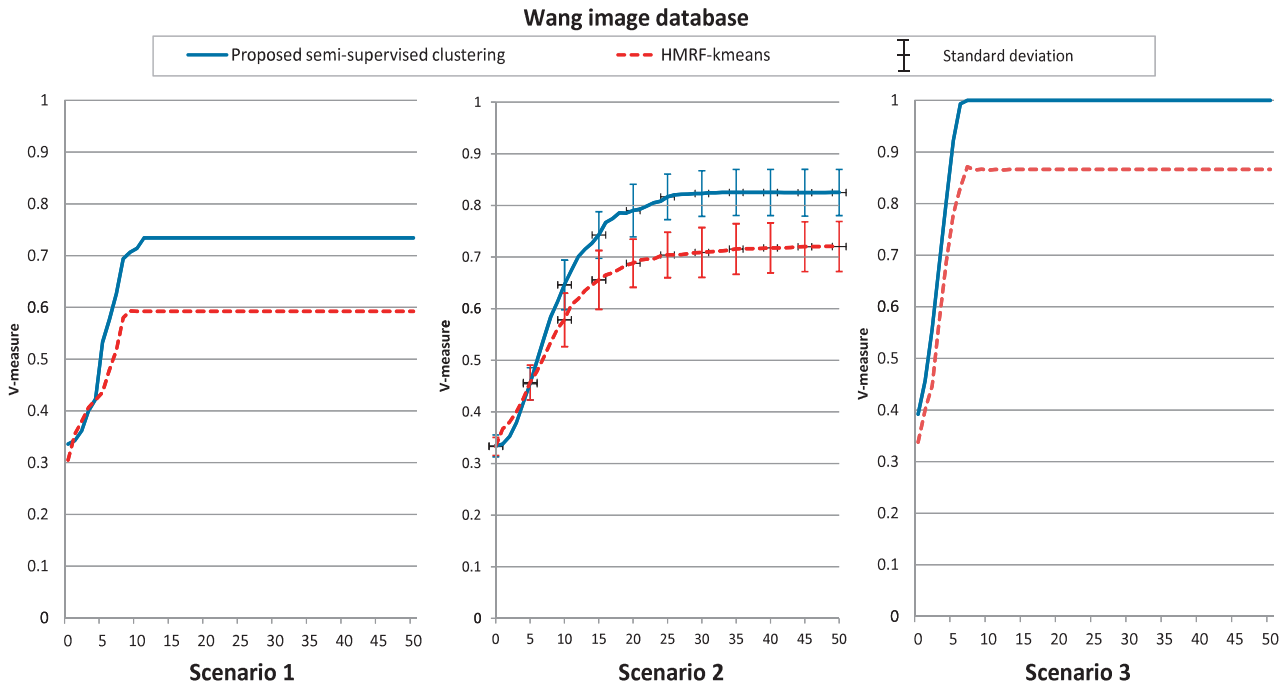
As regards different strategies for deducing pairwise constraints, we can see that for each database, the average standard deviations over 10 executions of the scenario 2 are similar for all scenarios. Therefore, we can compare different strategies based on the mean values shown on Figs. 3(a) and (b). We can see that:

- Strategy 1 shows, in general, very good performance but the processing time is huge because it uses all possible user constraints and deduced constraints created during all iterations.
- Strategy 2, the only strategy uniquely using user constraints, generally gives the worst results; thus deduced constraints are needed for better performance. Its processing time is also high due to the large number of user constraints.
- Strategy 3 shows good or very good performance but some oscillations exist between different iterations because, when overlooking previously deduced constraints, some important constraints may be omitted. Its processing time is high.
- Strategy 4 gives better results than strategy 2, but the results are unstable because this strategy also overlooks previously deduced constraints. It has good execution time while reducing the number of constraints.
- Strategy 5 generally gives good or very good results by keeping important constraints (must-links between the most distant images and cannot-links between the closest images), but its processing time is still high.
- Strategy 6, by reducing the deduced cannot-link constraints from strategy 5, gives in general very good results in low execution time.

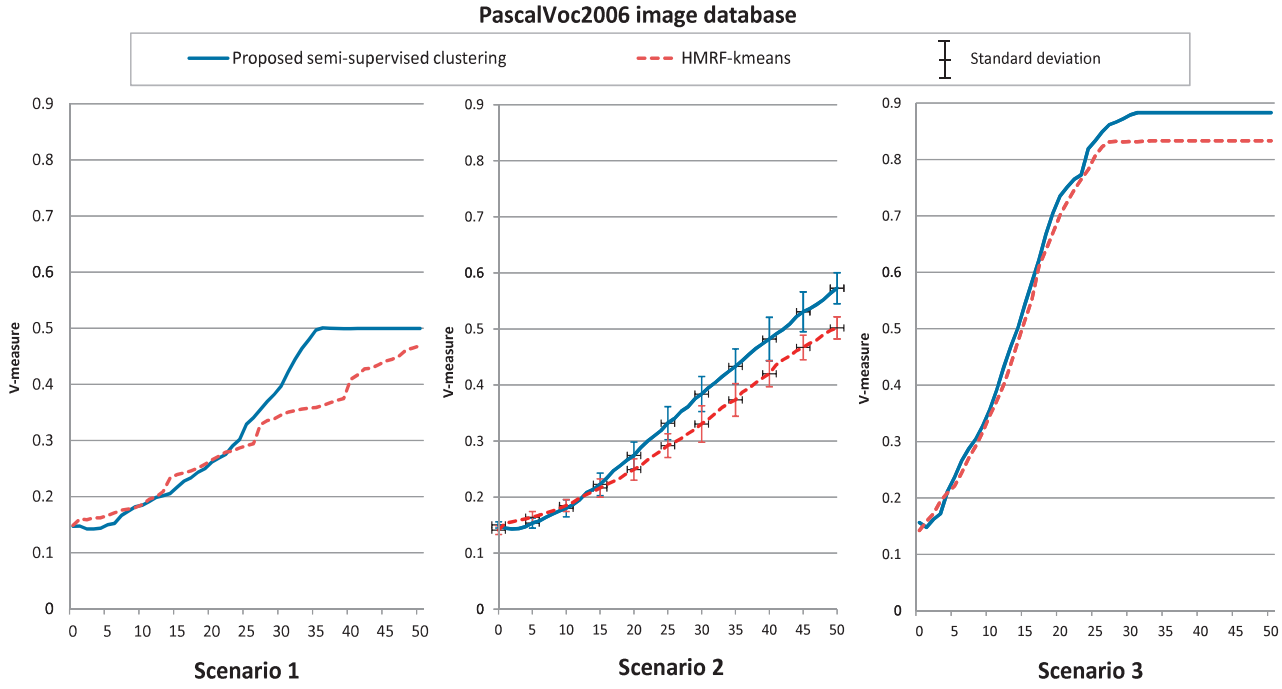
We can conclude, from this analysis, that strategy 6 shows the best trade-off between performance and processing time. This strategy will be used in further experiments.

4.2.2. Comparison of the proposed semi-supervised clustering model and the semi-supervised HMRF-kmeans

Figs. 4(a) and (b) represent, respectively, the clustering results for 50 interactive iterations on the Wang and the PascalVoc2006 image databases when using our proposed semi-supervised clustering and the semi-supervised HMRF-kmeans in the re-clustering phase. The three scenarios described in Section 4.2.1 and strategy 6, for deducing pairwise constraints between images, are used. Note that the results of scenario 2 represent the mean values and



(a) Comparison results on the Wang image database



(b) Comparison results on the PascalVoc2006 image database

Fig. 4. Comparison of the proposed semi-supervised clustering and the semi-supervised HMRF-kmeans with 50 interactive iterations using Strategy 6 in Table 1 for deducing the pairwise constraints between images. The horizontal axis represents the number of iterations.

also the standard deviations over 10 executions at each iteration. The corresponding processing time is presented in Table 4. We can see that in all scenarios, our proposed method gives better results, in lower processing time than the HMRF-kmeans. While the pairwise constraints between images are directly used by the HMRF-kmeans, they are deduced in pairwise constraints between CF entries for being used by our proposed semi-supervised clustering. A CF entry groups a list of similar images, thus many pairwise constraints between images can be represented by only one pairwise constraints between CF entries. Therefore, with a same set of user feedback, the number of pairwise constraints between images is generally greater than the number of the pairwise constraints between CF entries. Thus the processing time of the HMRF-kmeans is much higher than the processing time of our proposed method. Moreover, when a pairwise constraint (CF_i, CF_j) is deduced from the pairwise constraint of the corresponding images (x_k, x_l), $x_k \in CF_i$, $x_l \in CF_j$, the constraint (CF_i, CF_j) forces the grouping or separating of not only the two images x_i and x_j but also the

Table 4
Processing time after 50 interactive iterations corresponding to the experiments presented in Fig. 4(a) and (b) of the proposed semi-supervised clustering and of the semi-supervised HMRF-kmeans. Strategy 6 in Table 1 for deducing pairwise constraints is used.

	Scenario 1	Scenario 2	Scenario 3
Wang database			
Proposed semi-supervised clustering	6′	8′	8′
HMRF-kmeans	7′	11′	10′
PascalVoc2006 database			
Proposed semi-supervised clustering	1 h 3′	1 h 21′	2 h ′
HMRF-kmeans	2 h 16′	3 h 10′	2 h 49′

other images included in CF_i and CF_j. And therefore, the clustering results given by our proposed method are better than the ones given by the HMRF-kmeans. Moreover, similar to the experiments presented in Section 4.2.1, the scenario 2 in which the clusters are randomly chosen for interacting gives better results than the scenario 1 in which the closest clusters are chosen. In the following experiments on the Caltech101 image database, we present only the clustering results when the clusters are randomly chosen.

As the Caltech101 database has a large number of classes (101 classes), we do not show all clusters to the user on the principal plane but only a small number of clusters (we fix the maximum number of cluster that could be shown on the principal plane to 30). There are two strategies for choosing clusters to be shown on the principal plane: either clusters are randomly chosen or the closest clusters are chosen. The user agent randomly chooses, among shown clusters, c clusters for interacting. We use 4 scenarios for the experiments on the Caltech101 image database:

- Scenario 4: the closest clusters are chosen to be shown to the user, $c = 5$ clusters are chosen by the user agent for interacting.
- Scenario 5: clusters are randomly chosen to be shown to the user, $c = 5$ clusters are chosen by the user agent for interacting.
- Scenario 6: the closest clusters are chosen to be shown to the user, $c = 10$ clusters are chosen by the user agent for interacting.
- Scenario 7: clusters are randomly chosen to be shown to the user, $c = 10$ clusters are chosen by the user agent for interacting.

Fig. 5 compares our proposed semi-supervised clustering and the HMRF-kmeans during 50 interactive iterations on the Caltech101 image database. The corresponding processing time is

Table 5
Processing time after 50 interactive iterations corresponding to the experiments on the Caltech101 image database in Fig. 5 for the proposed semi-supervised clustering and for the semi-supervised HMRF-kmeans. Strategy 6 in Table 1 for deducing pairwise constraints is used.

	Proposed semi-supervised clustering	HMRF-kmeans
Scenario 4	13 h 26′	48 h 33′
Scenario 5	8 h 4′	33 h 45′
Scenario 6	33 h 34′	157 h 26′
Scenario 7	50 h 12′	101 h 11′

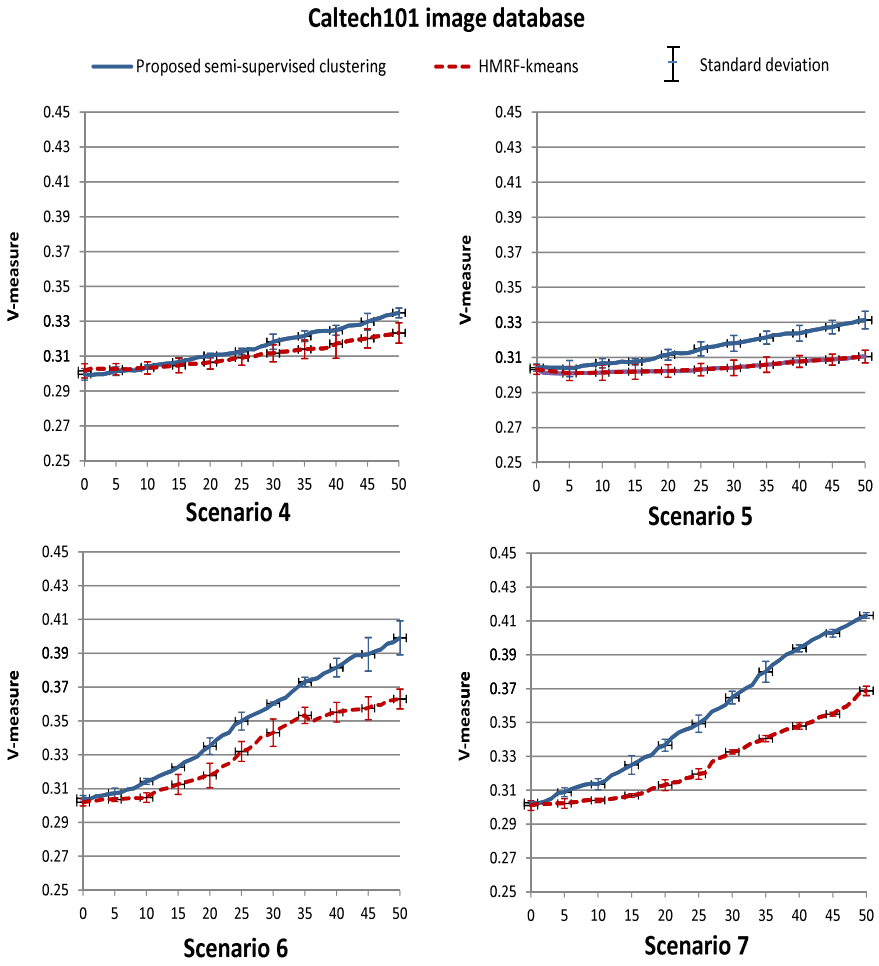


Fig. 5. Comparison of the proposed semi-supervised clustering and the semi-supervised HMRF-kmeans on the Caltech101 image database for 50 interactive iterations. The strategy 6 in Table 1 for deducing pairwise constraints are used. The horizontal axis represents the number of iterations.

presented in Table 5. As in all these four scenarios, the clusters are randomly chosen for interacting, we realize each scenario 5 times and present in Fig. 5 the mean values and also the standard deviations over 5 executions. The results shows that our proposed semi-supervised clustering outperforms the HMRF-kmeans in all four scenarios. Moreover, the clustering results are also better when the number of feedback for each iteration is high (scenarios 6 and 7 give better results than scenarios 4 and 5).

5. Conclusion

A new interactive semi-supervised clustering model for indexing image databases is presented in this article. After receiving user feedback for each interactive iteration, the proposed semi-supervised clustering re-organizes the dataset by considering the pairwise constraints between CF entries deduced from the user feedback. We present an interactive interface allowing the user to view, and to provide feedback. Experimental analysis, using a software user agent for simulating human user behavior, shows that our model improves the clustering results at each interactive iteration. Note that our experimental scenarios are realistic, they can be realized by a real user as the number of clicks required is tractable. The experiments on different image databases (Wang, PascalVoc2006, Caltech101), presented in this paper, also show that our semi-supervised clustering outperforms the semi-supervised HMRF-kmeans (Basu et al., 2004) in both performance and processing time.

Moreover, we propose and compare, experimentally, different strategies for deducing pairwise constraints from the user feedback accumulated from all interactive iterations. The experimental results show that strategy 6 in Table 1, which keeps only the most important constraints (must-links between the most distant images and cannot-links between the closest images), provides the best trade-off between the performance and the processing time. Strategy 6 is therefore the most suitable, in our context involving the user in the indexing phase by clustering.

Our future work aims to verify our proposed semi-supervised clustering model with larger image databases such as Corel30k, MIRFLICKR, to prove experimentally the convergence of our algorithm, and to look for different strategies for deducing the pairwise constraints or for representing the clustering results that could improve the performance of our model in the context of huge image databases.

References

Abdi, H., Williams, L.J., 2010. Principal component analysis.

Basu, S., Banerjee, A., Mooney, R.J., 2002. Semi-supervised clustering by seeding. In: Proceedings of the Nineteenth International Conference on Machine Learning, ICML '02. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 27–34.

Basu, S., Bilenko, M., Mooney, R.J., 2004. A probabilistic framework for semi-supervised clustering. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04. ACM, New York, NY, USA, pp. 59–68.

Bayer, R., McCreight, E.M., 1972. Organization and maintenance of large ordered indexes. Acta Informatica 1, 173–189.

Beckmann, N., Kriegel, H.-P., Schneider, R., Seeger, B., 1990. The r*-tree: an efficient and robust access method for points and rectangles. In: Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data, SIGMOD '90. ACM, New York, NY, USA, pp. 322–331..

Bentley, J.L., 1975. Multidimensional binary search trees used for associative searching. Commun. ACM 18 (9), 509–517.

Berchtold, S., Keim, D.A., Kriegel, H.-P., 1996. The x-tree: An index structure for high-dimensional data. In: Proceedings of the 22th International Conference on Very Large Data Bases VLDB '96. Morgan Kaufmann Publishers, San Francisco, CA, USA, pp. 28–39.

Dubey, A., Bhattacharya, I., Godbole, S., 2010. A cluster-level semi-supervision model for interactive clustering. In: Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part I. ECML PKDD'10. Springer-Verlag, Berlin, Heidelberg, pp. 409–424.

Guttman, A., 1984. R-trees: a dynamic index structure for spatial searching. In: International Conference on Management of Data. ACM, pp. 47–57.

Henrich, A., Six, H.W., Widmayer, P., 1989. The lsd tree: spatial access to multidimensional and non-point objects. In: Proceedings of the 15th International Conference on Very Large Data Bases, VLDB '89. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 45–53.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Comput. Surv. 31 (3), 264–323.

Katayama, N., Satoh, S., 1997. 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. ACM, New York, NY, USA, pp. 369–380.

Kulis, B., Basu, S., Dhillon, I., Mooney, R., 2005. Semi-supervised graph clustering: a kernel approach. In: ICML 05: Proceedings of the 22nd International Conference on Machine Learning. ACM Press, pp. 457–464.

Lai, H.P., Visani, M., Boucher, A., Ogier, J.-M., 2012a. An experimental comparison of clustering methods for content-based indexing of large image databases. Pattern Anal. Appl. 15, 345–366.

Lai, H.P., Visani, M., Boucher, A., Ogier, J.-M., 2012b. Unsupervised and semi-supervised clustering for large image database indexing and retrieval. In: IEEE International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the, Future (RIVF), pp. 1–6.

Lance, G.N., Williams, W.T., 1967. A general theory of classificatory sorting strategies II. Clustering systems. Comput. J. 10 (3), 271–277.

Likas, A., Vlassis, N., Verbeek, J.J., 2003. The global k-means clustering algorithm. Pattern Recognit. 36 (2), 451–461.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60 (2), 91–110.

Nievergelt, J., Hinterberger, H., Sevcik, K.C., 1988. Readings in database systems. The Grid File: An Adaptable, Symmetric Multikey File, Structure. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 582–598.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Philos. Mag. 6 (2), 559–572.

Robinson, J.T., 1981. The k-d-b-tree: a search structure for large multidimensional dynamic indexes. In: Proceedings of the 1981 ACM SIGMOD international conference on Management of data, SIGMOD '81. ACM, New York, NY, USA, pp. 10–18.

Rosenberg, A., Hirschberg, J., 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural, Language Learning (EMNLP-CoNLL), pp. 410–420.

Rousseeuw, P., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20 (1), 53–65.

Sellis, T.K., Roussopoulos, N., Faloutsos, C., 1987. The r+-tree: A dynamic index for multi-dimensional objects. In: Proceedings of the 13th International Conference on Very Large Data Bases, VLDB '87. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 507–518.

Sivic, J., Zisserman, A., 2003. Video google: a text retrieval approach to object matching in videos. In: . Proceedings of Ninth IEEE International Conference on Computer Vision 2003, pp. 1470–1477.

Van de Sande, K.E.A., Gevers, T., Snoek, C.G.M., 2008. Evaluation of color descriptors for object and scene recognition. In: IEEE Conference on Computer Vision and Pattern Recognition.

Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S., 2001. Constrained k-means clustering with background knowledge. In: Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 577–584.

White, D.A., Jain, R., 1996. Similarity indexing with the ss-tree. In: Proceedings of the Twelfth International Conference on Data Engineering, ICDE '96. IEEE Computer Society, Washington, DC, USA, pp. 516–523.

Xu, Rui, Wunsch II, Donald, 2005. Survey of clustering algorithms. IEEE Trans. Neural Networks 16 (3), 645–678.

Zhang, T., Ramakrishnan, R., Livny, M., 1996. Birch: an efficient data clustering method for very large databases. In: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, SIGMOD '96. ACM, New York, NY, USA, pp. 103–114.

International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 25, No. 4 (2011) 449–473

DOI: 10.1142/S0218001411008634

World Scientific
www.worldscientific.com

NAVIGALA: AN ORIGINAL SYMBOL CLASSIFIER BASED ON NAVIGATION THROUGH A GALOIS LATTICE

M. VISANI*, K. BERTET and J.-M. OGIER

Department of Computer Science — Laboratory L3I — University of La Rochelle
Ple Sciences et Technologie, Avenue Michel Crpeau
17042 La Rochelle Cedex 1, France
**muriel.visani@univ-lr.fr*

This paper deals with a supervised classification method, using Galois Lattices based on a navigation-based strategy. Coming from the field of data mining techniques, most literature on the subject using Galois lattices relies on selection-based strategies, which consists of selecting/choosing the concepts which encode the most relevant information from the huge amount of available data. Generally, the classification step is then processed by a classical classifier such as the k-nearest neighbors rule or the Bayesian classifier. Opposed to these selection-based strategies are navigation-based approaches which perform the classification stage by navigating through the complete lattice (similar to the navigation in a classification tree), without applying any selection operation. Our approach, named Navigala, proposes an original navigation-based approach for supervised classification, applied in the context of noisy symbol recognition. Based on a state of the art dealing with Galois Lattices classification based methods, including a comparison between possible selection and navigation strategies, this paper proposes a description of NAVIGALA and its implementation in the context of symbol recognition. Some objective quantitative and qualitative evaluations of the approach are proposed, in order to highlight the relevance of the method.

Keywords: Supervised classification; Galois (or concept) lattice; decision tree; graphical documents; graphical symbols recognition.

1. Introduction

Galois lattices (or *concept lattices*) were first introduced in a formal way in the graph and ordered structures theory.[3,9] Later, it was developed in the field of Formal Concept Analysis (FCA)[13] for data analysis and classification. The concept lattice structure, based on the notion of *concept*, enables data description while preserving its diversity.

Galois lattice is a graph with a structure similar to that of a tree. It provides a representation of all the possible correspondences between a set of *objects* (or examples) O and a set of *attributes* (or features) I. Whereas in decision trees the path from the root to a given leaf is unique, in Galois lattices there are multiple paths from

the maximal boundary to a given terminal concept. The technological improvements of the last decades enable the use of these structures for data mining problems though they are exponential in space/time (worst case). It has to be noted that in practice, in most cases, the size of the lattice remains reasonable. Recent studies realized by Mephu Nguifo *et al.*[22,20] provide a comprehensive review of some of the state-of-the-art classification approaches based on concept lattices, which are generally based on a selection of the most pertinent concepts in the lattice. This review shows that these methods are able to catch up with (and sometimes even outperform) more classical approaches such as decision trees. Multiple approaches have been proposed so far, confirming the relevance of using a Galois lattice for a classification task. Among these approaches, we can mention LEGAL and LEGAL-E,[19] Galois,[7] Zenou and Samuelides',[28] GRAND[25] and RULEARNER[27] which are based on a selection of the concepts directly, the CIBLe approach[21] which is based on object filtering and the CLNN and CLNB methods[34] where contextual rules are used.

The first objective of this paper is to introduce an original supervised classification method that does not rely on a selection step, named Navigala. Indeed, differing from the state-of-the-art approaches, Navigala relies on navigation through the lattice. The second objective of this paper is to compare Navigala (both formally and experimentally) to several other classification methods based on the Galois lattice. Navigala has been developed in the field of content-based graphical documents indexing; it is dedicated to noisy symbol recognition. These symbols, which are issued from digitized paper documents such as architectural or electrical plans, are most often noisy. In the proposed scheme, each symbol image is represented by a feature vector (signature), which may be statistical, structural or hybrid. The signatures are discretized to obtain discrete attributes and then classified using the Galois lattice.

This paper is organized as follows. In Sec. 2, we describe the Galois lattice structure and its properties and provide a review of the state-of-the-art classification approaches based on a Galois lattice. In Sec. 3, we present our navigation-based approach named Navigala. Then, Sec. 4 proposes various experimental results assessing the effectiveness of the proposed approach and an experimental comparative study towards *selection*-oriented approaches and decision trees. The conclusion and future works are presented in Sec. 5.

2. Description of a Galois Lattice

2.1. *Definition*

The *concept lattice* is a particular graph defined and generated from a relation R between objects O and attributes I. This graph is composed of a set of *concepts* ordered by a relation verifying the properties of a lattice, i.e. an order relation (transitive, reflexive and antisymmetric relation) such that, for each pair of concepts in the graph, there exists both a lower bound and an upper bound.

We associate to a set of objects $A \subseteq O$ the set $f(A)$ of attributes in relation R with the objects of A:

$$f(A) = \{x \in I \,|\, pRx \,\forall\, p \in A\}$$

Dually, to a set of attributes $B \subseteq I$, we define the set $g(B)$ of objects in relation with the attributes of B:

$$g(B) = \{p \in O \,|\, pRx \,\forall\, x \in B\}$$

These two functions f and g defined between objects and attributes form a *Galois correspondence*. The relations between the set of objects and the set of attributes are described by a *formal context*. A formal context C is a triplet $C = (O, I, R)$ (or $C = (O, I, (f, g))$) represented by a table. Table 1 gives an example of a formal context composed of a set of ten objects described by six attributes (a_1, a_2, b_1, b_2, c_1 and c_2). Additional information (class, feature and interval) is given in italics; for more details about this additional information please refer to Sec. 3.1.

A *formal concept* represents maximal objects-attributes correspondences (following relation R) by a pair (A, B) with $A \subseteq O$ and $B \subseteq I$, which verifies $f(A) = B$ and $g(B) = A$. The whole set of formal concepts thus corresponds to all the possible maximal correspondences between a set of objects O and a set of attributes I.

Two formal concepts (A_1, B_1) and (A_2, B_2) are in relation in the lattice when they verify the following inclusion property:

$$(A_1, B_1) \leq (A_2, B_2) \Longleftrightarrow \left\| \begin{array}{l} A_2 \subseteq A_1 \\ \text{(equivalent to } B_1 \subseteq B_2\text{)} \end{array} \right.$$

The whole set of formal concepts fitted out by the order relation $\leq$ is called *concept lattice* or *Galois lattice* because it verifies the lattice property: the relation $\leq$

Table 1. Example of formal context and (in italics) the classes of the objects and the features and intervals defining the attributes (for more details about the information in italics please refer to Sec. 3.1).

		Attributes					
		Feature f_1		*Feature* f_2		*Feature* f_3	
		$\mathbf{a_1}$	$\mathbf{a_2}$	$\mathbf{b_1}$	$\mathbf{b_2}$	$\mathbf{c_1}$	$\mathbf{c_2}$
Class	Id	*[0–3]*	*[6–20]*	*[0–4]*	*[12–20]*	*[0–2]*	*[11–20]*
1	1	X		X			X
	2	X		X			X
2	3	X			X		X
	4	X			X		X
	5	X			X		X
3	6		X		X		X
	7		X		X		X
	8		X		X		X
4	9		X	X		X	
	10		X		X	X	

is clearly an order relation, and for each pair of concepts (A_1, B_1) and (A_2, B_2), there exists a greatest lower bound (resp. a least upper bound) called *meet* (resp. *join*) denoted $(A_1, B_1) \wedge (A_2, B_2)$ (resp. $(A_1, B_1) \vee (A_2, B_2)$) defined by:

$$(A_1, B_1) \wedge (A_2, B_2) = (g(B_1 \cap B_2), \quad (B_1 \cap B_2)) \tag{1}$$

$$(A_1, B_1) \vee (A_2, B_2) = ((A_1 \cap A_2), \quad f(A_1 \cap A_2)) \tag{2}$$

Therefore, a lattice contains a minimum (resp. maximum) element according to the relation $\leq$ called the *bottom* (resp. *top*) of the lattice, and denoted $\bot = (O, f(O))$ (resp. $\top = (g(I), I)$.) The *Hasse diagram* of a graph[3] is the suppression on the graph of both transitivity and reflexivity edges.

Figure 1 shows an example of concept lattice (represented by its Hasse diagram) built from the formal context in Table 1. For more information, the reader can refer to Ref. 33.

2.2. *Generation algorithms*

Numerous generation algorithms for concept lattices have been proposed in the literature.[5,7,12,15,23,24,29,32] Although all these algorithms generate the same lattice, they propose different strategies. Some of these algorithms are incremental.[7,15,23] Ganter's NextClosure[12] is the reference algorithm that determines the concepts in lectical order (next, the concepts may be ordered by $\leq$ to form the concept lattice) while Bordat's algorithm[5] is the first algorithm that computes directly the Hasse diagram of the lattice. Recent work[14] proposed a generic algorithm unifying the existing algorithms in a unique framework, which facilitates the comparison of these algorithms. A formal and experimental comparative study of the different algorithms has been published.[18]

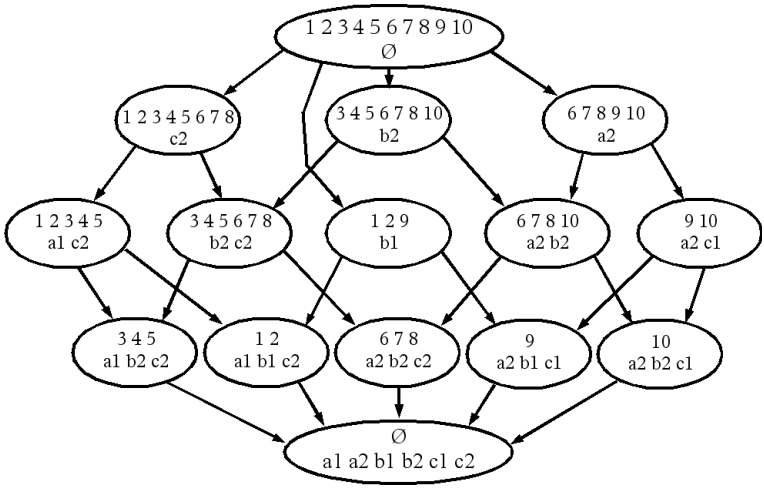


Fig. 1. Example of concept lattice (Hasse diagram).

All of these proposed algorithms have a polynomial complexity with respect to the number of concepts (at best quadratic in Ref. 24). The complexity is therefore determined by the size of the lattice, this size being bounded by $2^{|O+I|}$ in the worst case and by $|O + I|$ in the best case. Studies on average complexity are difficult to perform because the size of the concept lattice depends both on the dimensionality of the data to classify and on their organization and diversity. However, in practice, the size of the Galois lattice generally remains reasonable, as shown in various experiments.[20,22]

In Ref. 1, we introduced an extension of Bordat's algorithm, which has the advantage of enabling on-demand concept generation. With this algorithm, only the small portion of the lattice that is necessary for our particular classification task is constructed during the recognition stage. This leads to a drastic decrease in the complexity of the generation algorithm, as shown in Sec. 4.2.2, and is useful in many contexts where incrementality is needed, or where the learning set is different from the gallery. In the latter case, we can imagine a system where discretization is performed offline using a generic learning set, and then the set of symbols to recognize (gallery) is given online to the system during the recognition stage. For instance, we can imagine a generic symbol recognizer which has to be specialized online to a given set of symbols (architectural symbols, road-signs…), the specialization including the generation of the Galois lattice using the objects in the gallery and the attributes obtained after discretization (performed using the learning set), and the recognition of the query symbols itself.

2.3. *Application to data mining*

As most classification methods, both the selection-based and navigation-based approaches rely on the three following stages: data preparation, learning and classification itself, which are detailed in the following parts of this section.

2.3.1. *Data preparation*

The first step is feature extraction. In the context of graphic objects recognition, many different primitives may be extracted. The Galois lattice is defined only for primitives that can be organized using *formal contexts*, i.e. for discrete data.

Continuous-valued primitives must therefore be partitioned into a finite set of disjoint intervals (called attributes) which are referred to by using codes. This procedure is commonly called *discretization*. Discretization methods may be classified according to three criteria[11]:

- **supervised/unsupervised:** while unsupervised discretization techniques only use similarity between objects, supervised discretization methods also take into account the classes of the objects;
- **global/local:** the data space may be partitioned into intervals *before* the construction of the classifier (global discretization), or *as the construction of the classifier goes along* (local discretization);

- **mono dimensional/multidimensional:** while mono dimensional discretization processes each primitive independently from the others, multidimensional discretization simultaneously uses all the primitives to partition the data space into intervals. The main advantage of the latter technique is that it is capable of taking into account the interactions between primitives. However, mono dimensional discretization methods are the most widely used.

An experimental comparison of the effectiveness of various discretization techniques for classification is provided in Ref. 11. The experimental results show that supervised discretization techniques slightly outperform unsupervised discretization methods for a classification task.

Each of the global and local discretization methods has its own advantages and drawbacks. While local discretization has the advantage of taking into account the interactions between primitives, global techniques are more efficient because they process a feature space with lower dimensionality. The experimental comparisons provided in Refs. 11 and 26 do not settle the question of which strategy is the best for all circumstances; the choice of the technique is strongly related to the objective.

Depending on the application, the primitives may be continuous-valued and/or discrete so the discretization stage is not described for every method in the literature. In particular, the choice of the discretization strategy is not specified for the selection-based approaches described in Sec. 1. The discretization method we propose for the Navigala approach is detailed in Sec. 3.1.

2.3.2. *Learning*

During the learning stage, the Galois lattice will be constructed as a classifier from the set of discrete (or discretized) training data. For Galois lattice-based classification, the learning stage is *supervised* and therefore the training data consists of training objects primitives labeled by their associated classes (desired outputs). Preliminary to the training stage, we consider that the training data has been prepared (i.e. continuous-valued primitives have been discretized). The different steps that are carried out during the learning stage depend on the type of classification method.

For selection-based classification strategies, the learning stage includes three steps:

- The *lattice generation* step and, possibly, a pruning step. Different generation algorithms are described in Sec. 2.2;
- The *selection* step. The objective is to reduce the learning space using different relevance criteria, such as the occurrence frequencies of the different attributes. The selection step may lead to filtering out concepts,[7,19,25,27,28] objects[21] and/or contextual rules[34];
- Possibly the classifier's learning stage (e.g. the extraction of classification rules for the GRAND[25] and RULEARNER[27] methods).

The learning stage of navigation-based classification methods, detailed in Sec. 3.2, only involves two steps:

- The *lattice generation* step;
- The *labeling* step: the nodes which are pure enough are labeled with their corresponding class.

We can note that the lattice generation step may lead to a high complexity (exponential complexity in the worst case). That drawback is counterbalanced by the fact that the learning stage is offline and can be carried out before classification itself in most applications. In some applications where the learning step cannot be performed offline (for examples see Sec. 2.2), on-demand generation may be used (see Sec. 3.3.2). Changes in the training set generally lead to a new learning stage, even though some incremental solutions exist (see Sec. 2.2).

2.3.3. *Classification*

Once the classifier has been built from the training data, one can classify new samples. The aim is to classify these new elements on the basis of their description (primitives values). Differing from the learning stage which is generally performed offline, the classification stage is generally performed online.

Selection-based methods rely on classical classifiers such as the k-nearest neighbors or Bayesian classifiers.

Conversely, in navigation-based approaches, classification is based on the use of the whole Galois lattice. This step is of very low complexity (for more information about the computational times please refer to Sec. 4.2.2). Each object to be recognized (denoted by $p \in A$) progresses through the lattice from $\perp$ to $\top$ (see Sec. 2.1), moving from a formal concept to one of its successors (connected by an edge), until it reaches a labeled concept. At each concept $C_i = (A_i, B_i)$, the choice of the following concept $C_{i+1} = (A_{i+1}, B_{i+1})$ is made among its direct successors according to the set of attributes $x \in I$ where pRx and $x \in C_{i+1} \setminus C_i$. We must note that at each step, the choice of the successor concept is unique thanks to the inclusion property (see Sec. 2.1).

2.4. *Comparison with a classification tree*

At this point, the reader could naturally question the links between the decision tree and the Galois lattice. Indeed the navigation step is quite similar to the one proposed with a decision tree. The main difference lies in the existence of multiple paths to reach a given concept in the lattice, contrary to the decision tree where there is a unique path to reach a given node. This property confers flexibility to the recognition process using a lattice and therefore noise robustness is increased. Experiments (see Sec. 4.2.1) have shown that the navigation-based approach Navigala provides better recognition rates than decision trees in a context of noisy symbols recognition.

Moreover, we have recently shown the existence of structural links (inclusion and fusion) between a particular type of concept lattices and decision trees. For more details please refer to Ref. 2.

3. Description of the Proposed Approach: Navigala

We have developed a recognition system named Navigala (NAVIgation into GAlois LAttice), where classification is navigation-based. This method is fitted to recognize noisy graphical objects and especially symbol images. Such symbols appear in technical documents such as architectural plans or electrical diagrams. The possible origins of the noise are paper deterioration (stains, blotting out), scanning artefacts or vectorial distortions in the context of handwritten symbols (for examples of noisy symbols see Fig. 5).

Graphic objects may be described by *various types of primitives*. As statistical features describe the spatial distributions of the pixel values of the symbol, structural primitives describe the spatial or topological relations between certain subpatterns extracted from the symbol images. In the following, the primitive vector of each symbol is called the signature of this symbol.

Navigala is a supervised classification approach, whereas the discretization stage can be performed by using either a supervised or unsupervised criterion. In this section, we will describe the three steps of Navigala: data preparation, learning and classification. We will also provide a comparison of Navigala with the existing classification methods based on the use of a Galois lattice and mentioned in Sec. 1.

3.1. *Data preparation*

Firstly, several signatures are extracted from the symbol images: statistical signatures (Fourier−Mellin invariants,[10] Radon transform-based Radon transform,[30] Zernike moments[31]), and a structural signature named *flexible structural signature*.[8]

As presented in Sec. 2.3.1, the continuous valued primitives must be discretized in a preprocessing stage. Let us consider that the dataset is represented by an array of data where each row corresponds to the feature vector of one symbol image and every column corresponds to the (continuous) values of a given primitive f_i in the feature vector. The objective of the discretization stage is to obtain a formal context (as illustrated in Table 1) where each column (attribute) corresponds to an interval that separates the images corresponding to different classes (symbols). For example, in Table 1, the images of the first two symbols (classes 1 and 2) differ following the values of their second feature f_2: while for the images of the first symbol $f_2 \in [0, 4]$, the images of the second symbol verify $f_2 \in [12, 20]$.

Discretization is performed as follows. Initially, we consider that every column f_i in the data array is described by one interval V_i, the lower and upper bounds of which are respectively the minimum and maximum values in the corresponding column. At each iterative step of the discretization process, a criterion selects both the primitive

to split into intervals and the optimal cutting point. At iteration t, let $x \in I$ be a primitive interval, where $V_x = (v_1, \ldots, v_n)$ are the values of x observed in the training set and sorted by ascending order. The interval will be cut between the values v_j and v_{j+1}, where v_j maximizes a given "cutting" criterion $C(v_j)$. Numerous cutting criteria can be proposed; these criteria may be supervised or unsupervised, global or local and multidimensional or mono dimensional (see Sec. 2.3.1). The supervised, global and mono dimensional criteria are among the most widely used. We experimented three global and mono dimensional criteria (see Eqs. (3)−(5)): maximal distance, entropy and Hotelling's coefficient.[17] While maximal distance is an unsupervised criterion that aims at maximizing the gap between two consecutive values, entropy and Hotelling's coefficient are two supervised criteria which respectively minimize the degree of mixture of the classes and jointly maximize the scatter between classes while minimizing the within-class scatter.

- maximal distance:

$$C_{\mathrm{MD}}(v_j) = v_{j+1} - v_j \tag{3}$$

- entropy:

$$C_E(v_j) = E(V_x) - \left(\frac{j}{n} E(v_1, \ldots, v_j) + \frac{n-j}{n} E(v_{j+1}, \ldots, v_n)\right) \tag{4}$$

  with $E(V) = -\sum_{k=1}^{|c(V)|} \frac{n_k}{n} \log_2(\frac{n_k}{n})$, where n and $c(V)$ are respectively the number of images and the set of classes (symbols) corresponding to the values belonging to interval V (in the training set). n_k is the number of images, among the n images with values in V, which belong to class k.
- Hotelling's coefficient:

$$C_H(v_j) = H(V_x) - \left(\frac{j}{n} H(v_1, \ldots, v_j) + \frac{n-j}{n} H(v_{j+1}, \ldots, v_n)\right) \tag{5}$$

  where $H(V) = \frac{\sigma_B(V)}{\sigma_W(V)}$.

  With $\sigma_B(V) = \frac{1}{n} \sum_{k=1}^{|c(V)|} n_k (g_k - g)^2$ is the between-class variance and $\sigma_W(V) = \frac{1}{n} \sum_{k=1}^{|c(V)|} n_k (\sum_{i=1}^{n_k} (v_{k_i} - g_k)^2)$ is the within-class variance, where $g = \frac{1}{n} \sum_{j=1}^{n} v_j$ is the mean of the values belonging to V, v_{k_i} is the ith value in V corresponding to class k and $g_k = \frac{1}{n_k} \sum_{i=1}^{n_k} v_{k_i}$ is the mean of the values of images from class k and belonging to V.

The discretization process is iterated until a *stopping criterion* is met. In Navigala, the stopping criterion is class separation, which is met when each set of images sharing the same attributes is classified into one given class. In some cases where class separation cannot be achieved, we stop the discretization process when Hotelling's cutting criterion is less than a certain predefined threshold.

In Navigala, the obtained intervals are then extended as fuzzy intervals, to be more robust towards noise. During the classification stage, each query symbol image

will be considered as corresponding to its set of nearest fuzzy intervals in the feature space.

The distance $d(f_i, V)$ between the value f_i of the ith element in the query signature and an interval V obtained from the discretization of f_i can be expressed as:

$$d(S, V) = d(f_i, V) = 1 - \mu(f_i, V)$$

where $\mu(f_i, V)$ is the *membership functional* that specifies the level of membership of $f_i \in V$.

We propose the extension of an initial interval $[b, c]$ to a fuzzy number (described by a trapeze $[a, b, c, d]$ with $[a, d]$ as support and $[b, c]$ as kernel, see Fig. 2) by taking into account both the closest intervals and the objects distribution in the interval:

$$\begin{aligned} a &= b - \theta \times \min(d_{V^-}, d(g, c)) \\ d &= c + \theta \times \min(d_{V^+}, d(b, g)) \end{aligned}$$

where d_{V^-} (resp. d_{V^+}) is the distance with the closest previous interval (resp. closest next interval); g is the gravity center of the values in the initial interval $[b, c]$ and θ is a fuzzy parameter. Distances d_{V^-} and d_{V^+} are necessarily positive since intervals are disjoints. Each interval has at least one neighbor interval since undiscretized primitives are not selected. In the special case where the current interval has only one neighbor, we replicate that distance so as to obtain a symmetrical fuzzy interval.

3.2. *Learning*

The discretized data (issued from the data preparation stage) will then be used as a training set for the Galois lattice construction. Our algorithm[1] is an extension of Bordat's algorithm[5] which computes directly the Hasse diagram of the lattice (see Sec. 2.2). Indeed, during the classification, we use the successor relation to navigate through the graph, so we have to compute the successors of a given concept starting with the bottom concept.

Once the Hasse diagram of the discretized data is computed, we can label its concepts by using the classes in the training set. The terminal concepts (direct successors of the minimal boundary, located at the bottom of the Galois lattice) are labeled by using the formal context used for its generation (for an example of a formal context see Table 1). To each terminal concept, we associate the class that is most frequently associated to its objects (symbol images) in the training set. The labels associated to the terminal concepts will be used during the classification stage.

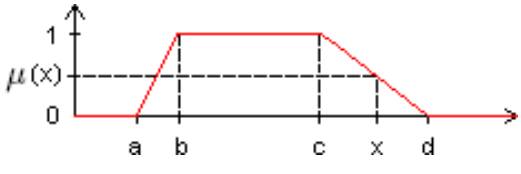


Fig. 2. Example of a fuzzy number.

3.3. *Classification*

3.3.1. *Navigation*

New symbols can be classified by using the Hasse diagram of the Galois lattice. Classification is performed by using the feature vector of the query symbol and navigating through the graph, from the minimum concept $\perp$ until a terminal (labeled) concept is reached. At each step of this navigation stage, the nearest fuzzy interval is selected (according to a fuzzy distance and a choice criterion). Intuitively, during the progression of a query image, the description of the query object is refined, until it is considered similar enough to a given set of objects belonging to the same terminal concept. When the query symbol reaches a terminal concept, it is labeled with the corresponding class.

More formally, at each step, given the current concept (A, B), one of its direct successors $(A_1, B_1), \ldots, (A_m, B_m)$ in the lattice is selected by validating one (or more) fuzzy intervals. Each set of attributes B_i corresponds to a set of intervals containing the set of intervals B: $B \subset B_i \forall i = \{1, \ldots, m\}$, because the concept (A_i, B_i) is a successor of the concept (A, B) in the lattice. At the current concept (A, B), all the intervals in the interval set B have been validated. Let us isolate, for each successor concept (A_i, B_i), the set of intervals $\tilde{B}_i$ that are candidates for validation (the set of intervals that have not been previously validated in (A, B)):

$$\tilde{B}_i = B_i - B$$

Let us further denote by $\tilde{B}$ the family of sets of intervals which are candidates for validation:

$$\tilde{B} = \bigcup_{i=1}^{m} \tilde{B}_i$$

The navigation elementary step therefore consists in selecting a set of intervals $\tilde{B}_i$ among the family of candidate interval sets $\tilde{B}$. This selection is performed according to a *choice criterion* defined using a fuzzy distance $d(S,\tilde{B}_i)$ between the signature S of the query object and the candidate sets of intervals $\tilde{B}_i$, for every candidate successor concept (A_i, B_i).

We have to define a choice criterion to select, among the candidate sets of intervals $\tilde{B} = \cup_{i=1}^{m} \tilde{B}_i$ (corresponding to the successors of the current concept), the set of intervals $\tilde{B}_i$ that best correspond to the signature S of the query object. The choice criterion relies on the use of the fuzzy distances $d(S, \tilde{B}_i)_{i=\{1,\ldots,m\}}$ between the signature S and the candidate sets of intervals $\tilde{B}_i$. Several choice criteria are possible, hereafter is a (nonexhaustive) list of these criteria:

(1) Choosing i where the sum of the distances between the signature S and the intervals V constituting the set of intervals $\tilde{B}_i$ is minimum. More formally,

$$i = \text{Argmin}_{i=1,\ldots,m}(\Sigma_{V \in \tilde{B}_i} d(S, V))$$

(2) Choosing i where the set of intervals $\tilde{B}_i$ contains the maximum number of intervals among the k nearest intervals from the signature S (according to the fuzzy distance measure d). More formally,

$$i = \text{Argmax}_{i=1,\ldots,m} \left| \tilde{B}_i \bigcap \{V^{(1)}, \ldots, V^{(k)}\} \right|$$

where the $V^{(j)}$ are the intervals in $\tilde{B}$, sorted by descending order following the distance $d(S, V)$ and k is a parameter of the choice criterion.

(3) Choosing i where the set of intervals $\tilde{B}_i$ contains the maximal number of intervals located at a distance inferior to the threshold c for the given query signature S. More formally,

$$i = \text{Argmax}_{i=1,\ldots,m} |\{V \in \tilde{B}_i \text{ such that } d(S, V) \leq c\}|$$

We can note that the first criterion, defined globally on all the intervals contained in $\tilde{B}_i$, has the drawback of "swallowing up" the noise. The second criterion relies on the principle of the k-nearest neighbor rule. We can also note that the third criterion is a particular case of the second criterion. All of these proposed criteria being local for each $i = 1, \ldots, m$, one can define more sophisticated criteria in order to benefit from the advantages of the different alternatives. In our case, we chose to use a combination of these criteria, which consists in:

- Applying criterion (3) with $c = 0$, which is equivalent to defining, for every interval V in $\tilde{B}_i$, a rectangular fuzzy number whose support is defined by the boundaries of V.
- Then, in case of an ambiguity, we apply criterion (3) with $0 < c < 1$. The support of the fuzzy number is extended to the fuzzy boundaries of the fuzzy interval V proportionally to its size.
- If the ambiguity remains, we apply criterion (1), which is equivalent to a symmetrical fuzzy number whose zero (center, gravity center or median) is the center of the interval.

3.3.2. *On-demand concepts generation*

The Galois lattice construction algorithm used for Navigala[1] presents several advantages: it is quite easy to implement, and it enables an *on-demand concepts generation* of the Galois lattice: concepts are generated only when they are proposed for selection during the recognition process. This is interesting, especially in some applicative contexts where the graph cannot be constructed offline (examples of such applications are given in Sec. 2.2). Indeed, it avoids the construction of the whole graph, which can be of an exponential complexity in the worst case. Indeed, recognition is performed by exploring only a small region of the lattice. As shown in Sec. 4.2.2, it leads to a slight increase in the complexity of the classification step but it considerably reduces the complexity of the learning stage.

3.3.3. *Iterative classification*

In the field of symbols classification, we also developed an *iterative recognition* system (see Ref. 16), which takes advantage of the complementarity of statistical and structural approaches. Indeed, this method can integrate several descriptions of various types for a more effective classification.

During navigation in the Galois lattice, in the case of uncertainty regarding the symbol to be recognized, it is possible to stop the progression and thus avoid certain classification errors. For example, let C_1 and C_2 be two successor concepts of the current concept C, where C_1 and C_2 contain objects of different classes whose descriptions are very similar to the query object. To avoid any doubt, the descriptions of the objects in C_1 and C_2 can be replaced by new descriptions issued from another type of feature extractor. In the iterative process, these new descriptions are then used to build a new Galois lattice especially designed to discriminate the objects from concepts C_1 and C_2 according to their classes.

3.4. *Comparison with other Galois lattice-based methods*

This section is dedicated to a comparison between selection-based methods and Navigala: a synthesis of the similarities and the differences between the various approaches is provided. For an experimental comparison see Sec. 4.3.

Figure 3 provides a comparison of the different classification methods based on a Galois lattice. Selection-based methods can be gathered depending on the elements used: concepts only, concepts and rules, concepts and prototypes or rules only. The Navigala approach is characterized by the use of the whole Galois lattice with an object classification by navigation.

Moreover, Table 2 proposes a comparison of the computational complexities of the construction stages of these different methods, and a synthesis of the experimental results obtained by the authors of those methods. In this table we have added the characteristics of Navigala. We can see that Navigala's complexity is very low compared to other lattice-based methods, especially when our applicative context enables the use of on-demand generation. We can see that its experimental results are encouraging.

In the following, we discuss the behavior of these eight methods when classes are weakly represented, in the presence of noise, and when the number of classes is large.

3.4.1. *Weakly represented classes*

In most selection-based approaches, the learning stage is limited to the most represented objects (in the learning set). That is why, with LEGAL-E for example, some objects may not be recognized even though they are very similar to a learning sample (if this learning sample is not representative enough of the learning set to be learnt). The opposite of this is Navigala, where the whole learning set of objects is learnt without favoring the most represented, which enables us to be exhaustive. However,

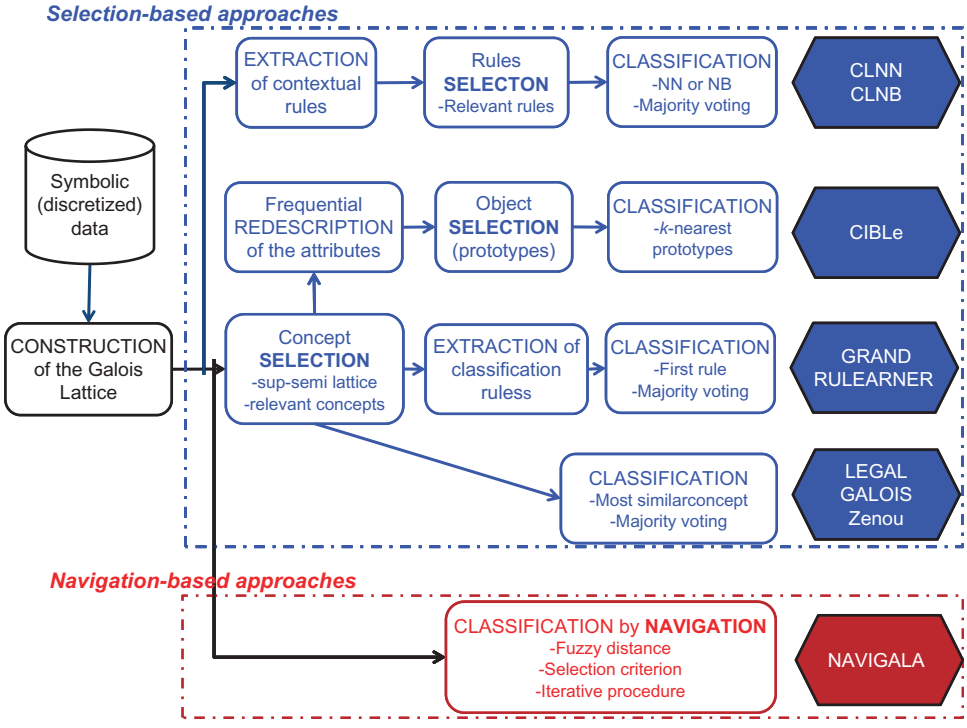


Fig. 3. Comparison of some Galois lattice-based classifiers. The acronym NN stands for "Nearest Neighbor", while NB stands for "Naive Bayes".

Table 2. Synthesis of the properties of the classification methods based on lattices.

Methods	Construction Complexity	Experimental Comparison
GRAND	$O(2^l l^4)$ with l the minimum between the number of examples and the number of attributes	Performances similar to Assistant, AQ15, AQR, Bayes and CN2
LEGAL	$O((\|L\|n(1-\alpha))$ with $\|L\|$ the number of concepts in the lattice	Performances similar to GLUE and superior to C4.5
GALOIS	$O(3^m 2^m n) < O(3^{2m} n)$ with m the number of attributes and n the number of examples	Performances similar to other methods in the literature
RULEARNER	idem GRAND	Performances similar to C4.5 and CN2, or even slightly better
CIBLe	$O(\min(n-1, m-1)^{h+1} m^3)$ (sup-semi lattice construction) $+ O(\min(n-1, m-1)^{h+1})$ (threshold search)	Better performances than IBi, K^* and Pebls
CLNN and CLNB	$O(\|L\|\|E\|^3 + \|L\|m + \|L'\|) + O(\|L'\|m).O(NN/NB)$ with $\|E\|$ the maximal number of successors of a concept $\|L'\| < \|L\|$ et $O(NN/NB)$ the complexities of NN or NB classifiers	Better performances than NBTree, CBA and C4.5 Rules

Table 2. (*Continued*)

Methods	Construction Complexity	Experimental Comparison
Zenou	$O(3^m 2^m n + \|L\|^2 m + \|L\| m)$	Encouraging performances
Navigala	($O(\|L\| n^3)$ optional lattice +) $O(nm^2)$ (classification)	Performances similar to Bayesian classifier and greater than CART (see Sec. 4.2.1)

while selection-based methods enable outlier detection (and further suppression), Navigala cannot detect them and they will be integrated in the Galois lattice. Nonetheless, Navigala is designed to be robust enough to accommodate these outliers, as detailed in the following section.

3.4.2. *Noise robustness*

The navigation enables avoiding the influence of a noise carried on several attributes. Indeed, the attributes are successively validated, as opposed to selection-based approaches where the validation is given by an average. Moreover, the validation order of the attributes is modifiable depending on their robustness to noise. The most represented attributes are proposed at the beginning of the navigation within the lattice and the frequency decreases during the progression within the graph. Finally, the fuzzy distance measure softens the interval boundaries and absorbs the disturbances due to noise. Noise robustness is a problem for a selection-based approach: while LEGAL-E resists quite well to noise using the validity quasi-coherence criteria, the thresholds' choice of validity and quasi-coherence can require considerable working time.[19]

3.4.3. *Large number of classes*

Some selection algorithms are not designed to manage a large number of classes. For instance, CIBLe has difficulties characterizing data containing a large number of classes especially with complex data.[21] With navigation, it is possible to perform classification at different levels, using different signature types (using iterative classification, see Sec. 3.3.3) and therefore to discriminate between a higher number of classes.

4. Experimental Results

In this section, we present various experimental results. Firstly, we study the effects of variations in the parameters required for Navigala (for a symbol recognition task). Second, we provide a comparative study of Galois lattice selection-based methods and Navigala.

4.1. *Setting the parameters of Navigala for symbol recognition*

The main objective of this first experimental study is to tune the parameters of the proposed approach for a symbol recognition task. For experimentations, we use two

different symbols image databases: GREC 2003[a] and GREC 2005 (Graphics RECognition).[b] These databases were developed for international symbol recognition contests organized by the IAPR TC 10 committee.[c]

Each database contains one *model* symbol image per class (where the model image does not contain any noise) and noisy versions of these model symbols. The noise mimics deteriorations generated when scanning or copying paper documents. GREC 2003 database contains 35,139 symbol images from 39 classes, with exactly 901 symbol images per class. Each class contains one model symbol and 900 noisy symbols (ten symbol images for each of the nine types of deterioration). GREC 2005 database contains 175 symbol images from 25 classes (one model symbol and a varying number of noisy symbols per class). The noisy symbols are distinguished among six types of deterioration. Figures 4 and 5 respectively show samples of model symbols and noisy symbols extracted from these two databases.

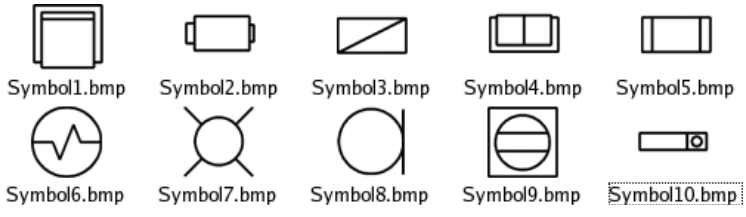


Fig. 4. Ten examples of model symbols (without noise) from GREC 2003.

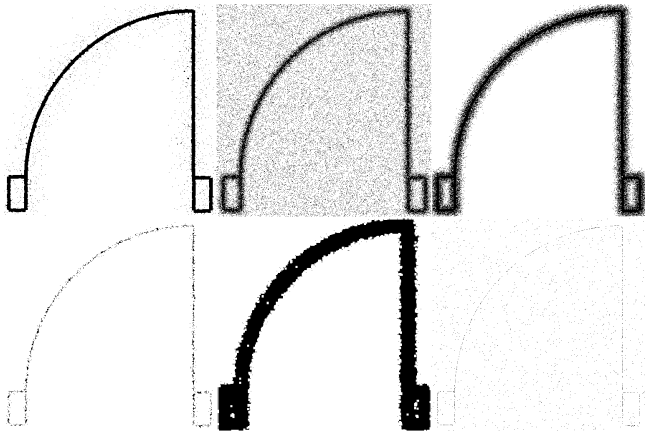


Fig. 5. Six examples of noisy symbols of the GREC 2005 database.

[a] www.cvc.uab.es/grec2003/SymRecContest/index.htm

[b] http://www.cs.cityu.edu.hk/ grec2005

[c] http://iapr-tc10.univ-lr.fr

4.1.1. *Cutting criterion*

In this experiment, we choose to test the adequacy of the cutting criteria (maximal distance, entropy or Hotelling's coefficient) to our recognition system (see Sec. 3.1).

To evaluate these criteria, we use a subset of the GREC 2003 database. Learning is performed by using only ten symbols per class and performance evaluation is made by using 90 noisy symbols per class. Figure 6 provides the recognition rates and Fig. 7 gives the size of the Galois lattice for each cutting criterion. The three statistical signatures presented in Sec. 3.1 are studied: Fourier−Mellin invariants, R-signature (Radon) and Zernike moments.

From these experimental results, we show that Hotelling's coefficient almost always provides the best results, no matter which signature is used. Moreover, we can see that the size of the lattice can explode using the maximal distance criterion as shown in Fig. 7. For the sake of effectiveness and efficiency, we therefore chose to use Hotelling's coefficient criterion.

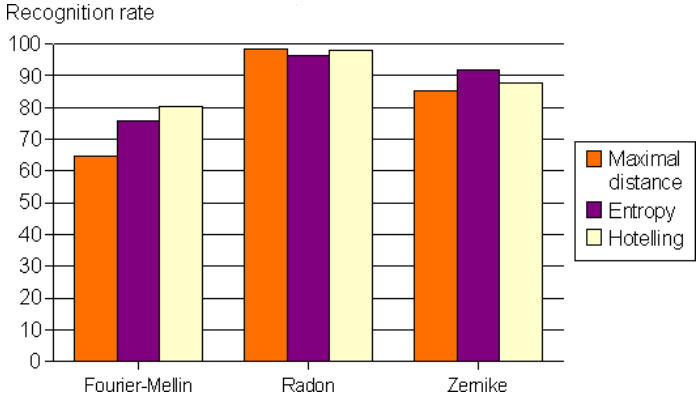


Fig. 6. Recognition rates depending on the cutting criterion.

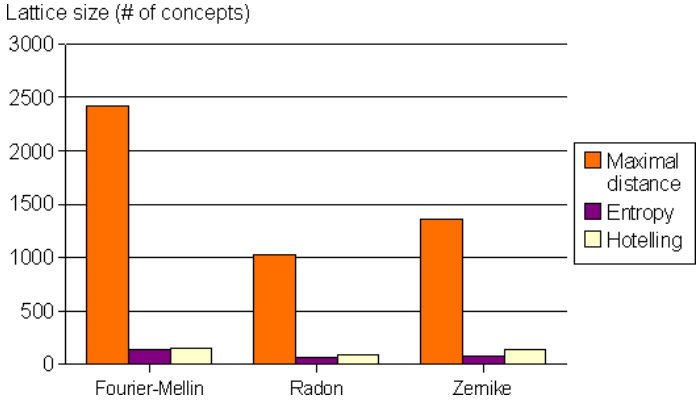


Fig. 7. Number of concepts in the Galois lattice depending on the cutting criterion.

4.1.2. *Signatures comparison*

In this subsection, we compare the effectiveness of certain statistical and structural signatures for our recognition system Navigala. We use a subset of the GREC 2003 database. The learning set is composed of eight classes (ten symbols per class) and performance characterization is performed by using a test set containing 90 noisy symbols per class. The recognition rates are presented in Fig. 8, and Galois lattice sizes in Fig. 9.

From these figures (and we can consider additionally Figs. 6 and 7) we can see that the R-signature (Radon) is the most interesting option both in terms of recognition rates and lattice size.

4.2. *Performance characterization for symbol recognition*

This section shows the results of different experiments.

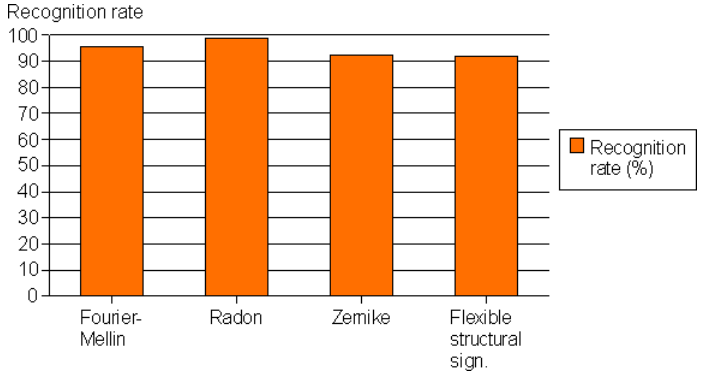


Fig. 8. Recognition rates depending on the signature.

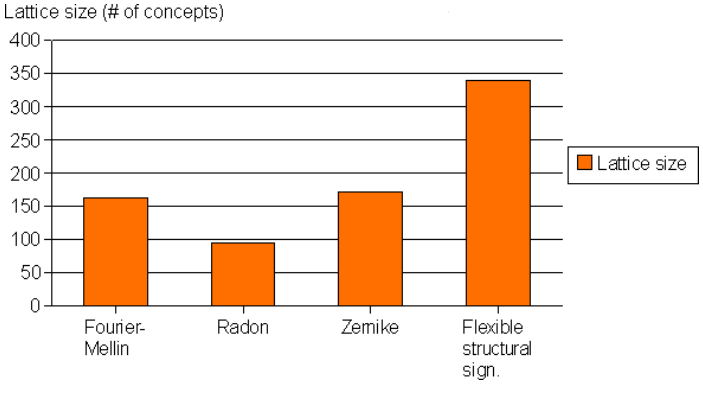


Fig. 9. Number of concepts in the Galois lattice depending on the signature (obtained by using the Hotelling cutting criterion).

Firstly, we compare the performances of Navigala and other standard classification approaches. Navigala is compared to a probabilistic classifier (Bayesian classifier), a statistical classifier (SVM) and a symbolic classifier (decision tree).

Secondly, we provide computational results when using on-demand generation algorithm.

4.2.1. *Comparison with other standard classification approaches*

In this subsection we show the results of two experiments where the performance of the proposed approach is compared to other standard classification approaches.

In the first experiment, we compare the recognition rates obtained using Navigala with those of the naive Bayesian classifier and a SVM classifier. We consider a dataset composed of symbol images from the two GREC databases: we use two sets of ten classes of symbols from the GREC 2003 database (named cl1-10 and cl11-20) and one set of 25 classes of symbols from the GREC 2005 database (named cl1-25), where the noise is stronger. The symbols are described by the statistical Radon signature (R-signature) composed of 50 values.

The classifiers are evaluated using cross-validation with varying sizes of the learning and test data: 5 blocks of 182 symbols from GREC 2003 (Test1), 10 blocks of 91 symbols from GREC 2003 (Test2), 26 blocks of 35 symbols from GREC 2003 (Test3), and 5 blocks of 35 symbols from GREC 2005 (Test4). The average recognition rates are given in Fig. 10.

While the naive Bayesian classifier is more effective than Navigala in the presence of only ten classes and of a limited noise (Test1 and Test2), Navigala outperforms the Bayesian classifier in more difficult situations where the number of images is increased (Test3) or where the noise is significant and the number of classes to be discriminated is increased (Test4). As a consequence, we can consider that our approach is more robust towards noise and towards an increase of the number of

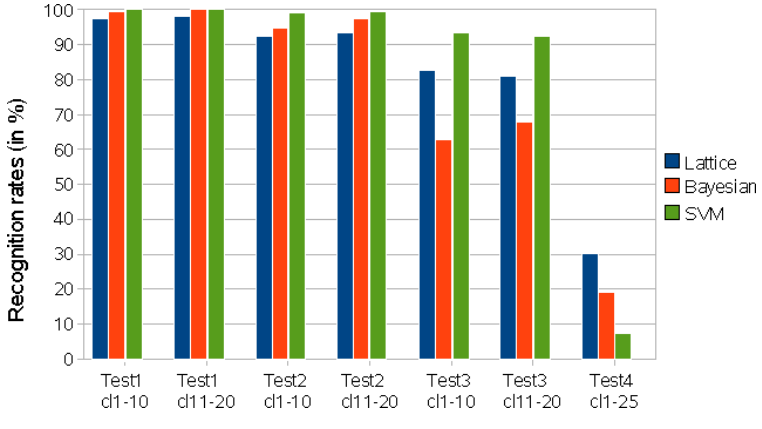


Fig. 10. Compared recognition rates of the four classifiers using cross-validation.

classes than the Bayesian classifier. In the experiment Test4, our approach even outperforms the SVM classifier.

Furthermore we can note that, Navigala performs feature selection prior to classification (during the discretization stage) and only uses 6 to 15 of the 50 elements of the feature vectors whereas the other classifiers use all of the 50 elements. For these reasons, we can consider that these results are encouraging.

In the second experiment, we compare the performances of Navigala and decision trees. The experimental protocol is the following. We consider ten classes from GREC 2003 dataset (10 model symbols for learning and 900 noisy symbols for test). This data is prepared as presented in Sec. 3.1: the Radon signature is extracted from the images, and then the signatures are discretized by using the proposed discretization approach (based on the Hotelling coefficient). From the discretized training signatures we generate both the concept lattice (the Navigala classifier is then built) and the decision tree using CART algorithm.[6]

The recognition rates we obtain are 57% for the decision tree versus 72% for the lattice. One of the main differences between a concept lattice and a decision tree is that in decision trees the path from the root to a given leaf is unique whereas in Galois lattices there are multiple paths from the maximal boundary to a given terminal concept (see Fig. 1). The improvement of the recognition rates when using the lattice shows that the existence of multiple paths gives the lattice better noise robustness.

In return, the size of the lattice is generally greater than the size of the decision tree. In our experiment, the number of discretization steps is 9 and the number of discretized intervals is 17. The number of concepts in the lattice is 70 against (only) 18 nodes for the decision tree. Thus, we can see that, when using the same discretized data, the size of the lattice is greater than the size of the decision tree. But this drawback may be counterbalanced by the possibility of generating the lattice on-demand.

4.2.2. *On-demand generation*

As presented in Sec. 3.3.2, the Galois lattice generation algorithm used for Navigala [1] enables an *on-demand concepts generation* of the Galois lattice. Recognition is performed by exploring only a small region of the lattice. The experimental results presented in Table 3 show the processing times for the learning and classification steps when using a 1.83 GHz processor with 512 MB RAM. It also gives the number of generated concepts. The learning set is composed of 25 model symbols (one per

Table 3. Processing times (using a 1.83 GHz processor with 512 MB RAM) and number of concepts generated.

	Learning	Classification	Number of Concepts
Whole lattice	430.2 sec	2 sec	3185
On-demand generation	0.5 sec	9.8 sec	282

each of the 25 classes) from GREC 2003. The test set is composed of 250 noisy symbols: ten symbols per class.

From this table we can see that the number of concepts generated on-demand (282) is significantly reduced compared to the construction of the whole lattice (3185), while the recognition (navigation) path is identical. Therefore, we can note that on-demand generation gives the same performances as offline generation of the lattice and reduces the size of the structure to be generated. Table 3 also shows that when using on-demand generation the computational cost of the generation step is partially moved from the learning stage to the classification stage (compared to offline generation). Nevertheless, the computational time is globally reduced while remaining reasonable.

4.3. *Comparison with other Galois lattice-based methods*

This section is dedicated to experimental comparisons between selection-based methods and Navigala on certain databases from the UCI Repository[4]: Breast-cancer (BC), Iris (IR), Soybean-small (SS) and Zoo (ZO). Table 4 provides a description of these databases: number of records, number of continuous attributes, number of discrete attributes and number of classes to distinguish.

We consider experimental results available in the papers describing the methods: RULEARNER,[27] CIBLe,[21] CLNB — CLNN and C4.5Rules.[34] This is why our experimentation results are not exhaustive. Table 5 gives the classification error rates obtained using cross-validation.

Table 4. UCI Repository databases.

		Number of Attributes		
Database	Number of Objects	Continuous	Discrete	Number of Classes
BC	699	9	0	2
IR	150	4	0	3
SS	47	0	35	4
ZO	101	1	15	7

Table 5. Results obtained using cross-validation on certain databases of the UCI Repository.

		Classification Error Rates				
DB	Cross-Validation	Navigala	Cible[21]	CLNB[34]	CLNN[34]	C4.5R[34]
BC	10 fold	5.4%		3.1%	3.4%	5.0%
	5 fold	5.5%	4.6%			
IR	10 fold	7.4%		5.3%	5.3%	4.7%
	5 fold	4.1%				
SS	10 fold	2.5%				
	5 fold	2.3%	8%			
ZO	10 fold	4.0%		3.9%	3.9%	7.8%
	5 fold	4.9%	6.1%			

In general, Navigala provides classification error rates relatively close to those obtained by other classifiers. It has to be noted that Navigala catches up with and even outperforms the other methods when the number of classes is increased, as in the Soybean-Small and Zoo databases. Therefore, we can note that Navigala is somewhat generic, as it has been designed for a very specific task of symbol recognition using statistical (continuous) signatures, and can be successfully applied to other types of data.

5. Conclusion and Discussion

The two main contributions of this paper are: firstly, the introduction of a classification method named Navigala dedicated to noisy symbol recognition and its experimental assessment and secondly, a comparative study (both formal and experimental) of eight classification methods based on Galois lattices (including Navigala).

Contrary to most of the previously proposed approaches, which use the Galois lattice as a selection tool, Navigala classifies the symbols by navigating through the lattice. While most selection-based approaches are well-suited for data mining applications with little noise and a limited number of classes, Navigala is dedicated to a task of noisy symbol image recognition, where the number of classes may be huge. By using the whole lattice as a classifier, Navigala has the advantage of being exhaustive and proposing multiple paths to reach a given class-labeled concept, which makes Navigala more robust towards noise. It has to be noted that the inherent complexity is limited thanks to our on-demand generation algorithm.

We are now working on the structural links between Galois lattices and classification trees in order to propose a new classification method based on a Galois lattice with local discretization, similar to the discretization stage of decision trees.

Aknowledgments

Grateful acknowledgement is made for financial support by the Poitou-Charentes Region (France). We would also like to aknowledge Dr. Stéphanie Guillas for her work on this subject in the context of her PhDs.

References

1. K. Bertet, S. Guillas and J. M. Ogier, Extensions of Bordat's algorithm for attributes, in *Fifth Int. Conf. Concept Lattices and Their Applications (CLA'2007)* (Montpellier, France, October 24−26, 2007), pp. 38−49.
2. K. Bertet, M. Visani, J. M. Ogier and N. Girard, Some links between decision trees and dichotomic lattices, *Sixth Int. Conf. Concept Lattices and Their Applications (CLA'2008)* (Olomouc, Czech Republic, October 2008), pp. 193−205.
3. G. Birkhoff, *Lattice Theory*, Vol. 25, 3rd edn. (American Mathematical Society, 1967).
4. C. Blake, E. Keogh and C. Merz, UCI repository of machine learning databases, 1998.
5. J. P. Bordat, Calcul pratique du treillis de Galois d'une correspondance, *Math. Sci. Hum.* **96** (1986) 31−47.

6. L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and Regression Trees* (Wadsworth Inc., Belmont, California, 1984).
7. C. Carpineto and G. Romano, Galois: An order-theoretic approach to conceptual clustering, *Proc. Int. Conf. Machine Learning (ICML'93)* (Amherst, July, 1993), pp. 33−40.
8. M. Coustaty, S. Guillas, M. Visani, K. Bertet and J. M. Ogier, Flexible structural signature for symbol recognition using a concept lattice classifier, *Seventh IAPR Int. Workshop on Graphics Recognition (GREC'07)* (Curitiba, Brazil, September, 2007), pp. 20−21.
9. B. A. Davey and H. A. Priestley, *Introduction to Lattices and Orders*, 2nd edn., (Cambridge University Press, 1991).
10. S. Derrode, M. Daoudi and F. Ghorbel, Invariant content-based image retrieval using a complete set of Fourier−Mellin descriptors, *Int. Conf. Multimedia Computing and Systems (ICMCS'99)* (1999), pp. 877−881.
11. J. Dougherty, R. Kohavi and M. Sahami, *Supervised and Unsupervised Discretization of Continuous Features* (Morgan Kaufman, 1995).
12. B. Ganter, Two basic algorithms in concept analysis, *Technische Hochschule Darmstadt* (Preprint 831) (1984).
13. B. Ganter and R. Wille, *Formal Concept Analysis, Mathematical Foundations* (Springer-Verlag, Berlin, 1999).
14. A. Gely, A generic algorithm for generating closed sets of binary relation, *Third Int. Conf. Formal Concept Analysis (ICFCA 2005)* (2005), pp. 223−234.
15. R. Godin, R. Missaoui and H. Alaoui, Learning algorithms using a Galois lattice structure, *Third Int. Conf. Tools for Artificial Intelligence* (San Jose, California, 1991) pp. 22−29.
16. S. Guillas, K. Bertet and J. M. Ogier, Concept lattice classifier: A first step towards an iterative process of recognition of noised graphic objects, *Fourth Int. Conf. Concept Lattices and their Applications (CLA'2006)* (2006), pp. 257−263.
17. H. Hotelling, Relations between two sets of variates, *Biometrika* **XX-VIII** (1936) 321−377.
18. S. Kuznetsov and S. Obiedkov, Comparing performance of algorithms for generating concept lattices, *J. Exper. Theor. Artif. Intell.* **14**(2−3) (2002) 189−216.
19. E. Mephu Nguifo, Galois lattice: A framework for concept learning. design, evaluation and refinement, *Proc. IEEE Int. Conf. Tools with Artificial Intelligence (IEEE-ICTAI-94)* (New-Orleans, November 1994), pp. 461−467.
20. E. Mephu Nguifo, V. Duquenne and M. Liquiere, Concept lattice-based knowledge discovery in databases: Introduction, *J. Exper. Theor. Artif. Intell.* **14**(2/3) (2002) 75−79.
21. E. Mephu Nguifo and P. Njiwoua, Using lattice-based framework as a tool for feature extraction, *Proc. European Conf. Machine Learning (ECML-98)* (Chemnitz, Allemagne, April 1998), LNCS 1398 (Springer Verlag), pp. 304−309.
22. E. Mephu Nguifo and P. Njiwoua, Iglue: A lattice-based constructive induction system, *Int. J. Intell. Data Anal. (IDA)* **4**(4) (2000) 1−49.
23. E. Norris, An algorithm for computing the maximal rectangles in a binary relation, *Revue Roumaine de Mathématiques Pures et Appliquées* **23**(2) (1978).
24. L. Nourine and O. Raynaud, A fast algorithm for building lattices, *Third Int. Conf. Orders, Algorithms and Applications* (Montpellier, France, August 1999).
25. G. Oosthuizen, *The Use of a Lattice in Knowledge Processing*, PhD thesis, University of Strathclyde, Glasgow, 1988.
26. J. R. Quinlan, *Bagging, Boosting and C4.5* (AAAI Press, Menlo Park, CA, 1996).

27. M. Sahami, Learning classification rules using lattices, eds. N. Lavrac and S. Wrobel, *Proc. European Conf. Machine Learning (ECML'95)* (Heraclion, Crete, Greece, April 1995), pp. 343−346.
28. M. Samuelides and E. Zenou, Learning-based visual localization using formal concept lattices, *2004 IEEE Workshop on Machine Learning for Signal Processing* (Sao Luis (Brasil), September 29−October 1st 2004), pp. 43−52.
29. G. Stumme, R. Taouil, Y. Bastide, N. Pasquier and L. Lakhal, Computing iceberg concept lattices with TITANIC, *Data Know. Engin.* **42**(2) (2002) 189−222.
30. S. Tabbone, L. Wendling and J. P. Salmon, A new shape descriptor defined on the Radon transform, *Comput. Vis. Imag. Underst.* **102**(1) (2006) 42−51.
31. M. Teague, Image analysis via the general theory of moments, *J. Opt. Soc. America (JOSA)* **70** (2003) 920−930.
32. P. Valtchev, R. Missaoui and P. Lebrun, A partition-based approach towards constructing Galois (concept) lattices, *Discr. Math.* **3**(256) (2002) 801−829.
33. R. Wille, Restructuring lattice theory: An approach based on hierarchies of concepts, *Ordered sets*, ed. I. Rival (Dordrecht-Boston, Reidel, 1982), pp. 445−470.
34. Z. Xie, W. Hsu, Z. Liu and M. Lee, Concept lattice based composite classifiers for high predictability, *J. Exper. Theoret. Artif. Intell.* **14**(2/3) (Taylor and Francis Ltd, 2002), pp. 143−156.

Muriel Visani received in 2005 her PhD in computer science from the Institut National des Sciences Appliques of Lyon, France. During this period, she worked on face recognition from images for Orange Labs Company. In 2006, she joined as an Associate Professor in the Computer Science Laboratory (L3i) of the University of La Rochelle (France).

Her research activities focus on image analysis and recognition, especially the recognition of complex objects in images, enriching image description for better indexing and retrieval. For more details please refer to http://perso.univ-lr.fr/mvisani.

Karell Bertet received her PhD degree in computer science from the University of Paris 7, France, in 1998. During her PhD thesis (1995−1998), she worked on some algorithmical and structural aspects of lattices. Since 1999, she is Assistant Professor at the University of La Rochelle, working on the links between fundamental aspects of lattice theory and their applications, such as symbolic classification methods and knowledge based representation for document images. Her present research interests focus on and knowledge based representation.

Jean-Marc Ogier received his PhD degree in computer science from the University of Rouen, France, in 1994. During this period (1991−1994), he worked on graphic recognition for Matra Ms & I Company. From 1994 to 2000, he was an Associate Professor at the University of Rennes 1 during the first period (1994−1998) and at the University of Rouen from 1998 to 2001. Currently he is full professor at the University of La Rochelle.

Dr Ogier's works in the L3i laboratory in which he manages a research group (12 permanent staff, 20 PhD) dealing with document analysis. He manages several French and European projects dealing with historical document analysis, either with public institutions, or with private companies. He is a Deputy Director of the GDR I3 of the French National Research Center (CNRS). He is also Chair of the Technical Committee 10 (Graphic Recognition) of the International Association for Pattern Recognition (IAPR). Finally, he is also Vice Rector of the University of La Rochelle.

IJDAR (2011) 14:87–100
DOI 10.1007/s10032-010-0125-5

ORIGINAL PAPER

A protocol to characterize the descriptive power and the complementarity of shape descriptors

Muriel Visani · Oriol Ramos Terrades · Salvatore Tabbone

Received: 21 November 2009 / Revised: 2 June 2010 / Accepted: 5 August 2010 / Published online: 23 September 2010

Abstract Most document analysis applications rely on the extraction of shape descriptors, which may be grouped into different categories, each category having its own advantages and drawbacks (O.R. Terrades et al. in Proceedings of ICDAR'07, pp. 227–231, 2007). In order to improve the richness of their description, many authors choose to combine multiple descriptors. Yet, most of the authors who propose a new descriptor content themselves with comparing its performance to the performance of a set of single state-of-the-art descriptors in a specific applicative context (*e.g.* symbol recognition, symbol spotting…). This results in a proliferation of the shape descriptors proposed in the literature. In this article, we propose an innovative protocol, the originality of which is to be as independent of the final application as possible and which relies on new quantitative and qualitative measures. We introduce two types of measures: while the measures of the first type are intended to characterize the descriptive power (in terms of uniqueness, distinctiveness and robustness towards noise) of a descriptor, the second type of measures characterizes the complementarity between multiple descriptors. Characterizing upstream the complementarity of shape descriptors is an alternative to the usual approach where the descriptors to be combined are selected by trial and error, considering the performance characteristics of the overall system. To illustrate the contribution of this protocol, we performed experimental studies using a set of descriptors and a set of symbols which are widely used by the community namely ART and SC descriptors and the GREC 2003 database.

Keywords Document analysis · Shape descriptors · Symbol description · Performance characterization · Complementarity analysis

1 Introduction

Over the last decades, there has been a growing interest about performance evaluation in the domain of graphics recognition. Many contests have been organized, concerning raster-to-vector conversion [2–4], arc segmentation [5] and symbol recognition [6,7]. Most symbol recognition methods rely on a two-step procedure: (1) symbol description (representation) by extracting a feature vector with one (or more) descriptor(s) and (2) supervised classification of the symbols to recognize, based on their feature vectors. Several shape descriptors have been proposed in the literature [8–10] and most of them have been applied to the task of symbol recognition.

This paper is focused on the first step of symbol description, which is a crucial step that may be used for many other tasks in the field of document analysis (symbol spotting…). Ramos et al. have introduced in [1] a taxonomy of the different shape descriptors frequently used for symbol representation. The new categorization they propose is made according to the properties of the different shape descriptors, pointing out their strengths and weaknesses. One of the main objectives of their work is to facilitate, for a given application, the choice of the best descriptor in that context.

Work supported by the Juan de la Cierva programme and the MITTRAL project (TIN2009-14633-C03-01).

M. Visani
L3I, University of La Rochelle, La Rochelle Cedex 1, France
e-mail: muriel.visani@univ-lr.fr

O. R. Terrades
Instituto Tecnológico de Informática,
Universidad Politécnica de Valencia, Valencia, Spain
e-mail: oriolrt@iti.upv.es

S. Tabbone (✉)
LORIA, University of Nancy 2, Vandoeuvre-les-Nancy, France
e-mail: salvatore.tabbone@loria.fr

However, when considering a problem of symbol recognition, selecting the descriptor which is best suited for a given type of symbol and/or noise can be a hard, or even an impossible, task. Instead, one may be interested in combining different descriptors from different categories in order to benefit from the advantages of all the descriptors to be combined. The combination may be performed at the level of the descriptor (early fusion) or at the level of the classifiers (late fusion). Early fusion is usually implicitly done with powerful classifiers like neural networks [11], boosting classifiers [12,13] and Support Vector Machines [14]. In those methods, descriptors are extracted and concatenated as a single feature vector. Later on, during the training process, each classifier combines the features from the different descriptors. However, for general applications where the number of classes is high and the symbols to recognize can be counted by thousands, these expert classifiers reach their limits as their performance may decrease drastically. In this case, late fusion schemes where the combination is performed at the level of the classifier are generally preferred [15]. Late fusion methods have been applied to shape descriptors for symbol recognition [16,17]. Even so, in these papers, the performance characteristics of the descriptors in terms of descriptive power were not evaluated (only the performance for recognition was studied). Additionally, the complementarity of the descriptors to be combined was not investigated upstream, even though it may be very useful when choosing the set of descriptors to be combined and the combination scheme which is best suited to this particular set of descriptors.

To the best of our knowledge, very few works have been proposed in the literature concerning the evaluation of the performance characteristics of symbol descriptors, most evaluations being focused on the final application. A methodology for characterizing the performance of shape descriptors for symbol recognition has been proposed in [18]. This paper additionally provides a general discussion concerning the main difficulties and problems one may be faced with when setting the data, evaluation metrics and evaluation protocol, to characterize the performance of a symbol recognition method. Delalandre et al. [19] propose a solution to generate ground-truth based on a system that builds synthetic graphical documents. In [20], two main performance characterization metrics have been proposed, but we will see that these measures have several drawbacks that need to be completed (see Sect. 3). Jouili et al. [21] propose a performance evaluation for symbol recognition based on graph matching measures. This evaluation is essentially quantitative, based on precision and recall rates.

In this paper, we propose an experimental protocol and both qualitative and quantitative measures for characterizing the descriptive power and the complementarity of different shape descriptors for symbol description. This methodology is as independent of the final application (symbol spotting, recognition) as possible. Contrary to the above-mentioned performance evaluation methodologies, we do not consider any classifier; at most we consider some dissimilarity or distance measure and the nearest neighbour rule, to characterize for instance the uniqueness and distinctiveness of a given descriptor. We introduce an innovative protocol and two types of measures: while the measures of the first type are intended to characterize the descriptive power (in terms of uniqueness, distinctiveness and robustness towards noise) of a descriptor, the second type of measures characterize the complementarity between multiple descriptors. Concerning the measures of the first type, we first recall the definitions of confusion matrices, recognition rate, precision, recall and Cumulative Match Characteristics (CMC) curves. Even though some of these measures are already used by many researchers in our community, our contribution here is that we link them to the notions of distinctiveness and uniqueness. Second, we introduce two measures that are original in the field of document analysis. These two measures are respectively the tolerance intervals, which characterize the robustness of descriptors towards noise, and a qualitative measure based on an analogy with Doddington's zoo, widely known in the field of biometrics, characterizing the symmetries in the confusions. Concerning the measures of the second type, we introduce original measures to characterize upstream the complementarity between multiple descriptors. These measures constitute an alternative to the usual approach where the descriptors to be combined are selected by trial and error, considering the performance characteristics of the overall system. It may also be helpful when choosing the best combination scheme for a given set of descriptors.

To illustrate our methodology, we present a case study using two well-known statistical descriptors: the Angular Radial Transform (ART) descriptor [22], based on region pixel values and the Shape Context (SC) [23] descriptor, based on contours. We use noisy versions of the GREC 2003 database (*cf.* Fig. 4), which is well known and widely used by researchers working in the field of document analysis. It has to be noted that with adequate dissimilarity or distance measures (*e.g.* edit distance) our methodology can also be applied to structural descriptors. Moreover, the proposed framework may be further used for characterizing the complexity of any symbol database.

The paper is organized as follows. In Sect. 2, some innovative measures for evaluating the descriptive power of different shape descriptors, their robustness towards noise and their complementarity are proposed. In Sect. 3, we propose an experimental protocol and perform experimental results using ART and SC descriptors on the GREC 2003 database. These results are analysed to highlight the interest of using the proposed protocol and measures. While Sect. 4 provides a discussion about the measures we propose, Sect. 5 concludes this paper and presents the future work.

2 Evaluating the descriptive power of the descriptors and the complementarity between descriptors

In many applications such as symbol spotting, symbol recognition, the richness of a descriptor is related to its ability to group the different occurrences of one given symbol (uniqueness of the representation) and to discriminate them from other symbols (distinctiveness). In this direction, Valveny et al. have proposed in [20] two measures characterizing respectively the uniqueness and the distinctiveness of a shape descriptor: homogeneity and separability. While homogeneity is based on the distances between the descriptions of different occurrences of one symbol, separability is based on the distances between descriptions of different symbols. In this work, a good descriptor is characterized by high values of both homogeneity and separability. These measures are generic and may be used in many applicative contexts where a distance matrix between all the elements of the database can be computed. However, they have three main drawbacks. First, it is difficult to fix the thresholds which are necessary to characterize high values of these measures, since the distributions of the distances between feature vectors vary a lot from one database to another. Second, in many applications, we have a model image (which may be considered as the original symbol) and noisy versions of this model that we need to confront to all the models in the database. These two categories of images (models and noisy symbols) have to be considered separately, which is not the case with homogeneity and separability measures. Indeed, they rely on a distance matrix computed between all the symbols in the database, whatever their type. Third, in general the confusions between symbols are not symmetric (*e.g.* symbol 1 may be confused with symbol 2 and not the opposite). And neither homogeneity nor separability characterizes the symmetry of the confusions. Therefore, homogeneity and separability, which provide a coarse characterization of the richness of the different descriptors, have to be completed by other measures which overcome the drawbacks listed earlier.

To conceive such measures (which will be defined in Sects. 2.3 to 2.8), we first need to define more precisely the concepts of uniqueness and distinctiveness (see Sect. 2.1) and further to propose a protocol which is independent of the final application (see Sect. 2.2).

2.1 Definitions

Let us focus on the very conventional case where, for each symbol i in the database (with $i = 1, \ldots, c$), we have in the descriptor's representation space a symbol model S_i and n_i noisy versions of this model $\hat{S_i}^j$ (with $j = 1, \ldots, n_i$). In this case we can consider that a descriptor provides a perfect representation of a given symbol i when:

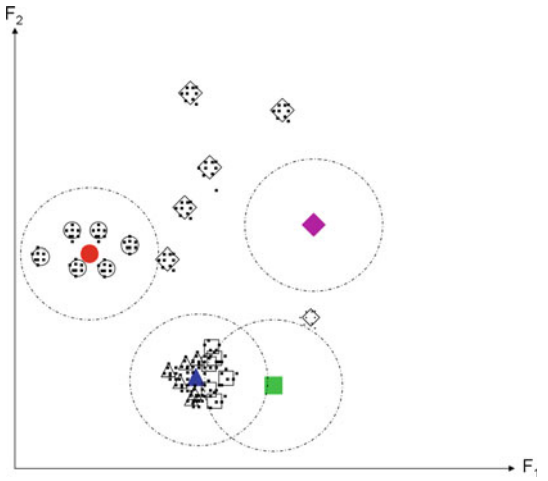


Fig. 1 Adaptation of Doddington's zoo (see Sect. 2.7) in our context. F_1 and F_2 are the two features of the descriptor's feature vector. In this example, we consider the Euclidean distance. The *filled shapes* are the models, each model having six noisy versions (*empty shapes* where noise is represented by *dots*). *Dashed circles* are situated at a distance θ from the models. The symbol *"circle"* is a sheep, the *"triangle"* is a lamb, the *"square"* is a wolf and the *"rhombus"* is a goat

– the representation of i is unique, *i.e.* feature vectors of noisy versions $\hat{S_i}^j$ of S_i are closer to S_i than to any other symbol model S_l (with $l \neq i$). Let us denote by d the distance (or dissimilarity measure) of interest. The representation of i may be considered as perfectly unique when:

$$\forall j = 1 \ldots n_i,\ \forall l = 1 \ldots c \ st\ l \neq i,\quad d(\hat{S_i}^j, S_l) > d(\hat{S_i}^j, S_i) \tag{1}$$

– the representation of i is distinctive, *i.e.* S_i is closer to its noisy versions $\hat{S_i}^j$ than to noisy versions $\hat{S_l}^m$ of other models S_l:

$$\forall l = 1 \ldots c \ st\ l \neq i,\ \forall m = 1 \ldots n_l,\quad d(\hat{S_l}^m, S_i) > \max_{j=1 \ldots n_i} d(\hat{S_i}^j, S_i) \tag{2}$$

Let us introduce the following notation: $NN(\hat{S}_i^j)$ being the nearest model of $\hat{S}_i^j$ in the descriptor's space (*i.e.* the result of the 1-nearest neighbour rule). The definitions of the uniqueness (see Eq. 1) and distinctiveness (see Eq. 2) of the symbol i may be respectively reformulated as follows:

$$\forall j = 1 \ldots n_i,\ NN(\hat{S}_i^j) = S_i \tag{3}$$

$$\forall l = 1 \ldots c \ st\ l \neq i,\ \forall m = 1 \ldots n_l,\ NN(\hat{S_l}^m) \neq S_i \tag{4}$$

For instance, the symbol "circle" in Fig. 1 is characterized by both perfect uniqueness and distinctiveness, while the symbol "triangle" is perfectly unique but not distinctive.

Equations 3 and 4 illustrate the notions of uniqueness and distinctiveness. However, they are not very useful in practice, because they only characterize perfect levels of uniqueness and distinctiveness. Relaxing them in order to allow the characterization of different levels of uniqueness and distinctiveness would require the introduction of additional parameters such as thresholds applied on the distance values. These parameters are difficult to settle in practice. In order to characterize the uniqueness and distinctiveness of a given descriptor, we therefore need to define a specific methodology.

2.2 Protocol

In order to characterize efficiently the descriptive power of different shape descriptors in terms of uniqueness and distinctiveness, we introduce the following protocol, which is independent of the final application (spotting, recognition…). First, the distances between all the noisy symbols and all the models are computed. Second, we apply the k nearest neighbour rule (kNN) in order to associate with each noisy symbol its k nearest models in the descriptor's representation space. Then, we can compute, for each descriptor, measures characterizing its descriptive power and robustness towards noise. Additionally, the complementarity between multiple descriptors may be measured.

In the remaining part of this section, we present two types of measures. The measures of the first type, introduced in Sects. 2.3 to 2.7, characterize indirectly the levels of uniqueness and/or distinctiveness (which are too parameter-dependent to be computed directly) of a single descriptor. We first recall in Sects. 2.3 to 2.5 the definitions of confusion matrices, recognition rate, precision, recall and Cumulative Match Characteristics (CMC) curves. Even though some of these measures are already used by many researchers in our community, our contribution here is that we link them to the notions of distinctiveness and uniqueness. Second, we introduce two measures that are original in the field of document analysis. These two measures are respectively the tolerance intervals (see Sect. 2.6), characterizing the robustness of descriptors towards noise, and a qualitative measure based on an analogy with Doddington's zoo, widely known in the field of biometrics, characterizing the symmetries in the confusions (see Sect. 2.7). Concerning the measures of the second type, we introduce in Sect. 2.8 original measures to characterize upstream the complementarity between multiple descriptors. These measures constitute an alternative to the usual approach where the descriptors to be combined are selected by trial and error, considering the performance characteristics of the overall system. It may also be helpful when choosing the best combination scheme for a given set of descriptors.

Table 1 A contingency matrix M. n_{il} is the number of noisy versions $\hat{S}_i^j$ of symbol i which nearest model is S_l in the representation space of the studied descriptor

$n_{i\cdot} - n_{\cdot l}$	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot c}$
$n_{1\cdot}$	n_{11}	n_{12}	…	n_{1c}
$n_{2\cdot}$	n_{21}	n_{22}	…	n_{2c}
⋮	⋮	⋮	⋱	⋮
$n_{c\cdot}$	n_{c1}	n_{c2}	…	n_{cc}

2.3 Confusion matrix

A confusion matrix M is a quantitative measure characterizing the descriptive power of a given descriptor. It is a contingency matrix computed from the array of distances between the descriptions of the noisy symbols and the models. This matrix contains c rows and c columns, where c is the number of models (see Table 1).The value in the cell n_{il} of the confusion matrix is the number of noisy versions $\hat{S}_i^j$ of symbol i which nearest model is S_l in the descriptor's representation space:

$$n_{il} = \sum_{j=1}^{n_i} \delta\left(NN(\hat{S}_i^j) = S_l\right) \tag{5}$$

with $NN(\hat{S}_i^j)$ being the model which is the nearest to $\hat{S}_i^j$ (*i.e.* the result of the $1NN$ rule following our protocol) and $\delta\left(NN(\hat{S}_i^j) = S_l\right) = 1$ if the nearest model of $\hat{S}_i^j$ is S_l, 0 otherwise. If we denote by $n = \sum_{i=1}^{c}\sum_{l=1}^{c} n_{il}$ the total number of noisy symbols, the descriptor may be considered as perfectly describing the database when the confusion matrix is diagonal (*i.e.* $\text{trace}(M) = n$).

Even if the confusion matrix is in general defined by using the 1-nearest neighbour rule (see Eq. 5), we can characterize the behaviour of the descriptor in an enlarged neighbourhood of the noisy symbol by considering confusion matrices $M(k)$ associated with higher ranks k, by defining:

$$n_{il}(k) = \sum_{j=1}^{n_i} \delta\left(kNN(\hat{S}_i^j) = S_l\right) \tag{6}$$

where $k\text{NN}(\hat{S}_i^j)$ is the kth nearest model to $\hat{S}_i^j$ (*i.e.* the result of the kNN rule following our protocol). We can note that the confusion matrix M shown in Table 1 is the same matrix as matrix $M(1)$, while matrix $M(k)$ with $k > 1$ may be obtained by replacing the values n_{il} in Table 1 with the $n_{il}(k)$ defined in Eq. 6. In the remaining part of this article, we consider by default confusion matrices at rank $k = 1$, unless we explicitly specify that we consider higher ranks $k > 1$ (for CMC curves for example, see Sect. 2.5).

2.4 Recognition rate, precision and recall

This section is dedicated to quantitative measures characterizing the richness of a given descriptor in terms of uniqueness and/or distinctiveness.

First of all, the recognition rate (RR) provides the percentage of noisy symbols such that their nearest model is the "good" one. It is computed from the confusion matrix (see Table 1 and Eq. 5) as follows:

$$RR = \frac{\text{trace}(M)}{n} \tag{7}$$

It has to be noted that we can also compute the recognition rates at any rank $k > 1$ by using the confusion matrix $M(k)$ (see Eq. 6):

$$RR(k) = \frac{\text{trace}(M(k))}{n} \tag{8}$$

Hereafter, we consider by default the recognition rate at rank $k = 1$, unless we explicitly specify that we consider higher ranks of k (for CMC curves for example, see Sect. 2.5).

While the recognition rate gives some information about the descriptive power of a descriptor on the whole database, one may be interested in the behaviour of the descriptor for a particular symbol. We can note that a symbol i which is badly described in the descriptor's representation space is associated with a large number of extra-diagonal elements n_{il} and/or n_{li} (with $l \neq i$) in the confusion matrix (see Eq. 5). A low distinctiveness for symbol i is characterized by high values of the column cells n_{li}. On the other hand, a low uniqueness for symbol i is characterized by high values of the row cells values n_{il} (see the definitions of distinctiveness and uniqueness given in Sect. 2.1). The level of distinctiveness and uniqueness for a given symbol i may therefore be respectively measured by using precision $P(i)$ and recall $R(i)$, where:

$$P(i) = \frac{n_{\text{ii}}}{\sum_{l=1}^{c} n_{\text{li}}} \tag{9}$$

$$R(i) = \frac{n_{\text{ii}}}{\sum_{l=1}^{c} n_{\text{il}}} \tag{10}$$

To characterize the distinctiveness and uniqueness for the whole dataset, one may consider only the scalar value corresponding to the average precision and/or recall among all the symbols $i = 1 \ldots c$:

$$P = \frac{1}{c} \sum_{i=1}^{c} \left(\frac{n_{\text{ii}}}{\sum_{l=1}^{c} n_{\text{li}}} \right) \tag{11}$$

$$R = \frac{1}{c} \sum_{i=1}^{c} \left(\frac{n_{\text{ii}}}{\sum_{l=1}^{c} n_{\text{il}}} \right) \tag{12}$$

We can note that, in the special case where the number n_i of noisy versions of symbol i is the same for all the c symbols i, the mean recall equals the recognition rate (given that $\sum_{l=1}^{c} n_{\text{il}} = n_i$, by construction).

2.5 Cumulative match characteristics curve

In order to characterize the behaviour of the descriptor in an enlarged neighbourhood of the symbols to describe (not only considering the nearest model), we may consider the recognition rates at ranks $k > 1$. The Cumulative Match Characteristic (CMC) curves are most widely used for evaluating the performance characteristics of semi-automated recognition systems where N candidates are proposed to a (often human) supervisor, the role of the supervisor being to select the good candidate. Such curves are useful to quickly visualize the cumulated recognition rates at different ranks k:

$$CMC(k) = \sum_{r=1}^{k} RR(r) \tag{13}$$

where $RR(k)$ is the recognition rate at rank $k \geq 1$ (see Eq. 8). For an example of a CMC curve see Fig. 7. If the CMC curve reaches a sufficiently high value at a rank k being tractable for the supervisor, then the semi-automated system is considered as effective.

2.6 Characterization of the Robustness towards noise

In this section, we present the Tolerance Interval, which has been defined in the context of face recognition in [25] in order to characterize the robustness of descriptors towards noise. Tolerance Intervals may be calculated in the case where the amount of noise is controlled by one parameter ω (*i.e.* in general when the noise is synthetically added to the images). For example, for Gaussian white noise the parameter is the standard deviation: $\omega = \sigma$. The recognition rate is computed as a function of the value of the noise parameter. A Tolerance Interval (TI) at $p\%$ may be defined as the range of values of parameter ω such that the recognition rate RR remains greater than $1 - p/100$. Examples of Tolerance Intervals are given in Table 3. For a fixed p, the larger is the Tolerance Interval, the more robust is the descriptor. Tolerance Intervals characterize in a compact way the robustness of a descriptor towards a given type of noise; they may be very helpful when choosing the descriptor which is best suited to a specific kind of noise.

2.7 The zoo qualitative characterization

All the previous measures are intrinsically quantitative. In particular, we have shown how the precision and recall measures may be used at the symbol level to characterize the asymmetries in the confusions of a single descriptor for a given symbol (*e.g.* symbol i may be confused with others

and not the opposite). However, when trying to compare the descriptive power of multiple descriptors for a given symbol, it is difficult to consider jointly multiple precision and recall values. That is why we introduce a qualitative measure based on the definition of categories of symbols. Our categorization is inspired from Doddington et al.'s terminology [26]. This terminology was first introduced in the field of speaker recognition for biometrics. Figure 1 provides examples of the different categories. We give the original definitions by Doddington et al., followed by their adaptations in our context.

- "In our model, sheep dominate the population and systems perform nominally well for them". In our context, sheep are the symbols which are well represented by the descriptor, with both high uniqueness and distinctiveness. Therefore we can define a sheep i as a symbol associated with high values of both precision and recall. In Fig. 1 the symbol "circle" is a sheep;
- "Lambs, in our model, are those speakers who are particularly easy to imitate". In our context, lambs are symbols characterized by a low distinctiveness and therefore associated with a low precision. In Fig. 1 the symbol "triangle" is a lamb;
- "Wolves, in our model, are those speakers who are particularly successful at imitating other speakers". In our context, lambs are symbols characterized by a low uniqueness and thus associated with a low recall. In Fig. 1 the symbol "square" is a wolf;
- "Goats tend to adversely affect the performance of systems by accounting for a disproportionate share of the missed detections". A goat is a model such that its noisy versions are in general farther than a given threshold θ to all the models (dotted circles in Fig. 1, in the case of an Euclidean distance). In Fig. 1 the symbol "rhombus" is a goat.

While in the case of sheep the descriptive power of the descriptor may be considered as satisfactory, in the case of goats the descriptive power is low. But, for a given symbol, the behaviours of two different descriptors may differ. For instance, symbol i may be a sheep with descriptor $D1$ and a wolf with descriptor $D2$. At the level of the whole database, these behaviours may be complementary, *e.g.* in the case where symbol i is a sheep with descriptor $D1$ and not a sheep with descriptor $D2$, and vice-versa for symbol j. In that case, the description may be improved by combining the two descriptors, instead of considering a single descriptor. That is the reason why, in the following section, we introduce measures to characterize the complementarity of different descriptors.

2.8 Ensemble measures characterizing the complementarity of different descriptors

In this section, we introduce quantitative measures of the complementarity between different descriptors. We can note that, in most cases, confronting the confusion matrices of different descriptors does not help to characterize their complementarity. Indeed, while the confusion matrix provides information at the symbol level, the complementarity occurs at the level of the noisy image. For instance, when the confusion matrix says that, for a given symbol i, only 15 noisy versions over 30 are well described by descriptor $D1$ and by descriptor $D2$, these 15 well-described images may be the same (in this case the two descriptors are not complementary at all for this symbol) or totally distinct (in this case the two descriptors are perfectly complementary for this symbol), but from the confusion matrix we cannot guess which case we are dealing with. That is why we introduce the following complementarity measures that may be computed at different ranks $k \geq 1$, with $\hat{S}_i^j$ being a noisy version of the original model S_i and $kNN_D(\hat{S}_i^j)$ the kth nearest model of $\hat{S}_i^j$ in the representation domain associated with the descriptor D:

$$U(k) = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \delta\left(\left\{kNN_{D1}(\hat{S}_i^j) = S_i\right\} \cup \left\{kNN_{D2}(\hat{S}_i^j) = S_i\right\}\right) \tag{14}$$

$$I(k) = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \delta\left(\left\{kNN_{D1}(\hat{S}_i^j) = S_i\right\} \cap \left\{kNN_{D2}(\hat{S}_i^j) = S_i\right\}\right) \tag{15}$$

$$I_{D1}(k) = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \delta\left(\left\{kNN_{D1}(\hat{S}_i^j) = S_i\right\} \cap \left\{kNN_{D2}(\hat{S}_i^j) \neq S_i\right\}\right) \tag{16}$$

$$I_{D2}(k) = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \delta\left(\left\{kNN_{D2}(\hat{S}_i^j) = S_i\right\} \cap \left\{kNN_{D1}(\hat{S}_i^j) \neq S_i\right\}\right) \tag{17}$$

$$C(k) = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \delta\left(\left\{kNN_{D1}(\hat{S}_i^j) \neq S_i\right\} \cap \left\{kNN_{D2}(\hat{S}_i^j) \neq S_i\right\}\right) \tag{18}$$

The measure $U(k)$ is the number of noisy images such that their kth nearest model is the "good" one for at least one of the two descriptors $D1$ or $D2$. The measure $I(k)$ is the number of noisy images such that their kth nearest model is the "good" one for both descriptors $D1$ and $D2$. We can note that, by construction, $I(k) \leq U(k)$. The measure I_{D1} is the total number of noisy images such that their kth nearest model is the "good" one for descriptor $D1$ but not for $D2$ (and vice-versa for I_{D2}). We can note that $U(k) = I(k) + I_{D1}(k) + I_{D2}(k)$. Finally, $C(k)$ is the total number of noisy images such that their kth nearest model is not the "good" one, neither for descriptor $D1$ nor for $D2$. We can note that

$C(k) + U(k) = n$, where n is the total number of noisy symbols in the database. Figure 8 provides a visual example of such measures with descriptors ART and SC. Let us note the following relationship: $\frac{I(k)+I_D(k)}{n} = RR_D(k)$, where $RR_D(k)$ is the recognition rate at rank k (see Eq. 8) associated with descriptor D. Of course the measures given in Eqs. (14–18) can be directly extended with more than two descriptors.

The numbers of images which are well represented by one descriptor but not by the other one (*i.e.* I_{D1} and I_{D2}) allow us to quantify the complementarity of the two descriptors. In particular, the value of $U(1)$ is the maximal number of symbols that may be well-described at rank 1 (*i.e.* the objective value) when conceiving a strategy for selecting, for each symbol, the best descriptor for this symbol. The more the objective value $\frac{U(1)}{n}$ exceeds the maximal recognition rate of the two descriptors, the more complementary are these two descriptors. Characterizing the complementarity between descriptors is very interesting, as considering a combination of complementary descriptors may improve the richness of the description of the symbol compared to considering a single descriptor.

3 Experimental study

In this section, we perform an experimental study to illustrate the effectiveness of the measures we define in Sect. 2. The objective is to show how to use these measures for (1) comparing multiple descriptors in terms of descriptive power and noise robustness and (2) measuring the complementarity of multiple descriptors. For this purpose, we consider two well-known shape descriptors (that will be described in Sect. 3.1): ART and SC. The main objective here is not to characterize the performance of these two descriptors, but rather to illustrate the contribution brought to the community by our innovative protocol and measures. We selected ART and SC among the large variety of available shape descriptors for two main reasons. First, because of the paper size limit, we could not consider more than two descriptors. Second, ART and SC belong to different categories of shape descriptors (ART is 2D while SC is a 1D descriptor based on contours). This fact certainly makes them complementary to some extent, which is interesting to illustrate our complementarity measure.

This section is organized as follows. A brief description of the considered descriptors is given in Sect. 3.1. Next, the databases we use and their features are detailed in Sect. 3.2. Then, in Sect. 3.3 we compute and analyse the measures proposed in Sect. 2.

3.1 Statistical shape descriptors

Among the various shape descriptors that have been proposed in the literature [8–10], we selected in this paper two well-known statistical descriptors which are essentially different in their primitives: Angular Radial Transform (ART) [22] and Shape Context (SC) [23]. While ART is based on a 2D primitive (region inside the image) and provides a polar-based feature vector, SC is a 1D primitive (relying on the extraction of shape contours) resulting in a histogram-based feature vector.

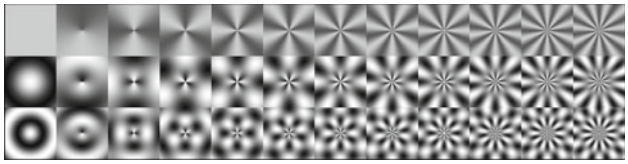


Fig. 2 Real parts of the basis functions of the descriptor ART, for n from 0 to $N = 2$ (*from top to bottom*) and m from 0 to $M = 11$ (*from left to right*). It has to be noted that their imaginary parts are similar except the quadrature phase difference

3.1.1 The ART descriptor

The ART descriptor [22] is the result of a complex 2D transform defined on a unit circle using polar coordinates. More precisely, ART coefficients are defined by the projection of the original image represented in polar coordinates on a basis of orthogonal complex functions $V_{n,m}(\rho, \theta) = A_m(\theta)R_n(\rho)$ (*cf.* Fig. 2). These basis functions are defined by multiplying a radial function R_n of parameter n by an angular function A_m of parameter m, the pair of parameters (n, m) defining the order of the coefficient $F_{n,m}$.

Invariance to similarity transforms is achieved by (1) using an exponential functional in the angular function (to get invariance towards rotations) and (2) centring and scaling the shape image before computing the coefficients (to achieve invariance towards scale and translation).

Finally, the distance between shapes is measured by the Manhattan (L_1) distance.

3.1.2 The SC descriptor

The SC descriptor [23] is based on relative spatial locations between some points extracted from the contours of the shape to analyse. Figure 3 illustrates its underlying principle. The shape to be described is represented by a discrete set of points extracted from the external and internal contours of the shape.

This descriptor can be considered invariant to scaling if the background is not too complex, since the radial distances are normalized by the average distance between all the pairs of points of the shape. In addition, it is invariant towards translation and can easily be made invariant towards rotation. And, given that the SC descriptor provides coarse information extracted from the whole shape, it is relatively robust towards occlusions.

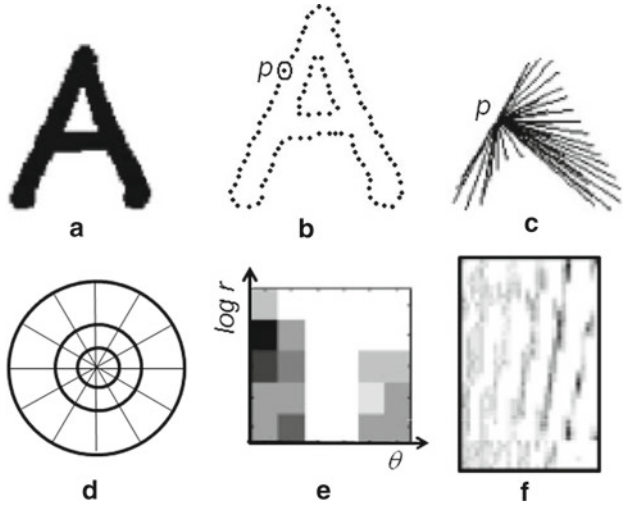


Fig. 3 Principle of the SC descriptor. **a** shape to be described, **b** contour points sampled from this shape, **c** set of vectors associated with the reference point p extracted from the contour, **d** classes (*bins*) used for the histograms, **e** histogram of the coordinates of the vectors shown in c (*i.e.* shape context of point p), **f** set of shape contexts of the shape shown in a

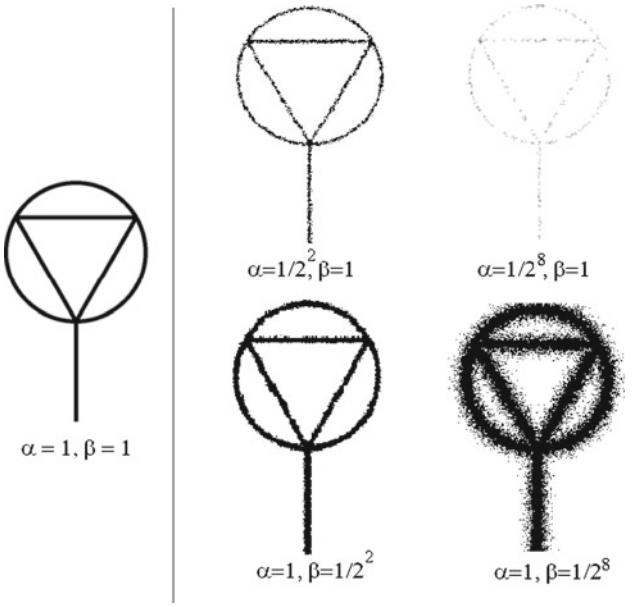


Fig. 4 Symbol # 86 with different levels of noise (from left to right, *unnoisy symbol*, first row: *noisy symbol* with noise α at levels $N = 2$ and $N = 8$, second row: noisy symbol with noise β at levels $N = 2$ and $N = 8$

Even though the χ^2 distance was initially used to compare shape contexts from different symbols [23], more recently numerous authors have chosen to consider shape contexts as feature vectors and to compare them by using the L_2 (Euclidean) distance [24].

3.2 Experimental protocol

In our experiments, we consider the GREC 2003 symbol database [6]. This database contains 150 models of symbols, which are used to generate noisy versions by applying the Kanungo algorithm [27]. The Kanungo noise is an additive noise applied to binary images; it is controlled by six parameters. Among these parameters, we chose to vary α and β, which simulate the presence of an ascending amount of noise in the image. When α decreases, the probability for a symbol pixel to be inverted and considered as a background pixel increases (which may be seen as some kind of "salt" noise), while when β decreases the probability to invert a background pixel as a symbol pixel increases ("pepper" noise). It has to be noted that these probabilities of inversion decrease according to the distance from the contour of the shape. Five databases, each one containing 30 random noisy versions of each of the 150 model symbols are generated for each $\alpha = \frac{1}{2^N}$, with N varying from 2 to 10 by a step of 2 (*cf.* Fig. 4). Five databases are constructed similarly by varying parameter β. At the end, we obtain a database containing the 150 model images (one image per symbol model) and 10 test databases, each one containing $30 \times 150 = 4,500$ noisy symbols.

In this experiment, we compare each noisy image from the 10 noisy databases to the database containing the model images. For this purpose, we use for each descriptor its associated distance (the Manhattan distance for ART and the Euclidean distance for SC) and the protocol presented in Sect. 2.2.

3.3 Experimental results

The analysis of the experimental results is in four steps. During the first step (Sect. 3.3.1), a coarse evaluation of the performance characteristics of the descriptors ART and SC is provided. For this coarse evaluation, we consider the usual performance measures (recognition rate, precision and recall, see Sect. 2.4). The second step (Sect. 3.3.2) is a comparison of the robustness of the two descriptors towards noise. For this comparison, we compute Tolerance Intervals (see Sect. 2.6) and we consider the two different types of Kanungo noise (α "salt" noise and β "pepper" noise). Then, a subset of databases (among the most noisy) are selected for the third step of the analysis (Sect. 3.3.3). The third step is a detailed analysis relying on the confusion matrices (see Sect. 2.4), the CMC curves (see Sect. 2.5), and the qualitative measures we introduce in Sect. 2.7. The fourth step, given in Sect. 3.3.4, is a study of the complementarity of ART and SC, based on the complementarity measures defined in Sect. 2.8.

Table 2 Mean recall and precision associated with the two descriptors on the databases with the levels of noise $\alpha = \frac{1}{2^N}$ and $\beta = \frac{1}{2^N}$ (with $N = 2, 4, 6, 8, 10$)

The values between brackets are the corresponding standard deviations. We can note that the mean recall equals the recognition rate because the number $n_i = 30$ of noisy versions is constant over all the symbols i (see Sect. 2.4)

N	Descriptor	Noise α		Noise β	
		Mean precision (SD)	Mean recall (SD)	Mean precision (SD)	Mean recall (SD)
2	ART	100% (0)	100% (0)	100% (0)	100% (0)
	SC	99.06% (0.059)	98.98% (0.076)	99.23% (0.05)	99.09% (0.066)
4	ART	94.69% (0.201)	96.09% (0.193)	100% (0)	100% (0)
	SC	99.14% (0.055)	98.87% (0.081)	98.99% (0.057)	98.80% (0.08)
6	ART	93.05% (0.232)	94.71% (0.217)	99.74% (0.032)	99.58% (0.052)
	SC	99.13% (0.059)	98.8% (0.092)	97.55% (0.09)	96.33% (0.148)
8	ART	90.60% (0.24)	90.53% (0.251)	89.46% (0.279)	91.87% (0.269)
	SC	98.814% (0.071)	98.76% (0.09)	71.96% (0.396)	66.33% (0.402)
10	ART	63.6% (0.335)	52.91% (0.339)	19.09% (0.355)	28.64% (0.447)
	SC	95.19% (0.089)	94.69% (0.121)	3.30% (0.151)	4.4% (0.168)

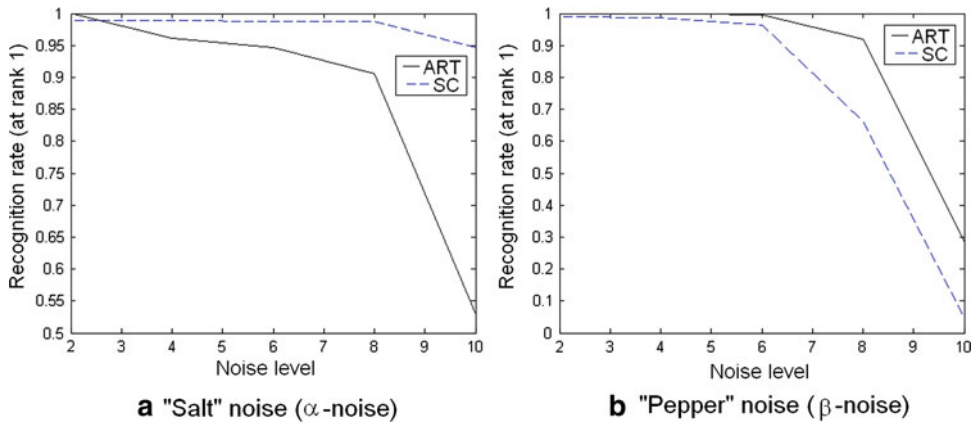


Fig. 5 Recognition rates RR (at rank 1) as a function of the level N of noise, with a noise of type α (*left*) and β (*right*)

3.3.1 *Overview of the performance characteristics of ART and SC*

The recognition rate (RR), depending on the type and amount of noise, is given in Table 2 and Fig. 5. We can note that, as explained in Sect. 2.4, the recognition rate equals the mean recall in our case where the number $n_i = 30$ of noisy versions is constant over all the symbols i. Table 2 also gives the mean precision (at rank $k = 1$ and the standard deviations of the recognition rate and mean precision. We can see that, for both descriptors and both types of noise, the quality of the description decreases when the amount of noise increases. We can also note that the decrease is more abrupt in the presence of β noise. For levels of noise $N > 2$, SC is superior to ART for "salt" noise (α-noise) while ART is superior to SC for "pepper" noise (β-noise). When $N \leq 2$ ART is always superior to SC, whatever kind of noise is applied.

We can easily understand why the performance of SC decreases drastically in the presence of "pepper" noise: SC is based on points sampled from the contour of the symbol (*cf.* Fig. 3). When pepper noise is added to the image, the shape contours are modified. In that case, the SC description is computed from inaccurate points and becomes imprecise. Conversely, "salt" noise has a thinning effect on the symbol. Therefore, when the α-noise increases, the amount of information available for computing ART is reduced, which makes ART description unstable (see the high standard deviation values in Table 2).

3.3.2 *Comparison of the Robustnesses of ART and SC towards noise*

From Table 2 and Fig. 5, we compute the Tolerance Intervals (TI) (described in Sect. 2.6) corresponding to descriptors ART and SC, towards α-noise and β-noise. The TIs at levels $p = 5\%$ and $p = 20\%$ are given in Table 3.

The results in Table 3 are consistent with the shapes of the curves in Fig. 5. In particular, in the presence of α-noise, the TI of ART is narrower than the TI of SC, which implies that SC is more robust than ART towards α-noise. For β-noise and descriptor SC, the fact that the TIs at $p = 5\%$ and $p = 20\%$ are both equal to $[1, 6]$ is due to a drastic drop

Table 3 Tolerance Intervals (TIs) of ART and SC towards Kanungo noise of type α and β

Type of noise	α-noise level N		β-noise level N	
	$p = 5\%$	$p = 20\%$	$p = 5\%$	$p = 20\%$
ART	[1, 4]	[1, 8]	[1, 6]	[1, 8]
SC	[1, 8]	[1, 10]	[1, 6]	[1, 6]

These TIs are computed by using the recognition rates given in Table 2.The noise levels N are such that $\alpha = \frac{1}{2^N}$, (respectively $\beta = \frac{1}{2^N}$)

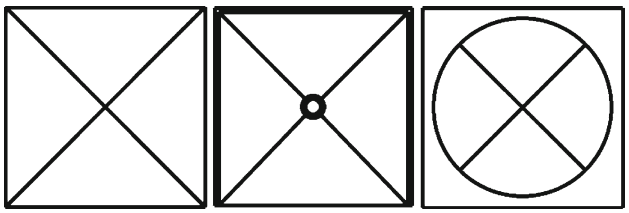


Fig. 6 From left to right: symbols 11, 87 and 125

in the recognition rates between the levels of noise $N = 6$ and $N = 8$. Table 3 shows the decrease in the quality of the description when the amount of noise is increased. In particular, none of the two descriptors is tolerant at $p = 5\%$ towards a noise of level $N = 10$ (neither for α-noise nor for β-noise). Consequently, the rest of this section is devoted to the detailed analysis of the results for the levels of noise $N = 6$ and $N = 8$ for α and β. Indeed, when applied on these databases, the behaviour of the descriptors is representative of the general case where the noise is not too strong and at least one of our two descriptors remains robust.

3.3.3 Detailed analysis of the performance characteristics of ART and SC

To provide a more detailed analysis of the descriptors' behaviours, we show in Table 4 some extracts of the confusion matrices (see Sect. 2.3) for symbols 11, 87 and 125 (see Fig. 6). We consider these particular symbols because, in the presence of β-noise (at levels 6 and 8), they are subject to confusions. Let us now focus on the database with noise β at level $N = 6$ (Table 4a and c). Among the 4,500 noisy symbols of the whole database, ART badly describes only 19 noisy symbols. All of these 19 poor descriptions come from confusions between symbols 87 and 125. On the other hand, the SC descriptor badly describes 165 noisy symbols over 4,500, among which 29 poor descriptions are due to confusions between symbols 11 and 87.

To go deeper into the analysis of the database with β-noise of level $N = 6$, let us consider additionally the precision and recall measures (see Sect. 2.4) associated with the symbols 11, 87 and 125 and the qualitative definitions introduced in Sect. 2.7. From this analysis, we conclude that

Table 4 Extracts of the confusion matrices of the descriptor ART on databases with noise β and levels (a) $N = 6$ and (b) $N = 8$ and of the descriptor SC on databases with noise β and levels (c) $N = 6$ and (d) $N = 8$

GT	Nearest model		
	11	87	125
(a) ART with noise β at level $N = 6$			
11	30	0	0
87	0	11	19
125	0	0	30
(b) ART with noise β at level $N = 8$			
11	30	0	0
87	0	0	30
125	0	0	3
(c) SC with noise β at level $N = 6$			
11	1	29	0
87	0	30	0
125	0	0	30
(d) SC with noise β at level $N = 8$			
11	0	22	8
87	0	12	18
125	0	2	28

- symbol 11 (respectively symbol 125) is a sheep for descriptor ART (resp. SC), as the quality of the description of this symbol is satisfying (indeed $P = 1$ and $R = 1$ for symbol 11 with ART and $P = 0.81$ and $R = 1$ for symbol 125 with SC);
- symbol 125 (respectively symbol 87) is a lamb for descriptor ART (resp. SC), as other symbols may be confused with it. Indeed, $P = 0.61$ for symbol 125 with ART and $P = 0.51$ for symbol 87 with SC;
- symbol 87 (respectively symbol 11) is a wolf for descriptor ART (resp. SC), as this symbol may be confused with other symbols. Indeed, $R = 0.37$ for symbol 87 with ART and $R = 0.03$ for symbol 11 with SC).

A preliminary statistical study has shown that the databases are very homogeneous for both descriptors [20]. This means that the number of goats is very reduced in this context. Therefore, we did not look for goats, which would have required the settlement of an additional parameter θ (see Sect. 2.7). The fact that symbol 11 is a wolf for symbol 87 in the presence of pepper noise for descriptor SC is easily understandable. Indeed, the pepper noise located at the centre of the cross in symbol 11 may be confused with the small circle at the centre of symbol 87 (see Fig. 6). With ART, 11 noisy occurrences of symbol 87 have the "good" model 87 as nearest neighbour, while the remaining 19 occurrences have the model 125 as nearest neighbour. This phenomenon

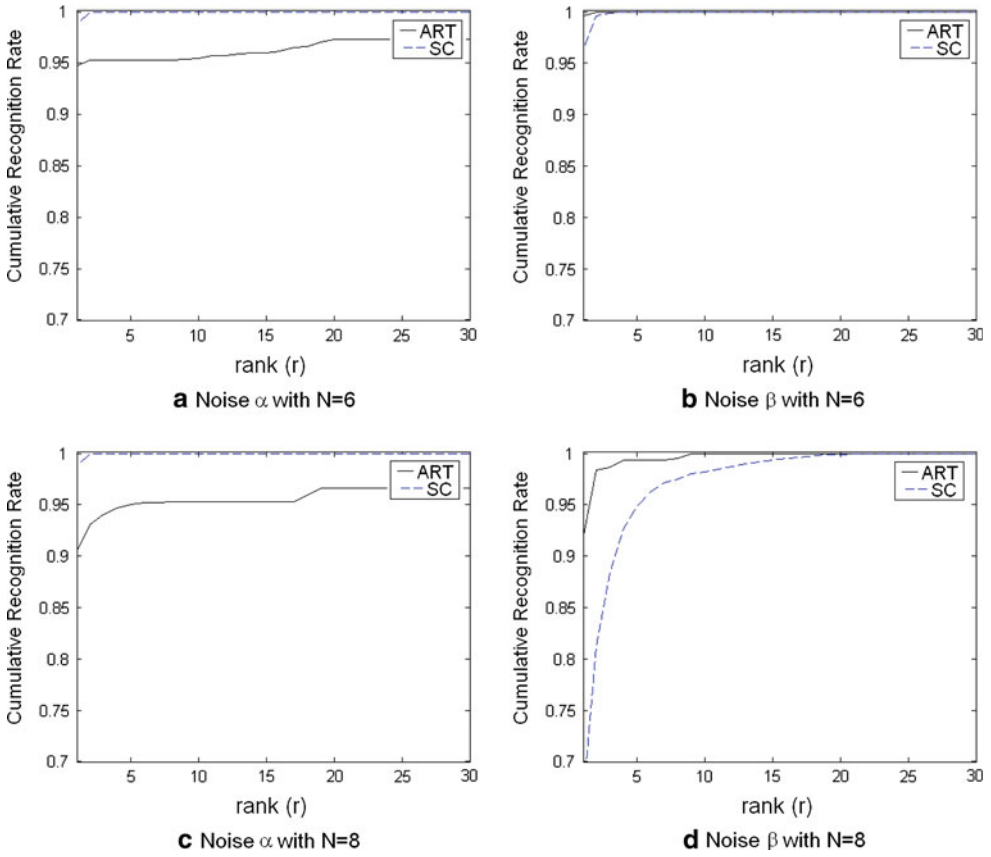


Fig. 7 The CMC curves associated with the databases (*first column*): $\alpha = \frac{1}{2^6}$ and $\alpha = \frac{1}{2^8}$ and (*second column*): $\beta = \frac{1}{2^6}$ and $\beta = \frac{1}{2^8}$

makes us guess that there is an overlap between the images of symbols 87 and 125 in the ART description space.

In addition, Table 4 shows that these dissymmetries in the confusion matrix increase when more pepper noise is added to the symbols (between levels $N = 6$ and $N = 8$). For instance, when the level of β-noise reaches $N = 8$, symbol 87 becomes a wolf for symbol 125 in the SC space, as the images of symbols 87 become closer to the model 125 than to the model 87.

From this point of view, characterizing the results of the descriptions not only at the first rank (*i.e.* using the nearest model), but at higher ranks (*i.e.* using the k nearest models) is very important. Indeed, in the case where there is an overlap between two symbols (*e.g.* in the case of symbols 87 and 125 for ART), the "good" model may be among the two nearest models but not necessarily the nearest one. And we can consider that, if the "good" model is among the two nearest models, the quality of the description is better than if the "good" model is farther. In other words, we can consider that a wolf model which is near its lamb (see Fig. 1) is better described than a goat. In order to take into account higher ranks $k > 1$, we consider the CMC curves (see Sect. 2.5) shown in Fig. 7. Figure 4a and c shows that most of the confusions of the descriptor SC at rank 1 with noise α (see Table 2) are solved at ranks 2 or 3. This means that the descriptive power of SC in the presence of α-noise is relatively good, because even the noisy symbols which are badly classified by using the $1NN$ rule are well classified when considering the $3NN$ rule. On the contrary, we can see from Fig. 4d that the descriptive power of SC is bad with β-noise of level $N = 8$, as the CMC curve of SC does not reach 100% recognition rate before rank $k > 20$.

From Table 4a and c, we can also note that, with noise β at level $N = 6$, while ART does not manage to discriminate between symbol 87 and symbol 125, SC perfectly discriminates between these two symbols. And vice-versa for symbols 11 and 87 in the same database. Therefore, we can consider that ART and SC are complementary for symbols 11, 87 and 125 in the particular database with noise β at level $N = 6$. Which means that combining ART and SC to obtain a single descriptor may improve (on average) the quality of our description, compared to a single descriptor. Now let us expand our study of the complementarity of ART and SC to the whole dataset.

3.3.4 Analysis of the complementarity of ART and SC

To end this experimental study, we measure the complementarity of ART and SC by the measures defined in

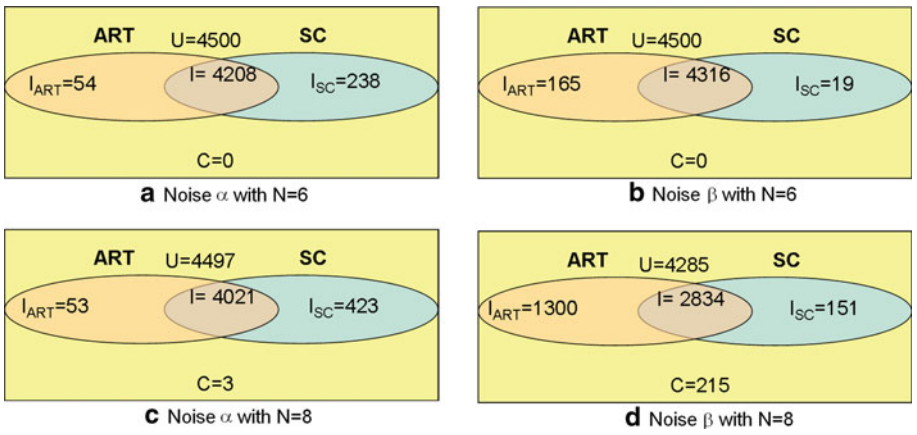


Fig. 8 Measures $U(k)$, $I(k)$, $I_{\mathrm{ART}}(k)$ and $I_{\mathrm{SC}}(k)$ (with $k = 1$, denoted as U, I, I_{ART} and I_{SC} for the sake of readability) with different levels of noise

Eqs. (14–18), where $D1$ = ART and $D2$ = SC. The results are given in Fig. 8.

From that figure we can see that both descriptors are really complementary on the four databases. For the two databases with a noise level $N = 6$, we can see that the two descriptors are perfectly complementary as $C = 0$ with the two types of noise (even if ART is superior to SC in the presence of β-noise while SC is superior to ART in the presence of α-noise). It means for instance that, for a task of recognition, any query symbol from these noisy databases may be correctly classified automatically by using the ART and SC descriptors. Provided an ideal adaptive descriptor selection method that would select, for each query symbol and each type of noise, the best performing descriptor for this particular symbol. With a noise level of $N = 8$, we can observe high values of both I_{ART} and I_{SC}, which means that combining ART and SC may enhance the average quality of description of the symbols, whatever type of noise is present in the images, compared to a single descriptor.

3.3.5 Conclusion of the experimental section

The results given in this section have shown that SC is superior to ART in the presence of "salt" noise while ART is superior to SC in the presence of "pepper" noise. Both descriptors are robust towards a small amount of noise but their performance decreases drastically when the amount of noise increases (especially with "pepper" noise). However, the SC descriptor remains more robust than ART with salt noise. We can also note that these two descriptors are complementary. Therefore, combining them may enhance the quality of description of a symbol compared to that of a single descriptor.

4 Discussion

Well-known and widely used evaluation measures such as the recognition rate (RR) and the Precision-Recall values (see Sect. 2.4) are very useful to measure whether a given descriptor is adapted to a particular context. They provide performance characteristics on a set of evaluation databases which are supposed to be representative of the images the system will find in a real environment. However, when no descriptor is superior to the others on all the databases (for different types of noise for example), then the issue of the usefulness of the evaluation measures proposed in Sects. 2.5 and 2.6 arises. Thereby, the Tolerance Interval quickly gives an idea of the robustness of descriptors towards noise, while the CMC curves characterize the quality of the description in the neighbourhood of the noisy symbols.

Nevertheless, the descriptive power of a given descriptor may vary from one symbol to another. Indeed, in any database and for any descriptor there are symbols which are well-described and others which are not (provided a database of sufficient complexity). In this context, using the qualitative measures introduced in Sect. 2.7 becomes useful, since it allows us to detect the overlapping symbols. The information given by these measures is equivalent to the information given by the CMC curves, but detailed for each symbol in the database, while the CMC remains general.

The complementarity measures (see Sect. 2.8) provide information about the benefit we can expect from the combination of multiple descriptors. Hence, the best configuration for a pair of descriptors is to be perfectly complementary, which is the case for ART and SC on the databases of noise level $N = 6$, for both α and β noise. Measuring upstream the complementarity of shape descriptors is an interesting alternative to the most widely used approach consisting in selecting the descriptors to be combined by trial and error, considering the performance characteristics of the overall system.

It has to be noted that the complementarity measures can also be used to characterize the complexity of a given database. Indeed, if we consider a well-chosen set of mutually complementary descriptors and that this set of descriptors gives poor results on a given database (*i.e.* the value of C is high), we can consider that this database is highly complex.

On the contrary, when (considering the same set of descriptors) the value of C is small (*i.e.* only a small number of samples are badly represented by all the descriptors), the database can be considered as less complex. When, in addition, the value of I almost equals the value of U (*i.e.* all the descriptors describe correctly almost all the samples), the complexity of the database may be considered as low.

To conclude this discussion, we can note that all the measures introduced in Sect. 2 may also be applied to several structural descriptors. Indeed, we have seen that what we only need in our protocol is a confusion matrix including distances or dissimilarity measures between noisy symbol descriptors and models. In this case, we can easily extend this framework to structural methods based on graph representation and graph similarity measures which quantify the effort needed to match one graph with a another [28–30].

5 Conclusion

In this paper, we introduced an experimental protocol and measures for characterizing the performance of descriptors in the context of symbol description. The measures we introduced are of two types. While the first type of measures is devoted to the descriptive power of each descriptor taken separately in terms of uniqueness, distinctiveness or robustness towards noise, the second type of measures aims at evaluating the complementarity of a set of descriptors. Concerning the first type of measures, we first recalled the definitions of confusion matrices, recognition rate, precision, recall and Cumulative Match Characteristics (CMC) curves. Although some of these measures are already known by many researchers in our community, our originality is that we linked them to the notions of distinctiveness and uniqueness. Second, we introduce two measures that are new in the field of document analysis. These two measures are respectively the tolerance intervals, characterizing the robustness towards noise, and a qualitative measure characterizing the symmetries in the confusions. Concerning the measures of the second type, we introduce original measures to characterize upstream the complementarity between multiple descriptors. These measures may assist the researchers when selecting the descriptors to be combined, instead of selecting them by trial and error downstream.

We analysed experimentally a didactic case study (considering the widely-known descriptors ART and SC), to illustrate the effectiveness of the measures we defined. Even if the main objective of this experimental part is didactic and not directly to draw conclusions about the performance characteristics of ART and SC, it highlights the relevance of combining SC and ART for describing symbols.

It has to be noted that the complementarity measures may be additionally used for characterizing the complexity of a given database: the basic idea behind this is that, when a well-chosen set of mutually complementary descriptors gives poor results on a given database, we can consider that this database is highly complex. As a conclusion, our measures may therefore be helpful for various purposes concerning performance evaluation, in the field of document description and analysis. We are currently working on an ambitious performance evaluation campaign relying on our protocol and measures, dedicated to symbol description by shape descriptors.

References

1. Terrades, O.R., Tabbone, S., Valveny, E.: A review of shape descriptors for document analysis. In: Proceedings of the International Conference on Document Analysis and Recognition—ICDAR'07, pp. 227–231 (2007)
2. Phillips, I., Chhabra, A.: Empirical performance evaluation of graphics recognition systems. IEEE Trans. PAMI **21**(9), 849–870 (1999)
3. Chhabra, A., Phillips, I.: The second international graphics recognition contest—raster to vector conversion: A report. In: Tombre, K., Chhabra, A.K. (eds.) Graphics recognition: Algorithms and Systems. LNCS, vol. 1389, pp. 390–410. Springer (1998)
4. Chhabra, A., Phillips, I.: Performance evaluation of line drawing recognition systems. In: Proceedings of 15th. International Conference on Pattern Recognition, vol. 4, pp. 864–869. Barcelona, Spain (2000)
5. Wenyin, L., Zhai, J., Dori, D.: Extended summary of the arc segmentation contest. In: Blostein, D., Kwon, Y.B. (eds.) Graphics Recognition: Algorithms and Applications. LNCS, vol. 2390, pp. 343–349. Springer (2002)
6. Valveny, E., Dosch, P.: Symbol recognition contest: a synthesis. In: Lladós, J., Kwon, Y.B. (eds.) Graphics Recognition Recent Advances and Perspectives. LNCS, vol. 3088, pp. 368–385, Springer (2004)
7. Dosch, P., Valveny, E.: Report on the second symbol recognition contest. In: Liu, W., Lladós, J. (eds.) Graphics Recognition. Ten Years Review and Future Perspectives. LNCS, vol. 3926, pp. 381–397. Springer (2006)
8. Trier, O.D., Jain, A.K., Taxt, T.: Feature extraction methods for character recognition—a survey. Pattern Recognit. **29**(4), 41–662 (1996)
9. da Fontoura Costa, L., Cesar, R.M. Jr.: Shape Analysis and Classification: Theory and Practice. pp. 685 CRC Press, Boca Raton (2001)
10. Zhang, D., Lu, G.: Review of shape representation and description techniques. Pattern Recognit. **37**, 1–19 (2004)
11. Tumer, K., Ghosh, J.: Analysis of decision boundaries in linearly combined neural classifiers. Pattern Recognit. **29**(2), 314–348 (1996)
12. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference on Machine Learning, pp. 148–156 (1996)
13. Skurichina, M., Duin, R.P.W.: Bagging, boosting and the random subspace method for linear classifiers. Int. J. Pattern Anal. Appl. **5**(2), 121–135 (2002)
14. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Int. J. Data. Min. Knowl. Discov. **2**(2), 1–43 (1998)
15. Kittler, J.: A framework for classifier fusion: is it still needed? In: Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition, pp. 45–56. Springer-Verlag (2000)

16. Ramos, O., Valveny, E., Tabbone, S.: Optimal classifiers fusion in a non-Bayesian probabilistic framework. IEEE Tran. PAMI **31**(9), 1630–1644 (2009)
17. Terrades, O.R., Valveny, E., Tabbone, S.: On the combination of ridgelets descriptors for symbol recognition. In: Liu, W., Lladós, J., Ogier, J.-M. (eds.) Graphics Recognition. Recent Advances and New Opportunities. LNCS, vol. 5046, pp. 40–50. Springer (2008)
18. Valveny, E., Dosch, P., Winstanley, A., Zhou, Y., Yang, S., Yan, L., Wenyin, L., Elliman, D., Delalandre, M., Trupin, E., Adam, S., Ogier, J.M.: A general framework for the evaluation of symbol recognition methods. Int. J. Doc. Anal. Recognit. **9**(1), 59–74 (2007)
19. Delalandre, M., Pridmore, T., Valveny, E., Locteau, H., Trupin, E.: Building synthetic graphical documents for performance evaluation. In: Liu, W., Lladós, J., Ogier, J.-M. (eds.) Graphics Recognition. Recent Advances and New Opportunities. LNCS vol. 5046, pp. 288–298. Springer (2008)
20. Valveny, E., Tabbone, S., Terrades, O.R., Philippot, E.: Performance characterization of shape descriptors for symbol representation. In: Liu, W., Lladós, J., Ogier, J.-M. (eds.) Graphics Recognition. Recent Advances and New Opportunities. LNCS vol. 5046, pp. 278–287. Springer (2008)
21. Jouili, S., Tabbone, S.: Evaluation of graph matching measures for documents retrieval. In: Eighth IAPR International Workshop on Graphics Recognition (GREC 09), La Rochelle (2009)
22. Kim, W.Y., Kim, Y.S.: A new region-based shape descriptor. ISO/IEC MPEG99/M5472 Maui, Hawaii (1999)
23. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. PAMI **24**(4), 509–522 (2002)
24. Mori, G., Belongie, S., Malik, J.: Efficient shape matching using shape contexts. IEEE Trans. PAMI **27**(11), 1832–1837 (2005)
25. Visani, M., Garcia, C., Laurent, C.: Comparing robustness of two-dimensional PCA and eigenfaces for face recognition. In: Proceedings of the International Conference on Image Analysis and Recognition (ICIAR 04) Springer LNCS 3212, 2:717–724. Porto, Portugal (2004)
26. Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D.: Sheep, Goats, Lambs and Wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In: International Conference on Spoken Language Processing (ICSLP), Sydney, USA (1998)
27. Kanungo, T., Haralick, R.M., Phillips, I.: Nonlinear global and local document degradation models. Int. J. Imaging Syst. Technol. **5**, 220–230 (1994)
28. Jouili, S., Tabbone, S.: Graph matching using node signatures. In: Proceedings of the 7th workshop on graph-based representations in pattern recognition—GbRPR 2009, pp. 154–163. Venice, Italy May (2009)
29. Robles-Kelly, A., Hancock, E.R.: Graph edit distance from spectral seriation. IEEE Trans. PAMI **27**(3), 365–378 (2005)
30. Papadopoulos, A.N., Manolopoulos, Y.: Structure-based similarity search with graph histograms. In: Proceedings of International Workshop on Similarity Search (DEXA IWOSS 99), pp. 174–178 Sep (1999)