Documents, Graphes et Optimisation Multi-Objectifs

Sébastien Adam

Documents, Graphes et Optimisation Multi-Objectifs

Sébastien Adam

Habilitation à Diriger les Recherches
de l'Université de Rouen

(Spécialité Génie Informatique, Automatique et Traitement du Signal)

Soutenue le 29/11/2011

Composition du jury

Rapporteurs :	Jean-Michel Jolion, INSA de Lyon
	Robert Sabourin, ETS, Université du Québec
	Karl Tombre, École des mines de Nancy
Examinateurs :	Jean-Marc Ogier, Université de La Rochelle
	Yves Lecourtier, Université de Rouen
	Laurent Heutte, Université de Rouen

Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes - EA 4108

Mis en page avec la classe thloria.

Remerciements

Je tiens à remercier vivement Jean-Michel Jolion, Robert Sabourin et Karl Tombre d'avoir accepté d'être les rapporteurs de ce document. Ils sont des références pour moi et j'ai beaucoup apprécié leur travail d'expertise.

Je remercie aussi Jean-Marc Ogier d'avoir accepté mon invitation et d'avoir présidé ce jury. Jean-Marc est la personne qui m'a donné le goût de la recherche et ses qualités humaines et scientifiques sont trop nombreuses pour les lister ici.

Mention spéciale aux collègues locaux de ce jury. Yves et Laurent ont pris le relai de Jean-Marc quand ce dernier est parti chercher ses fameuses 2250 heures de soleil par an sur la côte atlantique. J'apprécie énormément de travailler avec eux, et j'espère que ce n'est qu'un début.

Je remercie aussi vivement les nombreux doctorants et stagiaires avec qui j'ai travaillé ces dix dernières années. Les encadrer a été un véritable plaisir et je leur dois pour beaucoup les résultats obtenus.

Coté laboratoire, là encore les personnes auxquelles je voudrais témoigner ma reconnaissance sont très nombreuses. Je pense que travailler au LITIS est une chance, pour l'ambiance et la qualité des travaux qui y sont menés. Parmi tous les collègues, une mention particulière va à Pierrot et Clem. Ce sont mes binômes de travail et des amis, et j'espère qu'on va avoir l'occasion de travailler encore beaucoup ensemble. Merci également à Thierry avec qui c'est un réel plaisir de travailler. Une spéciale dédicace aussi à super Fabienne dont l'efficacité est impressionnante.

Enfin, merci à tous ceux qui ont fait que les choses se passent bien, que ce soit au niveau du laboratoire, au niveau du département ou ailleurs.

Table des matières

Première partie

Curriculum Vitæ

Chapitre 1

Synthèse de mes activités

1.1 Curriculum Vitæ

Sébastien Adam
Né le 27 novembre 1975
Nationalité Française, célibataire, 2 enfants

1.1.1 Situation actuelle

Maître de Conférences (61ème section)
Laboratoire d'**I**nformatique, de **T**raitement de l'**I**nformation, et des **S**ystèmes
LITIS - EA 4108
UFR des Sciences et Techniques, Université de Rouen
BP 12 - 76801 Saint-Etienne du Rouvray, FRANCE
Tel : 02.32.95.52.10 - Fax : 02.32.95.50.22
Courriel : Sebastien.Adam@litislab.eu

1.1.2 Formation

2001 Doctorat de l'Université de Rouen
Sujet : *Interprétation de Documents Techniques : des Outils à leur Intégration dans un Système à Base de Connaissances*
Jury : N. Vincent (rapporteur), J.M. Chassery (rapporteur), K. Tombre (examinateur), J. Gardes (examinateur), J. Labiche (Directeur), J.M. Ogier (Co-Directeur)
Mention très honorable avec Félicitations du jury

1998 DEA Instrumentation et Commande pour les Systèmes de Vision
Université de Rouen, mention Bien

1998 DESS Automatique et Informatique Industrielle
Université de Rouen, mention Bien

1.1.3 Dates importantes

2008 Bénéficiaire de la Prime d'Encadrement Doctoral et de Recherche (PEDR)

2003 Titularisation dans le corps des Maîtres de Conférences

2002 Nomination Maître de Conférences section 61 à l'Université de Rouen

1.2 Résumé des activités

1.2.1 Résumé des activités de recherche

Thèmes de recherche :

- représentations structurelles et analyse de documents : classification de graphes, isomorphismes de sous-graphes, reconnaissance et localisation de symboles, analyse de documents graphiques ;
- optimisation et analyse de documents : optimisation multiobjectif, approximation de courbes, sélection de modèles, forêts aléatoires ;
- personnalisation de la recherche d'information : retour de pertinence implicite, apprentissage de profils utilisateurs, sélection dynamique d'outils d'aide à la recherche.

Publications :

Type de publication	**total**
Revues internationales	**8**
Chapitres de livre	**2**
Ouvrages collectifs	**6**
Revues francophones	**2**
Conférences internationales de rang A	**11**
Autres conférences internationales	**21**
Conférences francophones	**26**

Encadrement doctoral :

Thèses soutenues	**4**
Thèses en cours	**2**
Jurys de thèse	**3**
Master Recherche	**13**

Relations avec la communauté scientifique nationale et internationale :

- *reviewer* pour les revues internationales *Pattern Recognition*, *Pattern Recognition Letters*, *International Journal of Document Analysis and Recognition* et *Electronic Letters on Computer Vision and Image Analysis*.
- relecteur pour la revue nationale *Traitement du Signal* ;
- membre du comité d'organisation de la conférence CIFED'08 ;
- membre régulier des comités de programme et/ou comité de sélection de différentes conférences nationales et internationales (ICPR, ICDAR, GREC, RFIA, CIFED, JFPDA...) ;
- membre du GRCE, de l'AFRIF, du GDR I3 au niveau national, et des TC-15 et TC-10 de l'IAPR au niveau international.

Valorisation et contrats industriels :

- coordinateur LITIS du projet technovision EPEIRES de 2005 à 2007 (plate-forme d'évaluation d'approche de reconnaissance et localisation de symboles) (15 k€) ;
- correspondant scientifique du LITIS avec la société Algo-tech informatique dans le cadre d'un stage de Master Recherche en 2006 ;

- responsable scientifique et administratif d'une convention CORTECH avec la PME haut-normande ITS-IAE en 2008 (3 k€) ;
- co-responsable scientifique et administratif de trois conventions de recherche accompagnant les thèses CIFRE de G. Dupont, N. Martin et A. Saint Réquier avec la société CASSIDIAN [1] ;

1.2.2 Résumé des activités d'enseignement

Filières concernées : j'interviens au sein du département de physique de l'Université de Rouen, dans les filières EEA (Électronique, Électrotechnique et Automatique), GEII (Génie Électrique et Informatique Industrielle) et STIM (Système de Traitement des Informations Multimédia), de la première année de licence jusqu'à la seconde année de master.

Matières enseignées : j'assure des cours, TD et/ou TP en traitement numérique de l'information, programmation C, génie informatique, microprocesseurs, programmation système, systèmes d'exploitation, outils pour le traitement du signal, traitement numérique du signal, filtrage numérique, reconnaissance de formes, traitement d'images, optimisation.

1.2.3 Résumé des activités administratives

Responsabilités pédagogiques :

- Responsable pédagogique et président de jury de la troisème année de licence (ex IUP-2) Génie Électrique et Informatique Industrielle (GEII) de 2002 à 2005 (environ 45 étudiants).
- Responsable pédagogique et président de jury de la première année du master Informatique, Génie de l'Information et des Systèmes (IGIS), spécialité Génie Électrique et Informatique Industrielle (GEII) depuis 2006 (environ 25 étudiants).
- Responsable de la gestion des projets étudiants (Travaux d'Etude et de Recherche) des différentes années de EEA/GEII/STIM (L3, M1, M2) depuis 2007 (environ 75 étudiants).

Responsabilités électives :

- Membre nommé du conseil de département de physique de l'UFR depuis 2008.
- Membre élu de la commission de spécialistes de l'Université de Rouen (61ème section et 27/61ème section - vice-président) de 2004 à 2008.
- Membre nommé des commissions de spécialistes de l'INSA de Rouen (27-61-63ème sections) de 2006 à 2008.
- Membre nommé d'un comité de sélection 61ème section de l'Université de Rouen en 2009.

1. CASSIDIAN est le nouveau nom de EADS Defense and Security

1.3 Activités de recherche

1.3.1 Contexte des travaux

Les travaux présentés dans ce mémoire se sont successivement déroulés au sein du laboratoire Perception Systèmes et Information (PSI) de l'Université de Rouen, puis au Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes (LITIS) qui est né de la fusion du PSI avec les laboratoires d'informatique des Universités de Rouen et du Havre. Ce laboratoire est ainsi devenu l'unité de recherche haut-normande dans le domaine des Sciences et Technologies de l'Information et de la Communication (STIC). Il implique les trois principaux établissements d'enseignement supérieur de la région : l'Université de Rouen, l'Université du Havre et l'Institut National des Sciences Appliquées (INSA) de Rouen. Le laboratoire développe des démarches cohérentes pour mieux comprendre et maîtriser la nature de « l'information » et son utilisation contextuelle. Les recherches portent à la fois sur des aspects théoriques, algorithmiques et sur la mise en œuvre de systèmes sensibles au contexte, allant du capteur à la base de données.

Le LITIS structure ses recherches autour de trois axes regroupant sept équipes de recherche : l'axe « Combinatoire et algorithmes » qui aborde les aspects formels de l'information dans l'équipe du même nom ; l'axe « Traitement des masses de données » qui associe les quatre équipes « Document et apprentissage », « Traitement de l'information en biologie santé », « Quantif » et « Systèmes de transport intelligents » ; et enfin l'axe « Interaction et systèmes complexes » composé des deux équipes « Modélisation, interactions et usages » et « Réseaux d'interactions et intelligence collective ». La démarche du LITIS est résolument pluridisciplinaire, associant praticiens et théoriciens à la confluence de l'informatique, de la reconnaissance des formes, du traitement du signal et des images, de la médecine et des mathématiques, tous associés dans de nombreux projets.

Mes activités de recherche s'intègrent dans l'équipe "Document et Apprentissage" (DocApp), composée de 18 enseignants-chercheurs (10 MCF, 2 MCF HDR et 6 PU) et d'une vingtaine de doctorants. L'équipe est localisée sur le campus du Madrillet (INSA et Université de Rouen) et intégrée à l'axe "Traitement des Masses de Données". Les recherches menées dans DocApp concernent le développement d'outils et de méthodes génériques permettant d'interpréter des données variées de par leur structure, leur dimensionnalité, leur stationnarité et issues de contextes hétérogènes (signaux, images, textes, web). Ces travaux sont abordés essentiellement sous l'angle de l'apprentissage à partir d'exemples et de connaissances *a priori* dans le cadre structurant de la reconnaissance de formes. Les compétences développées dans l'équipe sont essentiellement de nature théoriques et algorithmiques, et concernent les machines à noyaux (SVM, Kernel PCA, apprentissage de noyaux multiples), les modèles markoviens (HMM multi-streams, champs aléatoires, CRF), l'analyse de graphes (mise en correspondance de graphes, recherche d'isomorphismes de sous-graphes, classification de graphes) et la sélection de modèles (analyse des risques d'estimateurs, apprentissage avec coûts inconnus ou évolutifs, réglage d'hyper-paramètres dans le cadre des méthodes d'ensemble). Les domaines

dans lesquels ces travaux trouvent leurs applications sont principalement le traitement automatique de l'écrit et des images de documents (reconnaissance de l'écriture manuscrite, spotting de mots et de symboles, extraction d'information, analyse de documents manuscrits complexes, bibliothèques numériques) ; mais aussi le traitement du signal (diagnostic, supervision, interface cerveau-machine), le traitement d'images médicales (classification d'images, segmentation) et la recherche d'information sur Internet.

1.3.2 Parcours de recherche

Soutenue en décembre 2001, ma thèse de doctorat [81] [2] traitait de la problématique de l'interprétation de documents graphiques, appliquée aux plans de réseau de l'opérateur téléphonique France Telecom. Deux contributions principales furent proposées dans cette thèse. La première concernait la réalisation d'un système d'interprétation de documents à base de connaissances. Elle s'est concrétisée par la mise en œuvre d'un système orienté multi-agents nommé NATALI [75, 76, 72, 71, 45, 43], dont l'architecture s'adaptait en fonction d'une description explicite du modèle de document à traiter. La seconde concernait la reconnaissance de caractères et de symboles multi-orientés et multi-échelles par l'utilisation de la transformée de Fourier-Mellin [10, 70, 73, 74, 18, 29, 50, 49, 46, 8, 7, 14].

Lors de mon recrutement en tant qu'ATER, puis en tant que Maître de Conférences au PSI en septembre 2002, mes travaux se sont dans un premier temps inscrits dans la poursuite de ces axes de recherche.

Le premier axe a été poursuivi dans le cadre du projet DOCMINING, supporté de 2002 à 2003 par le Réseau National des Technologies Logicielles (RNTL), au sein d'un consortium réunissant France Telecom R&D de Lannion, le LORIA de Nancy, le L3I de La Rochelle et l'équipe de Rolf Ingold à Fribourg. Ce projet visait la réalisation d'une plate-forme d'acquisition de documents hétérogènes, adaptant la chaîne de traitement déclenchée au contenu du document. Ce projet a conduit à la réalisation d'une plate-forme logicielle basée sur des logiciels libres et décrite dans [27, 67]. Sur ce thème de recherche, j'ai également été amené à travailler avec Eric Trupin et Jacques Labiche dans le cadre de la thèse de Youssouf Saidali sur des aspects liés à la représentation des connaissances [13].

La poursuite du second axe, orientée reconnaissance de formes, s'est traduite par la mise en place, en 2003, des thèses d'Eugen Barbu et d'Hervé Locteau, tous deux allocataires de recherche. La thèse d'Eugen Barbu [80], co-encadrée avec Pierre Héroux et Éric Trupin, concernait l'application de techniques de fouille de données et d'apprentissage au domaine de l'analyse de documents graphiques. En réexploitant certaines propositions effectuées dans ma thèse pour représenter les symboles par des modèles statistico-structurels, le principal objectif était de rendre générique et apprenant un système de reconnaissance de symboles en conférant au système des capacités à extraire de façon non supervisée le dictionnaire des symboles présents dans les documents. Ceci nous a ensuite amené, à des fins de catégorisation de documents,

2. Réalisée en convention CIFRE entre le laboratoire PSI et France Telecom Recherche et Développement Belfort puis Lannion

à poursuivre des travaux que j'avais initiés dans le cadre du stage de Master Recherche de Romain Raveaux concernant la classification de graphes par apprentissage de prototypes [1]. Ces travaux se poursuivent encore aujourd'hui, en collaboration avec Pierre Héroux.

La thèse d'Hervé Locteau [79], co-encadrée avec Jacques Labiche et Eric Trupin, abordait quant à elle la problématique de la localisation de symboles dans des documents graphiques. Un tel problème dépasse le cadre déjà complexe de la reconnaissance de symboles isolés en y ajoutant une problématique de segmentation. Les résultats obtenus dans la thèse ont montré la pertinence des choix de modélisation retenus, mais ont également mis en exergue la nécessité de développer des travaux sur la recherche d'isomorphismes inexacts de sous-graphes pour localiser des symboles. Ces conclusions ont donné lieu à des travaux fondamentaux dans ce domaine, initiés par le stage de Master Recherche de Pierre Le Bodic sur l'utilisation de la programmation linéaire en nombres entiers pour la recherche d'isomorphismes tolérants aux erreurs d'étiquetage. Par ailleurs, la thèse d'Hervé Locteau a également permis d'initier, en collaboration avec Yves Lecourtier, des premiers travaux reposant sur l'utilisation d'algorithmes d'optimisation multiobjectif pour l'analyse de documents, et plus particulièrement pour l'approximation de courbes [25].

Ces travaux concernant l'utilisation du formalisme de l'optimisation multiobjectif ont ensuite été poursuivis suivant deux axes. Un axe était orienté applicatif, dans le cadre d'une collaboration avec Clément Chatelain à la fin de sa thèse, et lors du stage de Master Recherche de Yannick Oufella. Les travaux menés dans ce cadre visaient la proposition d'un environnement d'apprentissage de classifieurs reposant non pas sur l'optimisation d'un critère unique, mais sur un formalisme multiobjectif prenant en considération les deux critères de l'espace ROC [2]. L'autre axe était plus fondamental, dans le cadre du stage de Master Recherche de Gérard Dupont. Il a consisté à proposer un nouvel algorithme d'optimisation multiobjectif reposant sur les essaims particulaires [4].

À la suite de ces travaux, avec Yves Lecourtier, nous avons été à l'origine de la mise en place d'une collaboration sur le long terme entre le LITIS et le département IPCC (Information Processing Competence Center) de CASSIDIAN. Cette collaboration s'est d'abord concrétisée par la thèse CIFRE de Gérard Dupont (co-encadrée avec Yves Lecourtier), soutenue en juillet 2011 [77], puis par celles de Nicolas Martin (co-encadrée avec Thierry Paquet) qui sera soutenue en 2012 et d'Aurélien Saint Réquier (co-encadrée avec Yves Lecourtier) dont la soutenance est prévue en 2013. Dans chacune de ces thèses, nous apportons nos compétences et proposons des contributions dans les domaines de l'apprentissage et de l'optimisation en relation avec les problématiques d'IPCC dans le domaine de la recherche d'information. Plus récemment, ces échanges avec CASSIDIAN se sont concrétisés par un projet de grande envergure concernant l'analyse de performances de chaînes d'analyse de document. Il a débuté en juin 2011 et le LITIS est chargé, sous l'impulsion de Thierry Paquet, Clément Chatelain et moi-même, d'en assurer l'expertise scientifique. Pour le LITIS, une équipe composée d'un doctorant, de deux post-doctorants et d'un ingénieur de

recherche, a été constituée pour contribuer à la réussite de ce projet [3].

En parallèle de ces travaux, je me suis également intéressé à des aspects plus fondamentaux de l'apprentissage, en abordant la problématique de construction d'ensembles de classifieurs dans le cadre de la thèse de Simon Bernard [78], que nous avons co-encadrée avec Laurent Heutte. Dans cette thèse, nous avons proposé plusieurs améliorations de l'algorithme d'induction de forêts aléatoires initialement conçu par Léo Breiman. En particulier, la thèse a permis la mise en œuvre d'un nouvel algorithme d'induction dynamique qui s'est révélé particulièrement compétitif par rapport aux approches de la littérature. Nous avons également dans ce cadre mis en place une collaboration avec Pierre Geurst, de l'Université de Liège, où Simon Bernard a débuté un post-doctorat à compter de semptembre 2011.

De cette présentation synthétique, il ressort que l'ensemble de mes travaux sont à l'intersection de deux domaines de recherche que sont les représentations structurelles et l'optimisation multiobjectif, avec deux applications principales liées à l'analyse de document et la recherche d'information. La sous-section suivante dresse un bilan des contributions que nous avons proposées dans ces domaines. Les principales seront développées dans la seconde partie de ce manuscrit.

1.3.3 Contributions

Cette sous-section dresse une synthèse des principales contributions de mes travaux de recherche. Elles sont réparties suivant trois axes. Le premier axe regroupe les travaux liés aux recherches menées sur les représentations à base de graphes, que ce soit au niveau fondamental ou au niveau applicatif. Le second axe concerne des travaux liés à l'optimisation multiobjectif et plus particulièrement à l'apport de ces approches au domaine de l'analyse de documents. Le dernier axe est lié aux travaux, plus récents, menés dans le domaine de la personnalisation de la recherche d'information.

1.3.3.1 Représentations à base de graphes

Représentations structurelles pour la localisation et l'indexation de symboles Ces travaux constituent une suite naturelle de mes travaux de thèse. Ils ont été menés dans le cadre des thèses d'Hervé Locteau [79] et d'Eugen Barbu [80]. Les propositions faites dans ces travaux partent du constat que si une extraction de caractéristiques à partir d'images de symboles associée à une classification statistique offrent généralement un bon pouvoir discriminant pour reconnaître des objets isolés, une telle stratégie nécessite d'avoir au préalable résolu le problème de la segmentation de l'objet à reconnaître.

Dans la thèse d'Hervé Locteau, nous avons abordé la problématique de la localisation de symboles dans des documents complets en nous appuyant sur une approche statistico-structurelle. Nous avons proposé deux chaînes de traitement complémentaires permettant d'extraire de manière robuste des graphes pour représenter des symboles. Avec une telle modélisation, la détection des

3. Ce projet ayant des aspects confidentiels, je ne peux pas le développer davantage

symboles devient alors un problème de recherche d'isomorphismes de sous-graphes, dont le but est de trouver les occurrences d'un graphe modèle, appartenant à un alphabet de symboles, dans un graphe cible représentant un document complet. Cette tâche de recherche d'appariement était effectuée par un algorithme de la littérature. Les modèles ont été évalués sur des bases de données de référence, issues des travaux du projet EPEIRES auxquels j'ai participé [5], et distribuées lors des conférences Graphic RECognition (GREC) pour des concours de reconnaissance. Les résultats obtenus ont montré la validité des approches proposées [40, 64, 34]. Ils ont également mis en exergue la nécessité de développer de nouveaux algorithmes d'isomorphismes de sous-graphes autorisant une modification des étiquettes des nœuds et des arcs, ce qui a donné lieu aux travaux menés avec Pierre Le Bodic décrits ci-après.

Les propositions de la thèse d'Eugen Barbu [80] s'appuient également sur une modélisation statistico-structurelle, mais dans le contexte de l'indexation et de la classification de documents. Dans ce cadre, nous avons proposé un algorithme permettant d'extraire sans connaissance *a priori* un dictionnaire des symboles présents dans une collection de documents, par l'intermédiaire d'algorithmes de recherche de sous-graphes fréquents issus de la communauté de la fouille de données [6, 17, 37, 36, 26, 38, 65, 66]. Les symboles détectés sont ensuite utilisés pour représenter les documents sous la forme de sacs de symboles, à des fins d'indexation ou de classification. Les résultats obtenus pour différents cas d'usage ont montré la pertinence d'une telle description et indiquent ainsi que les symboles découverts automatiquement fournissent des caractéristiques intermédiaires intéressantes pour catégoriser des documents.

Classification de graphes Ces travaux ont été initiés lors du stage de Master Recherche de Romain Raveaux, puis poursuivis dans le cadre de la thèse d'Eugen Barbu, et enfin, du stage de Master Recherche d'Arnaud Levallois. Le problème abordé dans le cadre de ces travaux était celui de la classification supervisée de graphes. Nous avons proposé pour cela d'utiliser la règle des k plus proches voisins, mais appliquée à un ensemble de graphes prototypes qu'un algorithme d'apprentissage permet de générer, ceci afin de réduire la complexité combinatoire de cette règle. Quatre types de prototypes de graphes ont été proposés et comparés dans ce cadre : les graphes médians d'ensemble, les graphes médians généralisés, les graphes discriminants d'ensemble et les graphes discriminants généralisés. Ces différents types de prototypes diffèrent en fonction de (i) l'espace dans lequel ils sont recherchés et (ii) la fonction objectif qui est utilisée pour les calculer. Dans chacun des cas, la génération des prototypes est réalisée grâce à un algorithme génétique dédié. Une étude expérimentale menée sur différentes bases de données de graphes a permis de comparer l'efficacité des différents types de prototypes. Il en est ressorti une supériorité des prototypes discriminants, qui permettent d'obtenir de très bonnes performances en classification [1, 24, 35, 61, 59]. Les résultats ont été validés sur une application de reconnaissance de symboles.

Isomorphismes de sous-graphes Ces travaux ont fait suite aux résultats obtenus dans le cadre de la thèse d'Hervé Locteau. Ils ont été initiés par le stage

de Master Recherche de Pierre Le Bodic, en collaboration avec Arnaud Knippel du Laboratoire de Mathématiques de l'INSA (LMI) de Rouen, puis poursuivis dans le cadre du stage de Master Recherche de Jean-Noel Bilong. Le problème abordé dans ces travaux était celui de la recherche d'isomorphismes de sous-graphes tolérants aux substitutions d'étiquettes. Un tel problème consiste à chercher dans un graphe cible les occurrences d'un graphe modèle, en tolérant que les étiquettes (qui peuvent être numériques, voire vectorielles) des deux graphes diffèrent, ce qui permet de tolérer le bruit généré par des extracteurs de caractéristiques. L'approche permet ainsi d'aborder des problèmes que les approches de la littérature ne peuvent pas naturellement traiter. L'approche proposée repose sur une formulation du problème sous la forme d'un programme linéaire en nombres entiers. En utilisant un solveur dédié à la résolution de telles formulations, le système proposé est capable d'extraire toutes les occurrences du graphe modèle dans le graphe cible avec une garantie d'optimalité quant au coût d'édition des étiquettes. La technique proposée a été évaluée sur des ensembles de graphes synthétiques, et sur une application de localisation de symboles utilisant les modèles proposés dans la thèse d'Hervé Locteau. Les résultats obtenus ont montré l'intérêt de considérer le problème abordé comme un problème d'optimisation [52, 56, 20].

1.3.3.2 Optimisation multiobjectif et reconnaissance de formes

Optimisation multiobjectif pour l'approximation de courbes planaires Ces travaux ont été initiés dans le cadre de la thèse d'Hervé Locteau [79]. Ils visaient à appliquer le paradigme de l'optimisation multiobjectif dans le cadre de l'approximation de courbes planaires par des segments et/ou des arcs de cercles. Il s'agit d'une étape importante pour la reconnaissance de formes et le traitement d'images visant à fournir une description compacte, par exemple pour caractériser les formes en vue de leur reconnaissance. Dans ce cadre, nous avons proposé d'aborder le problème sous l'angle original de l'optimisation multiobjectif. Ainsi, plutôt que fixer un nombre de points d'approximation et chercher à minimiser une mesure d'erreur, ou inversement se fixer une erreur maximale et chercher à minimiser le nombre de points, nous avons proposé un algorithme qui recherche en une seule exécution l'ensemble des solutions Pareto optimales au sens des deux critères. En proposant un ensemble de solutions potentielles, l'utilisateur, ou une étape ultérieure de traitement du document, peut alors sélectionner la solution la plus pertinente au regard du cas d'usage. Les résultats obtenus ont montré que l'approche proposée permettait d'obtenir en une seule exécution un ensemble de solutions comparables à celles obtenues par les approches de la littérature qui fixent le nombre de points d'approximation [63, 25, 39].

Selection de modèles et Front ROC Ces travaux ont été initiés dans le cadre de la thèse de Clément Chatelain, encadrée par Laurent Heutte et Thierry Paquet. À la suite des travaux réalisés avec Hervé Locteau, j'ai en effet été amené à collaborer avec Clément Chatelain lors de sa dernière année de thèse à un projet lié à l'optimisation multiobjectif. La problématique abordée dans ces travaux concernait l'apprentissage de classifieurs dans des environne-

ments mal définis, pour lesquels les effectifs des classes sont déséquilibrés et les coûts de mauvaise classification sont inconnus. Il s'agit d'un contexte très fréquent dans les applications du monde réel, typiquement dans le domaine de la médecine pour lequel les exemples d'apprentissage de cas pathologiques sont rares, mais particulièrement critiques. Dans ce contexte, il est bien connu qu'un unique critère d'apprentissage ne permet pas de construire un classifieur adapté à toutes les situations. Nous avons dans ce cadre proposé un environnement d'apprentissage reposant sur l'optimisation de critères multiples. L'approche proposée permet ainsi d'entrainer un ensemble de classifieurs plutôt qu'un unique, chaque classifieur de l'ensemble optimisant un compromis particulier entre les objectifs de l'espace ROC. Nous avons dans ce travail introduit la notion de Front-ROC comme alternative à la courbe ROC, en y intégrant la notion d'optimalité. La stratégie générique proposée, qui peut s'appliquer à tout type de classifieur hyperparamétrique, a été dans ces travaux testée pour la sélection de modèles multiples de classifieurs SVM en utilisant un algorithme évolutionnaire. L'approche a été validée sur des bases de l'UCI et sur un problème applicatif de reconnaissance de l'écriture manuscrite. Les résultats obtenus ont été comparés favorablement à ceux qu'une approche basée sur l'optimisation de l'aire sous la courbe ROC permet d'obtenir [22, 57, 2, 3].

Sélection de modèle et induction dynamique de forêts aléatoires Ces travaux ont été initiés dans le cadre de la thèse de Simon Bernard, avec également des contributions apportées par les stages de Master Recherche de Émilie Oliveira, Yasser Alwan et Nhat Quang Doan. La problématique abordée dans ces travaux est celle de l'amélioration des algorithmes de forêts aléatoires, qui sont des ensembles de classifieurs à base d'arbres de décision dans lesquels est injectée une part d'aléatoire. Nous nous sommes d'abord intéressés dans cette thèse à la problèmatique de la sélection de modèles pour ces algorithmes, en analysant l'influence des deux hyperparamètres essentiels dans l'induction de forêts : le nombre de caractéristiques choisies aléatoirement à chaque nœud et le nombre d'arbres induits. Nous avons dans ce cadre montré que la valeur du premier hyperparamètre doit être choisie en fonction des propriétés de l'espace de description. Nous avons donc proposé un nouvel algorithme nommé Forest-RK qui adapte sa valeur en fonction du problème traité [54, 55, 31, 30, 32]. La seconde contribution de cette thèse a été de proposer un algorithme d'induction dynamique de forêts aléatoires, qui tient compte lors de l'induction de nouveaux arbres de la forêt préalablement construite [58, 19]. L'algorithme proposé s'est montré particulièrement performant en comparaison avec les procédures d'induction statique.

Optimisation multiobjectif par essaims particulaires Ces travaux ont été initiés lors du stage de Master Recherche de Gérard Dupont et poursuivis ensuite en filigrane pendant sa thèse. Le problème abordé dans ces travaux consistait à exploiter le formalisme des essaims particulaires dans le cadre de l'optimisation multiobjectif. Pour ce faire, nous avons proposé deux contributions liées à la transformation de l'algorithme des essaims particulaires proposé par Kennedy, Eberhart et Shi pour que celui-ci puisse appréhender des pro-

blèmes à objectifs multiples. La première contribution est relative à la gestion de l'archive contenant les solutions optimales courantes. Elle repose sur l'utilisation d'une variante de la méthode de l'ϵ-dominance. La seconde concerne le problème de la sélection de la particule "guide" qui doit être totalement revue dans un cadre multiobjectif. Ces contributions ont été validées sur des problèmes standard d'optimisation multiobjectif et sur le problème de sélection de modèles SVM évoqué ci-avant. Dans les deux cas, nous avons montré que l'algorithme proposé permettait d'obtenir des résultats comparables à ceux fournis par NSGA-II qui est, aujourd'hui, l'une des références dans le domaine de l'optimisation multiobjectif [60, 4].

1.3.3.3 Personnalisation de la recherche d'information

Ces travaux, en marge des précédents, ont été initiés dans le cadre de la thèse de Gérard Dupont [77], en collaboration avec CASSIDIAN. Ils ont constitué nos premières contributions à l'intersection des domaines de l'apprentissage, de l'optimisation et de la recherche d'information interactive. L'objectif de ces travaux était de créer le lien entre ces domaines par la mise en œuvre de principes d'apprentissage dans le but d'adapter les réponses d'un système de recherche d'information aux utilisateurs de celui-ci. Nous avons, dans ce cadre, proposé deux principales contributions. La première concerne la proposition d'un modèle de l'utilisateur prenant en compte ses interactions implicites de recherche avec le système (clic, navigation, impression, signets...). En exploitant ce modèle, nous avons proposé une approche d'apprentissage du besoin utilisateur, exploitée dans le cadre du retour de pertinence. Cette proposition a été opérationnalisée dans un outil de suggestion de requêtes qui a été évalué et comparé aux approches de la littérature dans une première série d'expérimentations interactives de recherche. Les résultats obtenus ont mis en exergue la variabilité importante des performances de différentes approches en cours de session et en fonction des utilisateurs.

Notre seconde contribution a donc consisté à introduire un cadre d'intégration dynamique optimisant le déclenchement d'outils d'aide à la recherche (suggestion de requête, de documents, filtrage...) au cours de sessions de recherche. Un algorithme d'apprentissage par renforcement permet d'apprendre à sélectionner la bonne approche au bon moment. Implantée dans un système complet, cette proposition a pu être validée par des expérimentations interactives pour la sélection d'outils de suggestion de requêtes [51, 53]. Ces travaux sont actuellement poursuivis par ceux de la thèse CIFRE d'Aurélien Saint Réquier, avec CASSIDIAN, dont le but est de proposer un agent personnel d'assistance à la recherche d'information.

1.3.4 Perspectives

Les travaux mentionnés dans la sous-section précédente offrent tous des perspectives intéressantes qui sont pour certaines en cours d'investigation. La plupart de ces perspectives seront évoquées dans la seconde partie de ce mémoire. Dans cette sous section, j'ai choisi de décrire les trois pistes que je considère comme prioritaires au regard des résultats prometteurs qu'elles offrent, et

de l'importance qu'elles revêtent, selon moi, pour la communauté scientifique concernée.

Sélection de modèles et apprentissage multiobjectif Ces perspectives de recherche font suite aux travaux menés en collaboration avec Clément Chatelain concernant le développement d'un cadre multi-critères pour l'apprentissage automatique. Elles ont fait l'objet d'une soumission nommée LeMOn (LEarning with Multi-objective OptimizatioN) lors de l'appel ANR Jeunes Chercheurs et Jeunes Chercheuses 2011 [4]. Dans le cadre de cette soumission, nous avons identifié deux aspects particuliers de l'apprentissage que nous souhaiterions aborder sous l'angle de l'optimisation multiobjectif et qui sont, naturellement, des perspectives pour mes recherches à venir.

Le premier aspect concerne l'exploitation de l'espace ROC lors de l'apprentissage des classifieurs. Dans [2], nous avons proposé un environnement de sélection de modèles basé sur une approche d'optimisation multiobjectif. Cet environnement permet de construire un ensemble de classifieurs à deux classes localement optimaux dans l'espace ROC, plutôt qu'un unique basé sur un critère scalaire. Les perspectives ouvertes par ce travail concernent deux axes. Le premier est le passage à l'échelle afin d'appréhender de très grands volumes de données, par l'intermédiaire d'un apprentissage en ligne. Le second axe est la généralisation de l'approche proposée à des problèmes multi-classes, pour lesquels le nombre de critères croît rapidement avec le nombre de classes.

Le second aspect de l'apprentissage que nous envisageons d'aborder sous l'angle de l'optimisation multiobjectif est celui de l'apprentissage multi-tâches, qui consiste à apprendre simultanément plusieurs modèles par des transferts de connaissances d'un modèle vers l'autre. Là encore, nous pensons que l'angle de l'optimisation multiobjectif pourrait apporter des pistes intéressantes. Dans le projet LeMOn, il est prévu d'appliquer ces différents travaux à deux domaines d'application : l'analyse d'images médicales, en collaboration avec l'équipe Quantif du LITIS ; et les interfaces cerveau-machine, en collaboration avec des chercheurs de l'équipe DocApp s'intéressant à cette problématique.

Isomorphismes de sous-graphes Ces perspectives de recherche font suite aux travaux menés avec Pierre Le Bodic concernant la recherche d'isomorphismes de sous-graphes, et à ceux concernant le cadre applicatif de la localisation de symboles menés dans le cadre des thèses d'Hervé Locteau et Eugen Barbu. Ces perspectives se déclinent suivant trois axes.

Le premier axe est lié à l'application de localisation de symboles. Dans [1], nous avons identifié des verrous relatifs aux modèles utilisés pour la détection de symboles. L'un d'eux est lié au modèle à base de régions actuellement exploité qui ne permet pas de distinguer certaines classes de symboles. L'une des perspectives pour dépasser ces limites consiste à enrichir le modèle orienté région par une description des symboles à partir de leur contour.

Le second axe est quant à lui orienté vers l'utilisation de la programmation linéaire en nombres entiers. Les résultats présentés dans [20] ont en effet

4. Le projet, dont je suis le porteur, est actuellement sur liste complémentaire

montré que les performances de l'approche proposée pour la recherche d'isomorphismes exacts étaient encore inférieures à l'état de l'art en termes de temps de traitement. Cette lacune pourrait être palliée à la fois en optimisant la formulation, mais aussi en tirant davantage parti des constantes améliorations des algorithmes de résolution proposés par les solveurs. Par ailleurs, nous travaillons également à la proposition d'une nouvelle formulation qui tolèrerait des modifications topologiques des graphes.

Enfin, le dernier axe de recherche que je souhaite aborder dans ce cadre, étroitement lié aux deux précédents, est celui de l'évaluation des performances d'algorithmes de recherche d'isomorphismes de sous-graphes par la proposition d'une base de graphes réels, étiquetés au niveau « application » pour comparer les algorithmes de recherche d'isomorphismes inexacts.

Personnalisation en recherche d'information Ces perspectives de recherche entrent dans le cadre de la collaboration avec la société CASSIDIAN sur les problématiques de recherche d'information, et plus particulièrement sur celles de la personnalisation des outils de recherche pour placer l'utilisateur au cœur du processus de recherche. Elles font suite aux travaux menés dans le cadre de la thèse de Gérard Dupont et à ceux en cours dans le cadre de la thèse d'Aurélien Saint Réquier. Elles concernent deux aspects principaux.

Le premier est lié à la modélisation de l'utilisateur et à l'élicitation de ses besoins d'information. Dans la thèse de Gérard Dupont, le modèle de besoin était construit à partir des interactions de l'utilisateur avec le système au cours d'une session de recherche. Si une telle analyse permet de dépasser le cadre classique de l'analyse orientée requêtes, l'intégration d'un modèle à plus long terme (issu par exemple de documents fournis par l'utilisateur ou de ses signets) et sa combinaison avec le modèle court terme proposé dans la thèse de Gérard Dupont offrent des perspectives indéniables d'amélioration. Cette perspective est en cours d'investigation dans le cadre de la thèse d'Aurélien Saint-Réquier. Par ailleurs, une autre perspective d'amélioration de cette modélisation repose sur le passage d'une représentation orientée « mots » à une représentation orientée « concepts » qui permettrait d'aller vers un moteur de recherche d'information sémantique.

Le second aspect concerne le cadre d'intégration dynamique proposé dans la thèse de Gérard Dupont. Là aussi, de nombreuses perspectives sont envisageables. À court terme, nous envisageons d'enrichir la plage d'actions à disposition de l'algorithme d'apprentissage par renforcement, pour multiplier les possibilités d'adaptation du système global. Au-delà des actions, la détermination des états peut également être améliorée. Actuellement, les états sont issus d'une segmentation effectuée par un algorithme de partitionnement pour lequel il est nécessaire de fixer le nombre d'états. Plusieurs approches alternatives pourraient être testées, comme celle par exemple consistant à s'appuyer sur une classification supervisée reposant sur une définition manuelle de micro-tâches de comportements issue de travaux en analyse du comportement. Il serait alors nécessaire d'adapter les algorithmes d'apprentissage des MDP (Markov Decision Process) pour y intégrer une notion d'incertitude (via les Partially Observable Markov Decision Process) et/ou une notion de hiérarchie

(*via* les Hierarchical Markov Decision Process). Par ailleurs, en lien avec les travaux mentionnés ci-dessus, des études complémentaires pourraient être menées quant à la mise en compétition de différents modèles d'apprentissage, passant ainsi d'un MDP mono-objectif à un MDP multiobjectif qui aurait pour finalité de maximiser un vecteur de récompense au lieu d'une récompense scalaire classique.

1.3.5 Encadrement doctoral

1.3.5.1 Encadrement de thèses soutenues

- Co-encadrement scientifique (25% avec P. Héroux et E. Trupin) de la thèse d'Eugen Barbu (Bourse MESR, 2003-2006)
  - Soutenue le 14/06/2006
  - Titre : *Fouille et classification de graphes : application à la reconnaissance de symboles dans les documents graphiques*
  - Jury : R. Ingold (rapporteur), R. Mullot (rapporteur), J. Llados, J.Y. Ramel, P. Héroux, E. Trupin
  - Publications associées :[23, 33, 24, 26, 38, 37, 62, 61, 65, 66, 6, 17]
- Co-encadrement scientifique (50% avec J. Labiche et E. Trupin) de la thèse d'Hervé Locteau (Bourse MESR, 2003-2008)
  - Soutenance le 27/10/2008
  - Titre : *Contributions à la localisation de symboles dans les documents graphiques*
  - Jury : J.Y. Ramel (rapporteur), J.M. Ogier (rapporteur), A. Tabbone, J. Labiche, E. Trupin, S. Adam
  - Publications associées : [33, 25, 24, 39, 40, 34, 61, 64, 63, 68]
- Co-encadrement scientifique (50% avec L. Heutte) de la thèse de Simon Bernard (Bourse MESR, 2006-2009)
  - Soutenue le 02/12/2009
  - Titre : *Forêts Aléatoires : De l'analyse des mécanismes de fonctionnement à la construction dynamique*
  - Jury : Y. Grandvalet (rapporteur), T. Artière (rapporteur), L. Wehenkel, M. Sebban, L. Heutte, S. Adam
  - Publications associées : [54, 58, 55, 19, 31, 21, 31]
- Co-encadrement scientifique (50% avec Y. Lecourtier) de la thèse de Gérard Dupont (Bourse CIFRE, 2006-2011)
  - Soutenue le 04/07/2011
  - Titre : *Apprentissage implicite pour la recherche d'information*
  - Jury : T. Artières (rapporteur), M. Boughanem (rapporteur), N. Vincent, S. Brunessaux, Y. Lecourtier, S. Adam
  - Publications associées : [60, 4, 53]

1.3.5.2 Encadrement de thèses en cours

- Co-encadrement (50% avec T. Paquet) de la thèse de Nicolas Martin (Bourse CIFRE EADS, 2009-2012)

  - Soutenance prévue en 2012
  - Titre : *Recherche et collecte d'informations sur les individus en sources ouvertes*
- Co-encadrement (50% avec Y. Lecourtier) de la thèse de Aurélien Saint Réquier (Bourse CIFRE EADS, 2010-2013).
  - Soutenance prévue en 2013
  - Titre : *Agent Personnel d'Aide à la Recherche d'Information*
  - Publication associée : [53]

1.3.5.3 Encadrement de stages de DEA et de Master Recherche

- Co-encadrement (50% avec Y. Lecourtier) du Master Recherche de S. Cognard. *Co-évolution et reconnaissance de formes*. 2005.
- Co-Encadrement (50% avec P. Héroux) du Master Recherche de R. Raveaux. *Reconnaissance de symboles à partir de schémas électriques*. 2006.
- Co-encadrement (50% avec Y. Lecourtier) du Master Recherche de G. Dupont. *Annotation sémantique et apprentissage implicite : vers une recherche d'information intelligente*. 2006.
- Co-encadrement (50% avec L. Heutte) du Master Recherche de E. Oliveira. *Construction dynamique de forêts aléatoires*. 2008.
- Co-encadrement (50% avec Y. Lecourtier) du Master Recherche de P. Le Bodic. *Isomorphisme inexact de sous-graphes*. 2008.
- Co-encadrement (50% avec L. Heutte) du Master Recherche de Y. Oufella. *Optimisation multiobjectif et apprentissage*. 2008.
- Co-encadrement (50% avec C. Lecomte) du Master Recherche de Nicolas Martin. *Extraction et recherche de concepts dans des images*. 2008.
- Co-encadrement (50% avec L. Heutte) du Master Recherche de Y. Alwan. *Classification One-Class avec les Forêts Aléatoires*. 2008.
- Co-encadrement (50% avec P. Héroux) du Master Recherche de A. Levallois. *Classification de graphes par algorithmes génétiques*. 2009.
- Co-encadrement (50% avec P. Héroux) du Master Recherche de J.N. Bilong. *Recherche d'isomorphismes exacts de sous-graphes par Programmation Linéaire en Nombre Entier (PLNE)*. 2009.
- Co-encadrement (50% avec Y. Lecourtier) du Master Recherche de A. Saint-Réquier. *Expérimentations utilisateur : étude comparative des performances d'un système de recherche d'information apprenant*. 2009.
- Co-encadrement (50% avec L. Heutte) du Master Recherche de N-Q. Doan. *One-Class random forests*. 2010.
- Co-encadrement (50% avec T. Paquet) du Master Recherche de F. Dewevre. *Recherche d'images par analyse du contenu*. 2011.

1.3.6 Activités contractuelles, projets ANR

Cette section précise le cadre contractuel dans lequel se sont développées certaines des actions de recherche présentées précédemment.

Responsable LITIS du projet Technovision EPEIRES Pendant les années 2005 et 2006, j'ai eu en charge la gestion et la responsabilité côté LITIS

du projet EPEIRES (Évaluation des PErformances de l'Interprétation et de la REconnaissance de Symboles) [5], retenu dans le cadre de l'appel à projet Technovision lancé conjointement par le Ministère de l'Enseignement Supérieur et de la Recheche et par la Direction Générale de l'Armement. Ce projet, d'une durée de deux ans, regroupait des membres d'Algo'Tech Informatique, de la City University of Hong Kong, du Laboratoire d'informatique de l'Université de Tours, de l'équipe QGAR du LORIA, du Laboratoire ONE de France Télécom R&D, du laboratoire PSI (devenu LITIS) de l'Université de Rouen et de l'équipe DAG du Computer Vision Center de l'Université Autonome de Barcelone. Le projet avait pour objectif la construction d'un environnement complet fournissant les outils et les ressources nécessaires à l'évaluation des performances de méthodes de localisation et de reconnaissance de symboles. Plus particulièrement, les membres de ce projet souhaitaient estimer de manière générique leurs capacités à reconnaître et localiser les symboles en fonction d'un certain nombre de critères : le domaine d'application, la modélisation, le nombre de symboles impliqués, la qualité du document... Le projet était centré sur deux points importants à évaluer : la reconnaissance et la localisation. L'environnement développé dans le cadre de ce projet était par ailleurs destiné à être utilisé par la communauté la plus large qui soit. Plusieurs campagnes de tests, ouvertes à tous les participants inscrits, ont été organisées après ce projet lors des conférences Graphic RECognition (GREC). Le site du projet est encore disponible aujourd'hui pour toute la communauté.

Responsable de contrats de recherche avec EADS Dans le cadre de mes activités de recherche liées à la recherche d'information, j'ai initié et développé, en collaboration avec Yves Lecourtier et Thierry Paquet, plusieurs opérations de recherche avec l'équipe IPCC de EADS (devenu CASSIDIAN depuis) dirigée par Stephan Brunessaux. Ces activités de recherche se traduisent par les activités contractuelles suivantes.

- Responsable scientifique et administratif de la convention "Apprentissage implicite pour la recherche d'information" de Novembre 2006 à Juillet 2011 (montant 30 k€). Ce contrat, initié dans le cadre de la thèse en convention CIFRE de Gérard Dupont, avait pour objet de concevoir un moteur de recherche d'information apprenant qui, en fonction des interactions avec l'utilisateur, l'assiste dans ses recherches.
- Responsable scientifique et administratif de la convention "Collecte intelligente des ressources du Web : application à la création de profils d'individus" de Mars 2009 à Mars 2012 (montant 30 k€). Ce contrat, initié dans le cadre de la thèse en convention CIFRE de Nicolas Martin, a pour objet de concevoir un système capable de créer des profils d'individus en collectant de manière ciblée des informations à partir de sources ouvertes.
- Responsable scientifique et administratif de la convention "Agent personnalisé de recherche d'information" de février 2010 à février 2013 (montant 30 k€). Ce contrat, initié dans le cadre de la thèse en convention CIFRE d'Aurélien Saint Réquier, a pour objet de concevoir un agent intelligent

5. `http://www.epeires.org/`

personnalisé de recherche d'information basé sur un système d'apprentissage sémantique du contexte des tâches de recherche et des centres d'intérêt de l'utilisateur.

Ces trois contrats sont la concrétisation d'une collaboration engagée sur le long terme avec l'équipe IPCC de CASSIDIAN. Celle-ci a débouché en juin 2011 sur la signature d'un nouveau contrat (montant 450 k€) ayant pour objet l'étude, le développement et la réalisation d'un démonstrateur de reconnaissance automatique de documents. Dans ce projet, le LITIS est le référent scientifique. Nos missions consistent, outre le développement de modules d'analyse d'images de documents, à assister CASSIDIAN en tant que référence scientifique. Thierry Paquet assure la responsabilité technique du projet et je suis, pour CASSIDIAN, le responsable recherche de ce projet.

Participation à des programmes nationaux À la suite de ma thèse, j'ai été impliqué dans le projet DOCMINING, qui est un projet exploratoire supporté par le Réseau National des Technologies Logicielles (RNTL). Ce projet a réuni, de janvier 2002 à décembre 2003, un consortium composé de France Telecom R&D Lannion, l'équipe QGAR de l'INRIA Lorraine de Nancy, le laboratoire L3i de l'Université de La Rochelle, le département d'informatique de l'Université de Fribourg et l'équipe Document du PSI. Ce projet visait la conception d'un système à base de connaissances et le développement d'un démonstrateur d'acquisition de documents hétérogènes représentant des plans d'accès à des bâtiments. Le système proposé avait pour objectif d'identifier les composantes contenues dans un document et d'adapter leurs modes de représentation aux besoins d'un service donné. Ce système couvre donc un large spectre d'utilisation. Il ne s'agit pas seulement de procéder à une rétroconversion systématique de documents entiers, mais de mettre en place une méthodologie de valorisation des objets contenus dans un document.

J'ai ensuite été impliqué dans une Action Concertée Incitative "MAsse de DOnnées issues de la Numérisation du patrimoiNE" (ACI MADONNE), fruit d'une collaboration entre les laboratoires PSI (Rouen), L3I (La Rochelle), LIRIS (Lyon), LORIA (Nancy), IRISA (Rennes) et LI (Tours). L'objectif des travaux de cette ACI était de permettre, à partir de l'extraction automatique d'indices dans les images, la navigation et la recherche d'informations dans les collections de documents patrimoniaux. Ces travaux se sont poursuivis dans le cadre du projet ANR Navidomass (NAVIgation into DOcument MASSes). Ce projet, labellisé par l'ANR de 2008 à 2011, a pour mission de mettre en valeur différents biens du patrimoine et plus particulièrement les ouvrages, les collections d'images et autres documents iconographiques. À court terme, ces nombreux documents constitueront une source gigantesque d'informations (masse de données). L'objectif de ce projet est de contribuer à la réalisation de systèmes d'indexation d'images de documents du patrimoine. Ce projet s'inscrit ainsi dans la volonté actuelle de préserver le patrimoine culturel et scientifique et d'assurer au plus grand nombre l'accès à celui-ci.

1.3.7 Relations avec la communauté scientifique nationale et internationale

Relecture d'articles pour revues et conférences J'expertise des articles soumis dans les revues internationales de référence *Pattern Recognition* (PR), *Pattern Recognition Letters* (PRL), *International Journal of Document Analysis and Recognition* (IJDAR) ainsi que dans la revue nationale *Traitement du Signal*.

Membre de comités de programmes et d'organisation J'ai été membre des comités de programme des conférences internationales *International Conference on Pattern Recognition* (ICPR 2008 à Tampa, et ICPR 2010 à Istambul) et *Graphic Recognition* (GREC 2007 à Curritiba, GREC 2009 à La Rochelle et GREC 2011 à Séoul). Au niveau national, j'ai participé à des comités de programme de la Conférence Internationale Francophone sur l'Ecrit et le Document (CIFED 2004 à La Rochelle, CIFED 2006 à Fribourg, CIFED 2008 à Rouen, CIFED 2010 à Sousse), à la conférence sur la Reconnaissance de Formes et l'Intelligence Artificielle (RFIA 2010 à Caen) et aux Journées Francophones de Planification, Décision et Apprentissage pour la conduite de systèmes (JFPDA 2011 à Rouen). J'ai également été sollicité pour présider la session "Graphes" de la conférence CIFED 2010 à Sousse. J'ai finalement fait partie du comité d'organisation de la conférence CIFED 2008 à Rouen.

Participation à des groupes de travail Je participe à différents groupes de travail au sein de la communauté scientifique française. Je suis membre d'associations rassemblant des chercheurs francophones de mon domaine de recherche telles que le *Groupe de Recherche en Communication Ecrite* (GRCE), l'*Association Française pour la Reconnaissance et l'Interprétation des Formes* (AFRIF).
Je participe également régulièrement au groupe de travail GT5.2 *Ecrit* du GDR I3 *Information-Interaction-Intelligence*. J'ai participé à l'Action Spécifique *Valorisation des Collections* dans le cadre du *Réseau Thématique Pluridisciplinaire Documents* (RTP-Doc) du CNRS.
En ce qui concerne mon implication dans la communauté internationale, je suis membre affilié des comités techniques TC15 (Graph-based Representations in the Pattern Recognition field) et TC10 (Graphic Recognition) de l'IAPR (International Association of Pattern Recognition).

1.3.8 Publications

Revues internationales avec comité de lecture

[1] R. Raveaux, S. Adam, P. Héroux, and E. Trupin. Learning graph prototypes for shape recognition. *Computer Vision and Image Understanding (CVIU)*, 115(7) :pages 905 – 918, 2011.

[2] C. Chatelain, S. Adam, Y. Lecourtier, L. Heutte, and T. Paquet. A multi-model selection framework for unknown and/or evolutive misclassification cost problems. *Pattern Recognition (PR)*, 43(3) :pages 815–823, 2010.

[3] C. Chatelain, S. Adam, Y. Lecourtier, L. Heutte, and T. Paquet. Non-cost sensitive SVM training using multiple model selection. *Journal of Circuits Systems, and Computers (JCSC)*, 19(1) :pages 231–242, 2010.

[4] G. Dupont, S. Adam, Y. Lecourtier, and B. Grilhère. Multi objective particle swarm optimization using enhanced dominance and guide selection. *International Journal of Computational Intelligence Research (IJCIR)*, 4(2) :pages 145–158, 2008.

[5] E. Valveny, P. Dosch, A. Winstanley, Y. Zhou, S. Yang, L. Yan, W. Liu, D. Elliman, M. Delalandre, E. Trupin, S. Adam, and J. Ogier. A general framework for the evaluation of symbol recognition methods. *International Journal of Document Analysis and Recognition (IJDAR)*, 9(1) :pages 59–74, 2007.

[6] E. Barbu, P. Héroux, S. Adam, and E. Trupin. Frequent graph discovery : Application to line drawing document images. *Electronic Letters on Computer Vision and Image Analysis (ELCVIA)*, 5(2) :pages 47–57, 2005.

[7] S. Adam, J. Ogier, C. Cariou, R. Mullot, J. Labiche, and J. Gardes. Symbol and character recognition : application to engineering drawings. *International Journal of Document Analysis and Recognition (IJDAR)*, 3(2) :pages 89–101, 2000.

[8] C. Cariou, J.-M. Ogier, S. Adam, R. Mullot, Y. Lecourtier, and J. Gardes. A multiscale and multiorientation recognition technique applied to document interpretation : Application to the French telephone network maps. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 13(8) :pages 1201–1218, 1999.

Chapitres de livres

[9] S. Adam and J. Ogier. Documents graphiques : de la rétroconversion à la recherche d'information. In R. Mullot, editor, *Les documents écrits : De la numérisation à l'indexation par le contenu*, pages 249–310. Hermès, 2006.

[10] S. Adam, J. Ogier, C. Cariou, R. Mullot, J. Gardes, and Y. Lecourtier. Fourier-mellin based invariants for the recognition of multi-oriented and multi-scaled shapes : Application to engineering drawings analysis, in invariants for pattern recognition and classification. In M. Rodrigues, editor, *Invariants for pattern recognition and classification*, pages 132–147. World Scientific, Singapore, 2000.

Contributions à des ouvrages collectifs

Les références mentionnées dans cette partie correspondent à des versions étendues de soumissions faites pour des conférences internationales et soumises à un second processus de relecture.

[11] E. Barbu, P. Héroux, S. Adam, and E. Trupin. Using bags of symbols for automatic indexing of graphical document image databases. In W. Liu and J. Lladós, editors, *Graphics Recognition. Ten Years Review and Future Perspectives*, volume 3926 of *Lecture Notes in Computer Science*, pages 195–205. Springer, 2006.

[12] H. Locteau, R. Raveaux, S. Adam, Y. Lecourtier, P. Héroux, and E. Trupin. Polygonal approximation of digital curves using a multi-objective genetic algorithm. In W. Liu and J. Lladós, editors, *Graphics Recognition. Ten Years Review and Future Perspectives*, volume 3926 of *Lecture Notes in Computer Science*, pages 300–311. Springer, 2006.

[13] Y. Saidali, S. Adam, J. Ogier, and E. Trupin. Knowledge representation and acquisition for engineering document analysis. In W. Liu and J. Lladós, editors, *Graphics Recognition : Recent Advances and Perspectives*, volume 3088 of *Lecture Notes in Computer Science*, pages 25–37. Springer, 2004.

[14] S. Adam, J. Ogier, C. Cariou, and J. Gardes. A scale and rotation parameters estimator application to technical document interpretation. In *GREC '01 : Selected Papers from the Fourth International Workshop on Graphics Recognition Algorithms and Applications*, volume 2390, pages 266–272. Springer-Verlag, London, UK, 2002. ISBN 3-540-44066-6.

[15] S. Adam, R. Mullot, J. Ogier, C. Cariou, J. Gardes, and Y. Lecourtier. Processing of the connected shapes in raster-to-vector conversion process. In *Selected Papers from the Third International Workshop on Graphics Recognition, Recent Advances*, pages 28–38. Springer-Verlag, London, UK, 2000.

[16] S. Adam, J. Ogier, C. Cariou, J. Gardes, R. Mullot, and Y. Lecourtier. Combination of invariant pattern recognition primitives on technical documents. In *Selected Papers from the Third International Workshop on Graphics Recognition, Recent Advances*, pages 238–245. Springer-Verlag, London, UK, 2000. ISBN 3-540-41222-0.

Revues nationales avec comité de lecture

[17] E. Barbu, P. Héroux, S. Adam, and E. Trupin. Fouille de graphes et découverte de règles d'association : application à l'analyse d'images de document. *Revue Nouvelles Technologies de l'Information (RNTI)*, E-3 :pages 463–468, 2005.

[18] S. Adam, J. Ogier, C. Cariou, R. Mullot, J. Gardes, and Y. Lecourtier. Utilisation de la transformée de Fourier-Mellin pour la reconnaissance de formes multi-orientées et multi-échelles : application à l'analyse automatique de documents techniques. *Revue Traitement du Signal (TS)*, 18(1) :pages 17–33, 2005.

Conférences internationales de rang A

Les références mentionnées dans cette partie correspondent à des communications dans des conférences majeures, considérées comme sélectives par la communauté (référencées A ou A+ par le site CORE `http://www.core.edu.au` par exemple).

[19] S. Bernard, L. Heutte, and S. Adam. On the selection of decision trees in random forests. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'09)*, pages 302–307. 2009.

[20] P. Le Bodic, H. Locteau, S. Adam, P. Héroux, Y. Lecourtier, and A. Knippel. Symbol detection using region adjacency graphs and integer linear programming. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'09)*, pages 1320–1324. 2009.

[21] S. Bernard, S. Adam, and L. Heutte. Using random forests for handwritten digit recognition. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'07)*, pages 1043–1047. 2007.

[22] C. Chatelain, S. Adam, Y. Lecourtier, L. Heutte, and T. Paquet. Multi-objective optimization for SVM model selection. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'07)*, pages 427–431. 2007.

[23] P. Héroux, E. Barbu, S. Adam, and E. Trupin. Automatic ground-truth generation for document image analysis and understanding,. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'07)*, pages 476–480. 2007.

[24] E. Barbu, R. Raveaux, H. Locteau, S. Adam, P. Héroux, and E. Trupin. Graph classification using genetic algorithm and graph probing : Application to symbol recognition. In *Proceedings of the International Conference on Pattern Recognition (ICPR'06)*, pages 296–299. 2006.

[25] H. Locteau, R. Raveaux, S. Adam, Y. Lecourtier, P. Héroux, and E. Trupin. Approximation of digital curves using a multi-objective genetic algorithm. In *Proceedings of the International Conference on Pattern Recognition (ICPR'06)*, pages 716–719. 2006.

[26] E. Barbu, P. Héroux, S. Adam, and E. Trupin. Clustering document images using a bag of symbols representation. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'05)*, pages 1216–1220. 2005.

[27] S. Adam, M. Rigamonti, E. Clavier, J.-M. Ogier, E. Trupin, and K. Tombre. DocMining : A Document Analysis System Builder. In S. Marinai and A. Dengel, editors, *Proceedings of the IAPR Workshop on Document Analysis Systems (DAS'04)*, volume 3163 of *Lecture Notes in Computer Science*, pages 472–483. 2004.

[28] M. Delalandre, P. Héroux, S. Adam, É. Trupin, and J.-M. Ogier. A statistical and structural approach for symbol recognition, using xml modelling. In T. Caelli, A. Amin, R. P. W. Duin, M. S. Kamel, and D. de Ridder, editors, *Proceedings of the International Workshop on Syntactical and Structural Pattern Recognition (SSPR'02)*, volume 2396 of *Lecture Notes in Computer Science*, pages 281–290. Springer, 2002.

[29] S. Adam, J. Gardes, Y. Lecourtier, J. Ogier, and R. Mullot. Multi-scaled and multi oriented character recognition : An original strategy. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'99)*, pages 45–48. 1999.

Autres conférences internationales avec actes et comité de sélection

[30] S. Bernard, L. Heutte, and S. Adam. Influence of hyperparameters on random forest accuracy. In J. A. Benediktsson, J. Kittler, and F. Roli,

editors, *Proceedings of Multiple Classifier Systems (MCS'09)*, volume 5519 of *Lecture Notes in Computer Science*, pages 171–180. Springer, 2009.

[31] S. Bernard, L. Heutte, and S. Adam. Towards a better understanding of random forests through the study of strength and correlation. In D.-S. Huang, K.-H. Jo, H.-H. Lee, H.-J. Kang, and V. Bevilacqua, editors, *Proceedings of the International Conference on Intelligent Computing (ICIC'09)*, volume 5755 of *Lecture Notes in Computer Science*, pages 536–545. Springer, 2009.

[32] S. Bernard, L. Heutte, and S. Adam. Forest-RK : A new random forest induction method. In D.-S. Huang, D. C. W. II, D. S. Levine, and K.-H. Jo, editors, *Proceedings of the International Conference on Intelligent Computing (ICIC'08)*, volume 5227 of *Lecture Notes in Computer Science*, pages 430–437. Springer, 2008.

[33] E. Barbu, C. Chatelain, S. Adam, P. Héroux, and E. Trupin. A simple one class classifier with rejection strategy : Application to symbol classification. In *Proceedings of the IAPR Workshop on Graphics Recognition (GREC'07)*, pages 35–36. 2007.

[34] H. Locteau, S. Adam, E. Trupin, J. Labiche, and P. Héroux. Symbol spotting using full visibility graph representation. In *Proceedings of the IAPR Workshop on Graphics Recognition (GREC'07)*. 2007.

[35] R. Raveaux, E. Barbu, H. Locteau, S. Adam, P. Héroux, and E. Trupin. A graph classification approach using a multi-objective genetic algorithm application to symbol recognition. In F. Escolano and M. Vento, editors, *Proceedings of the IAPR International Workshop on Graph Based Representations for Pattern Recognition (GbR-PR'07)*, volume 4538 of *Lecture Notes in Computer Science*, pages 361–370. Springer, 2007.

[36] E. Barbu, P. Héroux, S. Adam, and E. Trupin. Clustering of document images using graph summaries. In P. Perner and A. Imiya, editors, *Proceedings of Machine Learning and Data Mining in Pattern Recognition (MDLM'05)*, volume 3587 of *Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence*, pages 194–202. Springer, 2005.

[37] E. Barbu, P. Héroux, S. Adam, and É. Trupin. Indexation of document images using frequent items. In *Proceedings of the International Workshop on Pattern Recognition in Information System (PRIS'05)*, pages 164–173. 2005.

[38] E. Barbu, P. Héroux, S. Adam, and E. Trupin. Using bags of symbols for automatic indexing of graphical document image databases. In *Proceedings of the IAPR Workshop on Graphics Recognition (GREC'05)*, pages 195–205. 2005.

[39] H. Locteau, R. Raveaux, S. Adam, Y. Lecourtier, P. Héroux, and E. Trupin. Polygonal approximation of digital curves using a multi-objective genetic algorithm. In *Proceedings of the IAPR Workshop on Graphics Recognition (GREC'05)*, pages 300–311. 2005.

[40] H. Locteau, S. Adam, E. Trupin, J. Labiche, and P. Héroux. Symbol recognition combining vectorial and statistical features. In *Proceedings of*

the IAPR Workshop on Graphics Recognition (GREC'05), pages 76–87. 2005.

[41] Y. Saidali, S. Adam, J.-M. Ogier, É. Trupin, and J. Labiche. Knowledge representation and acquisition for engineering document analysis. In *Proceedings of the International Workshop on Graphics RECognition (GREC'03)*, pages 25–37. 2003.

[42] S. Adam, J.-M. Ogier, É. Trupin, and R. Mullot. A scale and rotation parameters estimator application to technical document interpretation. In *Proceedings of the International Workshop on Pattern Recognition in Information Systems (PRIS'03)*, pages 31–37. 2003.

[43] J. Gardes, J. Ogier, S. Adam, and R. Mullot. Caati - a system-based dynamic document interpretation device. In *Proceedings of the International Workshop on Graphics RECognition (GREC'01)*, pages 301–311. 2001.

[44] J. Ogier, S. Adam, A. Bessaid, and H. Bechar. Automatic topographic map analysis system : an overview. In *Proceedings of the International Workshop on Graphics RECognition (GREC'01)*, pages 229–244. 2001.

[45] E. Trupin, J. Ogier, S. Adam, and J. Gardes. Navigation into technical documents. In *Proceedings of the International Workshop on Graphics RECognition (GREC'01)*, pages 27–34. 2001.

[46] S. Adam, J.-M. Ogier, C. Cariou, and J. Gardes. A scale and rotation parameters estimator application to technical document interpretation. In *Proceedings of the International Workshop on Graphics RECognition (GREC'01)*, pages 27–34. 2001.

[47] S. Adam, F. Rousseau, J. Ogier, C. Cariou, R. Mullot, J. Labiche, and J. Gardes. A multi-scale and multi-orientation recognition technique applied to document interpretation application to french telephone network maps. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, pages 1509–1512. 2001.

[48] J.-M. Ogier, C. Cariou, S. Adam, J. Gardes, R. Mullot, and Y. Lecourtier. Similitude invariant pattern recognition on technical documents. In *Proceedings of the International Conference on Image Processing (ICIP'99)*, pages 570–574. 1999.

[49] S. Adam, R. Mullot, J.-M. Ogier, C. Cariou, J. Gardes, and Y. Lecourtier. Processing of the connected shapes in raster-to-vector conversion process. In *Proceedings of the International Workshop on Graphics RECognition (GREC'99)*, pages 28–38. 1999.

[50] S. Adam, J.-M. Ogier, C. Cariou, J. Gardes, R. Mullot, and Y. Lecourtier. Combination of invariant pattern recognition primitives on technical documents. In *Proceedings of the International Workshop on Graphics RECognition (GREC'99)*, pages 238–245. 1999.

Conférences nationales avec actes et comité de sélection

[51] G. Dupont, S. Adam, and Y. Lecourtier. Apprentissage par renforcement pour la recherche d'information interactive. In *Actes des 6emes Journées*

Francophones de Planification, Décision et Apprentissage pour la conduite de systèmes (JFPDA 2011). 2011.

[52] P. L. Bodic, P. Héroux, S. Adam, H. Locteau, J. Bilong, and Y. Lecourtier. Programmation linéaire en nombres entiers pour la recherche d'isomorphismes de sous-graphes. In *Actes du Colloque International Francophone sur l'Ecrit et le Document (CIFED'10)*, pages 153–168. 2010.

[53] A. S. Réquier, G. Dupont, S. Adam, and Y. Lecourtier. Évaluation d'outils de reformulation interactive de requêtes. In *Actes de la COnférence en Recherche d'Information et Applications (CORIA'10)*, pages 223–238. 2010.

[54] S. Bernard, L. Heutte, and S. Adam. Une Étude sur la paramétrisation des forêts aléatoires. In *Actes de la Conférence francophone sur l'Apprentissage Artificiel (CAP'09)*, pages 81–92. 2009.

[55] S. Bernard, L. Heutte, and S. Adam. Étude de l'influence des paramètres sur les performances des forêts aléatoires. In *Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'08)*, pages 207–208. 2008.

[56] P. L. Bodic, S. Adam, P. Héroux, A. Knippel, and Y. Lecourtier. Formulations linéaires en nombres entiers pour des problèmes d'isomorphisme exact et inexact. In *Actes électroniques des Journées Polyèdres et Optimisation Combinatoire (JPOC'08)*. 2008.

[57] C. Chatelain, S. Adam, Y. Lecourtier, L. Heutte, Y. Oufella, and T. Paquet. Optimisation multi-objectif pour la sélection de modèles SVM. In *Actes du congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'08)*, pages 67–72. 2008.

[58] L. Heutte, S. Bernard, S. Adam, and E. Oliveira. De la sélection d'arbres de décision dans les forêts aléatoires. In *Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'08)*, pages 163–168. 2008.

[59] R. Raveaux, E. Barbu, S. Adam, P. Héroux, and E. Trupin. Graphes prototypes vs. graphe médian généralisé pour la classification de données structurées. In *Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'08)*, pages 37–42. 2008.

[60] G. Dupont, S. Adam, Y. Lecourtier, and B. Grilhere. Multi objective particle swarm optimization using enhanced dominance and guide selection. In *Journées Optimisation par Essaims Particulaires (OEP'07) - actes électroniques*. 2007.

[61] E. Barbu, R. Raveaux, H. Locteau, S. Adam, P. Héroux, and E. Trupin. Classification de graphes par algorithmes génétiques et signatures de graphes : Application à la reconnaissance de symboles. In *Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'06)*, pages 91–96. 2006.

[62] P. Héroux, E. Barbu, S. Adam, and E. Trupin. Production de vérité terrain pour l'analyse et l'interprétation d'images de document. In *Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'06)*, pages 67–72. 2006.

[63] H. Locteau, R. Raveaux, S. Adam, Y. Lecourtier, P. Héroux, and E. Trupin. Approximation de courbes par algorithme génétique multi-objectif. In *Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'06)*, pages 37–42. 2006.

[64] H. Locteau, S. Adam, E. Trupin, J. Labiche, and P. Héroux. Reconnaissance de symbole guidée par une modélisation basée sur les graphes de régions adjacentes. In *Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'06)*, pages 151–156. 2006.

[65] E. Barbu, P. Héroux, S. Adam, and E. Trupin. Fouille de graphes et découverte de règles d'association : application à l'analyse d'images de document. In *Actes des journées Extraction et Gestion des Connaissances (EGC'05)*, pages 463–468. 2005.

[66] E. Barbu, P. Héroux, S. Adam, and E. Trupin. Découverte de motifs fréquents - application à l'analyse de documents graphiques. In *Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'04)*, pages 143–148. 2004.

[67] E. Clavier, S. Adam, P. Héroux, M. Rigamonti, and J.-M. Ogier. Docmining - une plate-forme de conception de systèmes d'analyse de document. In *Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'04*, pages 97–102. 2004.

[68] H. Locteau, S. Adam, E. Trupin, J. Labiche, and P. Héroux. Détection d'arcs de cercle par comparaison du tracé théorique de Bresenham. In *Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'04)*, pages 285–290. 2004.

[69] M. Delalandre, P. Héroux, S. Adam, E. Trupin, and J. Ogier. Une approche statistico-structurelle pour la reconnaissance de symboles exploitant une représentation xml des données. In *Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'02)*, pages 121–128. 2002.

[70] C. Cariou, J. Ogier, S. Adam, J. Gardes, R. Mullot, and Y. Lecourtier. Reconnaissance de formes multi-échelle sur documents techniques. In *Actes du Colloque du Groupe de Recherche et d'Études en Traitement du Signal et des Images (GRETSI'99)*, pages 283–286. 2000.

[71] V. Grenier, R. Mullot, J. Ogier, S. Adam, J. Gardes, and Y. Lecourtier. Distribution d'opérateurs pour l'analyse de documents techniques. In *Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'00)*, pages 151–160. 2000.

[72] V. Grenier, R. Mullot, J. Ogier, S. Adam, J. Gardes, and Y. Lecourtier. Une architecture distribuée pour l'interprétation de documents techniques. In *Actes du Congrès Reconnaissance de Formes et Intelligence Artificielle (RFIA'00)*, pages 427–436. 2000.

[73] S. Adam, R. Mullot, J. Ogier, C. Cariou, and J. Gardes. Interprétation de documents du réseau téléphonique : Approche multi-spécialistes. In *Actes du Congrès Reconnaissance de Formes et Intelligence Artificielle (RFIA'00)*, pages 357–364. 2000.

[74] S. Adam, J. M. Ogier, C. Cariou, R. Mullot, J. Gardes, and J. Labiche. Reconnaissance de formes multi-orientées et multi-échelle : Application à l'analyse automatique de documents techniques. In *Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'00)*, pages 21–30. 2000.

[75] V. Grenier, R. Mullot, J. Ogier, S. Adam, J. Gardes, and Y. Lecourtier. Proposition d'architecture distribuée pour un système de rétro-conversion de documents techniques. In *Actes du Colloque International sur le Document Electronique (CIDE'99)*, pages 139–153. 1999.

[76] S. Adam, R. Mullot, J. Ogier, C. Cariou, and J. Gardes. Stratégie multi-spécialistes d'extraction d'information sur des documents techniques. In *Actes du Colloque International sur le Document Electronique (CIDE'99)*, pages 139–153. 1999.

Thèses soutenues

[77] G. Dupont. *Apprentissage implicite pour la recherche d'information.* Ph.D. thesis, Université de Rouen, 2011.

[78] S. Bernard. *Forêts Aléatoires : de l'Analyse des Mécanismes de Fonctionnement à la Construction Dynamique.* Ph.D. thesis, Université de Rouen, 2009.

[79] H. Locteau. *Contributions à la localisation de symboles dans les documents graphiques.* Ph.D. thesis, Université de Rouen, 2008.

[80] E. Barbu. *Fouille et classification de graphes : application à la reconnaissance de symboles dans les documents graphiques.* Ph.D. thesis, Université de Rouen, 2007.

[81] S. Adam. *Interprétation de documents techniques : des outils à leur intégration dans un système à base de connaissances.* Ph.D. thesis, Université de Rouen, 2001.

1.4 Activités d'enseignement

1.4.1 Filières d'enseignement

Depuis ma nomination, j'interviens principalement en licence "Electronique Electrotechnique Automatique" (EEA), au niveau L1 et L3 et en Master "Informatique Génie de l'Information et des Systèmes" (IGIS) de l'UFR des Sciences. Dans ce dernier, j'assure des enseignements à la fois dans la spécialité professionnelle Génie Electrique et Informatique Industrielle (GEII) et dans la spécialité pro-recherche Système de Traitement des Informations Multimédia (STIM).

1.4.2 Enseignements dispensés

Traitement numérique de l'information : Logique combinatoire et séquentielle. Unité Arithmétique et Logique. Architecture des ordinateurs. Microprocesseurs.

Programmation C - Génie Informatique : Types de base, constantes, opérateurs, instructions de contrôle, pointeurs, fonctions, types composés, entrées-sorties, fichiers, listes chaînées, pile, files, arbres.

Traitement numérique du signal Outils mathématiques pour le signal (Transformées), Analyse Spectrale (Transformée de Fourier, TFD, FFT, Fenêtres spectrales), Filtrage numériques (filtres RIF, RII, cascades de filtres).

Traitement d'images et reconnaissance de formes : filtres, morphologie mathématique, analyse spectrale, extraction de caractéristiques, classifieurs, combinaisons de classifieurs.

Optimisation : moindres carrés, descente de gradient, Gauss Newton, algorithmes génétiques, essaims particulaires.

Programmation système : Gestion de processus, communication entre processus (tubes, signaux, mémoires partagées, sémaphores), Gestion de threads.

1.4.3 Volumes horaires

Année	2002 2003	2003 2004	2004 2005	2005 2006	2006 2007	2007 2008	2008 2009	2009 2010	2010 2011
Heures eq. TD	213	202	179	193	195	192	192	213	194

1.5 Activités administratives

1.5.1 Responsabilités administratives et pédagogiques

Direction d'études et présidence de jury Depuis ma nomination et jusqu'en septembre 2006, j'ai eu en charge la direction et l'organisation des enseignements de la seconde année de l'IUP Génie Electrique et Informatique Industrielle. Depuis septembre 2006, je suis directeur d'études et chargé de l'organisation des enseignements sur l'année (responsable pédagogique) de la première année de la spécialité Génie Electrique et Informatique Industrielle (GEII) du Master d'Informatique, de Génie de l'Information et des Systèmes. Je suis également président de jury de cette année.
Dans ce cadre, j'ai à ma charge la mise au point et la gestion des emplois du temps au cours des deux semestres, la recherche d'enseignants et de vacataires ainsi que l'organisation des pré-jurys, des oraux et du jury des deux sessions.

Responsable des Travaux d'Etudes et de Recherche Depuis septembre 2007, j'assure pour l'ensemble des années des filières EEA, GEII et STIM la responsabilité des projets annuels (Travaux d'Etudes et de Recherche). Dans ce cadre, je suis chargé de la collecte des sujets proposés par l'équipe pédagogique, de l'attribution de ces sujets aux étudiants (entre 50 et 100 en fonction des années), de la collecte et de l'examen des cahiers des charges rédigés par

les étudiants et des rapports finaux, de l'organisation des soutenances des étudiants, et de la présidence des jurys liés à ces soutenances.

1.5.2 Fonctions électives au sein de l'établissement

- Membre du conseil de département de physique de l'UFR de 2008 à 2011
- Membre de 2004 à 2008 des commissions de spécialistes de l'Université de Rouen (61ème section et 27/61ème section - vice président)
- Membre de 2006 à 2008 de la commission de spécialistes de l'INSA de Rouen (27-61-63ème sections).
- Membre en 2009 d'un comité de sélection 61ème section de l'Université de Rouen.

Deuxième partie

Contributions et Perspectives

Chapitre 2

Introduction générale

Au cours des vingt dernières années, le développement des Sciences et Technologies de l'Information et de la Communication (STIC) a bouleversé notre manière de vivre et de travailler. Aucun secteur n'est aujourd'hui « épargné » par cette émergence du numérique. Les STIC jouent désormais un rôle prépondérant dans la santé, l'éducation, la culture, la conservation des patrimoines, l'agriculture, les administrations, les médias, la finance, l'industrie... Ce bouleversement, que les historiens jugent aussi profond que celui de la révolution industrielle des $XVIII^e$ et XIX^e siècles, a engagé le monde sur la voie d'une société basée sur l'information et la connaissance.

Les progrès technologiques liés aux capacités de stockage et aux réseaux Internet et Intranet sont indéniablement les facteurs à l'origine de cette révolution. Toutefois, ces évolutions technologiques ont également fait émerger de nouvelles problématiques scientifiques indispensables pour permettre l'accès et le traitement de cette quantité phénoménale d'informations et qui, elles-mêmes, amplifient les besoins en nouvelles technologies. Parmi ces problématiques, on trouve celle de la Gestion Électronique de Document (GED), qui est l'une des solutions utilisées pour optimiser la gestion de l'information. Elle se définit comme l'ensemble des techniques et méthodes qui ont pour but de faciliter l'archivage, l'accès, la consultation, la diffusion des documents et des informations qu'ils contiennent.

Dans la chaîne électronique de gestion des documents, le traitement automatique des images de documents est l'un des maillons permettant d'alimenter les systèmes quand l'information initiale n'est disponible que sous la forme papier et quand une reprise manuelle est trop coûteuse. Même si initialement il ne s'agissait que de scanner le document papier et de le stocker sous forme d'image dans une archive afin d'en faciliter la circulation au sein des différents services d'une organisation, la problématique du traitement automatique de documents s'est étendue à la conception plus générale de méthodologies, outils et systèmes permettant de classer, trier, indexer, stocker et interpréter automatiquement des documents numériques rétroconvertis à partir de documents papiers (formulaires, plans, courriers, archives...).

C'est ainsi que s'est formée, il y a plus de vingt ans, une communauté travaillant autour de la problématique de l'analyse d'images de documents, dans le but de transformer de telles images en un contenu structuré et exploi-

table. Cette communauté s'est construite au niveau national autour de groupes tels que le GRCE (Groupe de Recherche en Communication Écrite) ou celui du thème « Documents Multimédia » du GDR I3 (Information-Interaction-Intelligence), et au niveau international autour des comités techniques TC10 *Graphic Recognition (GREC)* et TC11 *Reading Systems* de l'IAPR (*International Association for Pattern Recognition*). On a également assisté à l'émergence de nouvelles revues scientifiques dédiées à l'analyse de documents telles que l'*International Journal of Document Analysis and Recognition (IJDAR)*, à la tenue de congrès centrés sur cette thématique, comme le Colloque International Francophone sur l'Écrit et le Document (CIFED), ou au niveau international à l'organisation des conférences ou workshops tels que l'*International Conference on Document Analysis and Recognition (ICDAR)*, *Graphic Recognition (GREC)*, *Document Analysis System (DAS)* ou l'*International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pour ne citer que les plus anciens.

Une importante communauté scientifique s'est ainsi constituée autour de la problématique de la conception d'outils et de systèmes permettant d'interpréter le contenu d'images de documents. De tels systèmes d'analyse d'images de documents reposent sur de nombreuses étapes de traitement, allant du bas niveau (filtrage, restauration, redressement, binarisation, squelettisation, segmentation texte/graphique) à des processus d'interprétation de haut niveau sémantique (rétroconversion, extraction de méta-données, indexation de documents...), en passant par des étapes de reconnaissance des entités présentes dans le document (texte, symboles, lignes, arcs de cercle, logos...). La figure 2 illustre la complexité des chaînes d'analyse d'images de documents à travers trois systèmes de l'état de l'art respectivement dédiés à l'analyse de documents structurés [56], la reconnaissance de textes manuscrits [114] et la vectorisation de documents graphiques [92].

Les différents « niveaux » d'analyse qui apparaissent sur cette figure illustrent les interactions nécessaires entre la communauté de l'analyse de documents et des communautés connexes. On peut citer évidemment la communauté du traitement d'images pour améliorer les données brutes que sont les pixels, celle de la reconnaissance de formes statistique, syntaxique et structurelle pour transformer ces données en objets de plus haut niveau sémantique, celle de l'intelligence artificielle pour planifier les traitements, ou encore celle de l'optimisation pour régler les nombreux paramètres des chaînes d'analyse. Même s'ils n'apparaissent pas explicitement sur cette figure, il ne faut pas non plus négliger les aspects liés à l'ingénierie des connaissances, qui permettent de modéliser et de représenter les connaissances liées à la fois au domaine de l'analyse de documents et au(x) métier(s) concerné(s) par le document. Enfin, les systèmes complètement automatiques étant encore du domaine de la recherche à long terme pour des problématiques difficiles telles que l'analyse de plans ou de cartes, des interactions fortes avec la communauté de l'Interaction Homme Machine (IHM) sont fondamentales pour placer l'Homme au cœur des systèmes d'analyse. Sans aller jusqu'à une rétroconversion de grande masse de documents, une IHM adaptée sera également indispensable quand il s'agira de dialoguer avec un système qui cherchera à interpréter une image acquise avec un scanner ou un appareil photo.

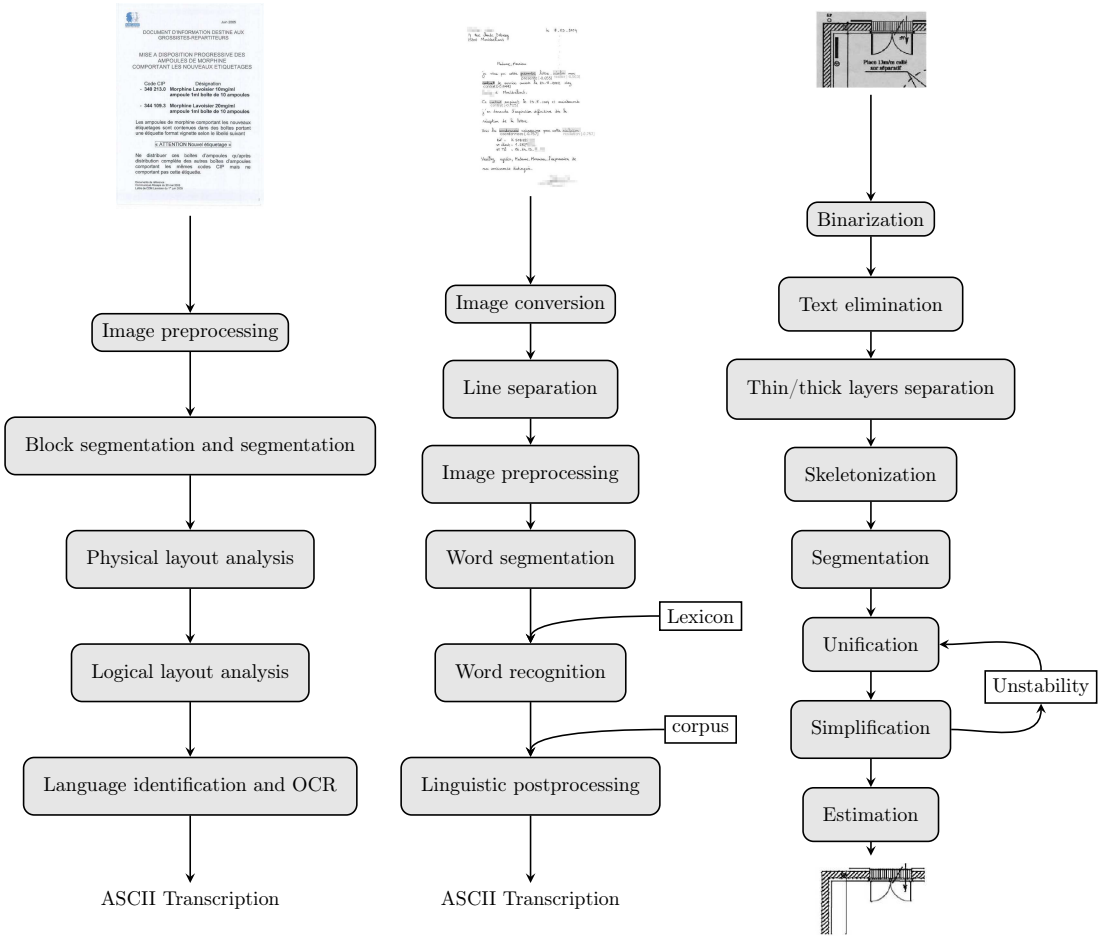


FIGURE 2.1 – Quelques chaînes typiques d'analyse d'images de documents respectivement dédiées à l'analyse de documents structurés [56], à la reconnaissance de textes manuscrits [114] et à la vectorisation de documents graphiques [92].

Les travaux qui sont abordés dans cette habilitation sont à la confluence de plusieurs de ces domaines de recherche. Ils concernent deux aspects principaux. Le premier est relatif à la reconnaissance structurelle de formes, en proposant deux contributions liées respectivement à la classification supervisée de graphes [170] et à la recherche d'isomorphismes de sous-graphes [31]. Le second concerne la prise en compte d'objectifs multiples en analyse d'images de documents, tant pour l'évaluation des performances des systèmes que pour leur optimisation [52, 130]. Dans les deux cas, les travaux sont appliqués à des problèmes d'analyse de documents, pour la reconnaissance et la localisation de symboles et pour la reconnaissance de courriers manuscrits.

Nous proposons ci-après de faire une synthèse de ces contributions et de nos

perspectives dans ces domaines, en positionnant celles-ci par rapport à l'état de l'art des différentes problématiques abordées. Après cette introduction, cette synthèse scientifique est organisée autour de trois chapitres.

Dans le chapitre 3, nous nous intéressons plus particulièrement à la reconnaissance structurelle de formes appliquée à l'analyse de documents graphiques. Deux problématiques fondamentales de ce domaine sont d'abord abordées indépendamment de l'application. La première concerne la classification supervisée de graphes, et plus particulièrement la définition de nouveaux prototypes permettant d'exploiter de façon efficace des méthodes du type k Plus Proches Voisins (kPPV). La seconde est celle de la recherche d'isomorphismes de sous-graphes tolérants aux erreurs, que nous avons abordée par la programmation linéaire en nombres entiers. Pour chacun de ces aspects, le problème est formalisé, l'état de l'art est décrit, et nos propositions sont discutées et positionnées par rapport à l'existant. Puis, dans la section suivante, nous présentons des applications de ces propositions à l'analyse d'images de documents graphiques, et plus particulièrement pour la reconnaissance et la localisation de symboles. La dernière section de ce chapitre dresse un bilan de ces contributions et présente des problèmes ouverts relatifs à ces travaux.

Dans le chapitre 4, nous partons du constat que la plupart des systèmes réels complexes, comme le sont les systèmes d'analyse de documents, mettent en jeu des objectifs multiples qui nécessitent le choix de compromis. Nous proposons donc d'illustrer à partir de quelques problèmes liés à l'analyse d'images de documents les apports de l'optimisation multiobjectif. Une présentation succincte du domaine de l'optimisation multiobjectif est d'abord proposée, et un état de l'art des approches permettant de résoudre de tels problèmes est présenté. Puis, trois contributions sont décrites. La première est une contribution propre au domaine de l'optimisation multiobjectif. Nous y proposons un algorithme pour aborder ces problèmes avec la technique des essaims particulaires. Puis, les deux contributions suivantes concernent des travaux pour lesquels nous avons tiré parti de l'intégration d'objectifs multiples en analyse de documents et en apprentissage. Le chapitre se termine par une discussion sur cet apport et sur les perspectives directement ouvertes par ces travaux.

Enfin, dans le chapitre 5, nous synthétisons d'abord les perspectives de recherche à court et moyen terme évoquées dans les chapitres 3 et 4. Puis, nous exposons nos perspectives à plus long terme, en soulignant la nécessaire interdisciplinarité en analyse de documents. Nous y abordons également la convergence de nos travaux entre analyse de documents et recherche d'information.

Chapitre 3

Documents et graphes

3.1 Introduction

Les graphes sont des structures de données fréquemment exploitées pour la représentation d'entités complexes. Dans une représentation à base de graphes, les nœuds et leurs étiquettes décrivent des objets et leurs propriétés, tandis que les arcs et leurs étiquettes décrivent les relations entre ces objets. Les graphes permettent ainsi de dépasser certaines limites inhérentes à une représentation vectorielle des données telles que (i) la taille fixe, généralement imposée par l'utilisation de classifieurs statistiques, (ii) l'impossibilité de modéliser naturellement des relations entre composants du vecteur. Un graphe permet au contraire de décrire non seulement les propriétés d'un objet, mais aussi les relations binaires (spatiales, temporelles, conceptuelles...) entre ses différentes parties. Parmi ces relations, citons le concept très important de sous-graphes, qui permet d'envisager la recherche de sous-structures au sein d'un graphe, et dont les implications importantes en analyse de documents seront soulignées en 3.4.1. Par ailleurs, comme nous le verrons dans ce chapitre, les graphes ne sont pas *a priori* contraints à une taille donnée, le nombre de nœuds et d'arcs n'étant théoriquement pas limité par les outils exploitant ces représentations.

Grâce à ce pouvoir représentationnel, couplé à l'augmentation de la puissance de calcul des ordinateurs, les représentations structurelles sont devenues de plus en plus populaires dans de nombreux domaines d'application comme la biologie, la chimie, la vision par ordinateur, l'analyse de textes ou encore la reconnaissance de formes. À titre d'illustration, en 2004, Conte *et al.* décrivaient dans [62] plus de 160 articles ayant trait aux outils d'appariement de graphes et à leur application dans le domaine de la reconnaissance de formes. Un comité technique de l'IAPR, le TC 15 [6], et une conférence internationale (Graph based Representations in Pattern Recognition - GbRPR) sont même spécifiquement dédiés aux représentations à base de graphe dans le domaine de la reconnaissance de formes. Dans ce contexte, les graphes ont trouvé un nombre considérable d'applications dans le domaine de l'analyse de documents, comme en témoigne l'état de l'art proposé tout récemment par Horst Bunke et Kaspar Riesen dans [42]. Ils sont par exemple exploités pour représenter des symboles [127, 206, 128, 125], des tableaux [167], la structure de docu-

6. http ://www.greyc.ensicaen.fr/iapr-tc15/

ments [132, 126], des caractères manuscrits [134, 48] ou encore des équations mathématiques [215]. La figure 3.1 illustre quelques applications d'analyse de documents s'appuyant sur des représentations à base de graphes.

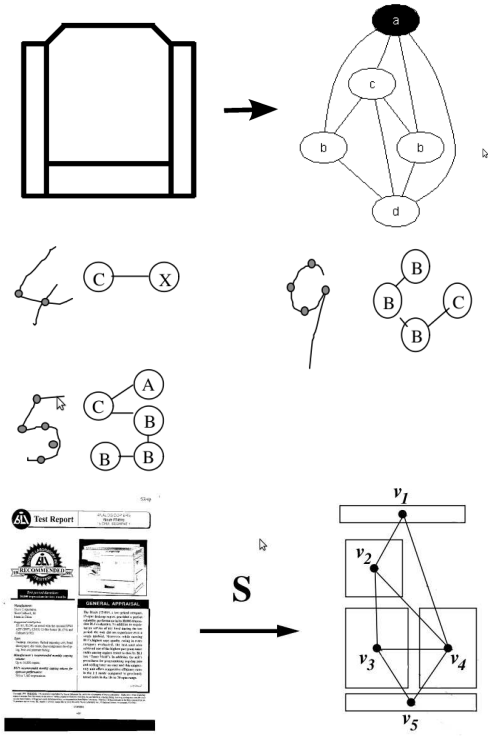


FIGURE 3.1 – Trois exemples de représentations à base de graphes extraites sur des images de documents pour des applications respectivement dédiées à la reconnaissance de symboles [18], la reconnaissance de chiffres manuscrits [48] et l'analyse de la structure physique de documents [16].

Avec cette émergence des représentations structurelles dans le domaine de l'analyse de documents ou dans celui de la reconnaissance de formes en général, les problématiques liées aux outils de traitement des graphes ont connu un essor ou un regain d'intérêt important. Parmi les sujets de recherche autour desquels de nombreux travaux sont en cours actuellement, on peut citer la fouille de graphe [118, 101], la classification supervisée [14, 153, 133], le clustering [216, 164] ou encore la recherche d'isomorphismes de graphes ou de sous-graphes [65].

Dans ce chapitre, nous abordons certaines de ces problématiques liées à l'analyse de graphes. Nos contributions dans ce domaine y sont présentées, justifiées et positionnées par rapport à l'existant. Dans les deux premières sections, nous présentons deux problèmes fondamentaux abordés aux cours de nos travaux. Le premier concerne la classification supervisée de graphes, et plus particulièrement la génération de nouveaux prototypes utilisés avec un classifieur de type k Plus Proches Voisins (kPPV) [170]. Un algorithme génétique manipulant des graphes pour générer ces prototypes est proposé. Le

second problème est celui de la recherche d'isomorphismes de sous-graphes tolérants aux erreurs, que nous avons abordé par la programmation linéaire en nombres entiers [31]. Pour chacun de ces aspects, nous décrivons le problème et dressons une synthèse de la littérature s'y reportant, avant de présenter synthétiquement notre contribution et de discuter des résultats obtenus. Dans la section suivante, nous présentons quelques applications de ces propositions pour l'analyse d'images de documents, plus particulièrement pour la reconnaissance et la localisation de symboles dans des documents graphiques. Enfin, la dernière section dresse un bilan de ces contributions et présente des problèmes ouverts relatifs à ces travaux.

3.2 Classification de graphes

3.2.1 Définition du problème et revue de l'existant

La classification supervisée de graphes est une problématique ayant émergé récemment avec le développement des représentations structurelles. Un algorithme de classification de graphes a pour but d'affecter une classe à un graphe inconnu, en utilisant une fonction généralement issue d'un processus d'apprentissage. Plus formellement, on peut définir l'apprentissage d'une telle fonction de la façon suivante :

Définition 1. Soit χ un ensemble de graphes étiquetés. Soit un ensemble d'apprentissage $L = \{\langle g_i, c_i \rangle\}_{i=1}^{M}$, où les $g_i \in \chi$ sont des graphes étiquetés et où $c_i \in C$ est la classe de g_i parmi les N classes présentes dans la base d'apprentissage. L'apprentissage d'un classifieur de graphes consiste à induire de L une fonction $f(g) : \chi \rightarrow C$ attribuant une classe à un graphe g inconnu.

Les algorithmes de classification de graphes sont utilisés dans différents domaines, allant de l'analyse de séquences biologiques (ADN, ARN) à l'analyse de données semi-structurées (XML, HTML), en passant par la prédiction de propriétés de composants chimiques et le traitement du langage naturel [116]. Dans la littérature, deux principales familles d'approches peuvent être distinguées pour résoudre un problème de classification de graphes. La première famille consiste à projeter les graphes dans un espace vectoriel, dans le but de bénéficier de la richesse et de la robustesse des méthodes d'apprentissage statistique. La seconde famille repose sur la règle des k Plus Proches Voisins (kPPV), en s'appuyant sur une mesure de dissimilarité spécifique aux graphes.

Les méthodes à base de projection Motivés par les progrès considérables réalisés en apprentissage statistique, un nombre important de travaux reposant sur des projections ont été publiés ces dix dernières années. Parmi les méthodes proposées, certaines calculent explicitement un vecteur de caractéristiques numériques décrivant le graphe (nombre de nœuds ou d'arcs d'un label donné, nombre de cycles, degrés des nœuds...). C'est le cas, par exemple, dans les travaux décrits dans [132]. Cette méthode, bien que très rapide, souffre de la

perte d'information structurelle générée par la projection et de sa non bijectivité. Deux graphes différents peuvent ainsi avoir la même description. Cette approche est étendue dans [191] par l'énumération de sous-graphes appartenant à un lexique exhaustif des graphes non isomorphes, ce qui permet de mieux prendre en compte la topologie du graphe. D'autres méthodes dans cette catégorie s'appuient sur la théorie spectrale des graphes, qui consiste à exploiter les valeurs propres et les vecteurs propres de la matrice d'adjacence [135] ou de la matrice laplacienne normalisée [212, 174]. Ces méthodes ont obtenu des succès importants dans le domaine de l'analyse d'images. Récemment sont également apparues des méthodes, parfois appelées *graph embedding*, consistant à représenter un graphe par un vecteur de mesures de dissimilarités calculées par rapport à un ensemble donné de graphes [177]. Ces vecteurs numériques sont ensuite utilisés pour l'apprentissage d'un classifieur statistique. De telles méthodes ont l'avantage de pouvoir traiter n'importe quel type de graphes, sous réserve de disposer d'une mesure de dissimilarité adéquate. Enfin, citons dans cette catégorie les méthodes à noyaux, sur lesquelles se concentrent beaucoup de travaux depuis quelques années. Initialement proposées par [85] et [108], ces approches ne projettent pas explicitement les graphes dans un espace vectoriel mais reposent sur le calcul d'un noyau qui exprime la similarité entre graphes et qui est ensuite utilisé comme un produit scalaire. De très nombreuses contributions relatives à la proposition de noyaux sont disponibles dans la littérature. Celles-ci se basent sur des marches aléatoires dans les graphes, des chemins, des cycles ou encore des sous-arbres. Une bonne revue de ces noyaux est disponible dans [42]. Ces noyaux sont ensuite exploités par des machines à noyaux telles que les SVM (*Support Vector Machine*) ou les KPCA (*Kernel Principal Component Analysis*) [108, 109, 197, 137, 138, 209].

Les méthodes à base de kPPV Ce type de méthode est souvent choisi pour sa simplicité de mise en œuvre et son bon comportement asymptotique. De telles approches consistent à classifier les graphes en appliquant la règle des k Plus Proches Voisins exploitant une mesure de dissimilarité entre graphes. Ces méthodes souffrent toutefois des limitations inhérentes à la méthode des kPPV, à savoir sa complexité combinatoire, son besoin de stockage important et sa sensibilité aux exemples bruités. Une solution souvent adoptée pour pallier ces défauts consiste, comme pour certaines des méthodes précédemment évoquées pour la projection, à réduire l'ensemble de graphes utilisés pour les kPPV par l'intermédiaire d'un processus d'extraction de prototypes (parfois appelés représentants). On parle alors de méthode des k plus proches prototypes (k *Nearest Prototype Classifier* - kNPC). Une telle stratégie n'est évidemment pas propre au problème de classification de graphes. Elle est également exploitée pour comparer des contours dans des applications de vision [58] ou pour la reconnaissance statistique de formes [69, 91, 50, 26]. Dans le domaine de la reconnaissance structurelle qui nous intéresse ici, on peut citer les travaux présentés dans [103] qui exploitent des prototypes basés sur la présence de sous-graphes communs, les approches proposées dans [32] et [40] qui crèent des représentations appelées *super-graphs* ou les travaux de [139] qui consistent à générer des *creative prototypes* en appliquant à un graphe germe une sé-

rie d'opérations d'édition pour générer les prototypes. La dernière approche à mentionner, probablement la plus fréquemment utilisée, est celle consistant à exploiter les graphes médians en tant que prototypes [79, 41, 104, 78, 93]. Le calcul de tels graphes repose sur la minimisation de la somme des distances du graphe recherché à l'ensemble des graphes d'une classe donnée. Deux types de graphes médians sont proposés dans la littérature : les graphes médians d'ensemble (Définition 2) et les graphes médians généralisés (Définition 3). Ils diffèrent en fonction de l'espace dans lequel ils sont calculés. Dans le premier cas, l'espace de recherche est limité à l'ensemble initial de graphes. On parle alors de sélection de prototypes. Dans le second cas, ils sont calculés dans un ensemble infini contenant tous les graphes pouvant être construits à partir de l'ensemble des labels des graphes initiaux. On parle alors de génération de prototypes. Les graphes médians généralisés se sont montrés particulièrement efficaces pour modéliser une classe de graphes et pour rejeter des exemples bruités [104].

Définition 2. Soit $d(.,.)$ une distance ou une mesure de dissimilarité entre deux graphes. Soit $\mathcal{S} = \{g_1, g_2, \ldots, g_n\}$ un ensemble de graphes. Le graphe médian d'ensemble (*set median graph - smg*) de $\mathcal{S}$ est défini par :

$$smg = \arg\min_{g \in \mathcal{S}} \sum_{i=1}^{n} d(g, g_i) \tag{3.1}$$

Définition 3. Soit $d(.,.)$ une distance ou une mesure de dissimilarité entre deux graphes. Soit $\mathcal{S} = \{g_1, g_2, \ldots, g_n\}$ un ensemble de graphes. Soit $\mathcal{U}$ l'ensemble infini des graphes qui peuvent être construits à partir des labels de $\mathcal{S}$. Le graphe médian généralisé (*generalized median graph - gmg*) du sous-ensemble $\mathcal{S}$ est défini par :

$$gmg = \arg\min_{g \in \mathcal{U}} \sum_{i=1}^{n} d(g, g_i) \tag{3.2}$$

Dans les deux cas, lorsqu'ils sont utilisés comme échantillons d'apprentissage pour un processus de classification, ces prototypes ne tiennent compte que de la distribution intra-classe des données. Ce sont ainsi des prototypes davantage modélisants que discriminants. Dans nos travaux, nous avons étendu la notion de graphe médian par la proposition de nouveaux types de prototypes appelés graphes discriminants. Les définitions de ces graphes, ainsi que l'algorithme permettant de les générer sont décrits dans la sous-section suivante.

3.2.2 Contributions

Pour pallier le défaut des approches modélisantes, nous avons proposé dans [170] l'utilisation de prototypes discriminants (*discriminative graphs - dg*) pour la classification de graphes. La différence principale avec les graphes médians réside dans le critère utilisé pour générer les prototypes. Dans le cas des *dg*, ce sont les performances de classification évaluées sur un ensemble de graphes de

validation qui sont utilisées pour optimiser les graphes prototypes. L'information exploitée pour générer les prototypes dépasse ainsi la simple connaissance de la distribution intra-classe des données. Par analogie avec la terminologie utilisée dans la communauté de la sélection de caractéristiques, nous proposons donc une approche de type *wrapper*, qui inclut le critère final de performance dans le processus de sélection. Les prototypes sont définis de la façon suivante :

Définition 4. Soit N le nombre de classes d'un ensemble d'apprentissage $\mathcal{L}$. Soit $\mathcal{T}$ un ensemble de validation et soit $\Delta\left(\mathcal{T}, \{g_i\}_{i=1}^{N}\right)$ une fonction calculant le taux d'erreur obtenu par un classifieur 1-PPV sur $\mathcal{T}$ en utilisant les graphes prototypes $\{g_i\}_{i=1}^{N} \subset \mathcal{L}$ comme échantillons d'apprentissage. L'ensemble des *Set Discriminative Graphs* (SDG), composé des sdg_i de chaque classe est donné par :

$$\begin{aligned} SDG &= \{sdg_1, sdg_2, \ldots, sdg_N\} \\ &= \underset{\{g_i\}_{i=1}^{N} \subset \mathcal{L}}{\arg\min}\ \Delta(\mathcal{T}, \{g_i\}_{i=1}^{N}) \end{aligned} \tag{3.3}$$

Définition 5. Soit N le nombre de classes d'un ensemble d'apprentissage $\mathcal{L}$. Soit $\mathcal{U}$ l'ensemble infini des graphes qui peuvent être construits à partir des labels de $\mathcal{L}$. Soit $\mathcal{T}$ un ensemble de validation et soit $\Delta\left(\mathcal{T}, \{g_i\}_{i=1}^{N}\right)$ une fonction calculant le taux d'erreur obtenu par un classifieur 1-PPV sur $\mathcal{T}$ en utilisant les graphes prototypes $\{g_i\}_{i=1}^{N} \subset \mathcal{U}$ comme échantillons d'apprentissage. Alors l'ensemble des *Generalized Discriminative Graphs* GDG composé des gdg_i de chaque classe est donné par :

$$\begin{aligned} GDG &= \{gdg_1, gdg_2, \ldots, gdg_N\} \\ &= \underset{\{g_i\}_{i=1}^{N} \subset \mathcal{U}}{\arg\min}\ \Delta(\mathcal{T}, \{g_i\}_{i=1}^{N}) \end{aligned} \tag{3.4}$$

Ces deux définitions ont été étendues à la possibilité de générer plusieurs prototypes par classes, afin de mieux décrire la distribution des données.

Définition 6. Soit N le nombre de classes d'un ensemble d'apprentissage $\mathcal{L}$. Soit $\mathcal{U}$ l'ensemble infini des graphes qui peuvent être construits à partir des labels de $\mathcal{L}$. Soit M le nombre de prototypes par classe. Soit $\mathcal{T}$ un ensemble de validation et soit $\Delta\left(\mathcal{T}, \{g_i\}_{i=1}^{N}\right)$ une fonction calculant le taux d'erreur obtenu par un classifieur 1-PPV [7] sur $\mathcal{T}$ en utilisant les graphes prototypes $\{g_{ik}\}_{i=1,k=1}^{N,M} \subset \mathcal{U}$ comme échantillons d'apprentissage. Alors l'ensemble $MGDG$ composé des gdg_{ik} de chaque classe est donné par :

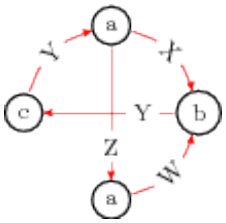
7. Dans ce cas, il est possible de considérer un classifieur des kPPV avec $k > 1$, et ainsi intégrer du rejet.

$$\begin{aligned} MGDG &= \{gdg_{11}, ..., gdg_{1M}, ..., gdg_{N1}, ..., gdg_{NM}\} \\ &= \underset{\{g_{ik}\}_{i=1,k=1}^{N,M} \subset \mathcal{U}}{\arg\min} \Delta\left(\mathcal{T}, \{g_{ik}\}_{i=1,k=1}^{N,M}\right) \end{aligned} \tag{3.5}$$

La recherche des prototypes ainsi définis est un processus d'optimisation. Dans [170], nous avons proposé de traiter ce problème d'optimisation par un Algorithme Génétique (AG) [87] dédié à la manipulation de graphes. Cette spécialisation originale a reposé sur les points suivants :

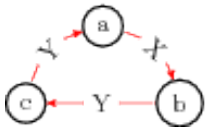
- le codage des individus représentant les solutions possibles du problème d'optimisation. Pour tous les types de prototypes proposés dans les définitions précédentes, un individu est représenté par un ensemble de $m \times N$ gènes correspondant aux graphes prototypes. Dans le cas de prototypes d'ensemble, les gènes sont simplement les indices des graphes sélectionnés dans l'ensemble d'apprentissage. Dans le cas des graphes généralisés, les gènes correspondent aux matrices d'adjacence des graphes ;
- une fonction évaluant le score d'un individu. Ces fonctions sont directement issues des définitions précédentes. Notons que quel que soit le type de prototype considéré, les calculs reposent sur un calcul de dissimilarité entre graphes. Dans [170], nous utilisons la distance proposée par [132], mais l'approche peut exploiter n'importe quel type de distance (la distance d'édition [39, 83] ou ses approximations [176], des distances basées sur le plus grand sous-graphe commun [43], des distances basées sur l'appariement de sous-graphes [172] ou des distances basées sur des unions de graphes [211]...) ;
- une stratégie de sélection. L'objectif de la sélection dans les AG est de sélectionner des individus pour former la génération suivante. Nous utilisons dans ce cadre une roue de loterie biaisée, en y ajoutant un mécanisme d'élitisme dans lequel les μ meilleurs individus sont préservés afin de garantir la convergence de l'algorithme ;
- des opérateurs génétiques dédiés. Le croisement utilisé pour tous les types de prototypes est un opérateur classique consistant à effectuer un échange de gènes entre individus à croiser, en respectant la distribution par classe. La mutation quant à elle, diffère en fonction des prototypes extraits. Dans le cas des prototypes d'ensemble, la mutation correspond simplement à changer l'indice d'un graphe prototype par un autre de la même classe. Dans le cas des prototypes généralisés, nous avons proposé un opérateur original consistant à appliquer aléatoirement un ensemble d'opérations d'édition (suppression, ajout ou modification des nœuds et des arcs) sur le graphe. Cet opérateur est détaillé dans [170] et illustré par la figure 3.2.

Toutes ces spécificités sont précisément décrites et illustrées dans l'annexe E. Les performances que permettent d'obtenir ces différents types de prototypes ont été évaluées sur quatre bases de graphes proposées dans la com-



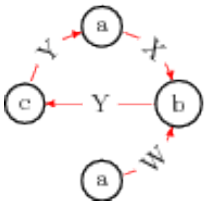
	1	2	3	4	
1		X		Z	a
2			Y		b
3	Y				c
4		W			a

(a) Initial graph



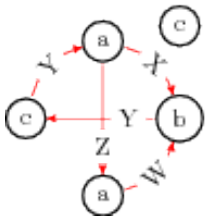
	1	2	3	
1		X		a
2			Y	b
3	Y			c

(b) Vertex deletion



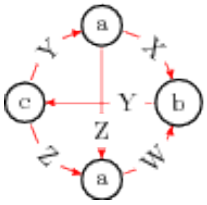
	1	2	3	4	
1		X			a
2			Y		b
3	Y				c
4		W			a

(c) Edge deletion



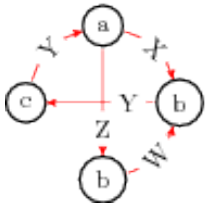
	1	2	3	4	5	
1		X		Z		a
2			Y			b
3	Y					c
4		W				a
5						c

(d) Vertex insertion



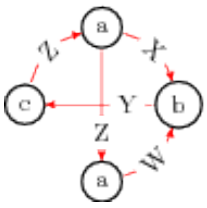
	1	2	3	4	
1		X		Z	a
2			Y		b
3	Y			Z	c
4		W			a

(e) Edge insertion



	1	2	3	4	
1		X		Z	a
2			Y		b
3	Y				c
4		W			b

(f) Vertex substitution



	1	2	3	4	
1		X		Z	a
2			Y		b
3	Z				c
4		W			a

(g) Edge substitution

FIGURE 3.2 – Illustration de l'opérateur de mutation proposé pour les graphes généralisés, tirée de [170]. La première ligne illustre un graphe de départ et sa matrice d'adjacence. La dernière colonne de la matrice contient le label des nœuds. Les sous-figures (b) à (g) présentent les différentes opérations d'édition possibles, avec leurs répercussions sur le graphe et sur la matrice.

%	*smg*		*gmg*		*sdg*		*gdg*	
	$\overline{Rec}$	σ	$\overline{Rec}$	σ	$\overline{Rec}$	σ	$\overline{Rec}$	σ
Base A	33.75	0.0	36.00	1.52	66.10	0.981	66.67	1.59
Base B	62.5	0.0	75	0.0	71.42	2.5	83.39	2.5
Base C	86.92	0.0	85.48	2.05	86.58	0.596	90.70	0.59
Base D	69.61	0.0	69.14	0.34	69.67	0.67	71.24	1.47

TABLE 3.1 – Résultats obtenus par les différents prototypes pour $M = 1$.

munauté. Les expériences ont d'abord concerné l'étude de la convergence de l'algorithme, en comparant les *smg* et *sdg* obtenus par l'algorithme proposé avec ceux déterminés par une recherche exhaustive. Les résultats ont montré que moins de 50 générations de l'AG étaient nécessaires pour obtenir des résultats identiques. Puis, les performances obtenues par les différents prototypes ont été comparées pour $M = 1$. Les résultats obtenus sont présentés dans le tableau 3.1. Ces résultats démontrent tout d'abord que les prototypes généralisés (*gmg*,*gdg*) sont plus performants que les prototypes d'ensemble (*smg*,*sdg*). Par ailleurs, ils confirment également la supériorité des prototypes discriminants sur les prototypes modélisants.

Nous avons également comparé les performances obtenues par les $MGDG$ à celles obtenues par un classifieur 1-PPV sur l'ensemble de la base d'apprentissage. Les résultats (tableau 3.2) ont prouvé qu'en augmentant le nombre de prototypes par classe, il était possible de dépasser les performances de l'algorithme de référence avec les $MGDG$. Cela démontre qu'exploiter les performances sur un ensemble de validation est un critère plus adapté pour la génération de prototypes que celui généralement utilisé pour calculer des graphes médians. Une analyse de la complexité temporelle est également proposée dans l'annexe E.

Enfin, cette contribution a aussi permis de mettre en exergue le fait que les algorithmes génétiques pouvaient être adaptés pour traiter des problèmes manipulant des graphes, ce qui, à notre connaissance, est rare dans la littérature [37, 136].

	BaseA		*BaseB*		*BaseC*		*BaseD*	
	gdg	1-NN	*gdg*	1-NN	*gdg*	1-NN	*gdg*	1-NN
Red (%)	92.52	0	50.71	0	86.67	0	76.3	0
Rec (%)	86.34	85.16	97.14	96.43	99.71	99.47	91.04	90.16

TABLE 3.2 – Taux de réduction ($Red = \left(100\left(1 - \frac{m \times N}{|Tr \cup Tv|}\right)\right)$ et taux de reconnaissance (Rec) obtenus par les *gdg* et un classifieur 1PPV sur la base d'apprentissage complète $Tr \cup Tv$.

L'approche proposée dans cette section permet ainsi de classifier des graphes. Toutefois, dans le contexte de l'analyse d'images de documents, un tel algorithme ne peut être appliqué que sur une entité isolée représentée par un graphe. Il faut donc avoir préalablement « segmenté » le graphe pour pouvoir

exploiter le classifieur. L'approche ne permet donc pas de résoudre l'un des principaux problèmes relatifs à l'analyse structurelle de documents, à savoir celui de la recherche d'occurrences d'objets présents dans un document complet. La section suivante aborde ce problème de localisation.

3.3 Isomorphismes de sous-graphes

3.3.1 Définition du problème et revue de l'existant

Dans le domaine de l'analyse de documents ou plus généralement de la vision par ordinateur, les graphes représentent généralement des objets à localiser dans un document ou une image. Pour ce faire, il est nécessaire d'avoir recours à des techniques d'appariement de graphes qui établissent une correspondance entre les sommets de deux graphes. Différents problèmes d'appariement de graphes existent, tels que l'isomorphisme de graphes, l'isomorphisme de sous-graphes, la recherche du plus grand sous-graphe commun ou la distance d'édition entre graphes. Nous nous intéressons ci-dessous au problème de l'isomorphisme de sous-graphes qui repose sur les définitions suivantes.

Définition 7. Un graphe attribué $\mathcal{G}$ est un *4-tuple* $\mathcal{G} = (V, E, \mu, \xi)$ tel que :
- V est l'ensemble des nœuds de $\mathcal{G}$;
- E est l'ensemble des arcs de $\mathcal{G}$, i.e. un ensemble de paires $e = (v_1, v_2)$ avec $v_1 \in V$ et $v_2 \in V$;
- $\mu : V \rightarrow L_V$ est une fonction affectant un *label* à un nœud, L_V étant l'ensemble des labels possibles pour les nœuds ;
- $\xi : E \rightarrow L_E$ est une fonction affectant un *label* à un arc, L_E étant l'ensemble des labels possibles pour les arcs.

Définition 8. Soit un graphe $\mathcal{G} = (V, E, \mu, \xi)$, un sous-graphe de $\mathcal{G}$ est un graphe $\mathcal{S} = (V_\mathcal{S}, E_\mathcal{S}, \mu_\mathcal{S}, \xi_\mathcal{S})$ tel que :
- $V_\mathcal{S} \subseteq V$;
- $E_\mathcal{S} \subseteq E$;
- $\mu_\mathcal{S}$ et $\xi_\mathcal{S}$ sont les restrictions de μ et ξ à $V_\mathcal{S}$ et $E_\mathcal{S}$, i.e. $\mu_\mathcal{S}(v) = \mu(v)$ et $\xi_\mathcal{S}(e) = \xi(e)$.

Notons qu'il existe des variantes de cette définition. En particulier, un sous-graphe $\mathcal{S}$ de $\mathcal{G}$ est appelé sous-graphe induit si $E_\mathcal{S} = E \cap (V_\mathcal{S} \times V_\mathcal{S})$. Cela implique que $\mathcal{S}$ contient tous les arcs $e \in E$ qui joignent des nœuds de $\mathcal{S}$.

Définition 9. Une fonction bijective $f : V \rightarrow V'$ est un isomorphisme entre un graphe $\mathcal{G} = (V, E, \mu, \xi)$ et un graphe $\mathcal{G}' = (V', E', \mu', \xi')$ si :
- $\mu(v) = \mu'(f(v))$ pour tout $v \in V$;
- pour tout arc $e = (v_1, v_2) \in E$, il existe un arc $e' = (f(v_1), f(v_2)) \in E'$ tel que $\xi(e) = \xi'(e')$, et pour tout $e' = (v'_1, v'_2) \in E'$, il existe un arc

$e = (f^{-1}(v'_1), f^{-1}(v'_2)) \in E'$ tel que $\xi(e) = \xi'(e')$.

Définition 10. Une fonction injective $f : V \to V'$ est un isomorphisme de sous-graphe d'un graphe $\mathcal{G} = (V, E, \mu, \xi)$ dans un graphe $\mathcal{G}' = (V', E', \mu', \xi')$ s'il existe un sous-graphe $\mathcal{S} \subseteq \mathcal{G}'$ tel que f est un isomorphisme de graphe de $\mathcal{G}$ vers $\mathcal{S}$:

- $\mu(v) = \mu(f(v))$ pour tout $v \in V$;
- pour tout arc $e = (v_1, v_2) \in E$, il existe un arc $e' = (f(v_1), f(v_2)) \in E'$ tel que $\xi(e) = \xi'(e')$.

Les définitions 9 et 10 caractérisent des appariements exacts, ainsi dénommés pour deux raisons. D'une part parce que la topologie des deux graphes (ou sous-graphes) doit être exactement la même, d'autre part parce que l'appariement nécessite une égalité stricte des labels.

De nombreux algorithmes ont été proposés dans la littérature pour résoudre de tels problèmes de recherche d'isomorphismes exacts. La plupart sont basés sur une recherche arborescente associée à une procédure de retour en arrière. Les approches diffèrent généralement (i) en fonction de l'ordre dans lequel les appariements partiels sont visités, par exemple en ajoutant une vérification de la cohérence des arcs [208] et (ii) en fonction des heuristiques qui sont utilisées pour élaguer l'arbre [64, 65, 193]. Ces heuristiques consistent à analyser les ensembles de nœuds adjacents à ceux contenus dans l'appariement courant. Il existe également quelques alternatives à ces recherches arborescentes, comme par exemple l'algorithme NAUTY proposé dans [144] qui s'appuie sur la théorie des groupes. Un état de l'art très complet et détaillé sur les méthodes d'appariement exact est proposé dans [62].

Toutes ces approches souffrent de deux limitations principales. La première est leur complexité algorithmique. À part dans le cas de l'isomorphisme exact de graphes, pour lequel il n'a pas été montré qu'il appartenait aux problèmes NP-*complets* ; et sauf pour des applications spécifiques, tous les problèmes d'appariement de graphes sont NP-*complets* et ont une complexité temporelle exponentielle dans le pire des cas [84]. Par ailleurs, le second et principal défaut de ces approches dans le cadre de nos travaux est leur sensibilité aux bruits et aux distorsions. En effet, dans le domaine de l'analyse de documents, les graphes représentent en général des objets qui peuvent être bruités ou déformés. Dans ce cas, les graphes résultants peuvent voir leur topologie ou leur étiquetage affectés par ce bruit. Vient alors la nécessité d'utiliser des algorithmes d'appariement tolérants aux erreurs, qui relâchent des contraintes sur *le matching*. Le problème de décision devient alors un problème d'optimisation dont le but est de trouver l'appariement qui minimise un coût tel que la distance d'édition [146]. Différentes alternatives ont été proposées dans la littérature pour résoudre ce problème. Certaines sont optimales, et assurent donc que l'appariement trouvé est optimal. C'est le cas des méthodes utilisant l'algorithme A^* combinées avec des heuristiques d'exploration de l'espace d'état. D'autres méthodes sont sous-optimales mais permettent de trouver des solutions dans un temps polynomial en exploitant par exemple la relaxation

probabiliste, les réseaux de neurones ou les algorithmes génétiques. Un état de l'art très complet de toutes ces méthodes est proposé dans [62].

Un cas particulier de problème d'appariement est la situation dans laquelle les graphes à apparier doivent être structurellement isomorphes, mais pour lesquels on tolère des différences entre attributs. Nous avons appelé un tel problème celui de l'isomorphisme de sous-graphes tolérant aux substitutions. Une approche permettant d'aborder ces problèmes consiste à modifier les algorithmes exacts, par une redéfinition de la fonction de compatibilité entre les nœuds et les arcs. Ainsi, en définissant deux seuils (un pour les nœuds, un pour les arcs), deux nœuds (resp. arcs) sont alors considérés comme compatibles si une distance entre leurs attributs est inférieure à ce seuil. Bien sûr, la difficulté principale est alors de définir la valeurs de ces seuils.

3.3.2 Contributions

Dans [31], nous avons abordé le problème de recherche d'isomorphisme de sous-graphes sous l'angle de la programmation mathématique (*Mathematical Programming* -MP) qui fournit un ensemble de solutions pour résoudre des problèmes d'optimisation. Plus précisément, la solution retenue a été celle de la Programmation Linéaire en Nombres Entiers (PLNE) [152, 187] qui est une restriction de la programmation mathématique permettant de modéliser des problèmes spécifiques, et pour laquelle de nombreux algorithmes de résolution existent et sont constamment améliorés par la communauté. Il existe ainsi de nombreux solveurs qui permettent de résoudre des problèmes de PLNE. Ce paradigme a été utilisé dans de très nombreux domaines, allant de l'énergie à la finance, en passant par les télécommunications ou la logistique. La PLNE est reconnue comme une des techniques les plus efficaces pour traiter des problèmes d'optimisation NP-complets [111, 11, 99].

Un programme mathématique est une modélisation d'un problème d'optimisation sous la forme d'une fonction objectif et d'un ensemble de contraintes. Dans le cas d'un programme linéaire, la fonction objectif et les contraintes sont des combinaisons linéaires des paramètres du problème d'optimisation. Le cas spécifique de la PLNE impose en plus que les solutions recherchées soient entières. La forme générale d'un programme linéaire en nombres entiers est donc la suivante :

$$\min_x \ c^t x \tag{3.6}$$

$$\text{sous la contrainte } Ax \leq b \tag{3.7}$$

$$x \in C \subseteq \mathbb{Z}^n \tag{3.8}$$

Dans cette formulation, $c \in \mathbb{R}^n, A \in \mathbb{R}^{m.n}, b \in \mathbb{R}^m$ sont les données décrivant le problème. Le vecteur x de n variables est la solution recherchée du problème, il appartient à $\mathbb{Z}^n$ dans le cas de la programmation en nombres entiers (3.8). Les variables de A permettent d'exprimer des contraintes linéaires (3.7). Une solution valide pour le problème est un vecteur x tel que les contraintes (3.7) et (3.8) sont respectées. Une telle solution est dite réalisable. Trouver

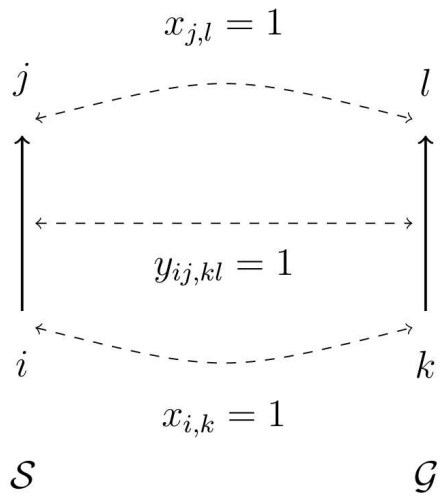


FIGURE 3.3 – Un exemple d'appariement. $\mathcal{S}$ et $\mathcal{G}$ contiennent chacun un arc unique, respectivement ij et kl. La solution suivante est représentée sur la figure : $x_{i,k} = 1$ (resp. $x_{j,l} = 1$, $y_{ij,kl} = 1$), i.e. i (resp. j, ij) est apparié avec k (resp. l, kl). Réciproquement, i (resp. j) n'est pas apparié avec l (resp. k), donc $x_{i,l} = 0$ (resp. $x_{j,k} = 0$).

une solution optimale consiste alors à minimiser la fonction objectif (3.6) sur l'ensemble des solutions réalisables.

La résolution d'un problème d'optimisation utilisant la programmation linéaire en nombres entiers repose donc essentiellement sur la formulation de celui-ci sous la forme d'une fonction objectif et d'un ensemble de contraintes. Pour modéliser le problème de la recherche d'isomorphisme de sous-graphes, nous avons proposé d'utiliser des variables binaires. La solution au problème prend donc ses valeurs dans $\{0,1\}^n$. Comme l'illustre la figure 3.3, deux types de variables sont définis :

- pour chaque nœud $i \in V_\mathcal{S}$ et pour chaque nœud $k \in V_\mathcal{G}$, une variable $x_{i,k}$ est définie telle que $x_{i,k} = 1$ si les nœuds i et k sont appariés, 0 s'ils ne le sont pas ;
- pour chaque arc $ij \in E_\mathcal{S}$ et pour chaque arc $kl \in E_\mathcal{G}$, une variable $y_{ij,kl}$ est définie telle que $y_{ij,kl} = 1$ si les arcs ij et kl sont appariés, 0 s'ils ne le sont pas.

Formellement, cette définition des variables du problème s'écrit donc :

$$x_{i,k} \in \{0,1\} \quad \forall i \in V_\mathcal{S}, \forall k \in V_\mathcal{G} \tag{3.9}$$
$$y_{ij,kl} \in \{0,1\} \quad \forall ij \in E_\mathcal{S}, \forall kl \in E_\mathcal{G} \tag{3.10}$$

Soient $\mathcal{S} = (V_\mathcal{S}, E_\mathcal{S})$ et $\mathcal{G} = (V_\mathcal{G}, E_\mathcal{G})$ les deux graphes à apparier. Supposons connues les fonctions de coût $c_V : V_\mathcal{S} \times V_\mathcal{G} \to \mathbb{R}^+$ et $c_E : E_\mathcal{S} \times E_\mathcal{G} \to \mathbb{R}^+$ donnant respectivement les coûts d'appariement des nœuds et des arcs telles que l'appariement de i et k (i.e. $x_{i,k} = 1$) a un coût $c_V(i,k)$, alors que leur non-appariement (i.e. $x_{i,k} = 0$) a un coût 0. Le coût global de l'appariement

peut alors s'écrire comme une combinaison linéaire $c_V(i,k) * x_{i,k}$. De façon similaire, le coût entre deux arcs $ij \in E_\mathcal{S}$ et $kl \in E_\mathcal{G}$ est $c_E(ij,kl) * y_{ij,kl}$. Dans ce cadre, la fonction objectif du problème d'appariement de $\mathcal{S} = (V_\mathcal{S}, E_\mathcal{S})$ avec un sous-graphe de $\mathcal{G} = (V_\mathcal{G}, E_\mathcal{G})$ peut s'écrire :

$$\min_{x,y} \left(\sum_{i \in V_\mathcal{S}} \sum_{k \in V_\mathcal{G}} c_V(i,k) * x_{i,k} + \sum_{ij \in E_\mathcal{S}} \sum_{kl \in E_\mathcal{G}} c_E(ij,kl) * y_{ij,kl} \right) \tag{3.11a}$$

Naturellement, la minimisation de cette expression n'est pas suffisante pour décrire le problème d'isomorphisme de sous-graphes, car aucune contrainte concernant le respect de la topologie n'est intégrée (le coût serait nul si aucun nœud et aucun arc n'était apparié). Cet aspect est géré par l'intermédiaire des contraintes. Celles-ci, qui sont illustrées dans [31], prennent la forme suivante :

$$\sum_{k \in V_\mathcal{G}} x_{i,k} = 1 \quad \forall i \in V_\mathcal{S} \tag{3.11b}$$

$$\sum_{kl \in E_\mathcal{G}} y_{ij,kl} = 1 \quad \forall ij \in E_\mathcal{S} \tag{3.11c}$$

$$\sum_{i \in V_\mathcal{S}} x_{i,k} \leq 1 \quad \forall k \in V_\mathcal{G} \tag{3.11d}$$

$$\sum_{kl \in E_\mathcal{G}} y_{ij,kl} = x_{i,k} \quad \forall k \in V_\mathcal{G}, \forall ij \in E_\mathcal{S} \tag{3.11e}$$

$$\sum_{kl \in E_\mathcal{G}} y_{ij,kl} = x_{j,l} \quad \forall l \in V_\mathcal{G}, \forall ij \in E_\mathcal{S} \tag{3.11f}$$

Les équations (3.9) à (3.11f) constituent le programme linéaire en nombres entiers qui est utilisé pour résoudre le problème d'isomorphisme de sous-graphes. Dès lors que la recherche d'isomorphisme est modélisée sous la forme d'un programme linéaire, il est possible de la résoudre en utilisant un solveur mathématique. Dans cette étude, nous utilisons un solveur disponible sous licence CPL appelé SYMPHONY et décrit dans [169]. Pour résoudre une instance du problème, le solveur dispose d'une batterie de méthodes qui améliorent ou s'inspirent du Branch and Bound (Séparation et Évaluation), et sont proposées par la communauté de la programmation mathématique. Notons que le problème ainsi modélisé peut être infaisable, s'il n'existe pas d'isomorphisme entre $\mathcal{S}$ et un sous-graphe de $\mathcal{G}$. Dans ce cas, le solveur ne retourne pas de solution. Si au moins un isomorphisme existe, le solveur retournera uniquement la solution de coût minimal, i.e. le meilleur isomorphisme. Or, il peut être intéressant, pour certains cas d'usage, de retourner une liste d'isomorphismes. Dans ce cadre, qui sera illustré en 3.4.1, nous avons proposé une stratégie consistant à appliquer itérativement le modèle, en supprimant les solutions trouvées de l'ensemble des solutions possibles.

L'évaluation des performances des algorithmes de recherche d'isomorphismes de sous-graphes est un problème difficile qui implique de disposer (i) de bases

de graphes et de sous-graphes et (ii) de la vérité terrain concernant les isomorphismes existants. Pour des applications du monde réel, générer cette vérité terrain est une tâche complexe et « chronophage », particulièrement dans le cas d'isomorphismes inexacts car il faut alors considérer simultanément les données brutes et leur représentation sous forme de graphes. Dans la littérature, la plupart des articles traitant du problème d'isomorphisme de sous-graphes proposent une évaluation reposant sur des bases de données synthétiques, comme la base VF décrite dans [63] ou les bases du TC 15 de l'IAPR [8]. Toutefois, toutes ces bases sont, à notre connaissance, dédiées à des appariements exacts, et reposent sur un étiquetage des nœuds et des arcs avec des attributs nominaux et non numériques ou vectoriels. Elles ne permettent donc pas d'évaluer des algorithmes tolérants aux substitutions. Dans ce contexte, nous avons choisi de mener nos expérimentations sur des bases de données synthétiques adaptées aux problèmes tolérants aux substitutions et sur une base issue d'une application réelle [9]. Nous avons donc implémenté un générateur de graphes synthétiques [10]. Pour les raisons évoquées précédemment, nous n'avons pu comparer nos résultats avec les deux algorithmes références de la littérature que sont VF2 [65] et LAD [193] que dans le cas d'appariements exacts. Les résultats obtenus ont montré que l'algorithme proposé n'était pas compétitif avec les approches de la littérature pour effectuer une recherche exacte. Ce résultat attendu s'explique, d'une part, par le fait que l'approche proposée n'a pas vocation première à traiter ce genre de problèmes, et d'autre part, parce que nous n'avons pas cherché à optimiser le fonctionnement du solveur dans ce cadre. En revanche, les résultats obtenus pour la recherche d'isomorphismes tolérants aux substitutions ont montré que l'approche permettait de résoudre le problème sans augmentation sensible des temps de calcul alors qu'à notre connaissance, il n'existe pas de solutions permettant naturellement de résoudre de tels problèmes.

3.4 Applications à l'analyse de documents graphiques

Dans cette section, deux systèmes reposant sur des représentations structurelles et dédiés au domaine de l'analyse d'images de documents graphiques sont présentés. Plus précisément, nous nous intéressons ici à la problématique du traitement des symboles sur de tels documents. Depuis quelques années, la reconnaissance de ces entités symboliques est devenue la problématique la plus prolifique en termes de littérature dans la communauté de l'analyse de documents graphiques. Cette évolution s'explique par la maturité des outils dédiés aux problématiques spécifiques à ces documents, telles que la séparation texte/graphique ou la vectorisation, qui ont longtemps été au cœur des préocupations des chercheurs du domaine, mais aussi et surtout par l'évolution des besoins et par l'importance que la reconnaissance des symboles revêt dans le cadre de l'indexation des documents graphiques. En effet, que le domaine d'application concerne l'architecture, la cartographie, l'électronique, la

8. http ://www.greyc.ensicaen.fr/iapr-tc15
9. Ces résultats seront présentés en 3.4.1.
10. Ce générateur et les bases utilisées dans ces travaux seront prochainement rendus disponibles à la communauté.

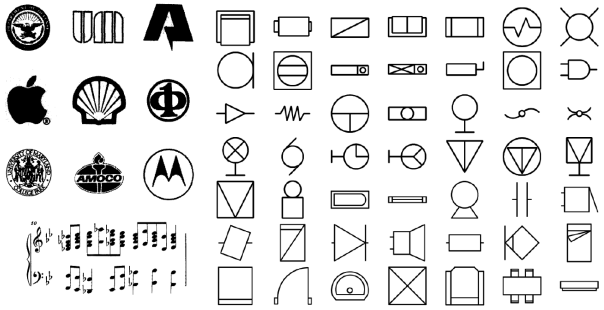


FIGURE 3.4 – Quelques exemples de symboles. Figure tirée de [3]

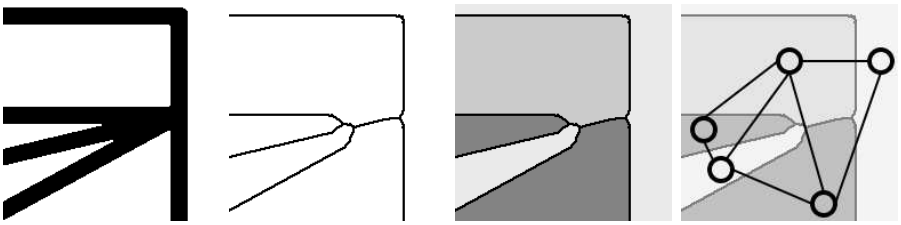
mécanique ou tout autre domaine d'ingénierie, voire même les documents du patrimoine, de nombreuses notations symboliques spécifiques au domaine sont présentes dans les documents (figure 3.4). Le haut niveau sémantique de l'information que ces symboles véhiculent rend leur reconnaissance indispensable dans le cadre d'un système d'analyse de documents graphiques.

Dans le domaine du traitement des données symboliques, comme dans celui de la recherche d'images par le contenu en général, on oppose généralement les approches statistiques, selon lesquelles les objets sont représentés par des vecteurs de caractéristiques, aux approches structurelles, selon lesquelles la modélisation des objets repose sur des graphes. Dans le domaine de la recherche de symboles, les approches structurelles ont toujours été très présentes de par la nature même des symboles, souvent constitués de sous-parties, et de par la capacité intrinsèque de telles modélisations à aider à la segmentation.

Dans la suite de cette section, les résultats obtenus en utilisant l'approche décrite en 3.3 dans le cadre d'un problème de localisation de symboles sont d'abord présentés. L'approche est appliquée à une représentation structurelle à base de graphes d'adjacence de régions proposée dans le cadre de la thèse d'Hervé Locteau [129]. Nous décrivons ensuite une autre application originale reposant sur la même modélisation et utilisant des techniques de fouilles de données pour construire une représentation des documents sous la forme de « sacs de symboles » [18]. Cette représentation est exploitée à des fins de classification ou d'indexation de documents.

3.4.1 Détection de symboles

La détection de symboles est un des problèmes relevant de l'analyse d'images de documents. Ce type de problème revêt une difficulté supérieure à celui de la reconnaissance de symboles isolés dans la mesure où il est nécessaire de simultanément segmenter et reconnaître le symbole. De fait, si la littérature abordant la reconnaissance de symboles est abondante, très peu d'approches sont proposées pour la détection de symboles [165, 181]. Dans cette sous-section, nous présentons un système dont l'objectif est la détection de symboles dans des images de documents graphiques. L'approche proposée s'appuie sur l'extraction d'une représentation structurelle à base de graphes d'adjacence de régions. De tels graphes sont reconnus pour être topologiquement plus stables en présence de bruit que les graphes exploitant le squelette des formes qui, eux, sont



(a) Image originale (b) Squelette de l'image (a) (c) Composantes connexes de l'image (b) (d) Graphe d'adjacence de régions

FIGURE 3.5 – Création du graphe d'adjacence de régions

fortement perturbés par le bruit. En exploitant une telle représentation, la recherche des occurrences d'un symbole modèle dans un document cible devient naturellement un problème de recherche d'isomorphisme de sous-graphes, avec la particularité de devoir être tolérant aux erreurs d'étiquettes, puisque dans un problème réel, les valeurs des labels seront altérées, tant pour les nœuds que pour les arcs. Notons que lorsque le bruit est trop important, la structure du graphe peut également être modifiée. Nous reviendrons sur ces aspects dans les perspectives de nos travaux.

La première étape du système proposé dans [31] consiste à construire une représentation structurelle du document. Les graphes d'adjacence de régions sont des structures de données adaptées dans ce cadre car elles permettent la modélisation des relations topologiques entre les régions extraites grâce à un processus de segmentation. Nous traitons des images de documents techniques (images binaires) où la composante blanche est associée au fond tandis que les composantes noires correspondent à la partie graphique. La segmentation de telles images peut être obtenue par étiquetage des composantes [49]. Cependant, afin d'obtenir une représentation fine des relations d'adjacence pour chaque paire de régions, l'image binaire est soumise à une squelettisation [72]. On fait alors correspondre à chaque composante blanche de cette image squelettisée un nœud dans le graphe en construction. Par ailleurs, un parcours des branches du squelette est exploité pour déterminer les relations d'adjacence entre les régions deux à deux. Cette relation d'adjacence est matérialisée par la création d'un arc entre les nœuds associés aux régions correspondantes. La figure 3.5 illustre, à partir d'un extrait d'image de document, le processus de construction du graphe d'adjacence de régions.

Afin de caractériser les nœuds représentant les régions et de préciser la nature des relations d'adjacence, le graphe est étiqueté. Plusieurs types de caractéristiques ont été proposés dans la littérature pour décrire les formes et les relations spatiales [202]. Parmi les nombreux descripteurs de formes proposés dans la littérature [2], les moments de Zernike [200] permettent d'atteindre de bonnes performances lors de la reconnaissance de formes soumises à des transformations affines ou des dégradations. Un vecteur de caractéristiques composé des 24 premiers moments de Zernike extraits de chaque composante connexe et caractérisant la forme est donc utilisé pour étiqueter les nœuds correspondants dans le graphe. Le graphe construit est dirigé et les attributs affectés aux arcs

FIGURE 3.6 – Exemples d'images de la base `floorplans` correspondant à différents fonds de plan.

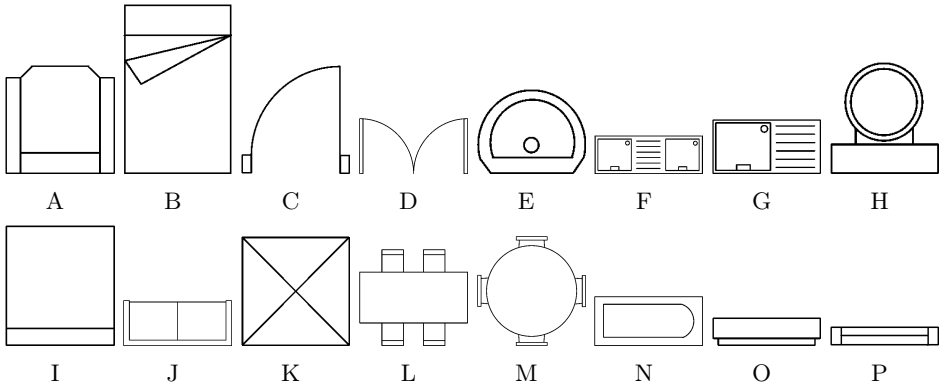


FIGURE 3.7 – Modèles des symboles recherchés.

(*source* → *destination*) sont :

- une caractéristique liée au rapport des surfaces des composantes associées aux nœuds source et destination ;
- une caractéristique liée à la distance entre les centres de gravité des régions associées aux nœuds origine et destination.

Les données utilisées pour évaluer l'approche proposée sont extraites de la base `floorplans`[11]. Cette base est constituée de données synthétiques représentant différentes dispositions de symboles placés sur 10 fonds de plans architecturaux. Notre évaluation se base sur 200 images synthétiques de plans architecturaux correspondant aux 20 premières dispositions proposées pour chacun des fonds. Des exemples de ces images sont proposés sur la figure 3.6.

La tâche associée à cette base de données consiste à retrouver les occurrences des 16 symboles modèles présentés sur la figure 3.7.

Grâce à une interface graphique développée pour l'occasion, il a été pos-

11. `http://mathieu.delalandre.free.fr/projects/sesyd/`

sible de constituer une vérité terrain pour la recherche d'isomorphismes de sous-graphes en identifiant au sein des 200 représentations structurelles les sous-graphes correspondant à des occurrences de symboles. Nous avons ainsi pu identifier que la base de plans contenait 5609 occurrences de symboles, soit environ 28 symboles par document en moyenne. Les sous-graphes correspondant aux symboles contenaient en moyenne 4 nœuds et 7 arcs. En comparaison, les représentations structurelles des plans complets contiennent en moyenne 121 nœuds et 525 arcs.

Dans une première expérimentation, nous avons recherché, dans chacun des plans, l'isomorphisme de coût minimal pour chacun des symboles modèles sur chacun des plans. Les résultats quantitatifs obtenus indiquent que sur les 3200 recherches d'occurrences (16×200), 1612 symboles ont été correctement détectés. 380 occurrences ont été partiellement détectées (au moins un nœud apparié à bon escient et au moins un nœud apparié à mauvais escient). Pour 453 recherches, le système a commis une erreur en ne trouvant pas d'occurrence du symbole. Enfin, pour 755 recherches, le symbole n'apparaissait pas dans le document.

Étant donné les résultats obtenus lors de la recherche d'une unique occurrence et considérant le fait qu'un même symbole est susceptible d'apparaître à plusieurs reprises sur un même document, nous avons souhaité évaluer la recherche de plusieurs isomorphismes. Compte tenu du fait qu'une composante connexe ne peut appartenir qu'à un seul symbole, nous avons, dans cette expérimentation, paramétré la recherche de telle sorte que soit exclu des solutions réalisables tout isomorphisme faisant apparaître un nœud déjà apparié dans un isomorphisme précédent.

Le tableau 3.3 présente les résultats d'une recherche de 50 occurrences de chaque symbole modèle pour chacun des 200 plans de la base `floorplans`. Même s'il persiste, comme dans le cas de la recherche d'une unique occurrence, des disparités entre classes, globalement, on retrouve exactement 62,7% et partiellement 29,2% des 5609 occurrences de symboles réellement présentes dans les documents. En estimant qu'une correspondance partielle suffit à considérer que l'occurrence du symbole est détectée, on atteint un rappel de 92% pour une précision de 7%.

La mauvaise précision obtenue est due au nombre important de détections imposé au système, qui est largement supérieur au nombre de symboles réellement présents (recherche de 50 occurrences par type de symbole pour chaque plan). Pour diminuer ce nombre de fausses détections, une stratégie de rejet a été mise en œuvre. Elle détermine, par un apprentissage supervisé, un seuil sur le coût d'appariement. Ce seuil est déterminé par classe, par une optimisation de la F-mesure obtenue sur une base de validation. Cette stratégie a permis d'augmenter la valeur de la précision à 71%, pour un rappel qui est maintenu à 83%. Le seuil appris sur chacune des classes, ainsi que les performances obtenues sont donnés dans le tableau 3.4.

3.4.2 Classification et indexation de documents

Dans [18], nous avons proposé une autre exploitation des représentations structurelles, à des fins de classification et d'indexation de documents gra-

Symbol	Recall (%)	Precision (%)
A	88	8
B	96	10
C	98	9
D	80	1
E	100	2
F	100	14
G	100	7
H	100	5
I	93	5
J	92	6
K	83	30
L	100	12
M	100	8
N	100	2
O	92	5
P	86	24
overall	92	7

TABLE 3.3 – Précision et rappel par classe de symboles lors de la recherche de 50 occurrences.

phiques complets. Le système proposé dans ce cadre est illustré sur la figure 3.8. L'approche repose sur l'utilisation de techniques de fouilles de graphes, dont le but est de faire émerger de nouvelles connaissances à partir d'un ensemble de données. Plus précisément, l'algorithme utilisé recherche des sous-structures avec l'objectif d'identifier, selon les informations encodées, des motifs fréquents ayant un rôle fonctionnel : par exemple une propriété des composés chimiques présentant un motif particulier, un gène responsable d'une pathologie dans une séquence, des motifs vecteurs de sens dans les documents. Les techniques reposent sur la satisfaction de contraintes telles qu'une fréquence d'apparition minimale et une confiance minimale dans le cas de règles d'association. Une revue de la littérature du domaine est proposée dans [18]. Les motifs fréquents sont alors utilisés comme lexique sur la base duquel les documents complets sont décrits sous la forme de « sacs de symboles ». Nous nous inspirons dans ce cadre des travaux utilisant des « sacs de mots » en analyse de texte [183], des « sacs de caractéristiques » [192, 77] ou encore des « paquets de chaînes de caractéristiques » [178] en indexation d'images. Ainsi, un document est caractérisé par un vecteur précisant la présence ou l'absence des différents motifs fréquents extraits de manière non supervisée dans la collection de documents. Une pondération *tf-idf* est utilisée pour enrichir la représentation.

Dans le cadre de l'application de ces techniques sur des documents graphiques, nous nous sommes appuyés, comme pour les travaux décrits en 3.4.1, sur la représentation à base de graphes d'adjacence de régions proposée dans la thèse d'Hervé Locteau [129]. Pour pouvoir utiliser les techniques de fouilles sur de telles représentations, les nœuds doivent être étiquetés avec des labels nominaux. Un algorithme de classification non supervisée s'appuyant sur les

Symbol	Matching cost threshold	Recall (%)	Precision (%)
A	2.706	80	77
B	3.041	90	81
C	0.489	70	49
D	0.827	80	5
E	1.136	100	100
F	2.215	100	100
G	1.959	53	41
H	2.418	90	100
I	0.857	90	25
J	1.249	84	82
K	2.442	74	89
L	3.499	100	86
M	2.590	99	100
N	0.970	84	70
O	0.279	56	45
P	3.079	86	86
overall		83	71

TABLE 3.4 – Précision et rappel par classe de symboles lors de la recherche de 50 occurrences avec la stratégie de rejet.

descripteurs de formes extraits des régions est utilisé pour ce faire. Puis, chaque cluster se voit affecté d'un label qui est utilisé pour étiqueter les nœuds. Un algorithme de fouille de graphe de la littérature [117] est ensuite exploité pour rechercher, dans le graphe, l'ensemble des sous-graphes fréquents. La figure 3.9 illustre quelques sous-graphes fréquents extraits d'un plan de réseau France Telecom. Elle montre que non seulement les symboles propres au domaine concerné sont extraits, mais en outre, l'algorithme a été en mesure d'extraire les chaînes de caractères associées à ces symboles. Cette première contribution permet ainsi d'extraire en partie la sémantique présente au sein du document.

L'approche a été évaluée sur un ensemble de documents techniques composé de 30 images de plans de réseau France Telecom, 25 images de schémas électroniques, et 5 images de plans architecturaux. Des tests ont été menés en classification supervisée, en exploitant un classifieur SVM. Les résultats, bien qu'obtenus sur des bases de petites tailles, ont atteint 90% de bonne classification, montrant la pertinence de l'approche, sachant qu'aucune connaissance *a priori* n'a été injectée dans le processus. C'est le système lui même qui découvre les caractéristiques à utiliser pour décrire le document, ce qui constitue une rupture par rapport aux approches classiques et qui permet clairement d'aller dans le sens de la généricité.

Des expérimentations préliminaires ont également été menées pour l'indexation de bases de documents utilisant la représentation en sacs de symboles, à des fins d'interrogation. Dans le système proposé, l'utilisateur peut rechercher dans la base de documents un ensemble de documents sur la base d'une

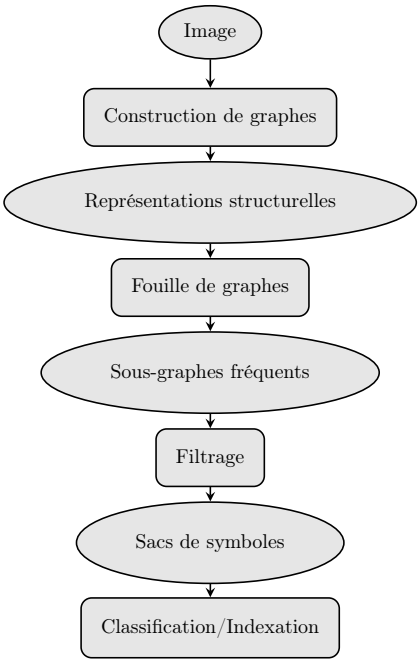


FIGURE 3.8 – Processus de fouille pour l'indexation et la classification d'images de documents.

requête exprimée par un extrait de documents, pour extraire des documents similaires au sens de leur contenu sémantique.

3.5 Discussion et problèmes ouverts

Dans ce chapitre, nous avons abordé des problématiques liées à la reconnaissance structurelle de formes avec des applications à l'analyse de documents graphiques. Deux contributions fondamentales ont été présentées et positionnées par rapport à la littérature du domaine. Elles concernent la classification supervisée de graphes et la recherche d'isomorphismes de sous-graphes tolérants aux erreurs de substitution. Deux applications liées à l'exploitation des graphes en analyse de documents graphiques ont ensuite été décrites. La première est une application de l'approche proposée pour la recherche d'isomorphisme à une tâche de localisation de symboles dans des documents graphiques, problème encore rarement abordé dans la communauté. La seconde concerne l'exploitation de techniques de fouille de graphes à des fins de classification supervisée ou d'indexation de base de documents.

Dans cette section, nous évoquons les perspectives directement issues de ces travaux. Une vision plus générale sera proposée dans le chapitre 5, avec en particulier les aspects liés au domaine applicatif de la reconnaissance de documents. Ces perspectives se déclinent donc ici essentiellement sur les aspects fondamentaux relatifs aux deux premières contributions.

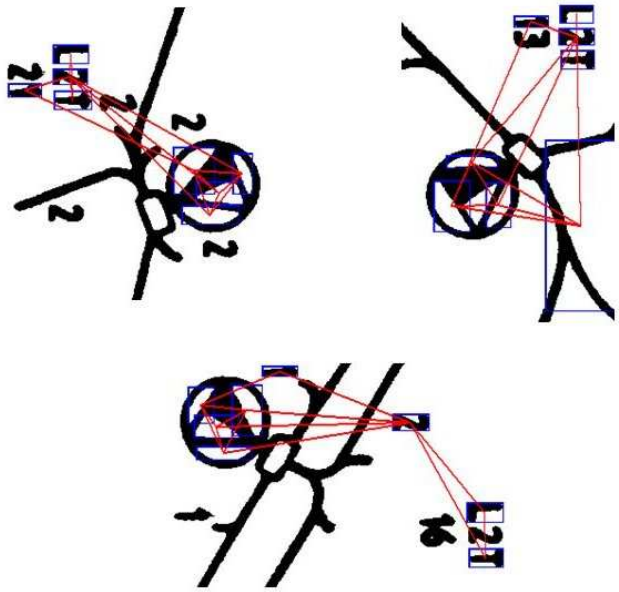


FIGURE 3.9 – Exemples d'occurrences d'un sous-graphe fréquent maximal.

3.5.1 Classification de graphes

Les résultats présentés dans [170] ont montré l'intérêt de prendre en considération un critère discriminant, l'espace des graphes généralisés et plusieurs représentants par classe dans le cadre de la génération de prototypes pour la classification de graphes. Toutefois, dans l'approche proposée, aucun mécanisme n'est mis en œuvre pour lui garantir de bonnes capacités en généralisation. Dans ce cadre, nous pensons qu'il pourrait être intéressant, lors de la génération des prototypes, d'intégrer un terme de régularisation dans le processus d'optimisation. Dans ce contexte, des approches d'optimisation multiobjectif telles que celles que nous évoquerons dans le chapitre suivant pourraient alors être considérées. Ces perspectives sont à l'intersection de ces travaux sur les graphes et de ceux qui sont proposés dans le cadre du projet LeMOn qui sera décrit dans le chapitre suivant.

L'intégration d'un critère de rejet dans l'approche est aussi une piste qu'il ne faut pas négliger. Le rejet est en effet très important dans les applications du monde réel. Le critère de génération des prototypes deviendrait là aussi multi-dimensionnel. Nous avons proposé de premiers travaux dans ce cadre dans [171]. Ceux-ci seront abordés en 4.4.1.

Une autre perspective liée à ces graphes prototypes repose sur l'utilisation de la technique du « Graph Embedding » [177] consistant à représenter un graphe par l'ensemble des distances entre ce graphe et un ensemble de graphes prototypes, en général les graphes médians.

3.5.2 Recherche d'isomorphisme

L'approche de recherche d'isomorphisme de sous-graphes décrite en 3.3 et appliquée au problème de localisation de symboles en 3.4.1 a prouvé qu'elle était capable de tolérer des modifications dans des étiquetages vectoriels et numériques des nœuds et des arcs. Elle permet ainsi de résoudre des problèmes que la littérature n'avait pas encore abordés directement, comme en témoigne d'ailleurs l'absence de bases de données avec de telles propriétés. L'une de nos premières perspectives consiste donc à mettre à disposition les données étiquetées que nous avons générées, par l'intermédiaire du site du TC 15 [12] par exemple. Ceci permettra à la communauté de comparer ses résultats à ceux que nous avons obtenus. Par ailleurs, nous planifions également de fournir l'environnement logiciel de recherche d'isomorphisme que nous avons proposé, celui-ci étant basé sur des logiciels libres.

Outre ces perspectives de diffusion à la communauté scientifique, d'autres axes de travail plus fondamentaux sont envisagés. Le premier consiste à modifier la version courante du solveur pour lui permettre de traiter des instances de tailles plus importantes (de l'ordre du millier de nœuds pour $\mathcal{G}$), ce qui constitue un challenge pour la communauté depuis longtemps [107]. En effet, dans la version courante du solveur, certaines instances atteignent des tailles telles que leur représentation en mémoire est trop volumineuse pour permettre la résolution. Une amélioration possible est d'utiliser la génération de colonnes, une technique de la communauté de programmation mathématique, qui consiste à considérer une formulation de départ dont le nombre de variables est largement réduit, puis de résoudre en ajoutant les variables qui manquent lorsqu'elles deviennent nécessaires pour explorer des solutions.

L'autre piste envisagée, la plus importante en termes d'applications mais aussi la plus complexe, consiste à généraliser la formulation proposée à des problèmes pour lesquels des modifications de la topologie des graphes pourraient être tolérées. Ce contexte nécessite de proposer une formulation robuste à l'absence dans $\mathcal{G}$ de sommets ou d'arcs pouvant être associés à ceux du graphe $\mathcal{S}$. La solution envisagée consiste à procéder par l'ajout direct dans $\mathcal{G}$ des éléments non appariés de $\mathcal{S}$. Cet ajout se traduit par de nouvelles variables de décision qui correspondent à ces ajouts de nœuds et d'arcs. La fonction objectif (equation 3.12) est alors modifiée pour prendre en compte ces modifications, auxquelles il faut alors attribuer des coûts (qu'il faudra idéalement apprendre).

$$\begin{aligned} \min \quad & \textstyle\sum_{i \in V_{\mathcal{S}}} \sum_{k \in V_{\mathcal{G}}} d(i,k) * x_{i,k} + \sum_{ij \in E_{\mathcal{S}}} \sum_{kl \in E_{\mathcal{G}}} d(ij,kl) * y_{ij,kl} \; + \\ & \textstyle\sum_{i \in V_{\mathcal{S}}} c(i) * u_i + \sum_{ij \in E_{\mathcal{S}}} c(ij) * e_{ij} \end{aligned} \tag{3.12}$$

Avec une telle modification, la fonction objectif intègre une tolérance aux modifications de structure. Évidemment, les contraintes du programme linéaire doivent également être revisitées pour intégrer cette tolérance. Une première proposition a été faite dans [30], elle doit maintenant être évaluée, ce qui pose de nouveau le problème des données et de leur vérité terrain associée tout comme celui de l'utilisation d'approches concurrentes. De plus, notons que ces travaux

12. http ://www.greyc.ensicaen.fr/iapr-tc15/

se rapprochent alors de la notion de distance d'édition entre graphes. L'outil pourrait d'ailleurs être utilisé pour calculer des dissimilarités entre graphes, sans considérer les sous-graphes.

Chapitre 4

Documents et optimisation multiobjectif

4.1 Introduction

L'introduction générale de ce mémoire a souligné la variabilité et la complexité des problématiques d'analyse d'images de documents. Elle a montré que la conception d'un système flexible et performant requiert le développement de nombreux composants logiciels inter-opérants dont il faut en outre maîtriser le réglage des paramètres et l'enchaînement (éventuellement en considérant des cycles) pour obtenir les meilleures performances possibles. Cette notion de performance suscite immédiatement la question du choix du(des) critère(s) utilisé(s) pour évaluer les composants du système et le système dans sa globalité. Par ailleurs, de tels critères sont également fondamentaux dans une optique d'optimisation des paramètres du système.

L'analyse de l'état de l'art du domaine de l'analyse de documents montre que la plupart des systèmes sont aujourd'hui conçus, évalués et réglés au regard d'un critère unique. Il peut s'agir du taux de reconnaissance pour un système de reconnaissance, de l'erreur quadratique en approximation polygonale ou de la F-mesure pour un système de *spotting* de mots ou de symboles. Pour illustrer ce constat, on peut mentionner les campagnes d'évaluation récentes menées en reconnaissance de l'écriture manuscrite (RIMES [88]), en reconnaissance de symboles (EPEIRES [73]) ou encore les concours de vectorisation menés à l'occasion des conférences internationales *Graphic RECognition* (GREC [6]). Dans chacun des cas, les métriques exploitées pour évaluer et comparer les approches sont aujourd'hui scalaires.

Or, plusieurs critères, souvent antagonistes, sont généralement importants au regard de l'utilisateur dans les applications du monde réel. On peut citer à titre d'illustration le rappel et la précision pour des problèmes de recherche d'information, le rejet et la confusion pour des tâches de reconnaissance, les performances en apprentissage et en généralisation pour les problèmes d'apprentissage, la qualité et le taux de compression pour des tâches de compression, ou de manière générale les performances qualitatives et le temps de traitement pour la plupart des problèmes.

Ainsi, la majorité des tâches d'un système d'analyse de documents peuvent,

de façon inhérente, être considérées comme des problèmes à objectifs multiples nécessitant le choix de compromis. Par ailleurs, les traitements impliqués sont généralement soumis à des paramètres dont le réglage permet de faire varier les valeurs de compromis entre les différents objectifs. Comparer deux algorithmes dans un cadre idéal revient alors à comparer des ensembles de points dans un espace à plusieurs dimensions. La figure 4.1 illustre ce contexte avec la comparaison de deux classifieurs SVM appris sur les mêmes données avec deux paramétrages différents et dont les performances sont représentées par des courbes ROC [34]. On constate qu'intrinsèquement, aucun des deux classifieurs n'est supérieur à l'autre mais que leur ordonnancement dépendra de la « zone » de fonctionnement choisie. Dans cet exemple, la comparaison des deux classifieurs tout comme le choix de la valeur des paramètres doivent idéalement prendre en considération cette nature multiobjectif.

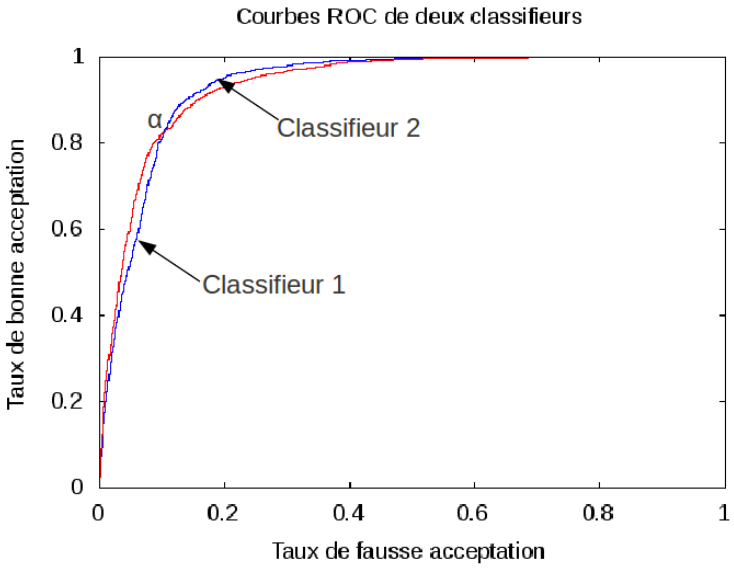


FIGURE 4.1 – Comparaison de deux classifieurs par leur courbe ROC. Pour des taux de fausse acceptation inférieurs à 0.1 (point α sur la courbe), le classifieur 2 est plus performant au regard des deux critères. Pour les taux de fausse acceptation supérieurs à 0.1, c'est le classifieur 1 qui est le plus performant.

Un problème d'optimisation pour lequel une décision doit être prise en présence de compromis entre des objectifs multiples est appelé un problème d'optimisation multiobjectif. Dans un tel contexte, il n'existe généralement pas de solution unique permettant d'optimiser simultanément tous les objectifs et il est nécessaire de chercher un ensemble de solutions qui correspondent aux compromis optimaux entre objectifs. Ce domaine n'est évidemment pas propre à l'analyse de documents et on trouve des problèmes d'optimisation multiobjectif dans de très nombreux domaines d'ingénierie tels que la réalisation de dispositifs électromagnétiques, la conception de circuits logiques, l'optimisation de tournées, en passant par l'analyse de promoteurs dans le domaine de la bio-informatique. L'ouvrage [61] propose un bon aperçu de ce large spectre

de domaines d'application. Cela se traduit par un nombre considérable de publications dans ce domaine ces deux dernières décennies, comme en atteste le site maintenu par Carlos A. Coello Coello [13] qui recense plusieurs milliers de contributions relatives à l'optimisation multiobjectif.

Paradoxalement, malgré l'émergence de cette problématique d'optimisation multiobjectif, on en trouve bien peu d'applications dans le domaine de l'analyse d'images de documents et même pour l'analyse d'images en général. On peut toutefois citer les travaux de Lazzerini ([124]) pour lesquels les tables de quantification JPEG sont optimisées au regard des deux critères que sont le taux de compression et la qualité de l'image décodée. Un certain nombre de contributions ont également été proposées pour la segmentation d'images. Dans [17] par exemple, les auteurs utilisent un algorithme génétique multiobjectif pour conserver une population de solutions plutôt qu'une unique pour les étapes de traitement ultérieures. Un état de l'art concernant ces problématiques de segmentation d'images utilisant des algorithmes d'optimisation multiobjectif est proposé dans [59], avec des applications essentiellement orientées vers le domaine médical et les images de scènes naturelles. À un niveau d'interprétation plus élevé, dans le domaine de la recherche d'information, on peut également citer des travaux utilisant des algorithmes d'optimisation multiobjectif pour améliorer des requêtes en utilisant comme critères le rappel et la précision [47, 131]. Dans le domaine de l'analyse de documents, il convient de souligner ici les travaux de l'équipe de Robert Sabourin sur l'analyse de l'écriture manuscrite qui sont, à notre connaissance, les seuls à soulever le problème de l'intégration d'objectifs multiples à la fois pour l'extraction de caractéristiques [157, 156] et pour la classification supervisée [166].

De cet état de l'art synthétique, on peut conclure que même si l'optimisation multiobjectif a été ponctuellement utilisée dans le contexte de l'analyse de documents ou pour l'analyse d'images, les bénéfices que permettent d'obtenir de telles approches n'ont pas été totalement explorés par la communauté de l'analyse de documents. Nous décrivons dans la suite de ce chapitre nos contributions dans ce domaine. Elles ont consisté à aborder différents problèmes d'analyse de documents sous l'angle de l'optimisation multiobjectif.

Ce chapitre est organisé de la façon suivante. Après un rappel de la problématique de l'optimisation multiobjectif et un état de l'art des approches permettant de résoudre de tels problèmes, trois contributions sont décrites. La première est une contribution propre au domaine de l'optimisation multiobjectif. Nous y décrivons un algorithme pour aborder ces problèmes avec la technique des essaims particulaires. Puis, les deux contributions suivantes concernent des travaux pour lesquels nous avons tiré parti de l'intégration d'objectifs multiples en analyse de documents et en apprentissage. Le chapitre se termine ensuite par une discussion sur cet apport et sur les perspectives directement ouvertes par ces travaux.

13. `http://www.lania.mx/~ccoello/EMOO/EMOObib.html`

4.2 Optimisation multiobjectif

Cette section rappelle la formulation d'un problème d'optimisation multiobjectif et donne un aperçu des méthodes de la littérature pour résoudre de tels problèmes.

4.2.1 Définition du problème

Un problème d'optimisation multiobjectif (parfois appelé optimisation vectorielle) contraint est un problème d'optimisation pour lequel K fonctions objectifs à minimiser (ou maximiser) sont définies, sous respect d'un certain nombre de contraintes d'inégalité ou d'égalité. Il se définit de la façon suivante :

Definition 1. La minimisation contrainte d'un vecteur de fonctions objectifs $\overrightarrow{f} = \{f_1, f_2, ..., f_K\}$ consiste à résoudre :

$$\left.\begin{array}{lll} \text{Minimiser} & f_k(\overrightarrow{x}) & k \in [1, K] \\ \text{sous contrainte de} & g_j(\overrightarrow{x}) \geq 0 & j \in [1, J] \\ & h_l(\overrightarrow{x}) = 0 & l \in [1, L] \\ & x_i^L \leq x_i \leq x_i^U & i \in [1, N] \end{array}\right\}$$

où $\overrightarrow{x}$ est un vecteur de N variables de décision. g_j and h_l sont respectivement les contraintes d'inégalité et d'égalité. Le dernier ensemble de contraintes définit l'espace de décision du problème, *i.e.* l'espace dans lequel les solutions sont recherchées. Les x_i^L et x_i^U désignent ici les bornes de cet espace.

Une différence fondamentale entre l'optimisation mono-objectif et l'optimisation multiobjectif repose sur le fait que pour la plupart des problèmes multiobjectif, les critères étant antagonistes, il n'existe pas de solution qui minimise tous les objectifs simultanément. Par conséquent, il n'existe plus de relation d'ordre total entre solutions. Celles-ci doivent être comparées par la relation de dominance de Pareto qui repose sur la définition suivante [14] :

Definition 2. Une solution $\overrightarrow{x}$ domine une autre solution $\overrightarrow{y}$ si et seulement si $\forall k \in [1, K], f_k(\overrightarrow{x}) \leq f_k(\overrightarrow{y})$ et si $\exists k \in [1, K] / f_k(\overrightarrow{x}) < f_k(\overrightarrow{y})$. Une telle relation est notée $\overrightarrow{x} \prec \overrightarrow{y}$

En utilisant le concept de dominance, une solution $x^\star$ est dite Pareto-optimale s'il n'existe pas de solution dans l'espace de décision qui domine $x^\star$. L'objectif d'un algorithme d'optimisation multiobjectif est de fournir une approximation de l'ensemble optimal de Pareto, défini par :

Definition 3. L'ensemble Pareto-optimal d'un problème d'optimisation multiobjectif est l'ensemble de toutes les solutions Pareto-optimales du problème :

$$POS = \left\{ \overrightarrow{x} \in \vartheta / \neg \exists \overrightarrow{y} \in \vartheta, \overrightarrow{f(y)} \prec \overrightarrow{f(x)} \right\}$$

14. Notons ici qu'il existe d'autres définitions de la dominance de Pareto telles que la dominance stricte, la dominance faible ou encore l'ϵ-dominance qui sont décrites dans [123]

4.2.2 Synthèse de la littérature

Dans la littérature, deux grandes familles d'approches peuvent être distinguées pour résoudre des problèmes d'optimisation multiobjectif. Elles diffèrent en fonction du fait que l'on intègre ou pas à la résolution mathématique une articulation *a priori* des préférences sur les objectifs. Lorsque de telles préférences peuvent être formulées, il est alors possible de combiner les différents objectifs pour obtenir une valeur scalaire. Le problème devient alors un problème d'optimisation mono-objectif qui peut être résolu avec des méthodes classiques. On parle alors d'approches « scalarisées ». Lorsque les préférences ne peuvent pas être exprimées *a priori*, l'algorithme d'optimisation doit alors fournir en sortie une population de solutions non dominées au sens de Pareto, parmi lesquelles l'utilisateur (ou éventuellement un autre traitement) doit choisir en intégrant cette fois des préférences *a posteriori*. On parle alors d'approche à base de Pareto. Notons que la littérature propose également quelques approches appelées progressives ou interactives, pour lesquelles le décideur intègre ses préférences au cours du processus d'optimisation. Nous n'abordons pas ici ces méthodes, mais le lecteur trouvera une bonne étude comparative dans [5].

Approches scalarisées De très nombreuses approches ont été proposées dans cette catégorie. Elles peuvent être classées en fonction de la formulation mathématique qui est utilisée pour combiner les objectifs en une valeur scalaire, mais aussi par la façon dont les préférences sont exprimées. Certaines approches, les plus nombreuses, imposent l'attribution d'un poids à chacun des objectifs. Ces derniers peuvent alors être combinés par différentes méthodes. La plus classique est celle de la somme pondérée [120], mais la littérature propose bien d'autres stratégies telles que la méthode Min-Max (ou méthode de Tchebycheff pondérée) [196], la méthode des exponentielles pondérées [13], la méthode du produit pondéré [86], avec de nombreuses variantes dans chacun des cas. Une alternative à l'affectation de poids à chaque objectif consiste à ne fournir qu'un ordonnancement des objectifs. C'est le cas par exemple de la méthode dite lexicographique [195]. Dans le cas de la programmation par buts [19], une valeur à atteindre est fixée pour chacun des objectifs et c'est la somme des écarts à ces valeurs qui est minimisée. Certaines approches proposent quant à elles de considérer uniquement l'objectif prioritaire comme critère à optimiser et de voir les autres objectifs comme des contraintes pour lesquelles il faut fixer des bornes. C'est le cas de la méthode dite des fonctions objectifs bornées [90], parfois appelée méthode ϵ-contrainte. Enfin, une méthode originale appelée programmation physique consiste à attribuer à chacun des critères une « classe d'objectif » en y affectant des paramètres [145, 57].

Naturellement, comme l'énonce le *No Free Lunch Theorem*[15] [213], aucune de ces méthodes ne se distingue réellement pour l'ensemble des problèmes et il faut choisir la méthode la plus adaptée au problème à traiter. Un état de l'art très complet de ces méthodes scalarisées est dressé dans [142]. En particulier, une discussion est proposée dans cet article sur le potentiel des différentes

15. Ce théorème classique en optimisation est généralisé aux problèmes multiobjectifs dans [189].

méthodes pour obtenir l'ensemble des points du front de Pareto, en faisant varier certains paramètres des méthodes.

Approches Pareto Même si des adaptations sont disponibles dans la littérature pour pallier ce problème, les méthodes décrites ci-avant sont par essence conçues pour calculer une solution unique aux problèmes d'optimisation multiobjectif. Or, comme évoqué précédemment, la solution d'un problème multiobjectif n'est généralement pas unique, mais est plutôt constituée d'un ensemble de solutions non dominées. Les méthodes à base de populations reposant sur un ensemble de solutions potentielles, telles que les algorithmes évolutionnaires (AE) [70], les essaims particulaires [175] ou les colonies de fourmis [7], sont donc de bonnes candidates à la résolution de ce type de problème [70, 60]. Depuis les travaux pionniers de Schaffer en 1985 [186] avec son algorithme VEGA, un nombre considérable d'approches évolutionnaires ont été proposées pour résoudre les problèmes d'optimisation multiobjectif (MOGA [82], NSGA [194], NPGA [94], SPEA [219], NSGA II [71], PESA [66], SPEA2 [218], pour ne citer que les plus connus). La figure 4.2 décrit la structure générale de tous ces algorithmes. Une population de solutions candidates est d'abord initialisée aléatoirement. Puis, celle-ci évolue au cours de générations successives par la combinaison d'opérateurs de sélection, de remplacement et de modification. Dans les approches élitistes, qui se sont montrées les plus performantes [113], une archive contenant les « meilleures » approximations de l'ensemble de Pareto est maintenue au cours de cette évolution. C'est cette archive qui constitue la sortie de l'algorithme.

Comparer les différents algorithmes d'optimisation existants est une tâche difficile. L'analyse de performance d'algorithmes d'optimisation multiobjectif est en effet encore du domaine de la recherche. Une synthèse des travaux existants dans ce domaine est disponible dans [198]. La difficulté provient du fait qu'un algorithme d'optimisation multiobjectif a lui-même plusieurs objectifs à atteindre. Il doit évidemment converger le plus rapidement possible vers l'ensemble optimal de Pareto, mais il doit également proposer des solutions diversifiées sur le front afin d'avoir un échantillon représentatif et ne pas se concentrer sur une zone de l'espace des objectifs. Ce dernier critère peut lui-même être scindé en deux sous-critères que sont l'étendue sur le front et la diversité. Dans une étude publiée dans [113], les performances des trois algorithmes les plus populaires (SPEA2, PESA et NSGA-II) sont comparées. La comparaison est menée sur différents problèmes de test en évaluant les algorithmes suivant les deux critères importants que sont la proximité au front de Pareto réel et la distribution des solutions. Les résultats obtenus, qui sont corroborés dans [218] et dans [38], montrent qu'aucun des trois algorithmes ne domine les autres au sens de Pareto sur ces deux objectifs. SPEA2 et NSGA-II se comportent de manière équivalente autant en termes de convergence qu'en termes de préservation de la diversité. Leur convergence vers le front de Pareto est inférieure à celle de PESA, mais la diversité est meilleure. L'étude montre également que NSGA-II est plus performant que SPEA2 en termes de temps de calcul, essentiellement à cause de la phase de clustering très chronophage de ce dernier. L'étude évoque également le fait que quelle que soit l'approche choisie,

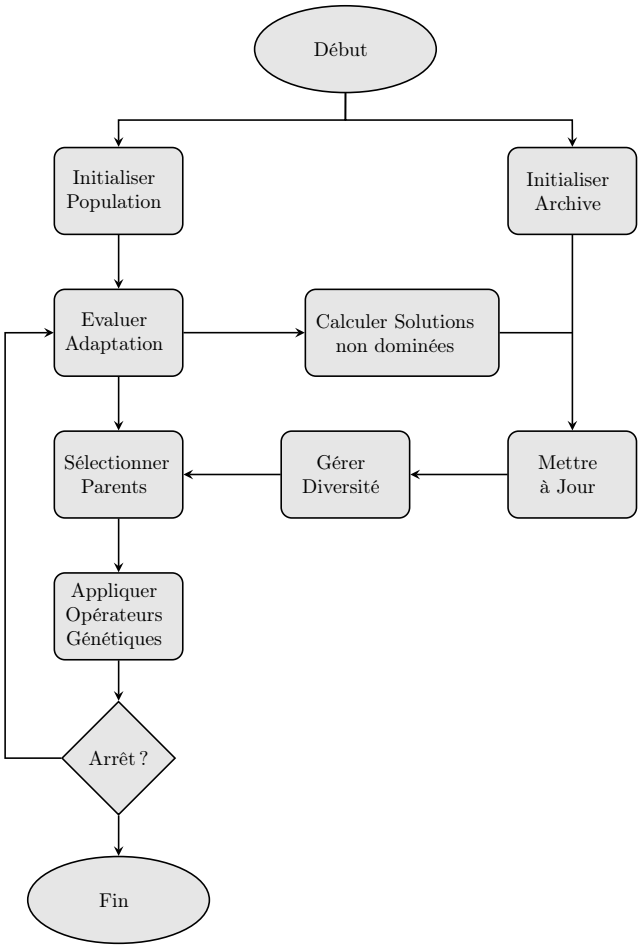


FIGURE 4.2 – Structure typique d'un algorithme évolutionnaire multiobjectif élitiste. L'archive est ici une population externe. Pour certains algorithmes tels que [71], une seule population est gérée et contient les éléments non dominés. La sortie de l'algorithme est le contenu de l'archive.

pour être performant, l'algorithme doit être adapté au problème à résoudre. Il est en particulier nécessaire de :

1. choisir une représentation adaptée des individus ;
2. concevoir une stratégie efficace d'initialisation des individus ;
3. concevoir une fonction d'évaluation des individus ;
4. concevoir des opérateurs de variation appropriés.

Dans la suite de ce chapitre, nous synthétisons les différents travaux que nous avons réalisés à l'intersection des domaines de l'analyse de documents, de la reconnaissance de formes et de l'optimisation multiobjectif.

4.3 Contributions

Dans cette section, trois contributions relatives à l'optimisation multiobjectif et son application en analyse de documents sont présentées. La première sous-section décrit une contribution propre au domaine de l'optimisation multiobjectif. Lors de notre phase d'analyse des divers algorithmes évolutionnaires et des limitations et difficultés de leur mise en œuvre, nous avons été conduits à proposer une variante d'algorithme d'optimisation multiobjectif utilisant les essaims particulaires. Cette proposition est exposée en 4.3.1. Puis, les deux sous-sections suivantes proposent une nouvelle façon de considérer deux problèmes classiques de l'analyse de documents et de l'apprentissage, en adoptant un point de vue multiobjectif. Pour chacune de ces applications, le problème est posé, les choix correspondants à la mise en œuvre de l'algorithme sont décrits et les résultats sont discutés.

4.3.1 Essaims particulaires et optimisation multiobjectif

4.3.1.1 Définition du problème et revue de l'existant

Au même titre que les algorithmes évolutionnaires évoqués en 4.2.2, l'Optimisation par Essaims Particulaires (OEP) est une métaheuristique d'optimisation reposant sur une population de solutions candidates. Proposée initialement dans [112], elle s'inspire de la nature en cherchant à copier le comportement social d'animaux évoluant en essaims. Dans un algorithme d'OEP, les particules sont des solutions potentielles du problème d'optimisation. Elles se déplacent dans un espace de dimension n, où n est le nombre de variables du problème. À chaque itération de l'algorithme, les positions des particules sont mises à jour en utilisant les équations simples de déplacement suivantes :

$$\begin{aligned} v_{i,t+1} = \quad & \omega.r_0.v_{i,t} + \\ & c_1.r_1.\left(p_{i,best} - x_{i,t}\right) + \\ & c_2.r_2.\left(p_{i,guide} - x_{i,t}\right) \end{aligned} \tag{4.1}$$

$$x_{i,t+1} = \quad x_{i,t} + \chi(v_{i,t+1}) \tag{4.2}$$

Dans ces équations, $x_{i,t}$ est la position de la i^{eme} particule à l'instant t. $v_{i,t}$ est sa vitesse. $p_{i,best}$ et $p_{i,guide}$ sont respectivement la meilleure position visitée par la particule i au regard de la fonction à optimiser et la position d'une autre particule de l'essaim choisie comme guide. Les poids appliqués à ces positions sont respectivement appelés facteurs individuel et social. Ils sont tous les deux calculés en multipliant un coefficient c_x fixé *a priori* par une valeur r_x aléatoirement tirée dans $[0, 1]$. En fonction des valeurs prises par ce produit, les particules auront tendance à explorer l'espace ou à affiner leur position dans un voisinage donné. Les valeurs du produit $r_x c_x$ ont donc un impact important sur la convergence de l'algorithme. ω est appelé facteur d'inertie. Un grand facteur d'inertie provoque une grande exploration de l'espace de recherche alors qu'un petit facteur d'inertie concentre la recherche sur un petit espace. La valeur de ω peut être constante ou évoluer au cours du temps comme dans [217]. Une valeur importante aura tendance à faire suivre à la particule

sa direction précédente, même si un facteur r_0 tiré aléatoirement dans $[0, 1]$ permet de nuancer cet aspect. La fonction $\chi()$ est généralement implémentée comme un simple facteur de turbulence [150], mais elle peut aussi correspondre à une fonction de normalisation ou une fonction de constriction, qui conserve la direction de la particule mais empêche une divergence de sa vitesse [159].

Pendant les dix dernières années, les algorithmes d'OEP ont été très largement étudiés et appliqués à de très nombreux domaines d'ingénierie. Les résultats obtenus ont montré qu'ils étaient compétitifs par rapport aux autres métaheuristiques d'optimisation telles que les algorithmes évolutionnaires ou les colonies de fourmis ([119, 151, 190]). Ces succès, couplés à l'émergence des problématiques d'optimisation multiobjectif, ont naturellement amené la communauté à s'intéresser à leur transformation pour appréhender des problèmes à objectifs multiples [175].

Le principal changement dans cet algorithme provient naturellement de l'absence de relation d'ordre total entre les solutions, si ce n'est pas le biais d'une aggrégation des critères. Ainsi, il n'existe plus réellement de meilleure particule, ni de meilleure position d'une particule. Dans ce contexte, les deux principales difficultés à surmonter sont [175] :

- la sauvegarde des solutions non dominées constituant l'estimation de l'ensemble de Pareto, qui impose la gestion d'une population externe, appelée archive, dont il est important de gérer la taille et la diversité pour éviter une explosion du nombre de comparaisons et pour fournir une solution exploitable à l'utilisateur ;
- la gestion de la mémoire de la particule (traditionnellement la meilleure position visitée) ainsi que la sélection de la particule guide dans l'essaim.

Ces modifications sont illustrées par l'algorithme 1. Les lignes 7, 11 et 13 illustrent respectivement l'intégration de la relation de dominance dans la gestion de la mémoire de la particule, la gestion de l'archive et la sélection du guide.

Dans [75], nous avons proposé des solutions originales à ces problèmes. Elles sont synthétisées dans les paragraphes suivants.

4.3.1.2 Approche proposée

Gestion de l'archive Le passage d'un problème mono-objectif à un problème multiobjectif basé sur la dominance de Pareto impose d'intégrer dans l'algorithme d'optimisation par essaims particulaires une archive contenant l'approximation courante de l'ensemble de Pareto du problème. Une approche simpliste consisterait à y intégrer toutes les solutions non dominées rencontrées lors de l'évolution des particules. Toutefois, il serait alors impossible de gérer la taille de l'archive et sa diversité. La solution généralement préconisée par la littérature pour résoudre ce problème consiste à remplacer la dominance de Pareto classique par l'ϵ-dominance, proposée dans [122] et évaluée dans [149]. Deux alternatives sont disponibles dans la littérature en termes de définition : l'ϵ-dominance additive proposée dans [122] ou l'ϵ-dominance multiplicative décrite dans [149]. Un consensus semble aujourd'hui se dégager pour la seconde solution pour laquelle le choix de la valeur d'ϵ est simplifié. Toutefois, les expériences que nous avons menées et décrites dans [75] en utilisant cette définition

Algorithm 1 Algorithme de l'implémentation des MOPSO.

```
DÉBUT
t ← 0
Initialisation aléatoire de l'essaim
répéter
   pour chaque particule i faire
      Mettre à jour la position x_{i,t+1} de la particule en utilisant l'eq. 4.2
      si p_{i,t+1} ≺ p_{i,best} alors
         p_{i,best} ← p_{i,t+1}
      finsi
   fin pour
   Mettre à jour l'archive
   pour chaque particule i faire
      Sélectionner un guide p_{i,guide}
   fin pour
   Évaluer les critères de fin
   t ← t + 1
tant que les critères de fin ne sont pas atteints
FIN
```

de ϵ-dominance multiplicative ont mis en exergue le fait que celle-ci ne permettait pas de décrire de façon homogène le front de Pareto. C'est pourquoi nous avons proposé une variante de cette dominance. Les équations 4.3 à 4.5 donnent respectivement les définitions de l'ϵ-dominance additive, de l'ϵ-dominance multiplicative et de notre proposition. La figure 4.9 illustre les différences entre ces différentes variantes.

$$\vec{x}_i \prec \vec{x}_j \quad \Leftrightarrow \quad \begin{cases} \forall k \in [1,N], f_k(\vec{x_i}) + \varepsilon \leq f_k(\vec{x_j}) \\ \exists k' \in [1,N] \mid f_{k'}(\vec{x_i}) + \varepsilon < f_{k'}(\vec{x_j}) \end{cases} \tag{4.3}$$

$$\vec{x}_i \prec \vec{x}_j \quad \Leftrightarrow \quad \begin{cases} \forall k \in [1,N], \frac{f_k(\vec{x_i})}{1+\varepsilon} \leq f_k(\vec{x_j}) \\ \exists k' \in [1,N] \mid \frac{f_{k'}(\vec{x_i})}{1+\varepsilon} < f_{k'}(\vec{x_j}) \end{cases} \tag{4.4}$$

$$\vec{x}_i \prec \vec{x}_j \Leftrightarrow \begin{cases} \begin{cases} \forall k \in [1,N], f_k(\vec{x_i}) \leq f_k(\vec{x_j}) \\ \exists k' \in [1,N] \mid f_{k'}(\vec{x_i}) < f_{k'}(\vec{x_j}) \end{cases} \\ OR \\ \begin{cases} \exists k' \in [1,N] \mid \\ f_{k'}(\vec{x_j}) < f_{k'}(\vec{x_i}) < \frac{1+2\varepsilon}{1+\varepsilon} f_{k'}(\vec{x_j}) \\ \forall k \in [1,N], \frac{f_k(\vec{x_i})}{1+\varepsilon} \leq f_k(\vec{x_j}) \end{cases} \end{cases} \tag{4.5}$$

Tout comme avec l'ε-dominance [148], la variante proposée permet de gérer simultanément la dominance et le voisinage dans l'espace des objectifs. Elle permet ainsi d'obtenir rapidement une approximation de l'ensemble de Pareto en modifiant la surface de dominance proportionnellement aux valeurs de critères. La variante proposée ajoute à ces propriétés le fait de mieux prendre en considération certaines formes particulières de front, en particulier les zones

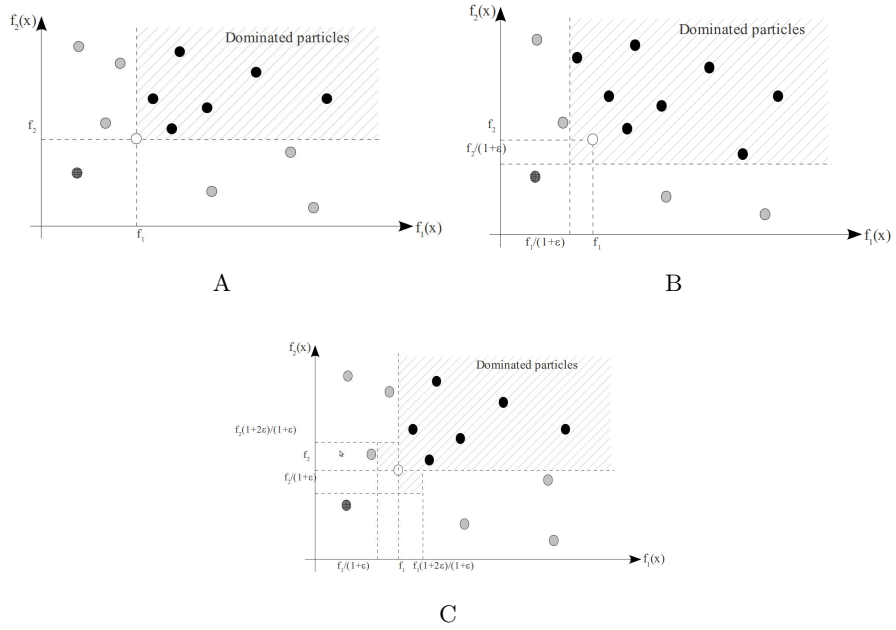


FIGURE 4.3 – Illustration des variantes proposées : dominance classique de Pareto dominance (a), ε-dominance relative (b) et notre variante de l'ε-dominance (c). Les zones hachurées correspondent aux zones dominées par la particule 'o'

pour lesquelles un seul objectif varie beaucoup. L'annexe C illustre cet avantage sur des problèmes de la littérature.

Gestion de la mémoire et du guide Les performances d'un algorithme d'optimisation par essaims particulaires dépendent fortement des choix qui sont faits pour la gestion de la mémoire des particules et pour le choix du guide. Ces choix sont évidemment impactés par le passage du mono-objectif au multiobjectif. Pour le choix de la mémoire, nous avons opté dans [75] (Annexe C) pour l'utilisation de la dernière position non dominée visitée. Ce choix permet de réduire considérablement les temps de calcul par rapport aux stratégies plus évoluées telles que celles proposées dans [35].

Pour la sélection du guide, nous avons proposé une méthode basée sur une approche stochastique, inspirée des processus de sélection utilisés dans les algorithmes génétiques. Il est prouvé dans [8] que de telles approches s'avèrent particulièrement efficaces. Ainsi, le guide est sélectionné par l'intermédiaire de la simulation d'un tirage par roue de loterie biaisée parmi les particules de l'archive. Les probabilités affectées à chaque particule dans ce cadre sont déterminées en fonction de la densité de leur voisinage sur l'estimation courante du front de Pareto, ce qui vise à améliorer la diversité sur le front. La métrique utilisée pour effectuer ce calcul de densité est détaillée dans [75].

Le dernier problème abordé dans nos travaux concerne la stratégie mise en place pour modifier le guide. Il est en effet important de ne pas modifier celui-ci à chaque itération pour que les particules aient le temps de converger vers celui-ci. Dans [75], nous avons proposé une approche consistant à (i) ne pas

utiliser de guide lorque la particule vient d'être intégrée à l'archive, lui laissant ainsi explorer librement l'espace des paramètres en fonction de sa mémoire individuelle et de son inertie, (ii) changer de guide en fonction d'un tirage aléatoire biaisé par le nombre d'itérations pour lesquelles le même guide a été utilisé, là encore pour améliorer l'exploration de l'espace des paramètres.

4.3.1.3 Résultats obtenus

Les propositions décrites ci-avant ont été évaluées sur différents problèmes standards de la littérature de difficultés variables ([29, 44, 199]), en utilisant les métriques proposées par la communauté de l'optimisation multiobjectif. Ces métriques recouvrent les deux objectifs principaux de l'optimisation multiobjectif, à savoir la convergence vers le front de Pareto et la diversité. Les expérimentations menées, qui sont précisément décrites dans l'annexe C, visaient à illustrer l'apport des contributions en comparant les performances avec et sans nos propositions. Concernant la proposition de dominance, les résultats obtenus prouvent que le front est décrit de manière beaucoup plus fine avec notre proposition, à la fois en termes de diversité (évaluée par la *Spacing Metric*) et d'extension (évaluée par la *Maximal Extension*). L'archive obtenue permet ainsi une bien meilleure description des solutions du problème. Concernant la stratégie de sélection de guide, là encore, les résultats obtenus ont montré une amélioration significative, en particulier pour les problèmes réputés les plus difficiles.

L'approche a également été comparée avec l'algorithme de référence NSGA-II [71] sur un problème d'analyse de documents. La figure 4.4 montre les résultats obtenus par les deux algorithmes sur ce problème qui sera précisément décrit en 4.3.3. Elle illustre le fait que l'algorithme proposé permet d'obtenir des résultats tout à fait compétitifs avec l'état de l'art et bien meilleurs que ceux que permet d'obtenir une approche « scalarisée ».

4.3.2 Approximation de courbes

4.3.2.1 Définition du problème et revue de l'existant

L'approximation de courbes planaires est un problème fréquemment abordé dans les communautés de l'analyse d'images et de l'analyse de documents. C'est en effet un moyen classiquement adopté pour représenter, stocker et traiter des courbes numériques. Les résultats d'une approximation peuvent par exemple être utilisés pour représenter des formes dans un processus de reconnaissance [147, 97, 163, 154].

L'approximation de courbes 2D peut être définie comme suit : soit une courbe décrite par une liste ordonnée de N points $C = \{p_i = (x_i, y_i)\}_{i=1}^{N}$. Le but d'un approximateur est de trouver une liste $B = \{b_i = (x_i, y_i)\}_{i=1}^{M} \subset C$ constituée de M points (souvent appelés point dominants) et un ensemble de paramètres $\Theta = \{\theta_i\}_{i=1}^{P}$ décrivant la courbe approximant les points entre les b_i consécutifs. Si C est une courbe ouverte (*i.e.* $p_1 \neq p_N$), p_1 et p_N sont généralement inclus dans l'ensemble B et par conséquent $P = M - 1$. Si au contraire la courbe est fermée, (*i.e.* $p_1 = p_N$), B ne contient *a priori* ni point initial ni point terminal et $P = M$ puisque la courbe entre b_P et b_1 doit

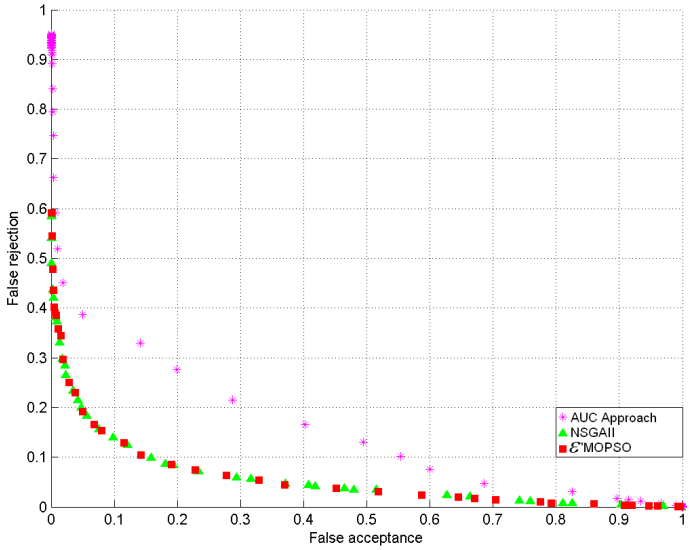


FIGURE 4.4 – Comparaison des estimations finales du front de Pareto d'un problème de sélection de modèle (NSAGII *vs.* MOPSO). La courbe marquée par des '*' correspond à une approche scalarisée à base d'aire sous la courbe ROC [168].

être approximée. Notons que si les courbes approximantes sont réduites à des segments, on parle alors d'approximation polygonale et la définition de Θ n'est pas nécessaire. Dans le cas où les arcs de cercle sont considérés, Θ est défini par $\Theta = \{\theta_i = (xc_i, yc_i)\}_{i=1}^{P}$ où (xc_i, yc_i) désignent les coordonnées du centre de l'arc de cercle.

De très nombreux algorithmes ont été proposés pour approximer des courbes dans ces différentes configurations. Parmi les approches existantes, deux paradigmes peuvent être distingués. Le premier consiste à approximer la courbe en détectant des points particuliers, en général sur la base d'heuristiques liées à la courbure. La littérature est extrêmement abondante pour ce type d'approches [179, 201, 10, 173, 67, 140, 143] et un état de l'art très complet est disponible dans [45]. Dans le second cas, l'approximation est considérée comme un processus d'optimisation [162, 182, 115, 96, 184, 160]. Un critère d'erreur est défini et l'algorithme cherche à optimiser l'approximation au regard de ce critère. On peut distinguer deux types de formulation d'une telle optimisation [115] :

- $min - \epsilon$: la valeur de M est fixée et le processus repose sur une minimisation de l'approximation de l'erreur. En général, le critère repose sur une erreur quadratique définie par $ISE = \sum_{i=1}^{P} e_i^2$ où e_i est la distance entre p_i et la courbe approximante.
- $min - \#$: une tolérance maximale sur l'erreur d'approximation est fixée et le processus minimise le nombre de points d'approximation P (*i.e.* maximise le taux de compression).

Or, minimiser l'ISE et le nombre de points P sont deux objectifs antagonistes. Dans ce cadre, des auteurs ont proposé des critères scalaires combinant les deux valeurs. Ainsi, Sarkar propose dans [185] un critère nommé *Figure*

Of Merit défini par $FOM = CR/ISE$. Markji et Syi proposent dans [141] un autre critère défini par $WE_2^x = ISE/CR^x$. Une bonne revue des critères existants est proposée dans [46]. Dans [130], nous avons proposé de traiter le problème de l'approximation de courbes sous l'angle de l'optimisation multiobjectif. L'approche est décrite dans la sous-section suivante.

4.3.2.2 Approche proposée

L'approche que nous avons proposée dans [130] pour aborder le problème d'approximation de courbes dans le cadre de l'optimisation multiobjectif repose sur l'utilisation d'un algorithme de la littérature, suivant la stratégie illustrée sur la figure 4.2. Pour appliquer cet algorithme au problème défini ci-dessus, celui-ci a été spécialisé. Cette spécialisation repose d'abord sur le codage des individus. Ainsi, un individu doit représenter une solution possible du problème d'approximation. Pour ce faire, un individu est simplement composé de N gènes. Un gène à '1' signifie que le point est conservé comme point dominant. Si sa valeur est '0', le point n'est pas retenu. Une seconde spécialisation concerne l'initialisation de la population. Pour réduire le nombre d'itérations de l'algorithme, un opérateur d'initialisation spécifique a été proposé. Il s'appuie sur une analyse préalable de la courbe à traiter en utilisant une fenêtre glissante de 3 points. Un histogramme des configurations est construit lors de cette analyse. Des probabilités en sont déduites et sont utilisées lors de l'initialisation des individus.

Les opérateurs utilisés pour faire évoluer la population sont des opérateurs génétiques classiques. Pour le croisement, une permutation à un point est utilisée. Elle permet de croiser les bonnes approximations de deux parties d'une courbe. Pour la mutation, un choix aléatoire est effectué entre deux possibilités. La première est une mutation classique consistant à changer la valeur d'un gène de 0 (resp. 1) à 1 (resp. 0). La seconde consiste à déplacer un point dominant d'une position à sa précédente ou à sa suivante. Il permet d'affiner une approximation.

L'évaluation d'un individu consiste à calculer (i) le nombre de points dominants, qui est simplement une somme de la valeur des gènes et (ii) l'ISE de l'approximation correspondante. Cette valeur est calculée en sommant les erreurs obtenues entre chaque paire de points dominants consécutifs. Pour chaque paire, l'algorithme compare l'ISE obtenue avec un segment avec celle obtenue avec un arc. Dans le cas d'un segment, nous utilisons $ISE = \sum_{i=1}^{n} d_i^2$, où d_i est la distance orthogonale du i^{eme} point au segment et où n est le nombre de points entre les extrémités de la courbe. Dans le cas d'arcs de cercle, l'évalutation repose sur deux étapes. La première consiste à estimer la position du centre de l'arc sous la contrainte de la position des points extrémités. Une telle estimation est très coûteuse en temps si une approche exacte est utilisée. Aussi, nous utilisons une approximation proposée dans [161], qui repose sur une fonction d'erreur définie dans [203]. Ainsi, le centre de l'arc approximant une séquence de point $(x_1, ..., x_n)$ est calculé par :

$$(x_c, y_c) = \left(-\frac{\sum_{i=1}^{n} K_1 K_2}{\sum_{i=1}^{n} K_1 K_3}, ax_c + b \right) \tag{4.6}$$

avec :
- $a = -(x_n - x_1)/(y_n - y_1)$,
- $b = ((y_1 + y_n)/2) - (a(x_1 + x_n)/2))$,
- $K_1 = -x_1 - ay_1 + x_i + ay_i$,
- $K_2 = x_1^2 + (y_1 - b)^2 - x_i^2 - (y_i - b)^2$,
- $K_3 = -2x_1 - 2a(y_1 - b) + 2x_i + 2a(y_i - b)$.

L'ISE peut ensuite être calculée par :

$$ISE = \sum_{i=1}^{n} \left[R - \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \right]^2 \tag{4.7}$$

avec $R^2 = (x_1 - x_c)^2 + (y_1 - y_c)^2$

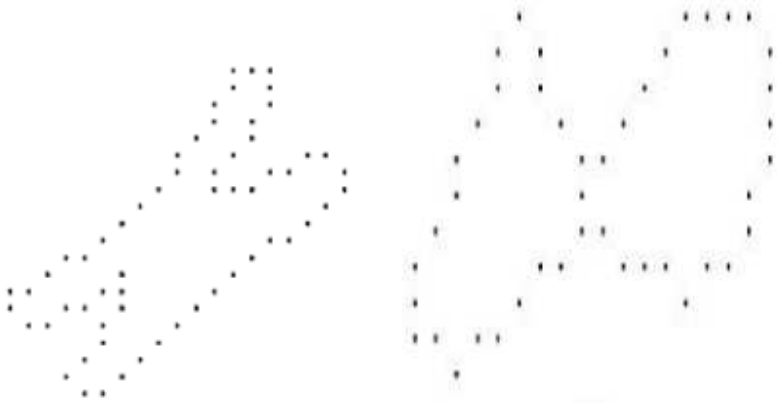
4.3.2.3 Résultats obtenus

Pour valider les performances de l'algorithme proposé, ce dernier a été évalué sur quatre courbes de référence proposées dans [201] et illustrées sur la figure 4.5.

Par définition, l'algorithme proposé consiste à estimer l'ensemble de Pareto du problème biobjectif correspondant. Aussi, le résultat est un ensemble de couples ISE/nombre de points dominants. Pour valider la convergence de l'algorithme, nous avons d'abord comparé les résultats qu'il permet d'obtenir avec une approche exhaustive recherchant les ISE optimales (en utilisant [162]) pour un nombre variable de points dominants. La figure 4.6 illustre le résultat obtenu. Elle montre que, grâce à la manipulation de populations de solutions à la base de l'approche, l'algorithme permet de trouver en une seule exécution un ensemble d'approximations proches des résultats optimaux, pour différents nombre de points dominants.

Les résultats obtenus ont également été comparés à ceux de la littérature. Une telle comparaison est une tâche difficile pour plusieurs raisons. D'abord, la littérature est assez pauvre concernant l'approximation de courbes par des segments et des arcs de cercle. À notre connaissance, moins de dix approches ont été proposées à ce jour pour résoudre une telle tâche ([180, 55, 96, 95, 184, 207, 98, 100, 155]). Par ailleurs, parmi les approches existantes, très peu d'articles donnent des résultats sur les courbes de références proposées par [201]. Enfin, pour ces quelques articles, comme souvent quand plusieurs objectifs sont considérés dans un problème d'optimisation, les résultats sont fournis pour uniquement quelques nombres de points dominants. Le tableau 4.1 résume ces résultats et les compare avec ceux obtenus par notre algorithme.

Ces résultats amènent plusieurs observations. La première est que, à nombre fixé de points, l'approche proposée n'est pas « meilleure » que celles de la littérature. Pour la plupart des configurations, l'une des approches de la littérature permet d'obtenir une ISE inférieure. Cependant, ce n'est pas toujours le même algorithme qui permet d'atteindre la performance optimale. Ainsi, il n'existe pas d'approche qui domine toutes les autres. Par ailleurs, l'avantage principal de l'approche proposée est qu'une seule exécution de l'algorithme suffit pour obtenir un ensemble de solutions parmi lesquelles l'utilisateur peut choisir le compromis qui lui convient. Enfin, l'approche est générique. Elle peut être



(a) Chromosome (b) Eight

(c) Semicircle (d) Leaf

FIGURE 4.5 – Les 4 courbes de test proposées dans [201]. (a) *chromosome-shaped* avec 60 points ; (b) *figure-of-eight* avec 45 points ; (c) *four-semicircle* avec 102 points et (d) *leaf-shaped* avec 120 points. .

adaptée à tout type de courbe paramétrique (ellipses, B-Splines), contrairement aux approches basées sur la détection de points dominants.

L'approche et les résultats présentés dans cette sous-section militent clairement, selon nous, pour la prise en compte des deux objectifs dans le contexte de l'approximation de courbes. Dans la section suivante, nous montrons que cela peut également être le cas dans le domaine de l'apprentissage.

4.3.3 Sélection de modèles

Cette sous-section synthétise les travaux que nous avons proposés dans le domaine de l'apprentissage multiobjectif, en particulier pour la sélection de modèles multiples de classifieurs SVM.

4.3.3.1 Définition du problème et revue de l'existant

Le réglage des hyperparamètres d'un classifieur est une étape critique de la construction d'un système de reconnaissance de formes. Cet aspect crucial

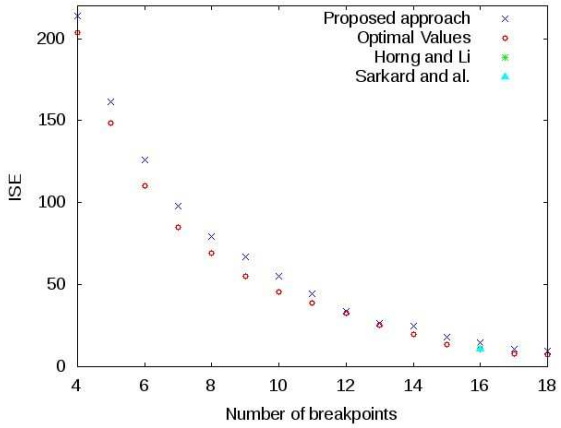


FIGURE 4.6 – Comparaison entre le front de Pareto réel du problème ('o') obtenu en utilisant une adaptation de l'approche proposée dans [162] et les résultats obtenus avec l'algorithme multiobjectif ('x') pour la courbe *leaf-shaped*

Reference	Chromosome		Figure-of-eight		Leaf		Semicirles	
	N	ISE	N	ISE	N	ISE	N	ISE
[96]	10	2,67	6	3,06	16	11,31	4	6,94
[184]	10	2,60	6	3,26	16	10,96	4	6,94
	11	2,18	8	2,36	18	7,40	6	5,83
	15	1,23	9	2,03	31	1,64	12	4,31
[98]	10	3,31	6	3,32	19	9,18	4	6,94
					20	6,27	8	6,83
Optimal Values	10	2,42	6	3,06	16	10,54	4	6,94
	11	1,94	8	2,27	19	6,18	6	5,77
	15	1,08	9	1,92	20	5,28	8	5,24
					31	1,34	12	3,98
Proposed Approach	10	2,68	6	3,23	16	14,73	4	6,94
	11	2,25	8	2,38	19	6,99	6	5,83
	15	1,27	9	1,99	20	6,69	8	5,25
					31	1,48	12	4,19

TABLE 4.1 – Comparaison des résultats obtenus par l'approche proposée et les résultats de la littérature pour les différentes courbes de test.

de la sélection de modèles a en effet un impact fort sur les performances en généralisation du système.

Les travaux menés dans le cadre de la thèse de Simon Bernard [21] sur le paramétrage des forêts aléatoires constituent une illustration parfaite de cette constatation. Nous y avons montré que la valeur du nombre K de caractéristiques choisies aléatoirement à chaque nœud lors de l'induction des arbres avait une influence importante sur les performances de l'ensemble [22, 24]. La figure 4.7 illustre cet aspect par les performances obtenues par l'algorithme Forest-RI [36] sur douze bases de l'*UCI Machine Learning repository* [1] en faisant va-

rier la valeur de K. On y constate la variabilité des performances en fonction de K, mais aussi le fait que la valeur optimale de K est variable en fonction des problèmes traités. Ces résultats ont d'ailleurs motivé la proposition d'un algorithme nommé Forest-RK qui adapte la valeur de K au problème traité [25, 23].

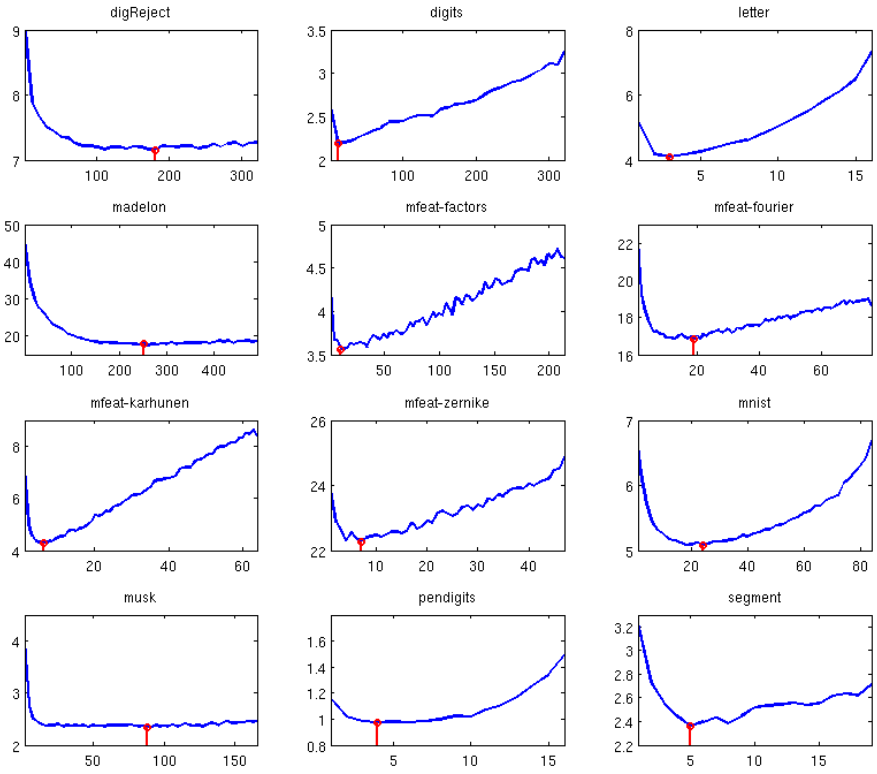


FIGURE 4.7 – Taux d'erreurs moyens obtenus en fonction de la valeur du paramètre K sur différentes bases de l'UCI. La valeur optimale de K est marquée sur chacune des courbes.

Ce problème du réglage des hyperparamètres n'est évidemment pas propre aux forêts aléatoires. On le retrouve pour tout type de classifieur. Dans la littérature, la plupart des contributions relatives à cette problématique concernent la proposition des critères à optimiser pour régler les hyperparamètres. Elles ont mené à de nombreux critères et stratégies visant à résoudre ce problème. On peut citer par exemple le *Xi-Alpha bound* de [106], la *Generalized Approximate Cross-Validation* de [210], l'*empirical error estimate* de [15], la *radius-margin bound* de [51] ou la *maximal-discrepancy* de [9]. Une revue des travaux dans ce domaine est proposée dans [89]. En exploitant ces critères, les valeurs des hyperparamètres sont généralement choisies en utilisant une recherche en grille, associée à une procédure de validation croisée. Quelques auteurs y adjoignent des techniques de descente de gradient, en rendant dérivable le critère, pour réduire la complexité combinatoire [20, 110].

Toutes ces approches, bien qu'efficaces, reposent sur un critère unique. Or, il est désormais admis qu'un critère unique n'est pas toujours un indicateur

de performances suffisant. En particulier, un critère scalaire n'est pas adapté lorsque les coûts de mauvaise classification sont (i) asymétriques (par exemple dans le domaine médical ou la biométrie), (ii) difficiles à estimer (par exemple quand le processus de classification est intégré dans un système plus complexe) et (iii) évolutifs au cours de la vie du système (par exemple pour des problématiques de détection de fraudes).

Dans de tels environnements généralement appelés « mal définis », les critères scalaires utilisés pour construire un classifieur unique sont inadaptés. Une alternative de plus en plus utilisée pour considérer ce problème est d'utiliser la courbe ROC (Receiver Operating Characteristics) proposée dans [34] pour évaluer les performances d'un classifieur. Dans le contexte d'un problème à deux classes, une courbe ROC (Figure 4.1) est une représentation synthétique des compromis entre les taux de vrais positifs et de faux positifs. Il existe des travaux en apprentissage s'appuyant sur l'espace ROC pour sélectionner le modèle du classifieur [80, 168, 33]. Toutefois, ils reposent en général sur une scalarisation en résumant la courbe ROC à une valeur telle que la F-Mesure, le *break even point*, ou l'aire sous la courbe ROC (*Area Under Curve*-AUC). Notons également l'existence de quelques travaux ([188, 76]) pour lesquels les deux critères de l'espace ROC sont intégrés dans le cadre de l'apprentissage de classifieurs.

4.3.3.2 Approche proposée

Dans [52], nous avons proposé de ne pas faire reposer le choix du modèle sur un critère scalaire visant à trouver le « meilleur » classifieur global, mais de construire une population de classifieurs localement optimaux. Le classifieur le plus adapté au contexte courant peut ainsi être sélectionné. L'environnement proposé peut donc être assimilé à une approche de sélection de modèles multiples qui s'inscrit naturellement dans le cadre de l'optimisation multiobjectif. Nous avons appelé « Front ROC » la sortie d'un tel système, par analogie avec la terminologie utilisée en optimisation multiobjectif. Ce concept est illustré sur la figure 4.8. Une telle vision du problème permet à un utilisateur (éventuellement une étape ultérieure de traitement), de déplacer le problème du choix du modèle à une étape ultérieure, évitant l'injection de connaissances *a priori* qui ne sont pas toujours disponibles au moment de la conception du système. Par ailleurs, le classifieur utilisé peut être modifié au cours de la vie du système si les conditions changent, sans nécessiter un réapprentissage des données.

L'approche, qui est généralisable à n'importe quel type de classifieur muni d'hypeparamètres, a été implémentée en utilisant un classifieur de type SVM. Ce type de classifieur permet en effet de bien prendre en charge les problèmes de classification à deux classes avec des coûts de mauvaise classification asymétriques, en introduisant, à la place du paramètre C classique, deux paramètres de pénalités différents C_- et C_+ [158]. Dans ce cas, étant donné un ensemble de m exemples d'apprentissage $x_i \in \Re^n$ appartenant à la classe y_i, la maximisation du lagrangien dual par rapport aux α_i devient :

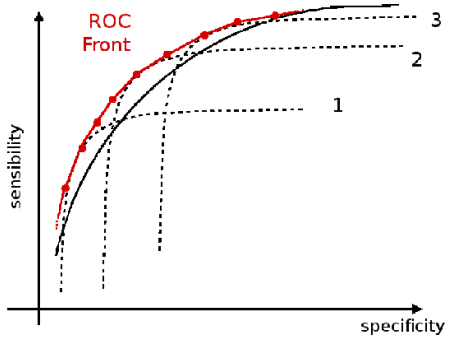


FIGURE 4.8 – Illustration synthétique du concept de Front ROC. La courbe continue est une courbe ROC correspondant à un classifieur pour lequel l'AUC a été optimisée. Les courbes 1,2 et 3 sont les courbes ROC de 3 classifieurs du Front ROC. Le Front ROC contient les parties non dominées de ces courbes.

$$Max_{\alpha}\Big\{\sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i,x_j)\Big\}$$

$$\text{sous les contraintes : } \begin{cases} 0 \leq \alpha_i \leq C_+ \quad pour \quad y_i = -1 \\ 0 \leq \alpha_i \leq C_- \quad pour \quad y_i = +1 \\ \sum_{i=1}^{m}\alpha_i y_i = 0 \end{cases}$$

où les α_i représentent les multiplicateurs de Lagrange et $K(.)$ représente la fonction noyau. Dans le cas d'un noyau gaussien, $K(.)$ est défini par :

$$K(x_i,x_j) = exp\left(-\gamma \times \|x_i - x_j\|^2\right)$$

Ainsi, dans le cas de coûts de mauvaise classification asymétriques, trois paramètres doivent être déterminés pour réaliser un apprentissage optimal de SVM :

- le paramètre du noyau, γ pour un le noyau gaussien ;
- les paramètres de pénalité introduits ci-dessus : C_- et C_+.

Dans [52], nous avons choisi l'algorithme NSGA-II proposé dans [71] pour optimiser la valeur de ces paramètres au regard des deux critères de l'espace ROC. Celui-ci est reconnu comme étant l'un des plus efficaces à la fois pour la convergence vers le front de Pareto du problème et pour la diversité des solutions. Un codage réel a été utilisé pour représenter les paramètres. Les opérateurs génétiques permettant de faire évoluer la population sont les opérateurs natifs proposés dans [71]. La stratégie utilisée est synthétisée sur la figure 4.9.

4.3.3.3 Résultats obtenus

L'approche proposée a été évaluée à la fois sur des bases de données publiques de l'*UCI Machine Learning repository* [1] et sur un problème d'analyse d'images de documents. Comme dans le cas de l'approximation de courbes présentée en 4.3.2, la comparaison avec des approches de la littérature a été

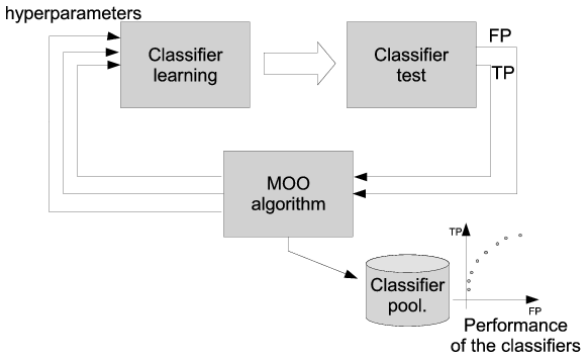


FIGURE 4.9 – Approche proposée pour la sélection de modèles multiples

rendue complexe par (i) le fait que les autres méthodes proposent généralement un classifieur unique et par (ii) la difficulté de la tâche de comparaison de sorties d'algorithmes d'optimisation muli-objectifs. Dans ce cadre, nous avons pris le parti de moyenner les performances locales des classifieurs sur le front ROC, afin d'obtenir une valeur comparable à l'AUC. En restant conscient que cette comparaison n'est théoriquement pas correcte puisque nous comparons un classifieur unique à une population de classifieurs, nous avons donc calculé une AUF (*Area Under Front*) qui peut être comparée à l'AUC obtenue par différentes approches, à savoir celles proposées dans [33] (*Decision lists* et *rules sets*), [68] (*Rankboost*), [81] (*Decision trees*), [168] (SVMs) and [214] (5 classifieurs différents). Une validation croisée sur 5 sous-ensembles a été réalisée pour attester de la stabilité des résultats.

Les résultats sont présentés dans le tableau 4.2. La première colonne contient les meilleures valeurs d'AUC trouvées dans la littérature et la seconde les valeurs d'AUF obtenues avec l'approche de sélection de modèles multiples.

problème UCI	AUC literature	ref.	AUF
australian	90.25 ± 0.6	[214]	96.22 ± 1.7
wdbc	94.7 ± 4.6	[81]	99.59 ± 0.4
breast cancer	99.13	[33]	99.78 ± 0.2
ionosphere	98.7 ± 3.3	[168]	99.00 ± 1.4
heart	92.60 ± 0.7	[214]	94.74 ± 1.9
pima	84.80 ± 6.5	[68]	87.42 ± 1.2

TABLE 4.2 – Comparaison entre l'AUC (*Area Under Curve*) obtenue par des approches de la littérature avec l'AUF (*Area Under Front*) de l'approche décrite dans [52]

Comme attendu, ces résultats montrent que le front ROC permet d'atteindre des performances qu'un classifieur unique ne permet pas d'obtenir. Même si cette comparaison est incorrecte, elle illustre toutefois le fait que l'approche proposée permet d'atteindre localement des compromis que les approches globales ne permettent pas d'atteindre.

Au vu de ces résultats, l'approche a également été testée dans le cadre de

la conception d'un système qui extrait les champs numériques (numéros de téléphone, code postal, code client ...) dans des images de courriers manuscrits [53, 54] (fig. 4.10). La principale difficulté d'une telle tâche vient du fait que les chiffres manuscrits peuvent être connectés à d'autres parties textuelles ou à des éléments graphiques du document. La figure 4.11 donne quelques exemples de composantes segmentées que le système doit reconnaître. Dans ce contexte, la détection des chiffres, leur segmentation et leur reconnaissance doivent être réalisées simultanément dans un système global. La première étape du système proposé dans [53, 54] consiste à filtrer d'abord les rejets évidents, pour éviter de leur appliquer une phase de reconnaissance coûteuse en temps de calcul. Cette étape repose sur une classification à deux classes pour laquelle les coûts de mauvaise classification sont asymétriques et inconnus. En effet, le rejet d'un chiffre peut avoir des conséquences importantes sur la détection et la reconnaissance d'un champ numérique complet mais ces conséquences ne sont pas évaluables *a priori*. Par ailleurs, ce composant de classification étant embarqué dans un système complet d'extraction de séquences numériques, il est difficile d'estimer ces coûts *a priori*.

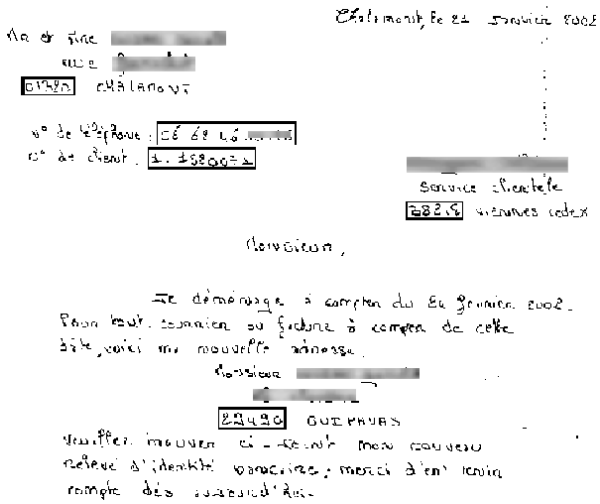


FIGURE 4.10 – Exemple d'image de courrier entrant. Les champs numériques à extraire sont surlignés.

Dans ce cadre, une base constituée de 19278 formes (1/3 digit, 2/3 outliers) a été constituée. L'approche a été évaluée en utilisant le même protocole expérimental que celui mis en œuvre pour les données de l'UCI. La courbe de la figure 4.12 illustre les résultats obtenus. Sur cette courbe, on constate que chacun des points obtenus par l'approche à base d'AUC est dominé par au moins un point du front ROC. L'approche a ainsi permis de construire un ensemble de classifieurs localement « meilleurs » que celui construit en utilisant l'approche proposée dans [168]. Chacun de ces classifieurs a ensuite été intégré dans le système complet afin d'évaluer l'influence de ces performances sur les performances en rappel et précision. Le tableau 4.3 décrit les résultats obtenus. Ils illustrent le fait que de petites différences sur les taux de vrais positifs peuvent avoir des conséquences importantes sur les performances fi-

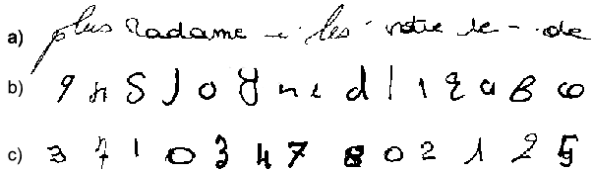


FIGURE 4.11 – Exemples de chiffres manuscrits et de rejets évidents. La première ligne (a) contient des formes qui peuvent être considérées comme des « rejets évidents ». La dernière ligne (c) contient des chiffres qui doivent être soumis au processus de reconnaissance. La ligne (b) contient les rejets ambigus, qui ressemblent à des chiffres mais qui doivent être rejetés par le système proposé.

nales du système, validant ainsi l'intérêt de ne pas avoir fait le choix d'un seul classifieur globalement bon.

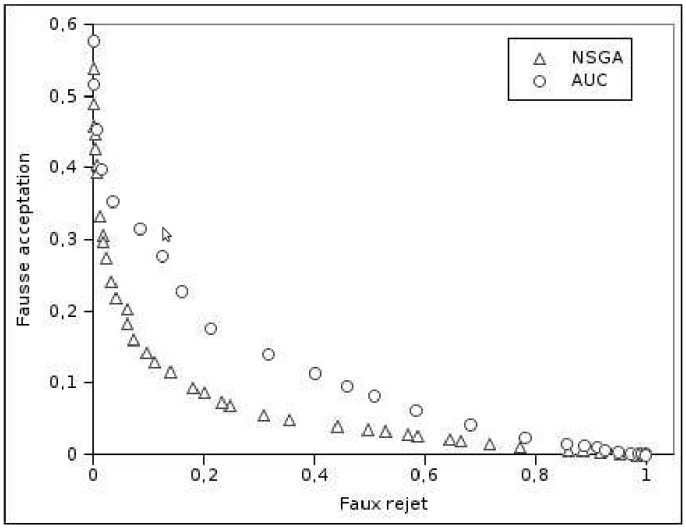


FIGURE 4.12 – Comparaison entre le Front ROC obtenu avec l'approche proposée et une courbe ROC obtenue en utilisant [168]. La courbe représente les compromis entre fausse acceptation et faux rejet de chiffres manuscrits.

Classifier TP rate	98.8	99.04	99.26	99.48	99.76	99.96	100
recall	0.370	0.410	0.440	0.458	0.462	0.481	0.488
precision	0.110	0.130	0.150	0.176	0.246	0.223	0.152
System F1-Measure	0.170	0.197	0.224	0.254	0.321	0.305	0.232

TABLE 4.3 – Précisions et Rappels obtenus pour le système complet en utilisant différents classifieurs du Front ROC, ici caractérisés par leur taux de vrais positifs.

4.4 Problèmes ouverts

Dans ce chapitre, nous avons discuté des liens qui existent entre analyse de documents, évaluation de performances et optimisation multiobjectif. Après une présentation synthétique du problème de l'optimisation multiobjectif et des solutions proposées dans la littérature, une contribution relative au domaine de l'optimisation a été proposée, au travers de l'amélioration d'un algorithme d'optimisation par essaims particulaires. Puis, les descriptions de deux contributions ont illustré le fait que la communauté de l'analyse de documents, et plus généralement de la reconnaissance de formes, pourrait tirer un grand bénéfice de la prise en compte de critères multiples, tant dans l'optique de l'évaluation de systèmes que pour le réglage et l'optimisation de ces derniers. Dans les deux cas, en dépit des difficultés liées à l'évaluation des approches, les résultats ont montré que la prise en compte d'objectifs multiples pouvait permettre de franchir un cap dans les performances par rapport à l'utilisation d'un critère scalaire.

Dans cette section, nous évoquons les perspectives directement issues de ces travaux. Une vision plus générale des pistes de recherche pour les années à venir dans la communauté sera proposée dans le chapitre 5. La première perspective importante à mentionner ici concerne la généralisation de la prise en compte d'objectifs multiples en analyse de documents, tant pour l'évaluation de performances que pour l'optimisation de systèmes. La seconde perspective importante concerne l'apprentissage multiobjectif avec une généralisation des travaux décrits en 4.3.3.

4.4.1 Analyse de documents et objectifs multiples

Les contributions proposées dans ce chapitre ont montré l'intérêt de la prise en compte de critères multiples pour l'évaluation et pour l'optimisation de certains composants de systèmes d'analyse d'images de documents. L'une de nos perspectives de recherche à court terme consiste naturellement à généraliser ce point de vue à d'autres outils utilisés dans les chaînes d'analyse. Dans ce contexte, les contributions décrites dans le chapitre 3, toutes deux basées sur des processus d'optimisation, constituent un très bon cadre d'étude.

Pour la classification de graphes, de premiers travaux relatifs à l'intégration d'un critère de rejet lors de la génération des prototypes sont en cours. Ces travaux permettent d'offrir en sortie un ensemble de solutions parmi lesquelles il est possible de choisir le compromis erreur/rejet qui convient le mieux. La figure 4.13 illustre les premiers résultats obtenus dans le cadre de ces travaux. Sur cette figure, une courbe donnée correspond aux différents compromis erreur/rejet obtenus par différents ensembles de prototypes optimisés avec un algorithme d'optimisation multiobjectif. Les différentes courbes correspondent à l'évolution de la population au cours des générations de l'algorithme d'optimisation. Ces premiers résultats sont très encourageants puisqu'ils montrent que l'algorithme permet d'une part d'améliorer les performances aux cours des générations et, d'autre part, de fournir des solutions diversifiées.

Pour la contribution relative à l'isomorphisme de sous-graphes et son application à des problèmes de localisation de symboles, le système proposé en 3.4.1

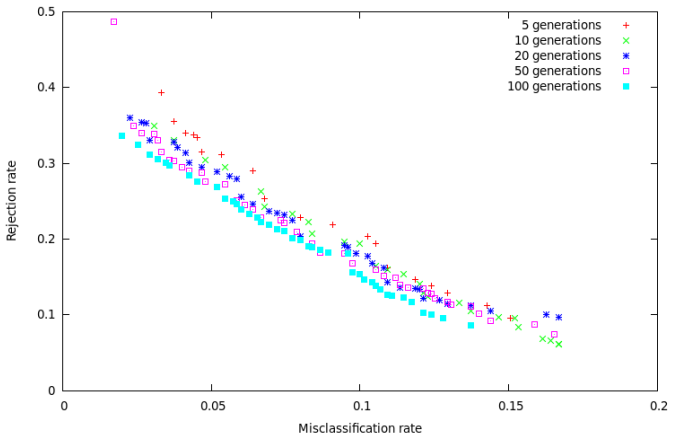


FIGURE 4.13 – Évolution des performances des différents ensembles de prototypes en fonction du nombre de générations de l'algorithme d'optimisation

est évalué par les deux critères classiques en recherche d'information que sont la précision et le rappel. L'optimisation d'une telle application pourrait donc naturellement bénéficier de la prise en compte de critères multiples. Dans le cas de la recherche d'isomorphismes tolérants aux substitutions, nous envisageons en particulier d'optimiser les fonctions de coûts c_V et c_E de l'équation 3.11a au regard de ces deux critères. Le bénéfice serait, là encore, de proposer un ensemble de compromis plutôt qu'une solution unique en sortie de l'algorithme.

4.4.2 Apprentissage multiobjectif

Ces perspectives de recherche font suite aux travaux menés en collaboration avec Clément Chatelain concernant le développement d'un cadre multicritère pour l'apprentissage automatique. Elles ont fait l'objet d'une soumission nommée LeMOn (LEarning with Multi-objective OptimizatioN) lors de l'appel ANR Jeunes Chercheurs et Jeunes Chercheuses 2011 [16]. Elles concernent deux aspects particuliers de l'apprentissage que nous souhaiterions aborder sous l'angle de l'optimisation multiobjectif.

Le premier aspect concerne l'exploitation de l'espace ROC lors de l'apprentissage des classifieurs. Dans [52], nous avons proposé un environnement de sélection de modèles basé sur une approche d'optimisation multiobjectif. Cet environnement permet de construire un ensemble de classifieurs à deux classes localement optimaux dans l'espace ROC, plutôt qu'un unique basé sur un critère scalaire. Les perspectives ouvertes par ce travail concernent deux verrous. Le premier est le passage à l'échelle de l'approche afin d'appréhender de très grands volumes de données. En effet, pour de tels volumes, la stratégie évolutionnaire proposée dans [52], qui repose sur de nombreux apprentissages de classifieurs, devient très coûteuse en temps de calcul. La piste envisagée dans le cadre du projet LeMOn pour pallier ce problème consiste à combiner l'ap-

16. Le projet, dont je suis le porteur, est, à l'heure de l'écriture de ce manuscrit, en seconde position sur liste complémentaire

proche évolutionnaire avec les travaux récemment proposés dans notre équipe pour l'apprentissage de SVM adapté aux problèmes de type Neyman-Person. Le second verrou concerne la généralisation de l'approche proposée à des problèmes multi-classes, pour lesquels le nombre de critères croît rapidement avec le nombre de classes ($N(N-1)$ critères sont à considérer pour un problème à N classes). Il sera alors nécessaire d'adapter l'approche proposée et en particulier d'envisager l'intégration d'opérateurs génétiques dédiés permettant d'accélérer la convergence de l'algorithme.

Le second aspect de l'apprentissage que nous envisageons d'aborder sous l'angle de l'optimisation multiobjectif est celui de l'apprentissage multi-tâches, qui consiste à apprendre simultanément plusieurs modèles par des transferts de connaissances d'un modèle vers l'autre. Ce paradigme a récemment permis d'obtenir de très bons résultats pour différentes applications [27, 12, 28, 102]. Dans la littérature, ce problème est aujourd'hui formulé comme un problème d'optimisation pour lequel les objectifs relatifs à chacune des tâches sont combinés, en y ajoutant un terme de régularisation tel que :

$$\min_{f_1,\cdots,f_T} \sum_{t,i} a_t \cdot L_t(f_t(x_{i,t}), y_{i,t}) + \lambda\Omega(f_1, \cdots, f_T). \tag{4.8}$$

où $L_t(f_t(x), y)$ est la fonction de perte, Ω est un terme de régularisation impliquant les fonctions de pertes liées à toutes les tâches f_t. Les $\{a_t\}$ et λ sont des paramètres de pondération de chacun des objectifs.

Nous envisageons dans le cadre du projet LeMOn d'explorer le potentiel de l'optimisation multiobjectif à base de Pareto pour traiter ce genre de problème, afin de fournir en sortie un ensemble de solutions.

Notons que dans le projet LeMOn, il est prévu d'appliquer ces différents travaux à deux domaines d'application : l'analyse d'images médicales, en collaboration avec l'équipe Quantif du LITIS, et les interfaces cerveau-machine, en collaboration avec des chercheurs de l'équipe DocApp s'intéressant à cette problématique.

Chapitre 5

Perspectives

L'année 2011 fut l'occasion pour la communauté de l'analyse d'images de documents de célébrer le vingtième anniversaire de la conférence internationale sur l'analyse et la reconnaissance de documents (*International Conference on Document Analysis and Recognition - ICDAR*). Depuis sa première occurrence en 1991 à Saint-Malo, les recherches menées dans le domaine ont été à l'origine de nombreux succès, dont certains ont même conduit à l'industrialisation de solutions logicielles : pour la lecture de chèques, d'adresses postales et de formulaires pour ne citer que ces exemples. Le constat est identique dans le domaine plus ciblé de l'analyse de documents graphiques qui m'a particulièrement intéressé dans mon parcours de chercheur. On peut en effet raisonnablement considérer aujourd'hui que certains outils, ceux qui sont réellement spécifiques aux documents graphiques (segmentation texte/graphique, vectorisation, reconnaissance de caractères multi-orientés et multi-échelles ...), ont atteint une maturité suffisante, en dépit des résultats imparfaits qu'ils permettent d'obtenir [204, 205].

Toutefois, ces succès ne doivent pas masquer le nombre et l'ampleur des défis qui restent encore à relever dans ce domaine. En effet, comme en témoignent les compétitions et les événements scientifiques nationaux et internationaux toujours plus nombreux, de nombreux verrous liés à l'analyse d'images de documents restent encore à lever. Les deux applications abordées dans ce mémoire, respectivement dédiées à la localisation de symboles dans des documents graphiques (§3.4.1) et à la détection de séquences numériques dans des courriers manuscrits (§4.3.3) ne sont que deux exemples des problématiques qui sont encore loin d'être résolues. Par ailleurs, de nouveaux usages émergent toujours et font eux-mêmes apparaître d'autres défis scientifiques. Les nombreux projets récents ayant trait à la valorisation de fonds documentaires anciens (NAVIDOMASS, IMPACT...) constituent une parfaite illustration de ces aspects. Dans ce cadre, les nouvelles problématiques concernent l'extraction de la structure de documents complexes, la reconnaissance de caractères dégradés ou l'analyse de lettrines. Les perspectives de recherche pour la communauté de l'analyse de documents sont donc encore extrêmement nombreuses et il y a fort à parier que ICDAR fêtera ses 40 ans en 2031.

Parmi ces perspectives, celles qui nous paraissent être les plus prometteuses au regard de nos travaux antérieurs sont décrites dans la suite de ce

chapitre. Plusieurs pistes de recherche ont déjà été présentées dans le corps de ce mémoire. En effet, pour en faciliter la lecture, nous avons fait le choix de développer les perspectives directement liées à nos contributions à l'issue de la présentation de celles-ci (cf. §3.5 et §4.4). Nous rappelons ici les deux propositions que nous considérons comme les plus ambitieuses en précisant le contexte dans lequel nous envisageons de mener ces travaux.

La première de ces pistes concerne la poursuite de nos travaux sur la recherche d'isomorphismes de sous-graphes, pour rendre l'approche proposée tolérante à des modifications de la topologie des graphes telles que l'absence dans le graphe cible de sommets ou d'arcs pouvant être associés à ceux du graphe modèle. Ces travaux sont menés dans le cadre d'une collaboration avec des chercheurs de la communauté de la recherche opérationnelle et plus particulièrement de la programmation mathématique (issus du LMI de Rouen et du LRI d'Orsay). Les échanges que nous avons dans le cadre de cette coopération, qui visent à optimiser l'utilisation des solveurs et à asseoir théoriquement les approches proposées, s'avèrent très prometteurs. Ces travaux constituent selon nous un challenge important, qui dépasse l'enjeu applicatif de la localisation de symboles, en permettant d'envisager de manière plus générale la localisation d'objets bruités et non segmentables dans des images.

La seconde piste importante que nous allons aborder dans les mois à venir concerne l'apprentissage multiobjectif, dans le cadre du projet LEMON (LEarning with Multiobjective OptimizatioN). Après s'être concentrée pendant plus de deux décennies sur des critères de performances scalaires, la communauté des chercheurs en apprentissage commence à examiner l'utilisation de critères multiples, comme en témoigne le récent ouvrage [105]. Ces travaux soulèvent de nouveaux problèmes théoriques et motivent la recherche de nouveaux algorithmes d'apprentissage. Nos perspectives dans ce cadre ont été développées en 4.4. De plus, ils permettent aussi d'envisager des avancées significatives dans les domaines de la reconnaissance de formes et de l'optimisation. Le consortium de chercheurs constitué dans le cadre de LEMON, avec ses compétences complémentaires (reconnaissance de formes, apprentissage statistique, optimisation multiobjectif, interfaces cerveaux machines, imagerie médicale), nous semble un excellent cadre de travail pour contribuer à ces avancées.

Outre ces travaux directement liés à nos contributions antérieures, nous souhaitons aussi profiter de cette conclusion pour mentionner un certain nombre de problématiques qui n'ont pas encore été abordées dans le corps de ce manuscrit et qui ouvrent elles aussi la voie à des travaux prometteurs pour les années à venir.

La première de ces problématiques est celle de l'évaluation de performances, qui fait depuis une dizaine d'années l'objet d'un vif intérêt de la communauté scientifique, comme en témoignent les très nombreuses campagnes en cours, que celles-ci concernent l'extraction d'information, la recherche d'information ou l'analyse d'images. Dans le domaine de l'analyse d'images de documents, on peut citer les campagnes RIMES, dédiée à la reconnaissance de l'écriture manuscrite, EPEIRES pour la reconnaissance et la localisation de symboles ainsi que les très nombreux concours qui sont organisés de façon récurrente lors des conférences ICDAR et GREC. Depuis juin 2011, nous participons à un projet

triennal qui réunit plusieurs industriels et un consortium de laboratoires de recherche et qui est dédié à l'évaluation de performances de systèmes de reconnaissance de documents écrits. Le projet vise deux objectifs ambitieux. Le premier consiste à mettre en place une campagne ouverte d'évaluation d'une chaîne complète d'analyse de documents. Le second vise la réalisation d'un démonstrateur intégrant une chaîne de traitement optimisée pour la reconnaissance de documents manuscrits et/ou dactylographiés.

Pour mener à bien ce projet, au delà du développement de modules de traitements, les différents aspects liés à la mise en place d'une campagne d'évaluation seront abordés. Le premier consiste naturellement à proposer un corpus d'un nombre conséquent de documents à la fois manuscrits et dactylographiés, libres de droit, très variés et surtout très réalistes. Ce corpus sera annoté pour établir une vérité terrain en détaillant, sur chaque document, les différents éléments à reconnaître par les outils évalués. Enfin, dans le cadre de l'évaluation des briques proposées par la communauté, une réflexion sur les métriques permettant d'évaluer les approches sera également menée. Notre contribution dans ce cadre consistera à prendre en considération certains aspects traités dans ce mémoire, en donnant une coloration résolument multiobjectif aux métriques. À titre d'illustration, dans un contexte de discrimination, il pourrait s'agir de demander aux participants aux campagnes d'évaluation de fournir les sorties des systèmes pour différents points de fonctionnement. Pour comparer de telles sorties, les métriques pourraient s'inspirer des travaux menés en évaluation de performances d'algorithmes d'optimisation multiobjectif [198].

Le second objectif de ce projet, qui consiste en la mise en œuvre d'une chaîne optimisée de traitement de documents, apporte lui aussi son lot de perspectives scientifiques. La première concerne l'interopérabilité des composants. Pour constituer la chaîne optimale mise en œuvre dans le démonstrateur, il est probable de devoir associer des composants issus de différents laboratoires. Dans ce cadre, l'approche envisagée pour surmonter cette difficulté est proche des travaux que nous avons proposés dans le cadre du projet Docmining [4], mais adaptés à un contexte de services Web. Comme dans les travaux proposés dans [121], elle repose sur l'utilisation d'une plateforme d'intégration orientée service nommée WebLab[17]. Cette plateforme, que nous utilisons par ailleurs dans nos travaux en recherche d'information, a été conçue pour construire des applications de traitement d'informations multimédia en faisant interopérer des composants logiciels spécialisés.

Outre l'intérêt que revêt à lui seul ce projet, il ouvre par ailleurs des perspectives à plus long terme particulièrement intéressantes. Il permet en effet d'envisager la constitution d'une bibliothèque d'outils divers et interopérants dédiés aux différentes tâches d'un système d'analyse de documents. La disponibilité d'une telle « batterie » d'outils pourrait alors servir de socle à des travaux dans le domaine de la planification, dont le but serait la génération automatique et adaptative de chaînes de traitements en fonction d'un but (segmenter, reconnaitre, localiser) et d'un contexte (le document). Il s'agirait alors d'apprendre, au regard de l'objectif fixé, la séquence d'outils permettant de maximiser les performances d'un système, éventuellement dans un cadre mul-

17. `http://weblab.ow2.org/`

tiobjectif pour laisser à l'utilisateur la possibilité de choisir parmi différentes options. Ces performances pourraient en effet être évaluées à partir de la vérité terrain fournie par le projet. L'une des pistes possibles pour optimiser de telles chaînes pourrait être l'apprentissage par renforcement qui propose un environnement particulièrement puissant pour l'optimisation de séquences, comme l'ont récemment montré nos travaux en recherche d'information [74]. Ces problèmes de planification de chaînes complètes d'analyse de documents constituent selon nous un véritable challenge pour l'avenir, et dont les résultats pourraient par ailleurs avoir des conséquences dans bien d'autres domaines d'application. Dans cet esprit, on pourrait même, à beaucoup plus long terme, envisager la coopération de systèmes divers tels que des systèmes d'analyse de documents ou d'images, des moteurs de recherches, des outils de traduction. . . .

Un dernier aspect que nous souhaitons aborder ici concerne la place de l'Homme dans ces systèmes coopérants de traitement de l'information au sens large. La prise en compte des interactions entre le système et l'humain est en effet indispensable à la réussite de tels projets, que ce soit pour leur conception ou pour l'utilisation des résultats qu'ils produisent. De ce point de vue, nous pensons que des collaborations avec les équipes travaillant dans le domaine de la recherche d'information seraient particulièrement enrichissantes. Cette communauté s'intéresse en effet depuis longtemps aux interactions, par l'intermédiaire des principes de retours de pertinence ou de personnalisation des moteurs de recherche par exemple. Ces toutes dernières années ont d'ailleurs été le cadre d'un rapprochement des communautés Françaises de l'analyse de documents et de la recherche d'information, comme en témoigne le regroupement en 2010 des conférences CIFED et CORIA. Un autre exemple de cette convergence est le projet fédérateur du LITIS nommé PlaIR [18] (Plateforme d'Indexation Régionale). Ce projet se donne pour objectif de mutualiser un ensemble de ressources documentaires numériques et numérisées et de bibliothèques logicielles d'analyse automatique ou semi-automatique pour constituer une plateforme d'indexation et de recherche multi-domaines et multi-usages. Dans ce contexte, des travaux ont été initiés dans le cadre de la thèse de Gérard Dupont [74], en collaboration avec CASSIDIAN. L'objectif de ces travaux était de créer le lien entre les domaines de la recherche d'information et de l'apprentissage par la mise en œuvre d'algorithmes pour adapter les réponses d'un système de recherche d'information aux utilisateurs de celui-ci. Ces travaux sont actuellement poursuivis par ceux de la thèse CIFRE d'Aurélien Saint Réquier, avec CASSIDIAN, dont le but est de proposer un agent personnel d'assistance à la recherche d'information.

Pour conclure ce manuscrit, s'il y a une chose primordiale que je retire de ces dix années de recherche et que j'ai souhaité faire transparaître au travers de ces quelques pages, c'est l'importance que revêtent le décloisement des disciplines, l'ouverture vers d'autres communautés, et les convergences entre recherches fondamentales et appliquées. Même si les évolutions actuelles de la recherche sont trop souvent orientées vers la « compétition » entre équipes, entre individus, je reste persuadé que la richesse vient et continuera à venir du partage.

18. `http://plair.org`

Chapitre 6

Bibliographie

[1] D.J. Newman A. Asuncion. UCI machine learning repository, 2007.

[2] S. Adam. *Interprétation de documents techniques : des outils à leur intégration dans un système à base de connaissances*. PhD thesis, Université de Rouen, 2001.

[3] S. Adam and J.M. Ogier. Documents graphiques : de la rétroconversion à la recherche d'information. In Rémy Mullot, editor, *Les documents écrits : De la numérisation à l'indexation par le contenu*, pages 249–310. Hermès, 2006.

[4] S. Adam, M. Rigamonti, E. Clavier, J-M. Ogier, E. Trupin, and K. Tombre. DocMining : A Document Analysis System Builder. In S. Marinai and A. Dengel, editors, *Proceedings of the Workshop on Document Analysis Systems (DAS'04)*, volume 3163 of *Lecture Notes in Computer Science*, pages 472–483, 2004.

[5] S. Adra, I. Griffin, and P. Fleming. A comparative study of progressive preference articulation techniques for multiobjective optimisation. In Shigeru Obayashi, Kalyanmoy Deb, Carlo Poloni, Tomoyuki Hiroyasu, and Tadahiko Murata, editors, *Evolutionary Multi-Criterion Optimization*, volume 4403 of *Lecture Notes in Computer Science*, pages 908–921. Springer Berlin / Heidelberg, 2007.

[6] H.S.M. Al-Khaffaf, A.Z. Talib, and M.A. Osman. GREC'11 arc segmentation contest : Performance evaluation on multi-resolution scanned documents. In *Proceedings of the IAPR Workshop on Graphics Recognition (GREC'11)*, 2007.

[7] I. Alaya, C. Solnon, and K. Ghedira. Ant colony optimization for multi-objective optimization problems. In *Proceedings of the International Conference on Tools with Artificial Intelligence (ICTAI'07)*, pages 450–457.

[8] J.E. Alvarez-Benitez, R.M. Everson, and J.E. Fieldsend. MOPSO algorithm based exclusively on pareto dominance concepts. *Proceedings of the International Conference on Evolutionary Mutli-Criterion Optimization (EMO'05)*, pages 726–732, 2005.

[9] D. Anguita, S. Ridella, F. Rivieccio, and R. Zunino. Hyperparameter de-

sign criteria for support vector classifiers. *Neurocomputing*, 55(1-2) :109–134, 2003.

[10] N. Ansari and K-W. Huang. Non-parametric dominant point detection. *Pattern Recognition (PR)*, 24(9) :849–862, 1991.

[11] D.L. Applegate, R.E. Bixby, V. Chvatal, and W.J. Cook. *The Traveling Salesman Problem : A Computational Study (Princeton Series in Applied Mathematics)*. Princeton University Press, 2007.

[12] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[13] T.W. Athan and P.Y. Papalambros. A note on weighted criteria methods for compromise solutions in multi-objective optimization. *Engineering Optimization*, 27(2) :155–176, 1996.

[14] S. Auwatanamongkol. Inexact graph matching using a genetic algorithm for image recognition. *Pattern Recognition Letters (PRL)*, 28(12) :1428–1437, 2007.

[15] N.E. Ayat, M. Cheriet, and C.Y. Suen. Automatic model selection for the optimization of SVM kernels. *Pattern Recognition (PR)*, 30(10) :1733–1745, 2004.

[16] A.D. Bagdanov and M. Worring. Fine-grained document genre classification using first order random graphs. In *Proceedings of the International Conference opn Document Analysis and Recognition (ICDAR'01)*, pages 79 – 83, 2001.

[17] J. Balicki. An adaptive quantum-based multiobjective evolutionary algorithm for efficient task assignment in distributed systems. *Proceedings of the WSEAES international conference on Computers (ICCOMP'09)*, pages 417–422, 2009.

[18] E. Barbu. *Fouille et classification de graphes : application à la reconnaissance de symboles dans les documents graphiques*. PhD thesis, Université de Rouen, 2007.

[19] A. Belaïd and K. Ossama. Goal programming model : A glorious history and a promising future. *European Journal of Operational Research (EJOR)*, 133(2) :225 – 231, 2001.

[20] Y. Bengio. Gradient-based optimization of hyperparameters. *Neural Computation*, 12(8) :1889–1900, 2000.

[21] S. Bernard. *Forêts Aléatoires : de l'Analyse des Mécanismes de Fonctionnement à la Construction Dynamique*. PhD thesis, Université de Rouen, 2009.

[22] S. Bernard, L. Heutte, and S. Adam. Etude de l'influence des paramètres sur les performances des forêts aléatoires. In *Actes du Colloque International Francophone sur l'Écrit et le Document (CIFED'08)*, pages 207–208, 2008.

[23] S. Bernard, L. Heutte, and S. Adam. Forest-RK : A new random forest induction method. In De-Shuang Huang, Donald C. Wunsch II,

Daniel S. Levine, and Kang-Hyun Jo, editors, *Proceedings of the International Conference on Intelligent Computing (ICIC'08)*, volume 5227 of *Lecture Notes in Computer Science*, pages 430–437. Springer, 2008.

[24] S. Bernard, L. Heutte, and S. Adam. Influence of hyperparameters on random forest accuracy. In Jon Atli Benediktsson, Josef Kittler, and Fabio Roli, editors, *Proceedings of Multiple Classifier Systems (MCS'09)*, volume 5519 of *Lecture Notes in Computer Science*, pages 171–180. Springer, 2009.

[25] S. Bernard, L. Heutte, and S. Adam. Une Étude sur la paramétrisation des forêts aléatoires. In *Actes de la Conférence francophone sur l'Apprentissage Artificiel (CAP'09)*, pages 81–92, 2009.

[26] J. C. Bezdek, T. R. Reichherzerand, G. S. Lim, and Y. Attikiouzel. Multiple-prototype classifier design. *IEEE Transaction on Systems, Man, and Cybernetics Part C (IEEE SMC)*, 28(1) :67–79, 1998.

[27] J. Bi, T. Xiong, S. Yi, M. Dundar, and B. Rao. An improved multi-task learning approach with applications in medical diagnosis. In *Proceedings of the European Conference on Machine Learning (ECML'08)*, 2008.

[28] S. Bickel, J. Bogojeska, T. Lengauers, and T. Scheffer. Multi-task learning for hiv therapy screening. In *Proceedings of the International Conference on Machine learning (ICML'08)*, pages 56–63, 2008.

[29] T.T. Binh and U. Korn. MOBES : A multiobjective evolution strategy for constrained optimization problems. In *Proceedings of the International Conference on Genetic Algorithms (ICGA'97)*, pages 176–182, 1997.

[30] P. Le Bodic, S. Adam, P. Héroux, A. Knippel, and Y. Lecourtier. Formulations linéaires en nombres entiers pour des problèmes d'isomorphisme exact et inexact. In *Actes des Journées Polyèdres et Optimisation Combinatoire (JPOC'08)*, 2008.

[31] P. Le Bodic, H. Locteau, S. Adam, P. Héroux, Y. Lecourtier, and A. Knippel. Symbol detection using region adjacency graphs and integer linear programming. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'09)*, pages 1320–1324, 2009.

[32] B. Bonev, F. Escolano, M.A. Lozano, P. Suau, M. Cazorla, and W. Aguilar. Constellations and the unsupervised learning of graphs. In *Proceedings of the Workshop on Graph-based Representations in Pattern Recognition (GBRPR'07)*, pages 340–350, 2007.

[33] H. Boström. Maximizing the area under the roc curve using incremental reduced error pruning. In *Proceedings of the Workschop of ROC Analysis in Machine Learning (ROCML'05)*, 2005.

[34] A.P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition (PR)*, 30(7) :1145–1159, 1997.

[35] J. Branke and S. Mostaghim. About selecting the personal best in multiobjective particle swarm optimization. In *Parallel Problem Solving from Nature*, volume 4193 of *Lecture Notes in Computer Science*, pages 523–532. Springer, 2006.

[36] L. Breiman. Random forests. *Machine Learning Journal (MLJ)*, 45(1) :5–32, 2001.

[37] N. Brown, BN. McKay, F. Gilardoni, and J. Gasteiger. A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *Journal of Chemical Information and Modeling (JCIM).*

[38] L.T. Bui, D. Essam, H.A. Abbass, and D. Green. Performance analyis of multiobjective evolutionary methods in noisy environnments. *Complexity International*, 11 :29–39, 2005.

[39] H. Bunke. On a relation between graph edit distance and maximum common subgraph. *Pattern Recognition Letters (PRL)*, 18(8) :689–694, 1997.

[40] H. Bunke, P. Foggia, C. Guidobaldi, and M. Vento. Graph clustering using the weighted minimum common supergraph. In *Proceedings of the Workshop on Graph-based Representations in Pattern Recognition (GBRPR'03)*, pages 235–246, 2003.

[41] H. Bunke, A. Münger, and X. Jiang. Combinatorial search versus genetic algorithms : A case study based on the generalized median graph problem. *Pattern Recognition Letters (PRL)*, 20(11) :1271–1277, 1999.

[42] H. Bunke and K. Riesen. Recent advances in graph-based pattern recognition with applications in document analysis. *Pattern Recognition (PR)*, 44(5) :1057–1067, 2011.

[43] H. Bunke and K. Shearer. A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters (PRL)*, 19(3-4) :255–259, 1998.

[44] L. Cagnina, S. Esquivel, and C.A.C. Coello. A particle swarm optimizer for multi-objective optimization. *Journal of Computer Science and Technology (JCST)*, 5(4), 2005.

[45] A. Carmona-Poyato, F.J. Madrid-Cuevas, R. Medina-Carnicer, and R. Munoz-Salinas. Polygonal approximation of digital planar curves through break point suppression. *Pattern Recognition (PR)*, 43(1) :14–25, 2010.

[46] A. Carmona-Poyato, R. Medina-Carnicer, F.J. Madrid-Cuevas, R. Muoz-Salinas, and N.L. Fernández-García. A new measurement for assessing polygonal approximation of curves. *Pattern Recognition (PR)*, 44(1) :45–54, 2011.

[47] L. Cecchini, C.M. Lorenzetti, A.G. Maguitman, and N.B. Brignole. Multiobjective evolutionary algorithms for context-based search. *Journal of the American Society for Information Science and Technology (JASIST)*, 61(6) :1258–1274, 2010.

[48] V. Srinivasa Chakravarthy and B. Kompella. The shape of handwritten characters. *Pattern Recognition Letters (PRL)*, 24(12) :1901 – 1913, 2003.

[49] C.-J. C. Fu Chang and C.-J. Lu. A linear-time component-labeling algorithm using contour tracing technique. *Computer Vision and Image Understanding (CVIU)*, 93 :206–220, 2004.

[50] C-L. Chang. Finding prototypes for nearest neighbor classifiers. *IEEE Transaction on Computers (IEEE TC)*, 23(11) :1179–1184, 1974.

[51] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning Journal (MLJ)*, 46(1) :131–159, 2002.

[52] C. Chatelain, S. Adam, Y. Lecourtier, L. Heutte, and T. Paquet. A multi-model selection framework for unknown and/or evolutive misclassification cost problems. *Pattern Recognition (PR)*, 43(3) :815–823, 2010.

[53] C. Chatelain, L. Heutte, and T. Paquet. Segmentation-driven recognition applied to numerical field extraction from handwritten incoming mail documents. *Proceedings of Document Analysis System (DAS'06)*, pages 564–575, 2006.

[54] C. Chatelain, L. Heutte, and T. Paquet. A two-stage outlier rejection strategy for numerical field extraction in handwritten documents. In *Proceedings of the International Conference on Pattern Recognition (ICPR'06)*, pages 224–227, 2006.

[55] J-M. Chen, J.A. Ventura, and C-H. Wu. Segmentation of planar curves into circular arcs and line segments. *Image and Vision Computing (IVC)*, 14(1) :71 – 83, 1996.

[56] N. Chen and D. Blostein. A survey of document image classification : problem statement, classifier architecture and performance evaluation. *International Journal on Document Analysis and Recognition (IJDAR)*, 10(1) :1–16, 2007.

[57] W. Chen, A. Sahai, A. Messac, and G.J. Sundararaj. Exploration of the effectiveness of physical programming in robust design. *Journal of Mechanical Design (JMD)*, 122(2) :155–163, 2000.

[58] W.Y. Chen, W.L. Hwang, and T.C. Lin. Planar-shape prototype generation using a tree-based random greedy algorithm. *IEEE Transaction on Systems, Man, and Cybernetics (IEEE SMC) Part B*, 36(3) :649–659, 2006.

[59] B. Chin-Wei and M. Rajeswari. Multiobjective optimization approaches in image segmentation - the directions and challenges. *International Journal in Advance in Soft Computing Application (IJASCA)*, 2(1) :40–65, 2010.

[60] C.A.C. Coello. Evolutionary multiobjective optimization. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 1(5) :444 – 447, 2011.

[61] C.A.C. Coello and G.B. Lamont. *Applications of Multi-Objective Evolutionary Algorithms*. World Scientific Publishing, 2004.

[62] D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 18(3) :266–298, 2004.

[63] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento. Performance evaluation of the VF graph matching algorithm. In *Proccedings of the International Conference on Image Analysis and Processing (ICIAP'99)*, pages 1172–1177, 1999.

[64] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento. Fast graph matching for detecting CAD image components. In *Proceedings of the International Conference on Pattern Recognition (ICPR'00)*, pages 6034–6037, 2000.

[65] L.P. Cordella, P. Foggia, C. Sansone, and M. Vento. A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Transaction on Pattern Analysis and Machine Intelligence (IEEE PAMI)*, 26(10) :1367–1372, 2004.

[66] D.W. Corne, J.D. Knowles, and M.J. Oates. The Pareto envelope-based selection algorithm for multiobjective optimization. In *Proceedings of the international conference on Parallel problem solving from nature (PPSN'00)*, pages 839–848, 2000.

[67] P. Cornic. Another look at the dominant point detection of digital curves. *Pattern Recognition Letters (PRL)*, 18(1) :13–25, 1997.

[68] C. Cortes and M. Mohri. AUC optimization vs. error rate minimization. In *Advances in NIPS*. MIT Press, 2004.

[69] B. V. Dasarathy. *Nearest neighbor (NN) norms : NN pattern classification techniques*. Los Alamitos : IEEE Computer Society Press, 1990, 1990.

[70] K. Deb. *Multi-Objective Optimization Using Evolutionary Algorithms*. Wiley, 2001.

[71] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan. A fast elitist nondominated sorting genetic algorithm for multiobjective optimization : NSGA-II. *IEEE Transactions on Evolutionary Computation (IEEE TEC)*, 6(2) :182–197, 2002.

[72] G. Sanniti di Baja and E. Thiel. Skeltonization algorithm running on path-based distance maps. *Image and Vision Computing (IVC)*, 14(1) :47–57, 1996.

[73] P. Dosch, E. Valveny, A. Fornes, and S. Escalera. Report on the Third Contest on Symbol Recognition. In Josep Lladós Wenyin Liu and Jean-Marc Ogier, editors, *Graphics Recognition. Recent Advances and New Opportunities*, volume 5046 of *Lecture Notes in Computer Science*, pages 321–328. Springer, 2008.

[74] G. Dupont. *Apprentissage implicite pour la recherche d'information*. PhD thesis, Université de Rouen, 2011.

[75] G. Dupont, S. Adam, Y. Lecourtier, and B. Grilhère. Multi objective particle swarm optimization using enhanced dominance and guide selection. *International Journal of Computational Intelligence Research (IJCIR)*, 4(2) :145–158, 2008.

[76] R.M. Everson and J.E. Fieldsend. Multi-objective optimisation for receiver operating characteristic analysis. In *Multi-Objective Machine Learning*, pages 533–556. 2006.

[77] L. Fei-Fei and P. Perona. A bayesian hierarcical model for learning natural scene categories. In *Proceedings of the international conference on Computer Vision and Pattern Recognition (CVPR)*, pages 524–531, 2005.

[78] M. Ferrer, F. Serratosa, and E. Valveny. On the relation between the median and the maximum common subgraph of a set of graphs. In *Proceedings of the Workshop on Graph-based Representations in Pattern Recognition (GBRPR'07)*, pages 351–360, 2007.

[79] M. Ferrer, E. Valveny, and F. Serratosa. Spectral median graphs applied to graphical symbol recognition. In *Proceedings of the Iberoamerican Congress on Pattern Recognition (CIARP'06)*, pages 774–783, 2006.

[80] C. Ferri and P. Flach. Learning decision trees using the area under the roc curve. In *Proceedings of the International Conference on Machine Learning (ICML'02)*, pages 139–146, 2002.

[81] C. Ferri, P. Flach, and J. Hernandez-Orallo. Learning decision trees using the area under the roc curve. In *Proceedings of the International Conference on Machine Learning (ICML'02)*, pages 139–146, 2002.

[82] C.M. Fonseca and P.J. Flemming. Genetic algorithm for multiobjective optimization : formulation, discussion and generalization. In *Proceedings of the International Conference on Genetic Algorithms (ICGA'93)*, pages 416–423, 1993.

[83] X. Gao, B. Xiao, D. Tao, and X. Li. A survey of graph edit distance. *Pattern Analysis & Applications (PAA)*, 13(1) :113–129, 2010.

[84] M. R. Garey and D. S. Johnson. *Computers and Intractability : A Guide to the Theory of NP-Completeness*. Freeman & co., 1979.

[85] T. Gartner, P. Flach, and S. Wrobel. On graph kernels : Hardness results and efficient alternatives. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *COLT*, volume 2777, pages 129–143. Springer-Verlag ; 1999, 2003.

[86] E.N. Gerasimov and V.N. Repko. Multicriterial optimization. *International Applied Mechanics*, 14(11) :1179–1184, 1978.

[87] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[88] E. Grosicki and H. El Abed. ICDAR 2011 - French handwriting recognition competition. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'11)*, pages 1459–1463, 2011.

[89] I. Guyon, A. Saffari, G. Dror, and G. Cawley. Model selection : Beyond the Bayesian/frequentist divide. *Journal of Machine Learning Research (JMLR)*, 11 :61–87, 2010.

[90] Y.Y. Haimes, L.S. Lasdon, and D.A. Wismer. On a Bicriterion Formulation of the Problems of Integrated System Identification and System Optimization. *IEEE Transactions on Systems, Man and Cybernetics (IEEE SMC)*, 1(3) :296–297, 1971.

[91] P. E. Hart. The condensed nearest neighbour rule. *IEEE Transaction on Information Theory (IEEE TIT)*, 14(5) :515–516, 1968.

[92] X. Hilaire and K. Tombre. Robust and accurate vectorization of line drawings. *IEEE Transaction on Pattern Analysis and Machine Intelligence (IEEE PAMI)*, 28(6) :890–904, 2006.

[93] A. Hlaoui and S. Wang. Median graph computation for graph clustering. *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 10(1) :47–53, 2005.

[94] J. Horn, N. Nafpliotis, and D.E. Goldberg. A niched Pareto genetic algorithm for multiobjective optimization. In *Proceedings of the IEEE World Congress on Computational Intelligence (WCCI'94)*, pages 82–87, 1994.

[95] J-H Horng. An adaptive smoothing approach for fitting digital planar curves with line segments and circular arcs. *Pattern Recognition Letters (PRL)*, 24(1-3) :565 – 577, 2003.

[96] J-H. Horng and J.T. Li. A dynamic programming approach for fitting digital planar curves with line segments and circular arcs. *Pattern Recognition Letters (PRL)*, 22(2) :183 – 197, 2001.

[97] J-S. Huang and H-C. Liu. Object recognition using genetic algorithms with a Hopfield's neural model. *Expert Systems with Applications (ESA)*, 13(3) :191 – 199, 1997.

[98] S-C. Huang and C-F. Wang. Genetic algorithm for approximation of digital curves with line segments and circular arcs. *Journal of the chinese institute of Engineers*, 32(4) :437 – 444, 2008.

[99] F. K. Hwang, D. S. Richards, and P. Winter. *The Steiner Tree Problem*, volume 53 of *Annals of Discrete Mathematics*. North-Holland, Amsterdam, Netherlands, 1992.

[100] C. Ichoku, B. Deffontaines, and J. Chorowicz. Segmentation of digital plane curves : A dynamic focusing approach. *Pattern Recognition Letters (PRL)*, 17(7) :741 – 750, 1996.

[101] A. Inokuchi, T. Washio, and H. Motoda. Complete mining of frequent patterns from graphs : Mining graph data. *Machine Learning Journal (MLJ)*, 50(3) :321–354, 2003.

[102] L. Jacob, F. Bach, and J.-P. Vert. Clustered multi-task learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[103] J. Jia and K. Abe. Automatic generation of prototypes in 3D structural object recognition. In *Proceedings of the International Conference on Pattern Recognition (ICPR'98)*, pages 697–700, 1998.

[104] X. Jiang, A. Münger, and H. Bunke. On median graphs : Properties, algorithms, and applications. *IEEE Transaction on Pattern Analysis and Machine Intelligence (IEEE PAMI)*, 23(10) :1144–1151, 2001.

[105] Y. Jin, editor. *Multi-Objective Machine Learning*, volume 16 of *Studies in Computational Intelligence*. Springer, 2006.

[106] T. Joachims. Making large-scale support vector machine learning practical. In A. Smola B. Scholkopf, C. Burges, editor, *Advances in Kernel Methods : Support Vector Machines*, pages 169–184. MIT Press, Cambridge, MA, 1998.

[107] J.M. Jolion. Graph matching : what are we really talking about ? In *Proceedings of the workshop on Graph-based Representations in Pattern Recognition (GbRPR'01)*, pages 170–175, 2001.

[108] H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. In *Proceedings of the International Conference on Machine Learning (ICML'03)*, pages 321–328, 2003.

[109] H. Kashima, K. Tsuda, and A. Inokuchi. *Kernels for graphs*, pages 155–170. MIT Press, 2004.

[110] S. Keerthi, V. Sindhwani, and O. Chapelle. An efficient method for gradient-based adaptation of hyperparameters in SVM models. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 673–680. MIT Press, Cambridge, MA, 2007.

[111] H. Kellerer, U. Pferschy, and D. Pisinger. *Knapsack Problems*. Springer, Berlin, Germany, 2004.

[112] J. Kennedy and R. Eberhart. Particle swarm optimization. *Proceedings of the IEEE International Conference on Neural Networks (ICNN'95)*, 4 :1942–1948, 1995.

[113] V. Khare, X. Yao, and K. Deb. Performance scaling of multiobjective evolutionary algorithm. In *Technical report - SCS, University of Birmingham*, pages 1–70, 2002.

[114] G.H. Kim, V. Govindaraju, and S.N. Srihari. An architecture for handwritten text recognition systems. *International Journal on Document Analysis and Recognition (IJDAR)*, 2(1) :37–44, 1999.

[115] A. Kolesnikov and P. Fränti. Polygonal approximation of closed discrete curves. *Pattern Recognition (PR)*, 40(4) :1282–1293, 2007.

[116] T. Kudo, E. Maeda, and Y. Matsumoto. An application of boosting to graph classification. In *NIPS*, 2004.

[117] M. Kuramochi and G. Karypis. Frequent subgraph discovery. In *Proceedings of the International Conference on Data Mining (ICDM'01)*, pages 313–320, 2001.

[118] M. Kuramochi and G. Karypis. Finding frequent patterns in a large sparse graph. *Data Mining and Knowledge Discovery (DMKD)*, 11(3) :243–271, 2005.

[119] N.M. Kwok, D.K. Liu, and G. Dissanayake. Evolutionary computing based mobile robot localization. *Engineering Applications of Artificial Intelligence (EAAI)*, 19(8) :857–868, 2006.

[120] l. Zadeh. Optimality and non-scalar-valued performance criteria. *IEEE Transactions on Automatic Control (IEEE TAC)*, 8(1) :59 – 60, 1963.

[121] B. Lamiroy and D. Lopresti. An open architecture for end-to-end document analysis benchmarking. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'11)*, pages 42–47, 2011.

[122] M. Laumanns, L. Thiele, K. Deb, and E. Zitzler. Combining convergence and diversity in evolutionary multiobjective optimization. *MIT Press in Evolutionary Computation*, 10(3) :263–282, 2002.

[123] M. Laumanns, L. Thiele, k. Deb, and e. Zitzler. Combining convergence and diversity in evolutionary multiobjective optimization. *Evolutionary Computation (EC)*, 10(3) :263–282, 2002.

[124] B. Lazzerini, F. Marcelloni, and M. Vecchio. A multi-objective evolutionary approach to image quality/compression trade-off in JPEG baseline algorithm. *Applied Soft Computing (ASC)*, 10(2) :548–561, 2010.

[125] W. Lee, L. K. Burak Kara, and T.F. Stahovich. An efficient graph-based recognizer for hand-drawn symbols. *Computers and Graphics (CG)*, 31(4) :554–567, 2007.

[126] J. Liang and D. S. Doermann. Logical labeling of document images using layout graph matching with adaptive learning. In *Proceedings of the International Workshop on Document Analysis Systems (DAS'02)*, pages 224–235, 2002.

[127] J. Lladoós, E. Martí, and J.J. Villanueva. Symbol recognition by error-tolerant subgraph matching between region adjacency graphs. *IEEE Transaction on Pattern Analysis and Machine Intelligence (IEEE PAMI)*, 23(10) :1137–1143, 2001.

[128] J. Lladós and G. Sánchez. Graph matching versus graph parsing in graphics recognition - a combined approach. *International Journal on Pattern Recognition and Articicial Intelligence (IJPRAI)*, 18(3) :455–473, 2004.

[129] H. Locteau. *Contributions à la localisation de symboles dans les documents graphiques*. PhD thesis, Université de Rouen, 2008.

[130] H. Locteau, R. Raveaux, S. Adam, Y. Lecourtier, P. Héroux, and É. Trupin. Approximation of digital curves using a multi-objective genetic algorithm. In *Proceedings of the International Conference on Pattern Recognition (ICPR'06)*, pages 716–719, 2006.

[131] A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera. Applying multi-objective evolutionary algorithms to the automatic learning of extended boolean queries in fuzzy ordinal linguistic information retrieval systems. *Fuzzy Sets and Systems (FSS)*, 160(15) :2192 – 2205, 2009.

[132] D. Lopresti and G. Wilfong. A fast technique for comparing graph representations with applications to performance evaluation. *International Journal of Document Analysis and Recognition (IJDAR)*, 6(4) :219–229, 2003.

[133] M.A. Lozano and F. Escolano. Protein classification by matching and clustering surface graphs. *Pattern Recognition (PR)*, 39(4) :539–551, 2006.

[134] S.W. Lu, Y. Ren, and C.Y. Suen. Hierarchical attributed graph representation and recognition of handwritten chinese characters. *Pattern Recognition (PR)*, 24(7) :617–632, 1991.

[135] B. Luo, R.C. Wilson, and E.R. Hancock. Spectral embedding of graphs. *Pattern Recognition*, pages 2213–2230, 2003.

[136] S. Mabu, K. Hirasawa, and J. Hu. A graph-based evolutionary algorithm : Genetic network programming (gnp) and its extension using reinforcement learning. *Evolutionary Computation (EC)*, 15(3) :369–398, 2007.

[137] P. Mahé, N. Ueda, T. Akutsu, J.-L. Perret, and J.-P. Vert. Extensions of marginalized graph kernels. In *Proceedings of the International Conference on Machine Learning (ICML'04)*, pages 552–559, 2004.

[138] P. Mahé, N. Ueda, T. Akutsu, J.-L. Perret, and J.-P. Vert. Graph kernels for molecular structure-activity relationship analysis with support vector machines. *Journal of Chemical Information and Modeling (JCIM)*, 45(4) :939–951, 2005.

[139] S. Marini, M. Spagnuolo, and B. Falcidieno. Structural shape prototypes for the automatic classification of 3d objects. *IEEE Computer Graphics and Applications (IEEE CGA)*, 27(4) :28–37, 2007.

[140] M. Marji and P. Siy. A new algorithm for dominant points detection and polygonization of digital curves. *Pattern Recognition (PR)*, 36(10) :2239 – 2251, 2003.

[141] M. Marji and P. Siy. Polygonal representation of digital planar curves through dominant point detection – a nonparametric algorithm. *Pattern Recognition (PR)*, 37(11) :2113 – 2130, 2004.

[142] R.T. Marler and J.S. Arora. Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization (SMO)*, 26(6) :369–395, 2004.

[143] A. Masood. Optimized polygonal approximation by dominant point deletion. *Pattern Recognition (PR)*, 41(1) :227–239, 2008.

[144] B. McKay. Practical graph isomorphism. In *Numerical mathematics and computing*, pages 45–87, 1981.

[145] A. Messac and P.D. Hattis. Physical programming design optimization for high speed civil transport. *Journal of aircraft*, 33(2) :446–449, 1966.

[146] B. T. Messmer and H. Bunke. A new algorithm for error-tolerant subgraph isomorphism detection. *IEEE Transactions on Pattern Analysis and Machine Inteligence (IEEE PAMI)*, 20(5) :493–504, 1998.

[147] D. A. Mitzias and B. G. Mertzios. Shape recognition with a neural classifier based on a fast polygon approximation technique. *Pattern Recognition (PR)*, 27(5) :627 – 636, 1994.

[148] S. Mostaghim and J. Teich. The role of ε-dominance in multi-objective particle swarm optimization. In *Proceedings of the Congress on Evolutionary Computation (CEC'03)*, volume 3, pages 1764–1771, 2003.

[149] S. Mostaghim and J. Teich. Strategies for finding good local guides in multi-objective particle swarm optimization. In *Swarm Intelligence Symposium*, 2003.

[150] S. Mostaghim and J. Teich. Covering pareto-optimal fronts by subswarms in multi-objective particle swarm optimization. In *IEEE Proceedings, World Congress on Computational Intelligence (CEC'04)*, volume 2, pages 1404–1411, 2004.

[151] C. R. Mouser and S. A. Dunn. Comparing genetic algorithms and particle swarm optimisation for an inverse problem exercise. In Rob May and A. J. Roberts, editors, *Proc. of the Computational Techniques and Applications Conference (CTAC'04)*, volume 46, pages 89–101, 2005.

[152] G. L. Nemhauser and L. A. Wolsey. *Integer and combinatorial optimization.* Wiley-Interscience, New York, NY, USA, 1988.

[153] M. Neuhaus and H. Bunke. Edit distance-based kernel functions for structural pattern classification. *Pattern Recognition (PR)*, 39(10) :1852–1863, 2006.

[154] R. Neumann and G. Teisseron. Extraction of dominant points by estimation of the contour fluctuations. *Pattern Recognition (PR)*, 35(7) :1447 – 1462, 2002.

[155] T.P. Nguyen and I. Debled Rennesson. Decomposition of a curve into arcs and line segments based on dominant point detection. In *Proceedings of the Scandinavian Conference on Image Analysis - (SCIA'11)*, pages 794–805, 2011.

[156] L. S. Oliveira, M. Morita, and R. Sabourin. Feature selection for ensembles applied to handwriting recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, 8(4) :262–279, 2006.

[157] L. S. Oliveira, R. Sabourin, F. Bortolozzi, and C. Y. Suen. Feature selection using multi-objective genetic algorithms for handwritten digit recognition. *Proceedings of the International Conference on Pattern Recognition (ICPR'02)*, 1 :10568–10571, 2002.

[158] E. Osuna, R. Freund, and F. Girosi. Support vector machines : Training and applications. Technical report, AI Memo 1602, Massachusetts Institute of Technology, 1997.

[159] E. Papageorgiou, K. Parsopoulos, C. Stylios, P. Groumpos, and M. Vrahatis. Fuzzy cognitive maps learning using particle swarm optimization. *Journal of Intelligent Information Systems (JIIS)*, 25(1) :95–121, 2005.

[160] M.T. Parvez and S.A. Mahmoud. Polygonal approximation of digital planar curves through adaptive optimizations. *Pattern Recognition Letters (PRL)*, 31(13) :1997–2005, 2010.

[161] S-C. Pei and J-H. Horng. Optimum approximation of digital planar curves using circular arcs. *Pattern Recognition (PR)*, 29(3) :383 – 388, 1996.

[162] J-C. Perez and E. Vidal. Optimum polygonal approximation of digitized curves. *Pattern Recognition Letters (PRL)*, 15(8) :743 – 750, 1994.

[163] A. M. G. Pinheiro and M. Ghanbari. Piecewise approximation of contours through scale-space selection of dominant points. *IEEE Transaction Image Processing (IEEE TIP)*, 19(6) :1442–1450, 2010.

[164] H. Qiu and E. R. Hancock. Graph matching and clustering using spectral partitions. *Pattern Recognition (PR)*, 39(1) :22–34, 2006.

[165] R. J. Queshri, J.-Y. Ramel, and H. Cardot. De l'appariement de graphes symboliques à l'appariements de graphes numériques : Application à la reconnaissance de symboles. In *Actes de la Conférence Internationale Francophone sur l'Écrit et le Document (CIFED)*, pages 31–36, 2006.

[166] P. V. Radtke, R. Sabourin, and T. Wong. Classification system optimization with multi-objective genetic algorithms. *Proceedings of the International Workshop on Frontiers in Handwriting Recognition (IWFHR'06)*, 2006.

[167] M. A. Rahgozar. Document table recognition by graph rewriting. In *Proceedings of the International Workshop on Applications of Graph Transformations with Industrial Relevance (AGTIVE '99)*, pages 279–295, 2000.

[168] A. Rakotomamonjy. Optimizing AUC with support vector machine. *Proceedings of ECAI Workshop on ROC Curve and AI (ROCAI'04)*, pages 469–478, 2004.

[169] T. K. Ralphs and M. Gzelsüoy. *The Next Wave in Computing, Optimization, and Decision Technologies*, volume 29 of *Operations Research/Computer Science Interfaces Series*, chapter The Symphony Callable Library for Mixed Integer Programming, pages 61–76. Springer US, 2005.

[170] R. Raveaux, S. Adam, P. Héroux, and E. Trupin. Learning graph prototypes for shape recognition. *Computer Vision and Image Understanding (CVIU)*, 115(7) :905 – 918, 2011.

[171] R. Raveaux, E. Barbu, H. Locteau, S. Adam, P. Héroux, and É. Trupin. A graph classification approach using a multi-objective genetic algorithm application to symbol recognition. In Francisco Escolano and Mario Vento, editors, *Proceedings of the IAPR International Workshop on Graph Based Representations for Pattern Recognition (GbR-PR'07)*, volume 4538 of *Lecture Notes in Computer Science*, pages 361–370. Springer, 2007.

[172] R. Raveaux, J.C. Burie, and J.M. Ogier. A graph matching method and a graph matching distance based on subgraph assignments. *Pattern Recognition Letters (PRL)*, 31(5) :394–406, 2010.

[173] B.K. Ray and K.S. Ray. An algorithm for detection of dominant points and polygonal approximation of digitized curves. *Pattern Recognition Letters (PRL)*, 13(12) :849 – 856, 1992.

[174] P. Ren, R. C. Wilson, and E. R. Hancock. Graph characterization via Ihara coefficients. *IEEE Transactions on Neural Networks (IEEE TNN)*, 22(2) :233–245, 2011.

[175] M. Reyes-sierra and C.A.C. Coello. Multi-objective particle swarm optimizers : A survey of the state-of-the-art. *International journal of computational intelligence research (IJCIR)*, 2(3) :287–308, 2006.

[176] K. Riesen and H. Bunke. Approximate graph edit distance computation by means of bipartite graph matching. *Image Vision Computing (IVC)*, 27(7) :950–959, 2009.

[177] K. Riesen and H. Bunke. Graph classification based on vector space embedding. *International Journal on Pattern Recognition and Articicial Intelligence (IJPRAI)*, 23(6) :1053–1081, 2009.

[178] J. Ros, C. Laurent, and J-M. Jolion. A Bag of Strings representation for Image Categorization. *International Journal of Mathematical Imaging and Vision (JMIV)*, 35(1) :51–67, 2009.

[179] A. Rosenfeld and J.S. Weszka. An improved method of angle detection on digital curves. *IEEE Transaction on Computers (IEEE TC)*, 24(9) :940–941, 1975.

[180] P.L. Rosin and G.A.W. West. Segmentation of edges into lines and arcs. *Image and Vision Computing (IVC)*, 7(2) :109 – 114, 1989.

[181] M. Rusiñol, J. Lladós, and G. Sánchez. Symbol spotting in vectorized technical drawings throug a lookup table of region strings. *Pattern Analysis and Applications (PAA)*, 33(3) :321–331, 2009.

[182] M. Salotti. An efficient algorithm for the optimal polygonal approximation of digitized curves. *Pattern Recognition Letters (PRL)*, 22(2) :215 – 221, 2001.

[183] G. Salton. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall Inc., Englewood Cliffs, 1971.

[184] B. Sarkar, L.K. Singh, and D. Sarkar. Approximation of digital curves with line segments and circular arcs using genetic algorithms. *Pattern Recognition Letters (PRL)*, 24(15) :2585–2595, 2003.

[185] D. Sarkar. A simple algorithm for detection of significant vertices for polygonal approximation of chain-coded curves. *Pattern Recognition Letters (PRL)*, 14(12) :959–964, 1993.

[186] J.D. Schaffer and J.J. Grefenstette. Multiobjective learning via genetic algorithms. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI'85)*, pages 593–595, 1985.

[187] A. Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, New York, NY, USA, 1998.

[188] M. Sebag, J.Azé, and N. Lucas. Roc-based evolutionary learning : Application to medical data mining. *Proceedings of the International Conference on Artificial Evolution (ICAI'03)*, pages 384–396, 2003.

[189] T.C. Service. A no free lunch theorem for multi-objective optimization. *Information Processing Letters (IPL)*, 110(21) :917–923, 2010.

[190] M. Settles, B. Rodebaugh, and T. Soule. Comparison of genetic algorithm and particle swarm optimizer when evolving a recurrent neural network. In Springer Berlin / Heidelberg, editor, *Genetic and Evolutionary Computation - GECCO 2003*, volume 2723/2003 of *Lecture Notes in Computer Science*, pages 148–149, 2003.

[191] N. Sidère, P. Héroux, and J-Y. Ramel. Vector representation of graphs : Application to the classification of symbols and letters. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'09)*, pages 681–685, 2009.

[192] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2005.

[193] C. Solnon. Alldifferent-based filtering for subgraph isomorphism. *Artificial Intelligence (AI)*, 174(12-13) :850 – 864, 2010.

[194] N. Srinivas and K. Deb. Multiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary Computation*, 2(3) :221–248, 1994.

[195] W. Stadler. Fundamentals of Multicriteria Optimization. pages 1–25. Plenum Press, 1988.

[196] R. Steuer and E-U. Choo. An interactive weighted tchebycheff procedure for multiple objective programming. *Mathematical Programming*, 26(3) :326–344, 1983.

[197] F. Suard, V. Guigue, A. Rakotomamonjy, and A. Bensrhair. Pedestrian detection using stereovision and graph kernels. In *Proceedings of the IEEE Intelligent Vehicle Sysposium (IVS'05)*, pages 267–272, 2005.

[198] K.C. Tan, T.H. Lee, and E.F. Evolutionary algorithms for multi-objective optimization : Performance assessments and comparisons. *Artificial Intelligence Review*, 17(4) :251–290, 2002.

[199] M. Tanaka, H. Watanabe, Y. Furukawa, and T. Tanino. GA-based decision support system for multicriteria optimization. In *Proceedings of the International Conference on Systems, Man and Cybernetics (ICSMC'95)*, volume 2, pages 1556–61, 1995.

[200] M. Teague. Image analysis via the general theory of moments. *Journal of the Optical Society of America (JOSA)*, 70(8) :920–930, 1980.

[201] C.H. Teh and R.T. Chin. On the detection of dominant points on digital curves. *IEEE Transaction on Pattern Analysis and Machine Intelligence (IEEE PAMI)*, 11(8) :859–872, 1989.

[202] O. R. Terrades, S. Tabbone, and E. Valveny. A review of shape descriptors for document analysis. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 227–231, 2007.

[203] S. M. Thomas and Y. T. Chan. A simple approach for the estimation of circular arc center and its radius. *Computer Vision, Graphics, and Image Processing (CVGIP)*, 45(3) :362 – 370, 1989.

[204] K. Tombre. Is graphics recognition an unidentified scientific object ? In Wenyin Liu, Josep Lladós, and Jean-Marc Ogier, editors, *Graphics Recognition. Recent Advances and New Opportunities*, pages 329–334. Springer-Verlag, Berlin, Heidelberg, 2008.

[205] K. Tombre. Graphics Recognition – What Else ? In Jean-Marc Ogier, Wenyin Liu, and Josep Llados, editors, *Graphics Recognition - Achievements, Challenges and Evolution. Selected Paper from 8th International Workshop GREC 2009, La Rochelle, July 2009*, volume 6020 of *Lecture Notes in Computer Science*, pages 272–277. Springer Verlag, 2010.

[206] K. Tombre, S. Tabbone, and Ph. Dosch. Musings on Symbol Recognition. In Wenyin Liu and Josep Lladós, editors, *Graphics Recognition—Ten Years Review and Future Pespectives*, volume 3926 of *Lecture Notes in Computer Science*, pages 23–34. Springer Verlag, 2006.

[207] F. Tortorella, R. Patraccone, and M. Molinara. A dynamic programming approach for segmenting digital planar curves into line segments and circular arcs. In *Proceedings of the International Conference on Pattern Recognition (ICPR'08)*, pages 1–4, 2008.

[208] J. R. Ullmann. An algorithm for subgraph isomorphism. *Journal of the ACM (JACM)*, 23(1) :31–42, 1976.

[209] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. Borgwardt. Graph kernels. *Journal of Machine Learning Research (JMLR)*, 11 :1201–1242, 2010.

[210] G. Wahba, X. Lin, F. Gao, D. Xiang, R. Klein, and B. Klein. The bias-variance tradeoff and the randomized gacv. In *Proceedings of NIPS*, pages 620–626, 1999.

[211] W. D. Wallis, P. Shoubridge, M. Kraetz, and D. Ray. Graph distances using graph union. *Pattern Recognition Letters*, 22(6-7) :701–704, 2001.

[212] R. C. Wilson, E. R. Hancock, and B. Luo. Pattern vectors from algebraic graph theory. *IEEE Transaction on Pattern Analysis and Machine Intelligence (IEEE PAMI)*, 27(7) :1112–1124, 2005.

[213] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transaction on evolutionary computation (IEEE TEC)*, 1(1) :67–82, 1997.

[214] S. Wu and P. Flach. A scored AUC metric for classifier evaluation and selection. In *Proceedings of the workshop on ROC analysis in Machine Learning at ICML (ROCML'05)*, 2005.

[215] S. Yu and F. K. Soong. A symbol graph based handwritten math expression recognition. In *Proceedings of the International Conference on Pattern Recognition (ICPR'08)*, pages 1–4, 2008.

[216] H. Zanghi, C. Ambroise, and V. Miele. Fast online graph clustering via Erdös-Rényi mixture. *Pattern Recognition (PR)*, 41(12) :3592–3599, 2008.

[217] H. Zhang, C.M. Tam, and H. Li. Multimode project scheduling based on particle swarm optimization. *Computer Aided Civil and Infrastructure Engineering (CACIE)*, 21(2) :93–103, 2006.

[218] E. Zitzler, M. Laumanns, and L. Thiele. SPEA2 : Improving the strength Pareto evolutionary algorithm. Technical report, Computer Engineering and Networks Laboratory (TIK), ETH Zurich, 2001.

[219] E. Zitzler and L. Thiele. Multiobjective evolutionary algorithms : A comparison case study and the strength Pareto approach. *IEEE Transactions on Evolutionary Computation (IEEE TEC)*, 3(4) :257–271, 1999.

Troisième partie

Recueil de publications

Annexe A

Réference CV : 6

E. Barbu, P. Héroux, S. Adam, and E. Trupin. Frequent graph discovery : Application to line drawing document images. Electronic Letters on Computer Vision and Image Analysis (ELCVIA), 5(2) :47-57, 2005.

Electronic Letters on Computer Vision and Image Analysis 5(2):47-57, 2005

Frequent Graph Discovery: Application to Line Drawing Document Images

Eugen Barbu, Pierre Héroux, Sébastien Adam, and Éric Trupin

Laboratoire PSI
CNRS FRE 2645 - Université de Rouen
76 821 Mont-Saint-Aignan cedex - France

Received 16 July 2004; accepted 16 November 2004

Abstract

In this paper a sequence of steps is applied to a graph representation of line drawings using concepts from data mining. This process finds frequent subgraphs and then association rules between these subgraphs.

The distant aim is the automatic discovery of symbols and their relations, which are parts of the document model. The main outcome of our work is firstly an algorithm that finds frequent subgraphs in a single graph setting and secondly a modality to find rules and meta-rules between the discovered subgraphs. The searched structures are closed [1] and disjunct subgraphs. One aim of this study is to use the discovered symbols for classification and indexation of document images when a supervised approach is not at hand. The relations found between symbols can be used in segmentation of noisy and occluded document images. The results show that this approach is suitable for patterns, symbols or relation discovery.

Key Words: Computer Vision, Image Analysis, Pattern Recognition, Graph Mining, Line Drawings, Association Rules.

1 Introduction

A symbol encodes a message into the form of an arbitrary sign. This sign has acquired a conventional significance. According to the document model, the symbol conveys graphical and semantic information. In this paper we try to discover both the representation as a written sign, and the relations (rules) that a symbol respects. The graphical representation and the rules found can be considered as an approximation of the message carried by the symbol. Automatic symbol extraction on document images without any prior domain knowledge is an appealing task. This approach has been pursued by Altamura [2] and Messmer [3]. In the context of line drawings document, one way to detect symbols is to consider the frequent occurrences of included entities. The entities can be graphs, geometric shapes or image parts depending at which processing level (segmentation) we apply this method [4], [5], [6]. A possible extension of this approach is to find relations between symbols. Such a relation can be viewed as a new entity that can be frequent and participates on its own right in other more complex relations. The standard for mining frequent item sets is the A priori algorithm [7]. However if the objects are graphs, some modifications to the basic algorithm

Correspondence to: eugen.barbu@univ-rouen.fr

Recommended for acceptance by J.M. Ogier, T. Paquet, G. Sanchez
ELCVIA ISSN: 1577-5097
Published by Computer Vision Center / Universitat Autonoma de Barcelona, Barcelona, Spain

should be made. Several papers describe A priori-like algorithms for mining frequent graph substructures [8], [9], [10].

This paper presents an algorithm that finds frequent subgraphs in a graph, a modality of creating rules and meta-rules between the discovered symbols and some possible utilization for the detected rules.

The principle of our approach is described on Fig.1.

A document image is characterised in a certain extent by the set of symbols that are frequent. Using this incomplete description of a document, generated in an unsupervised manner, we can use techniques from Information Retrieval in order to index [11] and classify [12] document images.

A good example for using the rules between objects can be to cluster a set of document images. If the symbols are described in the common graph language, the rules can also be shared. Two documents are from the same class if they respect the same rules. The distance between two documents can be evaluated using the extent to which one document conforms to the rules of the other.

Another application of the rules between symbols is to apply these rules in the segmentation process when noise or occluded symbols are present.

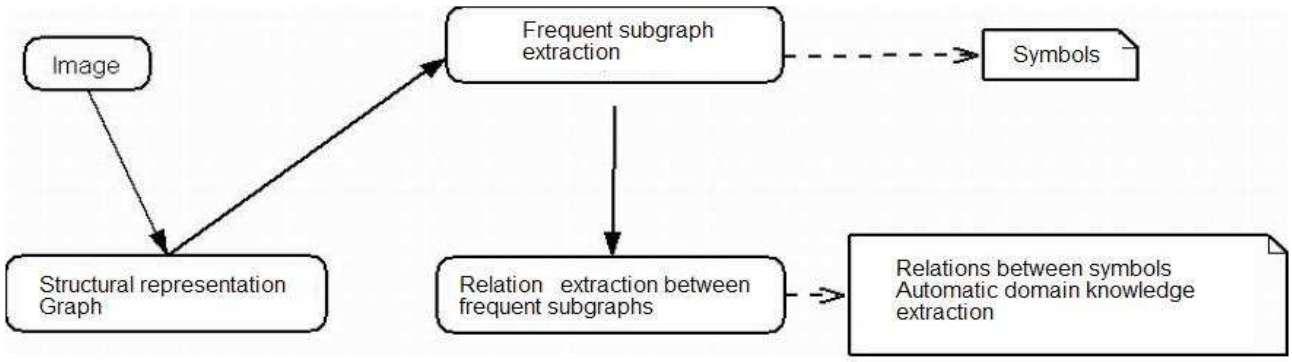


Fig. 1. Approach principle

This paper is organized as follows. Section 2 addresses the algorithm for finding frequent subgraphs. Section 3 emphasizes the ways we can find association rules between symbols. Section 4 presents an example of the proposed method. Section 5 elaborates several conclusions.

2 An algorithm that finds frequent subgraphs

The proposed approach is based on the fact that symbols on technical drawings graphically encode message elements according to a certain convention. So, in several document images sharing the same document model, a pattern always describes the same entity. The symbols of a document class appear with a certain frequency.

The purpose of this algorithm is to find the frequent subgraphs from a graph that describes the neighbourhood relations between shapes in a line drawing document. The subgraphs which represent symbols are closed graphs (a graph is closed if it does not have a super-graph with the same number of apparitions in the dataset) [1].

In the process of document image analysis, different graph based representations can be used. These representations can be constructed depending on the understanding level of the document when the graph is generated or according to the type of document that one tries to model (mostly textual, mostly graphical, mixed)

In this paper we extract a graph from the document image at a low level of document understanding. We only use connected components and their neighbouring relations to construct the graph. The documents analysed are mostly graphical documents called line drawings. From a semantic point of view, a line drawing document is a document that does not lose information when the morphological operation of skeletonisation is applied on it.

The document graph is obtained from a line drawing considering:

- the regions (closed loops, two-dimensional shapes) or one-dimensional shapes as nodes.
- the neighbouring relations between these shapes as edges.

Two shapes are neighbours if they share a common frontier (see Fig. 2). This relation of neighbourhood can also be computed using a distance between node regions. One example can be: two occlusions are neighbours if the distance between their centers is less than a fixed or relative threshold. This representation is more robust than the binary relation of neighbourhood computed using the existence or not of a common frontier but has the disadvantage of using a more or less arbitrary threshold.

In order to label each node we extract a vector of features called Zernike moments for every part of the image that represents a node of the representation graph. These features are rotation invariant. More properties on these features can be found in [13].

We apply an unsupervised clustering algorithm on the nodes of the representation and each node has the class it belongs to as label. The clustering algorithm used is hierarchical ascendant, clustering using the Euclidean distance as dissimilarity, complete-linkage distance between clusters, and the Calinsky-Harabasz index to obtain the number of clusters. This algorithm has been chosen after a comparison with a hierarchical descendant clustering using the Duda-Hart index as stopping criterion and based on the conclusions from [14].

Two graphs represent the same symbol if they are isomorphic and if each pair of nodes (associated by the isomorphism function) has the same label.

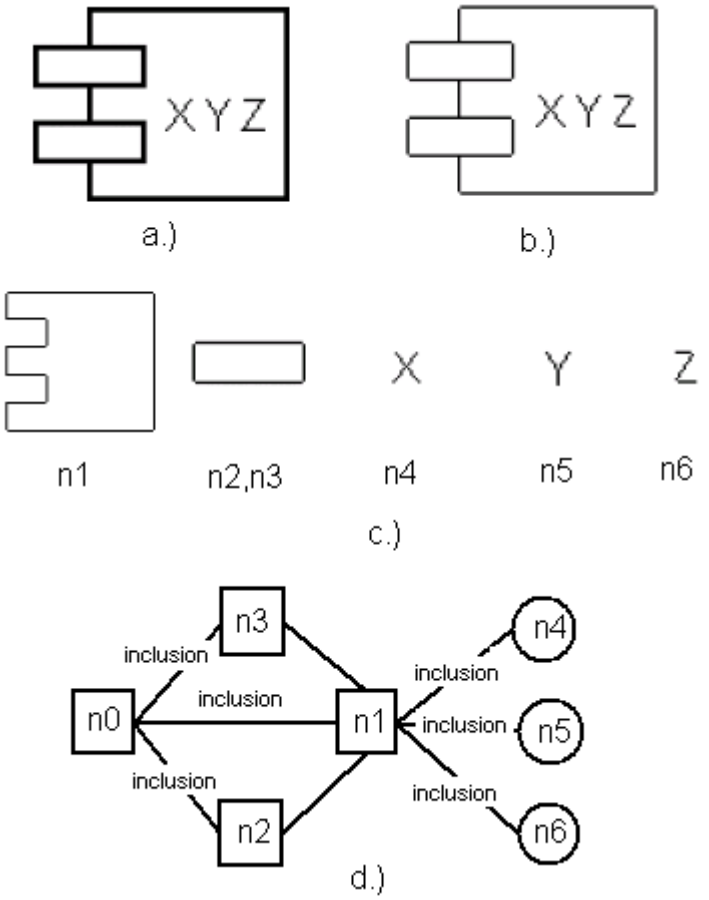


Fig. 2. A drawing a.) and its associated graph d.), considering the background region *n0*. The *1-dimensional* shapes are represented by circles. The *2-dimensional* shapes are represented by rectangles.

In this context a subgraph is considered frequent if its number of apparitions as non-included in other subgraphs is greater than a certain threshold *s*.

The way the threshold is defined can be linked to two possible settings: single or multiple graphs. In multiple graphs setting, i.e. we have a set of graphs and each graph is called a "transaction", we can say a subgraph is frequent if it appears in more than *s%* transactions. In our case we are interested in the frequent occurrences of a subgraph in the same graph, so we are in a single graph setting.

Because the number of subgraphs of the same class (any two subgraphs from the same class are isomorphic) is considered for a single graph, the threshold cannot be defined in relation with the number of transactions as it is done in other similar algorithms ([9], [10]). Considering a single transaction, we are interested in symbol occurrences included in that transaction. Here the threshold s is computed considering an approximation of the maximum possible number of subgraphs, with disjoint node sets and fixed number of edges and nodes, contained in the document graph.

The proposed algorithm uses the principle behind "A priori"-like algorithms combined with two simplifying hypotheses:

- the symbols are rarely expressed by graphs with a large number of nodes (10)
- occurrences for the same symbol are subgraphs with disjoint node sets

The idea behind all A priori-like algorithms is that we can construct the frequent sets of objects by adding objects to a set that is frequent until it is not frequent anymore. When objects are graphs, a graph is frequent if all its subgraphs are also frequent. In the general case this last proposition is not true but if we are in the context of disjoint node sets for subgraphs, this proposition is true. On Fig. 3, the graph c) has only one occurrence in the graph a). If we consider that subgraphs can have common nodes, three occurrences of graph b) can be found in graph a). In our case, nodes only participate in the representation of a single symbol. Hence, subgraphs must have distinct nodes. Then, only one occurrence of graph b) can be found graph a).

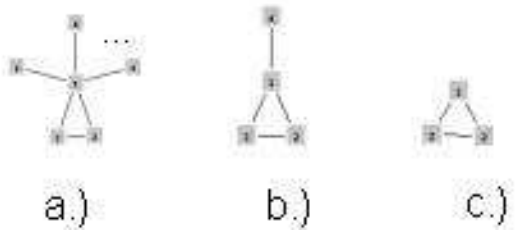


Fig. 3. Illustration for frequent subgraph search

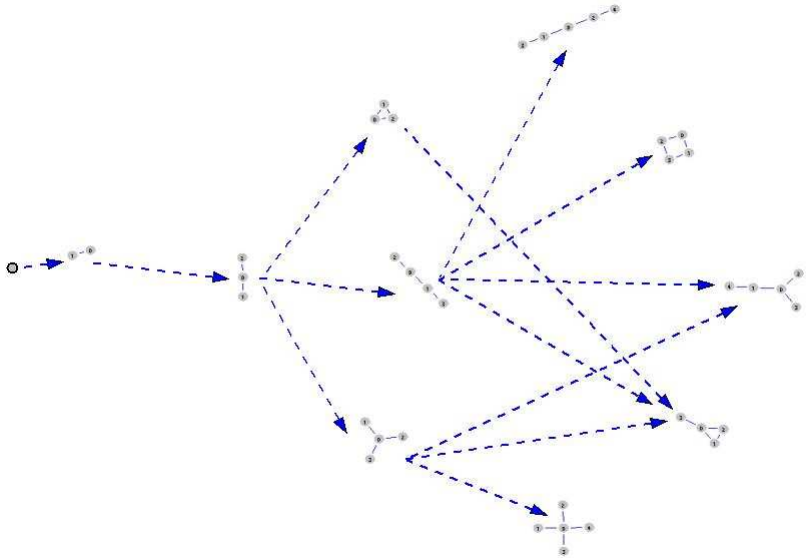


Fig. 4. Non-isomorphic graph network

In the algorithm used here, in order to reduce time complexity, we compute a network of non-isomorphic graphs off-line.

The network is used to guide the search for frequent subgraphs and to avoid isomorphism related computations (exponential in time) during this procedure. The network contains all graphs that have less than MAX edges. The graphs and their relations of inclusion are generated using the method presented in [15]. This method generates all non isomorphic subgraphs of a particular size. The complexity of this method is exponential.

Based on the relation of inclusion between these graphs the network is an acyclic oriented graph, whose nodes are all non-isomorphic graphs with less than MAX edges, where MAX is an input parameter. Fig. 4 presents how a search for frequent subgraphs is done. If at a certain stage a graph is not frequent, all of its descendants, with more edges, cannot be frequent. This network was computed with *MAX=9* in our application. Two reasons sustain this choice: the size of the network increases more than exponentially with the number of graph edges and the symbols are rarely expressed with graphs that have a bigger number of edges. The algorithm uses the information contained in the network of non-isomorphic graphs (the inclusion relations and automorphisms for each graph) to efficiently search for frequent subgraphs. Based on the non-isomorphic graph network, the search for frequent subgraphs is done in polynomial time.

2.1 Algorithm

Network initialisation till level *MAX*

begin
Input An undirected labelled graph
Output A list of frequent subgraphs and for each one the apparition list
k:=1
while *k<=MAX*
for all graphs that can be frequent
let *G* be the current graph
using the apparition lists of his predecessors the apparition list of G is computed
if the apparition list contains more entries than a *threshold*
then graph *G* is considered frequent
if *G* is frequent

then update the list of predecessor setting the (inclusion in a frequent graph) flag on *true*
else update the successors of G setting the flag, for the possibility to be frequent, on *false*
for all frequent graphs from level *k-1*
update the list of apparitions taking into account the inclusion in other frequent graphs
update accordingly the frequent flag
k:=k+1
end while
end.

The threshold is computed using the following formula:

$$threshold = p * \min(\frac{e}{e'}, \frac{n}{n'}) \quad (1)$$

This formula represents an approximation of the maximum number of subgraphs that can be found in a graph. We consider that a subgraph is frequent if the number of occurrences is bigger than p% out of the maximum (possible) total number of subgraphs having e edges and n nodes. This algorithm can be applied to a graph or a set of graphs associated to a document or a collection of documents.

3 Rules and meta-rules

After some symbols were found using the above algorithm, relations between those symbols can be considered. The search for association rules between symbols is made using the A priori algorithm [7]. In the subsequent paragraphs the setting of this algorithm is presented. If we consider a set of symbols all having a common property, for example being on the same level in the inclusion tree (this tree models the inclusions between shapes), we may say this set of symbols participates in a transaction. All transactions are considered when relations between symbols are computed. An example for a set of transactions that describes how the objects are related can be:

$$T_1(o_1, o_2, o_3); T_2(o_1, o_2); T_3(o_2, o_3); T_4(o_1, o_2, o_4)$$

From this set of transactions one can extract a rule as the following if the object *o1* participates in a transaction then the object *o2* will probably be there too.

The transactions can be defined using other criterions such as: a document represents a single transaction. The relations found have the meaning that if a set of symbols appears in a document then it is highly probable that the consequent set of symbols will appear as well.

In the single graph setting we can relate transactions to graph partitioning or subgraph clustering. However, in the present paper only transactions based on the inclusion relation are used.

Applying the A priori algorithm in this context (i.e. using the above described transactions) we find relations of the following type:

$$(o_{i1}, o_{i2}, \ldots, o_{in}) \Rightarrow (o_{j1}, o_{j2}, \ldots, o_{jm}) \quad (2)$$

Where

$$(o_{i1}, o_{i2}, \ldots, o_{in}) \cap (o_{j1}, o_{j2}, \ldots, o_{jm}) = \varnothing$$

If we consider a rule R obtained by the "A priori" algorithm, we can compute for each transaction whether R is confirmed or not. The confirmation is verified using the logical definition of the implication relation.

This computation has the following meaning: a rule is considered in its own right as a pattern and we consider that this particular rule appears in the transaction if it is confirmed in that transaction.

When in a given document we find a relation between some symbols then this fact implies the existence of a relation between some other symbols in the document.

Considering rules as patterns can be recursively applied in order to obtain meta-rules of type:

$$((o_{i1},...,ok_1) \Rightarrow (o_{i2},...,ok_2)) \Rightarrow ((o_{i3},...,ok_3) \Rightarrow (o_{i4},...,ok_4)) \quad (3)$$

or

$$(o_{i1},...,ok_1) \Rightarrow ((o_{i2},...,ok_2) \Rightarrow (o_{i3},...,ok_3))$$

or

$$((o_{i1},...,ok_1) \Rightarrow (o_{i2},...,ok_2)) \Rightarrow (o_{i3},...,ok_3)$$

The meta-rules found add knowledge to the associations and are not equivalent with simple rules. To support this assertion, we present an example where a meta-rule is not reducible to a simple rule (like Eq. 2.). The meta-rule $(o_1 \Rightarrow o_2) \Rightarrow (o_3 \Rightarrow o_4)$ is written in a disjunctive normal form as: $\bar{o}_1 o_2 + \bar{o}_3 + o_4$ but no simple rule such as $(o_1, o_2) \Rightarrow (o_3, o_4)$ or $o_1 \Rightarrow (o_2, o_3, o_4)$ written in a disjunctive normal form will contain a conjunction of a statement letter and a negation of other letter as it is the case for the meta-rule.

These types of meta-rules are more difficult to be expressed in informal language but are closer to the domain knowledge rules. One can describe a relation $R_1 \Rightarrow R_2$ between rules as follows: all transactions that contain a certain rule will probably contain the second rule as well.

4 Examples

4.1 Tutorial example

This section presents a didactic example of our approach applied on a synthetic document (Fig. 5.) containing architectural symbols. First, connected components, loops and neighbouring relations are extracted. After that, the neighbouring graph is built (Fig. 6(a)). Inclusion of shapes can be obtained from the graph [17]. Then, the corresponding inclusion tree is obtained (Fig. 6(b)). The threshold s is computed (s = 6) by applying equation (1) with p = 0.2. Then a subgraph is considered frequent if we can find 6 occurrences at least. The results of frequent subgraph search are shown on Fig. 7. In this search the inclusion relation is not considered as a neighbouring relation. Using the discovered symbols, transactions that contain these symbols can be obtained. Each transaction represents a leaf of the inclusion tree.

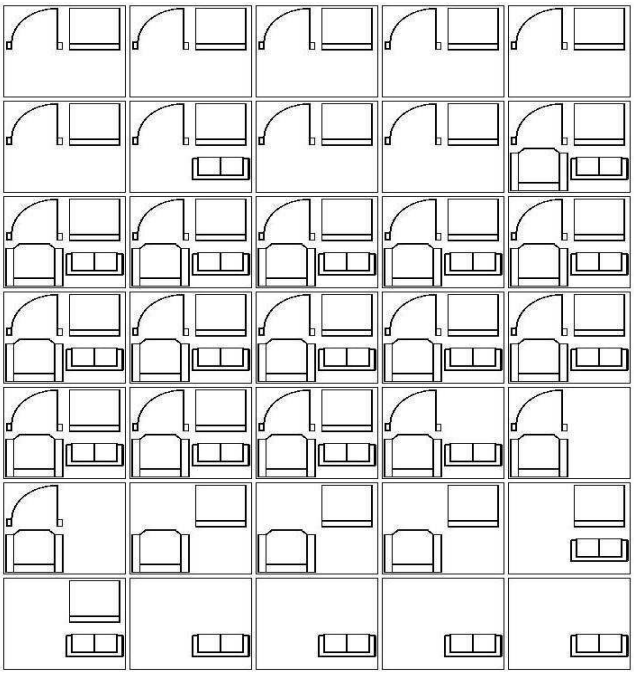


Fig. 5. A technical drawing

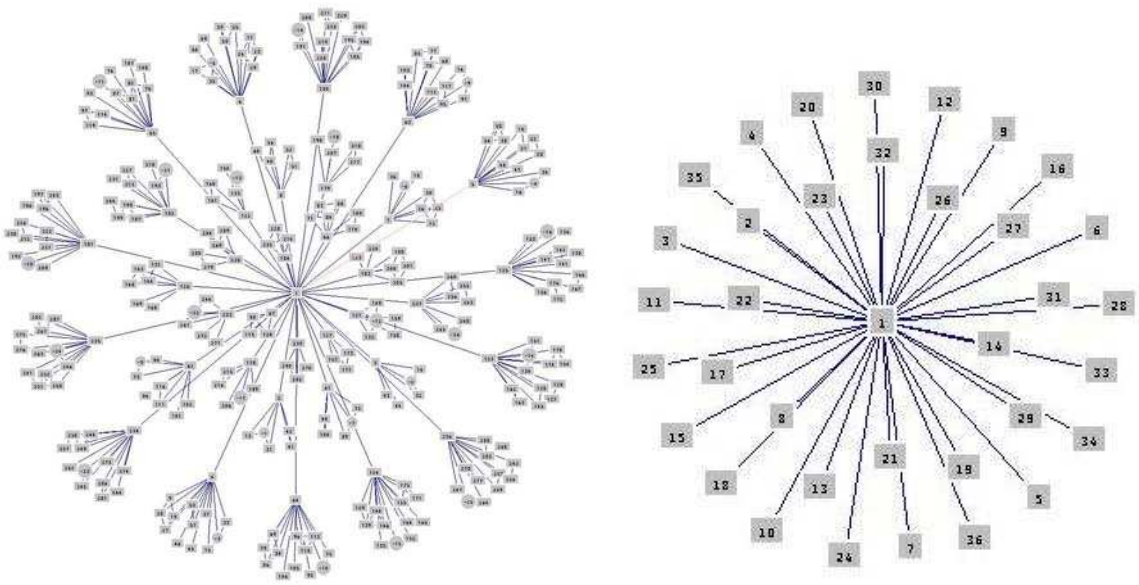


Fig. 6. Neighbourhood graph and inclusion tree

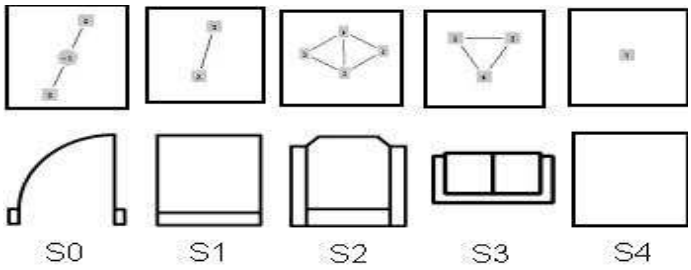


Fig. 7. Frequent subgraphs and corresponding symbols

The symbols are named S_0, S_1, S_2, and S_3. Considering the above assumptions the transactions are:

$T_1(S_0,S_1),T_2(S_0,S_1),T_3(S_0,S_1),T_4(S_0,S_1),T_5(S_0,S_1),T_6(S_0,S_1),$
$T_7(S_0,S_1,S_3),T_8(S_0,S_1),T_9(S_0,S_1),T_{10}(S_0,S_1,S_2,S_3),T_{11}(S_0,S_1,S_2,S_3),T_{12}(S_0,S_1,S_2,S_3),$
$T_{13}(S_0,S_1,S_2,S_3),T_{14}(S_0,S_1,S_2,S_3),T_{15}(S_0,S_1,S_2,S_3),T_{16}(S_0,S_1,S_2,S_3),T_{17}(S_0,S_1,S_2,S_3),$
$T_{18}(S_0,S_1,S_2,S_3),T_{19}(S_0,S_1,S_2,S_3),T_{20}(S_0,S_1,S_2,S_3),T_{21}(S_0,S_1,S_2,S_3),T_{22}(S_0,S_1,S_2,S_3),$
$T_{23}(S_0,S_1,S_2,S_3),T_{24}(S_0,S_2,S_3),T_{25}(S_0,S_2),T_{26}(S_0,S_2),T_{27}(S_1,S_2),T_{28}(S_1,S_2),T_{29}(S_1,S_2),$
$T_{30}(S_1,S_3),T_{31}(S_1,S_3),T_{32}(S_3),T_{33}(S_3),T_{34}(S_3),T_{35}(S_3).$

The support and the confidence are often used to qualify association rules. For a rule $a \Rightarrow b$, these are defined by:

$$Support=\frac{n_a}{n} \qquad Confidence=\frac{n_{ab}}{n_a}$$

where n is the number of transactions, n_a is the number of transactions which satisfy a and n_{ab} is the number of transaction which satisfy $a \wedge b$.

Based on these transactions the following rules and meta-rules were obtained:

R_1:$(S_0 \Rightarrow S_1)$ *support=0.74 confidence=0.88*

R_2:$(S_2 \Rightarrow S_0)$ *support=0.57 confidence=0.85*

R_3:$(S_3 \Rightarrow (S_2 \Rightarrow S_0))$ *support=0.62 confidence=1.0*

The rules were found considering a threshold of *0.8* for confidence and *0.5* for support in the "A priori" algorithm.

The meta-rule found using the above thresholds has a significance (in the context of these artificially created document image) equivalent with a logo in a real document image. When we find a certain logo we expect rules between symbols which are specific to that document.

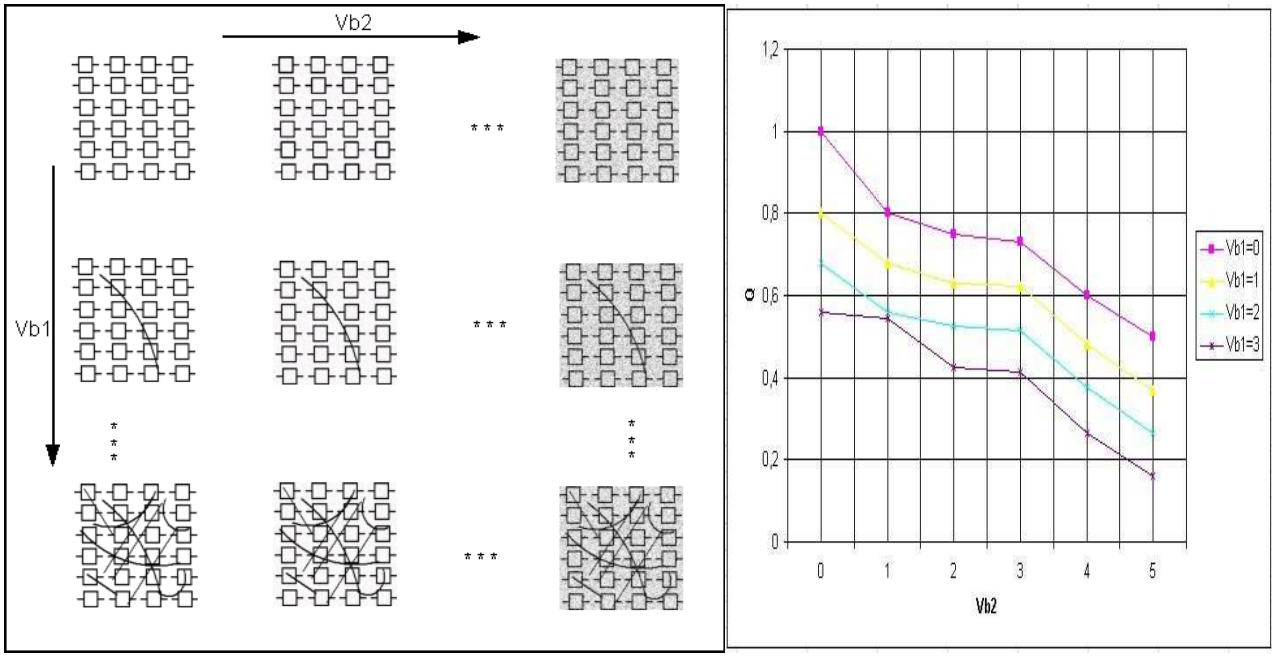
4.2 Robustness

This section presents an experiment which aims at assessing the robustness of our approach. Fig. 8(a) represents several occurrences of the same symbol with different levels of noise. Two kinds of noise have been introduced :

- *Vb1* models the connectivity of several graphic information,
- *Vb2* is a gaussian noise on the grey level image.

The Vb1 noise highlights the capacity of the method to deal with connected and distorted symbols. Even when some symbols are unrecognisable the property of being frequent is kept.

Fig. 8(b) gives for each noise level of *Vb1*, the proportion of found symbols in relation to *Vb2*. Even if this proportion decreases with the noise, our objective is not to extract all symbols but rather to find redundancies that qualify the document. However, we can conclude that the thresholds have to be adapted to the noise on the document image.



(a)Different noise levels (b)Robustness evaluation

Fig. 8. Robustness to noise

5 Conclusions

The research undertaken represents a novel approach for finding symbols in line drawing documents as well as for discovering relations between automatically mined symbols. The approach uses data mining concepts for knowledge extraction. It aims at finding frequent symbols and relations. These frequent patterns are part of the document model and can be put in relation with the domain knowledge. The exposed method can be applied to other graph representations of a document. The only condition is that the document graph should contain symbols as disjoint graphs. In our future works, we will apply this approach to layout structures of textual document images to extract formatting rules. Some follow-up activities could be:

- post-processing of the neighbourhood graph in order to attenuate the noise influence;
- employment of error tolerant graph matching;
- utilization, at a semantic level, of more powerful indices for association rules;
- creation of a hierarchy of rules, probably a similar approach with Gras et al. [17].

References

[1] Yan, X., Han, J.: Closegraph: mining closed frequent graph patterns In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press (2003) 286 295

[2] Altamura, O., Esposito, F., Malerba, D.: Transforming paper documents into xml format with Wisdom++ *International Journal on Document Analysis and Recognition* 4 (2001) 2 17

[3] Messmer, B.: *Efficient Graph Matching Algorithms for Preprocessed Model Graphs*. PhD thesis, University of Bern, CH, Institute of Applied Mathematics (1995)

[4] Berardi, M., Ceci, M., Malerba, D.: Mining spatial association rules from document layout structures. In: *Proceedings of the Third International Workshop on Document Layout Interpretation and its Applications*. (2003)

[5] Cornuéjols, A., Mary, J., Sebag, M.: « Classification d'images à l'aide d'un codage par motifs fréquents ». In: *Actes de la Journée analyse de données, statistique et apprentissage pour la fouille d'image du Congrès RFIA*. (2004) 11–16

[6] Ordonez, C., Omiecinski, E.: Discovering association rules based on image content. In: *Proceeding of the IEEE Advances in Digital Libraries Conference*. (1999)

[7] Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In Bocca, J.B., Jarke, M., Zaniolo, C., eds.: *Proc. 20th Int. Conf. Very Large Data Bases*, VLDB, Morgan Kaufmann (1994) 487–499

[8] Washio, T., Motoda, H.: State of the art of graph-based data mining. *SIGKDD Explor. Newsl*. 5 (2003) 59–68

[9] Kuramochi, M., Karypis, G.: Frequent subgraph discovery. In: *Proceedings of the International Conference on Data Mining*. (2001)

[10] Inokuchi, A., Washio, T., Motoda, H.: An apriori-based algorithm for mining frequent substructures from graph data. In: *Proceedings of the Conference on Principle and Practice of Knowledge Discovery in Databases*. (2000)

[11] Gupta, A., Jain, R.: Visual information retrieval. *Comm. Assoc. Comp. Mach.*, 40 (May 1997) 70-79

[12] Barbar D., Domeniconi C., Kang N., Classifying Documents Without Labels, In : *Proceedings of the Fourth SIAM International Conference on Data Mining*, Lake Buena Vista, Florida, USA, April 22-24,2004

[13] Khotanzad, A. and Hong, Y.H. Invariant Image Recognition by Zernike Moments. IEEE Trans. on PAMI, 12 (5). 289-497, 1990

[14] Milligan, G. W., Cooper, M.C.: An Examination of Procedures for Determining the Number of Clusters in a Data Set. Psychometrika, 58(2),(1985)159-179.

[15] Skvoretz J., An algorithm to generate connected graphs, In: *Current research in social psychology*, Vol. 1, No. 5, 1996

[16] Pavlidis, T., Algorithms or Graphics and Image Processing, Computer Science Press, 1982.

[17] Gras, R., Kuntz, P., Briand, H.: « Hiérarchie orientée de règles généralisées en analyse implicative ». In: *Actes des journées francophones d'extraction et de gestion des connaissances*. (2003)

Annexe B

Réference CV : 5

E. Valveny, P. Dosch, A. Winstanley, Y. Zhou, S. Yang, L. Yan, W. Liu, D. Elliman, M. Delalandre, É. Trupin, S. Adam, and JM. Ogier. A general framework for the evaluation of symbol recognition methods. International Journal of Document Analysis and Recognition (IJDAR), 9(1) :59-74, 2007.

IJDAR (2007) 9:59–74
DOI 10.1007/s10032-006-0033-x

ORIGINAL PAPER

A general framework for the evaluation of symbol recognition methods

E. Valveny · P. Dosch · Adam Winstanley · Yu Zhou · Su Yang · Luo Yan · Liu Wenyin · Dave Elliman · Mathieu Delalandre · Eric Trupin · Sébastien Adam · Jean-Marc Ogier

Received: 1 April 2005 / Accepted: 22 September 2006 / Published online: 18 November 2006

Abstract Performance evaluation is receiving increasing interest in graphics recognition. In this paper, we discuss some questions regarding the definition of a general framework for evaluation of symbol recognition methods. The discussion is centered on three key elements in performance evaluation: test data, evaluation metrics and protocols of evaluation. As a result of this discussion we state some general principles to be taken into account for the definition of such a framework. Finally, we describe the application of this framework to the organization of the first contest on symbol recognition in GREC'03, along with the results obtained by the participants.

Keywords Performance evaluation · Symbol recognition

1 Introduction

Performance evaluation has become an important research interest in pattern recognition during the last years. As the number of methods increases there is a need for standard protocols to compare and evaluate all these methods. The goal of evaluation should be to establish a solid knowledge of the state of the art in a given research problem, i.e., to determine the weaknesses and strengths of the proposed methods on a common and general set of input data. Performance evaluation should allow the selection of the best-suited method for a given application of the methodology under evaluation.

E. Valveny (✉)
Centre de Visió per Computador, Edifici O, Campus UAB, Bellaterra (Cerdanyola), 08193 Barcelona, Spain
e-mail: ernest@cvc.uab.es

P. Dosch
LORIA, 615, rue du jardin botanique, B.P. 101, 54602 Villers-lès-Nancy Cedex, France
e-mail: Philippe.Dosch@loria.fr

A. Winstanley · Y. Zhou
National University of Ireland, Maynooth, County Kildare, Ireland
e-mail: adam.winstanley@nuim.ie

Y. Zhou
e-mail: yuzhou@cs.nuim.ie

S. Yang
Department of Computer Science and Engineering, Fudan University, Shanghai 200433, China
e-mail: suyang@fudan.edu.cn

L. Yan · L. Wenyin
Department of Computer Science, City University of Hong Kong, Honk Kong, China
e-mail: luoyan@cs.cityu.edu.hk

L. Wenyin
e-mail: csliuwy@cityu.edu.hk

D. Elliman
University of Nottingham, Nottingham, UK
e-mail: dge@cs.nott.ac.uk

E. Trupin · S. Adam
LITIS Laboratory, Rouen University, Rouen, France
e-mail: Sebastien.Adam@univ-rouen.fr

M. Delalandre · J.-M. Ogier
L3i Laboratory, La Rochelle University, Rochelle, France
e-mail: mathieu.delalandre@univ-lr.fr

J.-M. Ogier
e-mail: jean-marc.ogier@univ-lr.fr

Following these criteria, image databases have been collected and performance metrics have been proposed for several domains and applications [6,12,18,21,29]. Several of these works deal with the evaluation of processes involved in document analysis systems, such as thinning [13], page segmentation [2], OCR [28], vectorization [22,26,27] or symbol recognition [1], among others. In fact, the general performance evaluation framework proposed in this paper is based on the work carried out for the contest on symbol recognition organized during GREC'03 [25].

Although in any domain there are always some specific constraints, we can identify three main issues that must be taken into account in the definition of any framework for performance evaluation: a common dataset, standard evaluation metrics and a protocol to handle the evaluation process. The common dataset should be as general as possible, including all kinds of variability that could be found in real data. It must contain a large number of images, each of them annotated with its corresponding ground-truth. Metrics must be objective, quantitative and accepted by the research community as a good estimate of the real performance. They must help to determine the weaknesses and strengths of each method. In many cases, it is not possible to define a single metric, but several metrics have to be defined according to different evaluation goals. The protocol must define the set of rules and formats required to run the evaluation process.

In this paper, we propose a general framework for performance evaluation of symbol recognition. For each of these issues (data, metrics and protocol), we describe the main problems and difficulties that we must face and we state the general guidelines that we have followed for the development of such a framework. Finally, we show how we have applied this framework to the organization of the GREC'03 contest on symbol recognition.

Symbol recognition is one of the main tasks in many graphics recognition systems. Symbols are key elements in all kinds of graphic documents, as they usually convey a particular meaning in the context of the application domain. Therefore, identifying and recognizing the symbols in a drawing is essential for its analysis and interpretation and a great variety of methods and approaches have been developed (see some of the surveys on symbol recognition [5,8,17] to get an overview of the current state of the art).

In fact, symbol recognition could be regarded as a particular case of shape recognition. However, there are some specific issues that should be taken into account in the definition of an evaluation framework. First, symbol recognition is not a stand-alone process. Usually, it is embedded in a whole graphics recognition system where the final goal is not only to recognize perfectly segmented images of symbols, but to *recognize and localize* the symbols in the whole document. Sometimes segmentation and recognition are completely independent processes, but sometimes they are related and performed in a single step. For evaluation, that means that we must consider two different sub-problems: recognition of segmented images of symbols and localization and recognition of symbols in a non-segmented image of a document. These two different sub-problems will be referred to as *symbol recognition* and *symbol localization*, respectively, throughout the paper. Second, sometimes, symbol recognition depends on other tasks in the graphics recognition chain (for example, binarization or vectorization). The performance of these processes can also influence the performance of symbol recognition. We should try to make the evaluation of symbol recognition independent of these other tasks. At least, the analysis of the results should be made taking into account their influence. Third, symbol recognition is applied to a wide variety of domains (architecture, electronics, engineering, flowcharts, geographic maps, music, etc.). Some methods have been designed to work only in some of these domains and have been only tested using very specific data.

Finally, if the goal of performance evaluation is to help to determine the current state-of-art of research, then, any proposal should give response to the needs of the whole research community and should be accepted by it. Therefore, in our proposal, a key point is the idea of collaborative framework. The initial proposal must be validated by the users and must be easily extended as research advances and new needs or requirements appear. Thus, our proposal relies on four desirable properties:

- public availability of data, ground-truth and metrics
- adaptability to user needs: each person must be able to select a subset of the framework to work with
- extensibility the framework must allow for new kinds of images or metrics to be easily added
- collaborative validation of data, metrics and ground-truth.

The paper is organized as follows: Sects. 2 and 3 are devoted to discuss each of the main aspects in performance evaluation, data and evaluation metrics, respectively. In Sect. 4 we describe the protocol and implementation issues of the framework. In Sect. 5 we show the application of this framework to the GREC'03 contest. Finally, in Sect. 6 we state the main conclusions and discuss the future work.

2 Data

One of the key issues in any performance evaluation scheme is the definition of a common set of test data. Running all methods on this common set will permit to obtain comparable results. This set should be generic, large, and should contain all kinds of variability of real data.

In symbol recognition, generality means including all different kinds of symbols, i.e., symbols from all applications (architecture, electronics, engineering, flowcharts, geographic maps, music, etc.) and symbols containing all types of features or primitives (lines, arcs, dashed-lines, solid regions, compound symbols, etc.). In this way, we will be able to evaluate the ability of recognition methods to work properly in any application.

On the other hand, variability can be originated by multiple sources: acquisition, degradation or manipulation of the document, handwriting, etc. All of them should be taken into account, when collecting test data in order to evaluate the robustness of recognition methods.

However, in symbol recognition many methods are specifically designed for a particular application or a particular kind of symbols under specific constraints. Therefore, it is not possible to define a single dataset containing all kinds of images. Then, following the general principle of adaptability, stated in the previous section, we propose to define several datasets, instead of a single one. Each dataset will be labeled according to the kind of images contained in it. In this way, users can select the datasets they want to use according to the properties of their method. In addition, we can generate as many datasets as required, combining all kinds of symbols and criteria of variability.

Therefore, we need to establish some criteria to classify and organize all kinds of symbols (Sect. 2.1). Then, we must also identify and categorize all kinds of variability of real images (Sect. 2.2). Finally, we will be able to discuss how to collect and generate a large amount of data and organize it according to these criteria of classification (Sect. 2.3).

2.1 Classification of symbols

In general, there are two points of view for classifying evaluation tests and their associated data [9]: technological and application. The technological point of view refers to the evaluation of methods as stand-alone processes trying to measure their response to varying methodological properties of input data and execution parameters. Datasets must be independent of the application and must differ on the kind of image features. For symbol recognition this point of view corresponds to the generic evaluation of performance independently of the application domain. Image features will be the different shape primitives that can be found in the symbols. According to the data used in the contest, we have identified three shape primitives: straight lines, arcs and solid regions. However, new primitives (for example, dashed lines, text, textured areas) could be added to the dataset if required.

On the other hand, the application point of view refers to the evaluation of methods in a particular application scenario. Different datasets will correspond to different application domains of a given method, and each dataset will only include specific data for the given application. In symbol recognition, categories refer to the different domains of application: architecture, electronics, geographic maps, engineering drawings or whatever domain we should consider.

We have used this double criteria to classify symbols in our framework. The support for it is that algorithms are usually designed using these two points of view too. Some methods are intended to be as general as possible, and work well with symbols in a wide range of applications. On the other hand, some other methods are intended to be part of a complete chain of a graphics recognition system in a particular application domain. They are specifically designed to recognize the symbols in that application.

These are the two main criteria for classifying test data. But from a more general viewpoint, we can use labels corresponding to property/value pairs. The property can refer to the application domain, primitives, origin, etc., while values are occurrences of these properties (respectively, architecture/electronic/..., segments/arcs and segments/..., CAD design/sketch/...). This provides a general labeling system which can be easily extended, allowing to define as much data as needed.

Therefore, we will assign at least two categories of labels to each symbol: one with the domain of the symbol and the other with the set of primitives composing it. Each dataset is also labeled in the same way according to the symbols included in it. With this organization each user can select those datasets that fit the features of the method under evaluation. In addition, new categories of data can be easily added or modified and therefore, the framework can evolve according to research needs. In Fig. 1 we can see several examples of images classified according to both points of view. Note that each symbol can be included in several categories.

2.2 Variability of symbol images

Robustness to image degradation is essential for the development of generic algorithms. Then, a framework

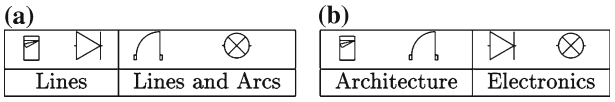


Fig. 1 Classification of the same images according to the two points of view: **a** technological, **b** application

for performance evaluation must include all kinds of degradation in the test data. Besides, images should be ranked according to the degree of degradation in order to be able to determine whether the performance decreases as the difficulty of images increases.

In general, we can distinguish four sources of variability in symbol recognition:

- acquisition parameters: acquisition device (scanner, camera or online device) and acquisition resolution
- global transformations: global skew of the document, rotation and scaling of symbols
- binary noise: degradation of old documents, photocopies, faxes and binarization errors.
- Shape transformations: missing or extra primitives (due to segmentation errors) and shape deformations due to hand-drawing.

We need to guarantee that all these types of degradations are included in the common dataset. We will generate different datasets corresponding to each kind and degree of transformation and to selected combinations of them. Each dataset will be labeled accordingly too.

2.3 Generation of test data

According to the principles stated in previous sections we need to collect a large number of images. These images will be organized into several datasets, including all kinds of symbols described in Sect. 2.1 and all types of variability identified in Sect. 2.2. In addition, images must be labeled with the ground-truth, i.e., the expected result. We have to collect segmented images of isolated symbols, but also non-segmented images of documents in order to evaluate both symbol recognition and symbol localization, as stated in Sect. 1.

There are basically two possibilities for collecting test data: to use real data or to generate synthetic data. In the following of this section, first, we will discuss the advantages and drawbacks of each approach and how we use them in our framework. Then, we will consider some other specific issues related to the generation of data for evaluation of symbol recognition.

2.3.1 Real data

Clearly, the main advantage of using real data is that it permits to evaluate the algorithms with the same kind of images as for real applications. Then, evaluation will be a very good estimate of performance in real situations. However, manually collecting a large number of real images is a great effort, unaffordable in many cases. The task of annotating images with their corresponding ground-truth is also time-consuming, and errors can easily be introduced. Another disadvantage is the difficulty of collecting images with all kinds of transformations and noise. Besides, it is not easy to quantify the degree of noise in a real image. Then, it is not possible to define a ranking of difficulty of images according to the degree of noise.

2.3.2 Synthetic data

As an alternative, we can develop automatic methods to generate synthetic data. Clearly, the main advantage is that it allows to generate as many images as necessary, and the annotation of images with the ground-truth is also automatic. Then, manual effort is reduced. However, we need to devote research effort to the development of models and methods able to generate images resembling real ones with all possibilities of noise and transformations. This is not an straightforward task in many cases although several works have been done in related fields of document analysis [3,11,15,16]. Images generated using these methods will be easily classified according to the type and degree of noise or degradation applied, permitting to assess the reduction in performance with increasing degrees of image degradation.

We argue that both types of images are useful in a general framework for performance evaluation of symbol recognition. We believe that real images are the best test for assessing performance in symbol localization. It is really difficult to develop automatic methods to generate non-segmented images of complete graphic documents. Besides, as we can find many symbols in a single graphic document, not many images are required. The problem can be the annotation of images with the ground-truth. We discuss it in Sect. 3.3.

On the other hand, synthetic images are the only way to perform evaluation tests with large sets of segmented images taking into account all degrees of degradation and variation. In this case, many images are required and it is easier to develop methods for their generation. In our framework we have developed methods for the generation of global transformations, binary noise (based on Kanungo's method [15] and shape transformation (based on active shape models [25]).

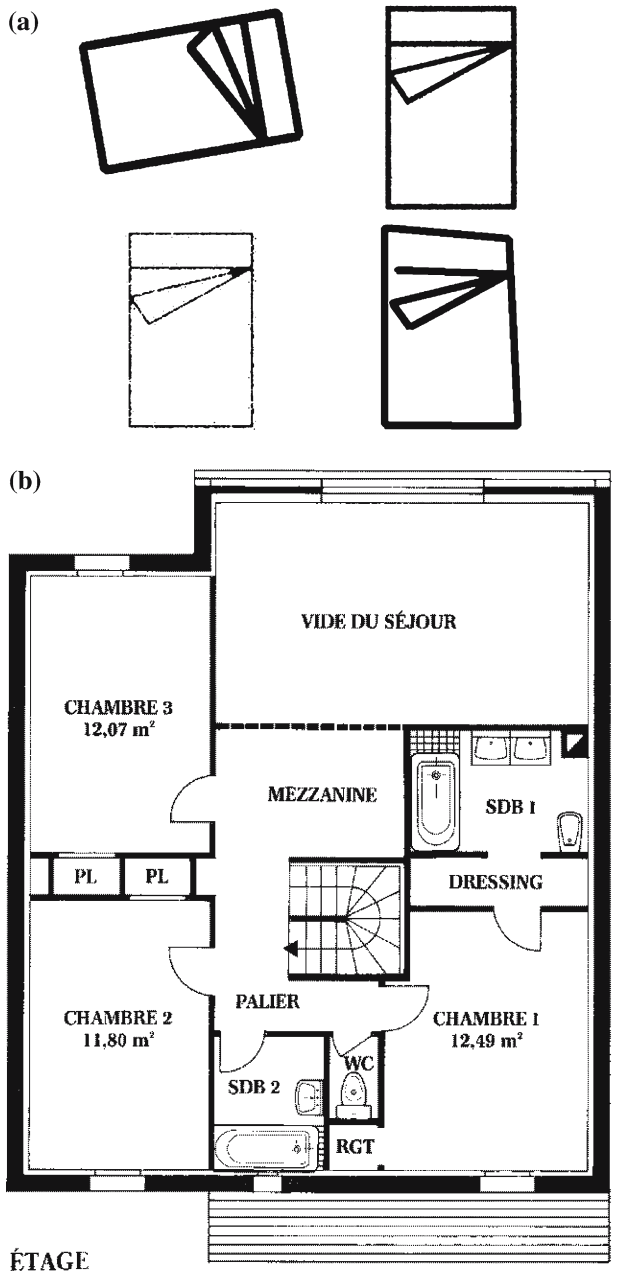


Fig. 2 Generation of data: **a** synthetic images, **b** real images

Figure 2 shows both synthetic and real images for symbol recognition.

2.3.3 Specific issues

In addition, we have to take into account two other specific issues of symbol recognition when generating test data.

- *Relation to vectorization*: As explained in Sect. 1 symbol recognition is simply one task in the graphics recognition chain. Vectorization is usually performed as a previous step for recognition and then, many symbol recognition methods work directly on the vectorial representation of the image. The problem is that, although there is not an optimal vectorization method, the result of vectorization can influence the performance of recognition. Then, apart from a raster representation of images, we must also provide images in a common vectorial format so that all methods can use the same vectorial data and recognition results are not influenced by the selected vectorization method. For images that can be automatically generated in vectorial format, we can provide images in their ideal vectorial representation, without need for applying any vectorization method. If not possible (for example, for real images of for synthetic images with binary degradations), we should apply different standard vectorization methods to the raster image.
- *The problem of scalability*: One of the problems in symbol recognition [17] concerns scalability: many methods work well with a limited number of symbol models, but their performance decrease when the number of symbols is very large (hundreds or thousands of symbols). One of the goals of the evaluation of symbol recognition must be to assess the robustness of methods with a large number of symbols. Then, for each kind of test several datasets with an increasing number of symbols will be generated.

3 Performance evaluation

3.1 Objectives

In some pattern recognition fields, the main goal of evaluation is the definition of a global measure that permits to determine the "best" method on a standard and common dataset. However, it seems difficult to follow the same approach for symbol recognition. As we have stated in previous sections, performance of symbol recognition depends on many factors and it is not realistic trying to define a single measure and dataset taking into account all of them. Then, as symbol recognition remains an active research domain, it seems more interesting to focus on analyzing and understanding the strengths and the weaknesses of the existing methods. This will be the main goal of the proposed evaluation framework.

In this context, evaluation relies on three issues: first, the definition of a number of standard datasets, covering the full range of variability, as discussed in Sect. 2. Second, the definition of a set of measures, each of them aiming at evaluating a specific aspect of performance.

This will be discussed in Sect. 3.2. The definition of metrics is highly related to the definition of the ground-truth. This point will be developed in Sect. 3.3. Third, the analysis of the results after calculating all the measures over all the datasets, in order to draw conclusions on the strengths and weaknesses of each method (Sect. 3.4).

3.2 Metrics

In the last years, several graphics recognition contests have been organized, notably in the framework of the International Workshop on Graphics Recognition (GREC). As a result of this effort, several metrics and protocols have been developed [14,22,26], with more or less success, as sometimes, they favor the properties of some of the contestant methods.

A similar work has to be done for symbol recognition: what is the measure that permits to say that a given symbol recognition method is good? Clearly, the answer will be different for each of the two sub-problems identified in Sect. 1: *symbol recognition* and *symbol localization*. In the first case, for the recognition of isolated symbols, it can be enough to count the number of correctly recognized symbols. But, in the second case, other information, such as location, orientation and scale of symbols should also be considered. Thus, in the following, we will discuss different metrics for each of these sub-problems.

3.2.1 Symbol recognition

It seems clear that the basic metric for symbol recognition should be to test if the recognized symbol matches the test symbol according to the ground-truth. Thus, the recognition rate is the main evaluation criteria. This was the simple approach used in the GREC'03 contest. Because of the wide number of open questions regarding performance evaluation of symbol recognition, we decided, in a first time, to consider only the basic features in order to advance in a better understanding of all issues involved in it.

However, we believe that this criteria could be complemented with other measures, in order to get a deeper analysis of recognition methods, taking into account other evaluation aspects. For example,

- The recognition rate, considering second or third candidates, if this information is provided by some methods.
- The orientation and scale of the symbol: we could complete the recognition rate with a measure of the accuracy in recovering the orientation and scale of the symbol. This measure can be based on the difference between the orientation and scale provided by the recognition method and the ground-truth.
- The computation time: we propose to use the average time per image. This metric will allow to compare the results on tests with different number of images or symbols. However, to be comparable, all recognition methods should be run on the same machine under the same conditions. That should be considered in the definition of the protocol (Sect. 4.2).
- Scalability, i.e., how the performance degrades as the number of symbol models increases. We can measure it according to the degradation of recognition rates or according to the computation time.

3.2.2 Symbol localization

In the best of our knowledge, no performance evaluation has ever been organized on symbol localization. For this task, the problem of defining accurate metrics is harder than in the case of symbol recognition. We have to face two issues: the representation of the symbols, and the definition of the metric itself.

The representation of a symbol (in the ground-truth as well as in the recognition result) must include not only an identifying label (as in the case of symbol recognition), but also the location of the symbol. The problem is that it is not easy to define a single representation of the location of a symbol. The best representation will depend on the kind of method. For example, if a recognition method works on the raster representation of a symbol, the symbol location has to be computed with respect to the related set of pixels. But if a recognition method works on the vectorial representation of the symbol, its location has to be computed with respect to the involved set of vectorial primitives, maybe taking into account some attributes of these primitives, such as thickness. Clearly, both representations do not have to be equal.

In fact, we argue that the representation of the location of a symbol must be unique and independent of the kind of method or image format, as the definition of multiple representations arise the following issues:

- Multiple metrics have to be defined as the definition of the metric depends on the representation of the symbols. This can permit to define more accurate metrics but also requires to take into account all possibilities.
- Multiple representations also lead to the definition of multiple ground-truth for the same data.

- Multiple metrics and multiple ground-truth then lead to multiple performance analysis as it will be difficult to compare results evaluated with different metrics.

As a first approach for representing the location of a symbol, we propose the use of basic including rectangles, that enclose symbols, as described by Mariano et al. [20]. This representation seems to be simple and efficient. These rectangles can even be defined as bounding-boxes.

Then, the metric between a ground-truth symbol and a result symbol can be based on the percentage of overlapping between their including rectangles, in the case that their associated labels match. Otherwise, the similarity value will be 0. This metric permits to work at the desired level of accuracy. We can fix a threshold so that only symbols with a percentage of overlapping above this threshold are considered as recognized. In this way, defining several thresholds, we can obtain different recognition results at different levels of accuracy.

In order to combine the results of the metric obtained for every symbol in the image, we propose to adopt a metric similar to the one used during the ICDAR'03 conference on the robust reading competition [19] for the text recognition in everyday scenes. The definition principles are based on the fact that the metric must favor the most pertinent applications, and penalize trivial solutions, like the definition of a single bounding-box which fully overlaps the image, or the definition of an excessive large number of bounding-boxes.

So the proposed metric is based on the notions of *precision* and *recall*. For a given test, let T be the number of targets belonging to the ground-truth, and R the set of results supplied by an application. The number of exact results is called e. The precision p is then defined as the number of exact results divided by the number of results:

$$p = \frac{e}{|R|}.$$

Thus, the applications that overestimate the number of results are penalized by a little precision score. The recall r is defined as the number of exact results divided by the number of targets:

$$r = \frac{e}{|T|}.$$

Thus, the applications that underestimate the number of results are penalized by a little recall score. The precision and the recall may then be combined, if needed, to determine the global score s, expressing the recognition rate:

$$s = \frac{2}{(1/p) + (1/r)}.$$

3.3 Ground-truth

As said above, the definition of the ground-truth depends basically on the representation of the symbols. Once again, we have to distinguish between the definition of the ground-truth for symbol recognition and for symbol localization.

If we consider symbol recognition, where only segmented symbols are involved, ground-truthing can be a simple task. It basically consists of determining the label of the symbol and this can be easily done by an human operator and even, more easily by an automatic method of image generation. If we also want to take into account the accuracy in orientation and scale, we must include this information in the labeling of the symbol too. But this can be easily done with an automatic method of image generation.

However, if we consider symbol localization, ground-truthing is more difficult. In this case, both the label and the location of the symbol have to be defined. According to the single proposed metric (see Sect. 3.2), the definition of the ground-truth is also unique, and then easier and more realistic to manage.

Although the representation of the symbol gives a theoretical and concrete framework for the definition of the ground-truth, some differences can exist between the theoretical definition and the real definition of a given ground-truth. Indeed, the bounding-box defined by one person for a given symbol could appear misplaced to another person. Thus, there is a part of personal and subjective interpretation in the definition of the ground-truth.

This point can be a serious problem, as the ground-truth has to be accepted by the whole community to be fully considered as a reference. To address this issue, we are fully convinced that a collaborative framework is required, as already pointed out in Sect. 1.

The basic idea is to involve a ground-truth designer and some ground-truth validators for a given ground-truth. Meanwhile, a ground-truth definition can be modified if it is not satisfactory. Of course, a ground-truth designer of some test data cannot be the ground-truth validator of the same test data too. Once a ground-truth is validated by some people, say two or three, then, it can be considered valid. This organization could be compared to a review process for a scientific conference. Obviously, this organization is easier to implement if a collaborative tool is available, as the associated workflow is crucial. This tool includes the following features:

- General ground-truthing functionalities: images visualization (raster, vectorial), bounding-box definition, label definition ...
- Directly interfaces with the database implementing the information system containing all information required for performance evaluation:
  - information about the data: models of symbols, test data and related ground-truthing.
  - information about users involved in the evaluation: their role and corresponding access privileges (ground-truth design and validation, data contributor ...)
- The collaborative tool must be unique, in order to be used in good conditions by all people involved in the ground-truthing process. This implies that it has to be available for a sufficient number of platforms and ensures that all people work with the same environment or references.

We want to point out that these principles and this framework are a priori necessary in order to ensure that test data, as well as their associated ground-truth, are considered as valid by the whole community, and not by only one person. All the performance evaluation process relies on this assertion.

3.4 Analysis of the results

The results of the participants have to be analyzed in order to determine the objectives of such a performance evaluation campaign: the understanding of the strengths and the weaknesses of the existing methods. This analysis must be done with respect to the considered categories of data, the number of model symbols involved and several other interesting criteria.

Independently of this large number of criteria, we would point out that basically the analysis can be led from the data point of view (data based), as well as from the methods point of view (methods based). Indeed, if it is interesting to understand what are the methods giving good results with a lot of data, it is also interesting to understand what are the data difficult to recognize with respect to the several recognition approaches. The interest of a performance evaluation campaign is guided by these two points of view.

Based on the metric that has been defined for symbol recognition, we propose to define an index that permits to perform the analysis of the results from different points of view. This index is a measure of the degradation of the performance along a set of tests with an increasing level of difficulty. Let r_0 be the recognition rate for the test acting as the reference test (it should be the "easiest" test in the series). Then the degradation of performance for a given test i is defined as

$$d_i = \frac{r_0 - r_i}{r_0}.$$

This index gives the measure of how the original performance degrades when some kind of degradation is applied to the original images. As the index is normalized by the original recognition rate it provides a good estimate of the loss of performance as it does not depend on the recognition rate for ideal images.

In this way, we can measure the robustness of recognition methods to several properties, such as scalability or degradation. We simply need to define a series of tests with an increasing number of symbols (for scalability) or with different levels of degradation and compute the degradation index for every test. Some examples of the application of this index to the analysis of the results will be shown in Sect. 5.

4 Implementation

4.1 Introduction

The implementation of any performance evaluation system requires the definition of a set of tools and protocols in order to execute the tests, exchange information between the participants and the organizers and manage all the information about test data and results. This set of tools and protocols must rely on the general concepts stated in Sect. 1, such as the public availability of data, the adaptation to user requirements and the simplicity of management.

Among all these issues, in the remainder of this section we will discuss the main ideas regarding protocols and formats (Sect. 4.2), the organization of datasets (Sect. 4.3) and the general architecture of the system (Sect. 4.4).

4.2 Protocols and formats

Whatever the evaluation criteria and data, an evaluation framework must provide formats and tools allowing to exchange information about models, tests and results [24]. In performance evaluation of symbol recognition, the first issue is about the format of images. One basic assumption to be made is that the format of images must not degrade the original image and must be freely available for all participants. As there are methods working on raster binary images and methods working on vectorial images, whenever it is possible, we have to

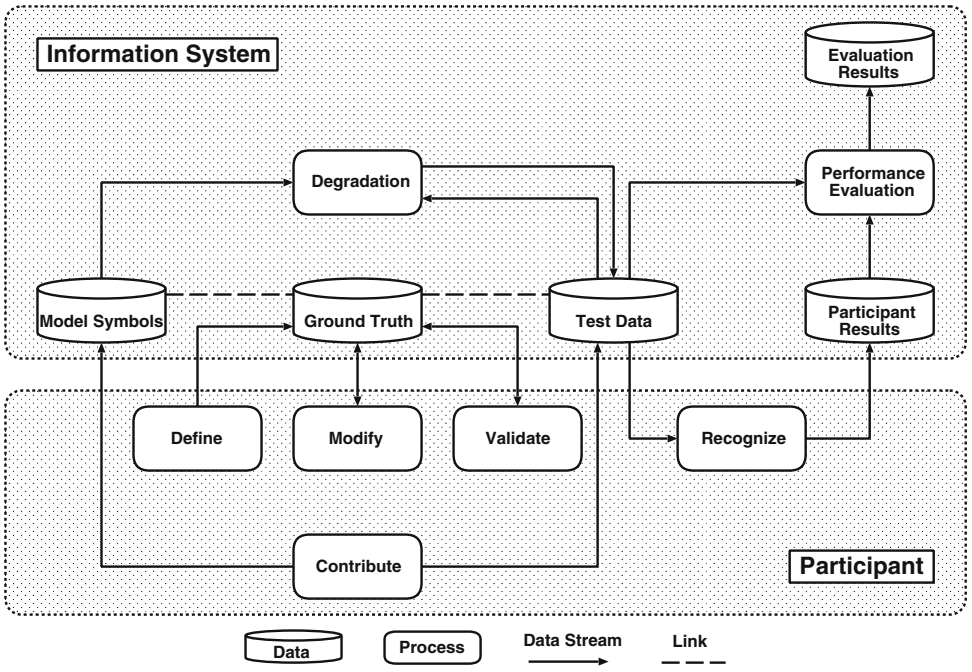


Fig. 3 Overview of the discribed performance evalution system

provide test images in both formats. Raster images are not a big problem as there are a lot of very popular solutions (such as TIFF, BMP and PNG). On the vectorial side, some "standard" formats exist, such as DXF or more recently SVG, but they are complex to manage. Thus, we have decided to use a simpler vectorial representation, the VEC format proposed by Chhabra and Phillips [4]. This simple format have already been used in other contests on graphics recognition (vectorization and arc detection) and therefore, it is already known by the symbol recognition community. Moreover, the simplicity of its definition would permit to eventually extend it, if required.

To manage the contest, several other file formats are required to precisely describe the tests, the results and the ground-truth. In this case, the choice of the format is a question of finding the best compromise that permits to express all the information that is required without obliging the participant methods to interface with too complex formats. We have found that XML fulfills these requirements as it is a flexible and standard format, allowing to easily describe complex information. Moreover, the use of a DTD or a scheme can help to normalize the data, avoiding description problems or confusions, and associated with the XSLT style-sheets, it allows the extraction and filtering of data that can be automatically processed, both for participants and organizers. Examples of these XML files can be seen in Figs. 3 and 4.

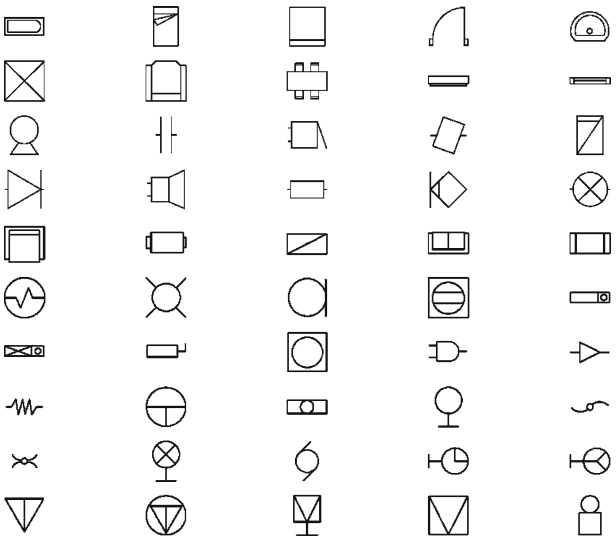


Fig. 4 Fifty symbols used in Contest

Another important issue is the protocol for execution of the tests. Following the principle of adaptability to user requirements, the basic idea must be to give each participant the possibility to choose which tests he want to compete in, according to the features of his method. To achieve this point, each test has to be considered as a stand-alone part and described with an independent XML file as explained in the next section. This principle is useful in some other situations. Thus, if a program crashes during a test, it is able to run the other tests.

The model that we have selected for the execution of the tests is a distributed model: each participant can take a file describing a test, execute it locally and then, provide the XML file with the results to the organizers. This option gives the maximum freedom to the users, for example regarding the platform of development or the interface of the recognition method. This is coherent with the general principles of the framework, but it can also have some drawbacks as the organizers do not have complete control on the development of evaluation and on some of the results. For example results regarding computation time are not fully comparable.

Finally, we want to point out that the availability of the framework (formats, data, etc.) is very important. In the context of performance evaluation, information about formats and data is required to prepare the methods for running the tests and for learning purposes.

4.3 Organization of datasets

A general framework for performance evaluation must include a very large number of datasets, taking into account all the variability described in previous sections. In order to manage this volume of datasets, we have to organize and classify them according to their properties. We will achieve this goal in a double way. On one hand, internally, we will store all information of every test in the information system that supports the evaluation framework and is described in the next section. On the other hand, externally, we will make it public to the participants by providing an XML description file for every test, as can be seen in Fig. 3. This file contains all the information that a participant has to know about a test:

- the name of images
- the ground-truth for each image (for training sets only)
- the category of symbols (as described in Sect. 2.1) from technological and application point of view
- the number of symbols involved in the dataset (for scalability issues)
- supported formats for images in the test
- whether the test corresponds to segmented or non-segmented images
- whether the test includes real or synthetic images
- whether the image acquisition is online or offline
- the type and degree of degradation applied to the data.

This organization allows to describe each test, so its associated properties are known. In this way, each participant can select the tests with the properties that fit to the method being evaluated. Moreover, it facilitates the analysis of the results, as it allows to organize the analysis according to the properties of the tests.

4.4 Information system

In order to manage all this framework, we propose to implement an information system supporting all required features. This information system must be implemented on the organizer's side, but it must be of public access and available through the Web with standard navigation tools. It plays the role of a public repository where any user (participant, organizer, ground-truth validator) can find all the required information about the evaluation process. However, the users are not tied to the implementation of the information system as the access is done through the web and all the exchange of information through the XML files that have been described in Sect. 4.2. Providing public access to all the information about data stored in the information system permits to set up a continuous evaluation framework. Evaluation does not depend on some predefined milestones, such as the organization of specific contests, but any user can, at any moment, download a set of tests, run a given method on them and provide the results back to the organizers. In this way we obtain the maximum flexibility for evaluation of current research.

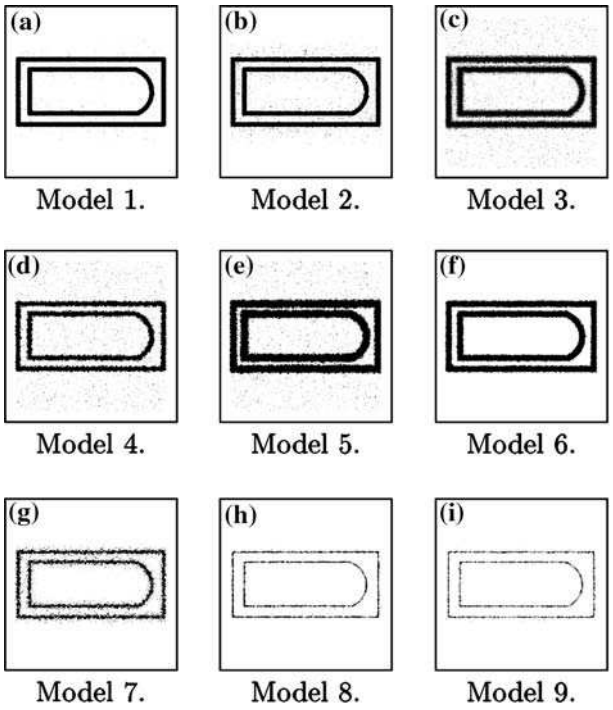


Fig. 5 Samples of some degraded images generated using the kanungo method for each model of degradation used

An overview of the system is presented in Fig. 5. Of course, the processes associated to the "participants" are related to all kinds of participants (contributors, ground-truth designers, contest participants ...) and some constraints are associated to the system. In particular, a participant cannot validate a ground-truth he has defined before, he cannot get his own test data (at least if it has not been degraded before), etc. Our aim is to point out that collaborative aspects must be taken into account from the beginning of the design of such a system.

5 Application of the framework: contest on symbol recognition at GREC'03

In this section we will show an example of application of the general framework presented before used in the *First Contest on Symbol Recognition* held during GREC'03. In this section we will explain how we have defined the two main issues involved in evaluation systems: data and metrics. We will also show the results obtained by the participants in the contest.

5.1 Data

The first decision concerned which symbols we were going to use in the contest and how to classify and organize them. For this first edition of the contest, we selected 50 symbols from two domains: architecture and electronics. All symbols were composed of at most two graphical primitives: lines and arcs. Then, according to the classification introduced in Sect. 2.1 we have used two features at the technological level (lines and arcs) and two categories at the application level (architecture and electronics) which have been used to classify test data. In Fig. 6 we can see all the symbols used in the contest.

We decided to use only synthetic data since it was easier to have a lot of well-organized images. Regarding the variability of data we worked with five categories of images: ideal data, images with aspect transformation (rotation and scaling), images with binary noise, images with shape distortions and images combining binary noise and shape distortion. We used the degradation model of Kanungo et al. [15] to generate nine different models of binary noise, and we defined a shape-distortion model based on *Active Shape Models* [7] to simulate hand-drawn images. Figures 7 and 8 show some examples of images with binary noise and shape degradation, respectively.

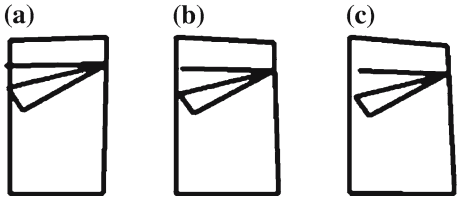


Fig. 6 Examples of increasing levels of vectorial distontion

Concerning specific issues of symbol recognition, we only used segmented images, so that only recognition was evaluated and not the ability to segment. Whenever possible, we provided both binary and vectorial versions of images. We used ideal vectorial representation when it could be automatically generated by the generation model. Therefore, for images with binary noise, only the binary representation was available as we did not apply any vectorization method to noisy binary images. Finally, we defined three different sets of symbols, with 5, 20 and 50 symbols each, to test the robustness of methods to scalability.

With all these combinations we generated a total number of 72 different tests of data. For each test, we provided a description file to the participants with the specification of symbols and images included in the test. Besides, we generated an XML file (Fig. 3) for each test, describing all the properties of the test, along with the ground-truth. Finally, participants generated an XML file (Fig. 4) with the description of the results obtained by their method for each test. Both kinds of XML files were imported to the contest database allowing for automatic comparison of the results with the ground-truth and automatic generation of recognition rates for each method and test.

5.2 Metrics

In this case, the definition of the metrics was very simple. We only worked with non-segmented images and, therefore, the only result of the application of a symbol recognition method was the label of the symbol identified in the image. Then, the metric simply consists of a recognition rate for each method and test, without taking into account the rejection.

5.3 Results

Five methods took part in the contest, although not all of them could run all the tests, due to the properties of their methods. The five participants were groups from the following institutions: University of Rouen—La Rochelle, National University of Ireland—Maynooth, City University of Hong Kong, University of Nottingham and Fudan Universty.

Fig. 7 Examples of XML file for test description

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE test SYSTEM "GRECTestSpecifications.dtd">
<?xml-stylesheet type="text/xsl" href="GRECOrgan2Particp.xsl"?>
<test format="bitmap"
      segmentation="true"
      applicationdomain="architecture">
  <testname>degrad-level1-m1</testname>
  <modelspath>models</modelspath>
  <imagespath>degrad-level1-m1</imagespath>
  <noise>
    <degradation type="kanungo">
      <noiseparam name="a0" value="0.5"/>
      <noiseparam name="a" value="0.5"/>
      <noiseparam name="b0" value="0.005"/>
      <noiseparam name="b" value="0.001"/>
      <noiseparam name="eta" value="0.0"/>
      <noiseparam name="ero" value="1.0"/>
    </degradation>
  </noise>
  <model name="ArchitecturalB.vec" id="m1"/>
  <model name="ArchitecturalC.vec" id="m2"/>
  <model name="ArchitecturalF.vec" id="m3"/>
  <model name="ArchitecturalG.vec" id="m4"/>
  <model name="ArchitecturalJ.vec" id="m5"/>
  <testimage name="image1.tif">
    <refmodel ref="m1"/>
  </testimage>
  <testimage name="image2.tif">
    <refmodel ref="m5"/>
  </testimage>
  <testimage name="image3.tif">
    <refmodel ref="m1"/>
  </testimage>
  <testimage name="image4.tif">
    <refmodel ref="m3"/>
  </testimage>
  ...
</test>
```

In Figs. 9, 10, 11, 12, 13, 14, 15, 16, we can see the results obtained by each of the methods in the tests they took part in. Figure 9 shows the results with ideal images of the symbols for the sets of 5, 20 and 50 symbols. It shows how the methods are able to discriminate among a large number of symbols. In Fig. 10 we can find the results for rotated and scaled images (for the set of 5, 20 and 50 symbols too).

Figure 11 contains the results with binary degraded images. In this case, only two methods were run on all the images and, therefore, only the results for these two methods are included. For each of the nine models of degradation the results with 5, 20 and 50 symbols are shown. In order to provide a more detailed analysis of the results with degradation we have also generated Fig. 12. In this figure we apply the degradation index defined in Sect. 3.4 to the nine models of binary degradation with the set of 50 symbols. The reference recognition rate for computing the index is the recognition rate for ideal images. This index clearly shows that for all models of degradation the method by the Fudan University is more robust to degradation than the method by the City University of Hong Kong.

Figures 13 and 14 show the results for images with vectorial distortion (for three levels of distortion) and with a combination of vectorial distortion and binary degradation.

In order to evaluate more precisely the scalability of methods we have included Fig. 15. This figure has been generated taking, for each method, the mean of recognition rates for all tests with 5 symbols, for all tests with 20 symbols and for all tests with 50 symbols. In this way, we can get a measure of the global scalability of each method. In Fig. 15a we can see the absolute recognition rates, while in Fig. 15b we have the degradation index defined in Sect. 3.4 applied to scalability. It is clear that this index helps to see the robustness of each method as the number of symbol increases.

Fig. 8 Examples of XML file for discription of results

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE testresult SYSTEM "GRECParticipantResults.dtd">
<testresult>
  <testname>testgrec</testname>
  <participant>participant1</participant>
  <imageresult>
    <imagename>
      testgrec-image1.tif
    </imagename>
    <symbol>
      <symbolname>ArchitecturalB.vec</symbolname>
      <location x1="35" y1="65" x2="67" y2="108"/>
      <orientation>34</orientation>
      <confidencerate>0.9885</confidencerate>
    </symbol>
  </imageresult>
  <imageresult>
    <imagename>
      testgrec-image2.tif
    </imagename>
    <symbol>
      <symbolname>ElectricalC.vec</symbolname>
      <location x1="108" y1="2003" x2="54" y2="21"/>
    </symbol>
  </imageresult>
  <imageresult>
    <imagename>
      testgrec-image4.tif
    </imagename>
    <symbol>
      <symbolname>ArchitecturalF.vec</symbolname>
      <location x1="30" y1="77" x2="165" y2="245"/>
    </symbol>
    <symbol>
      <symbolname>ArchitecturalF.vec</symbolname>
      <location x1="450" y1="479" x2="35" y2="88"/>
    </symbol>
  </imageresult>
</testresult>
```

Finally, in Fig. 16 we can see the computation time for every kind of test for sets with 5, 20 and 50 symbols. Only the method by the City University of Hong Kong reported results about the computation time. As expected, computation time increases as the number of symbols in the dataset increases too.

From these results we can draw some general conclusions:

- As expected, performance decreases when the number of symbols increase, even with ideal images.
- In general, methods can handle well the images with rotation or scaling. However, the performance degrades when both transformations are combined.
- There are no significant differences in the performance for the nine models of binary degradation.
- Methods are robust to the kind of shape deformations generated by the model of deformation.

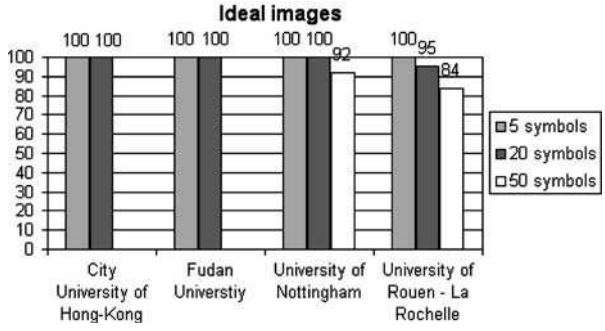


Fig. 9 Recognition rates (in the *y*-axis) of each participant method (in *x*-axis) for ideal tests

A more detailed discussion of these results can be found in the report on the GREC'03 contest [25].

Later, some of the groups have done further work on their methods and have obtained and published improved results [10].

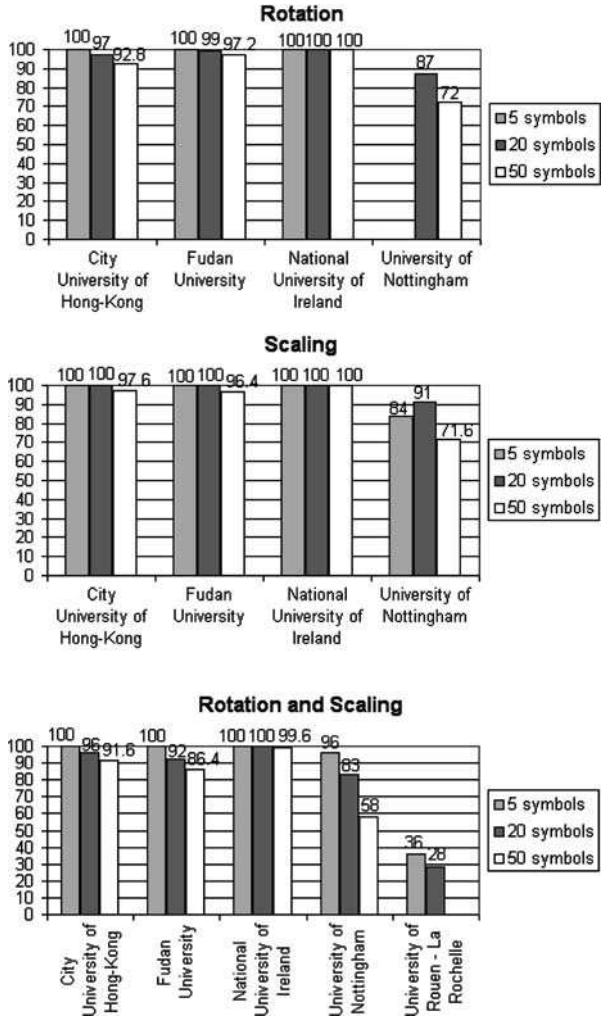


Fig. 10 Recognition rates (in the *y*-axis) of each participant method (in *x*-axis) for tests with rotation, scaling and combination of rotation and scaling

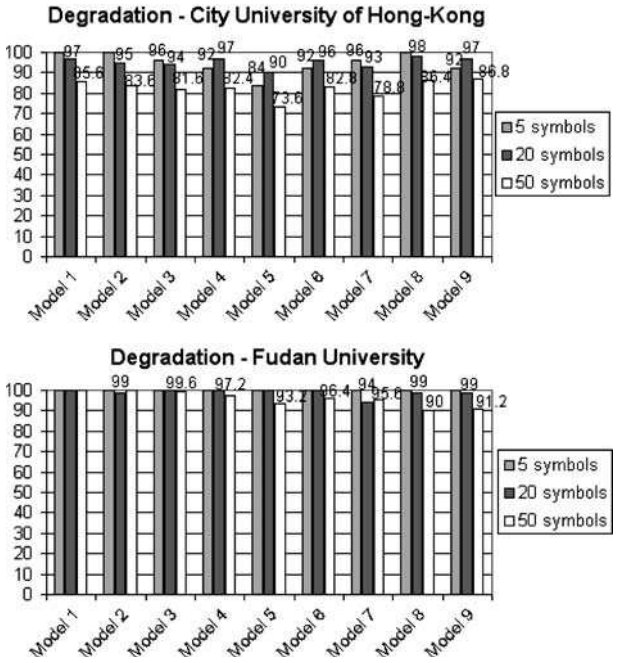


Fig. 11 Recognition rates (in the *y*-axis) for tests with the nine models of degradation (in *x*-axis) for methods by the City University of Hong Kong and the Fudan University

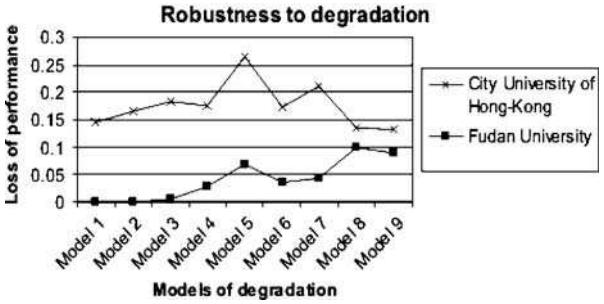


Fig. 12 Measure of robustness to degradation for the nine models of degradation with 50 symbols

6 Conclusion and future work

We have presented a general framework for performance evaluation of symbol recognition methods. This framework relies on some general principles that could also be applied to other similar performance evaluation tasks in the domain of graphics recognition and pattern recognition. These general principles arise from the discussion about the two main issues concerning any performance evaluation task: data and evaluation.

Concerning data, the framework relies on the classification of input data according to two different points of view: methodological—based on image features and application—based on the application scenario. This classification permits to define many different datasets for all possible kinds of input data. Regarding data generation we have stated the importance of using both real and synthetic images, including all types of noise and distortion. We have introduced a possible classification of distortion types and remarked the importance of including in the framework models and methods for automatic generation of degraded images.

Concerning evaluation, we have defined several metrics for symbol recognition and symbol location. Each metric gives response to different goals of performance evaluation.

In addition, one of the key ideas in the proposed framework is that of collaborative work so that the framework can be validated by the research community, and evolve according to its needs. Following this idea, a public and collaborative environment for performance evaluation of symbol recognition methods, ÉPEIRES,[1]

[1] http://www.epeires.org

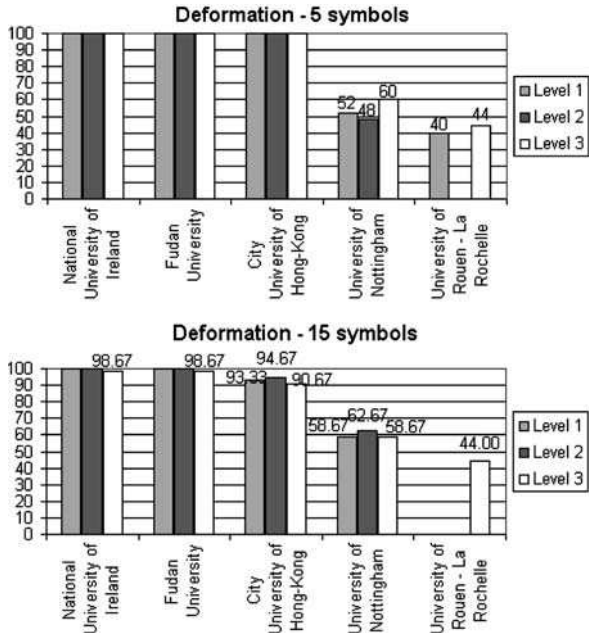


Fig. 13 Recognition rates (in the *y*-axis) of each participant method (in *x*-axis) for tests with deformation for both sets of symbols

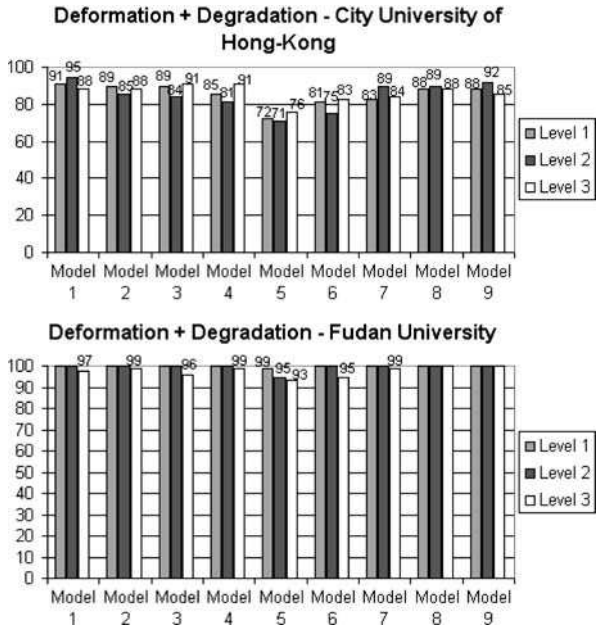


Fig. 14 Recognition rates (in the *y*-axis) for tests with the nine models of degradation (in *x*-axis) and three levels of degradation for methods by the City University of Hong Kong and the Fudan University

is currently under development. We hope that this environment will supply all data and resources needed by the symbol recognition community for evaluation purposes. All interested people are urged to use and to contribute to this environment.

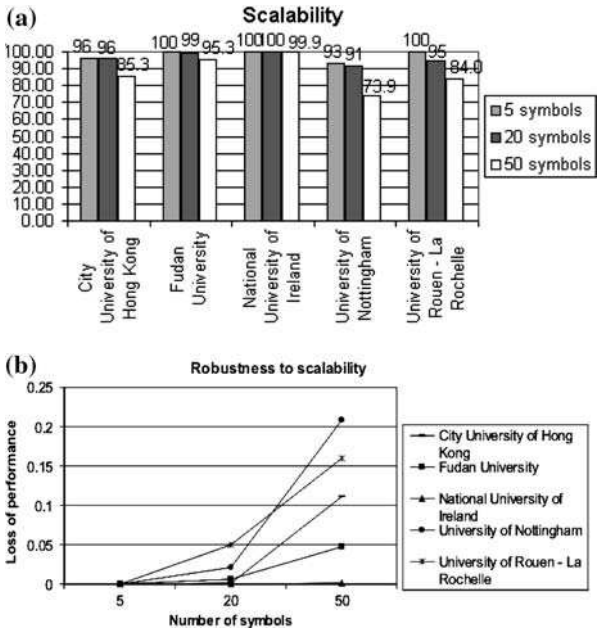


Fig. 15 **a** Evolution of recognition rates (in the *y*-axis) of each participant method (in *x*-axis) for tests with increasing number of symbols (5,20 and 50). **b** Measure of robustness to scalability for each participant method

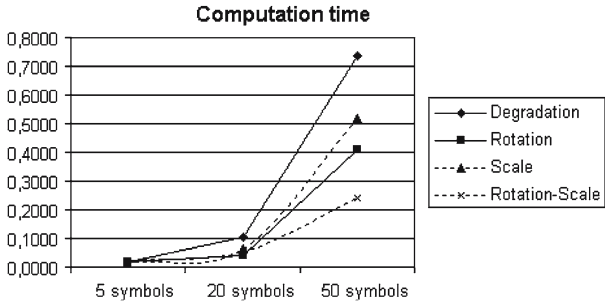


Fig. 16 Evolution of the computation time with the method by the City University of Hong Kong with an increasing number of symbol for each kind of test

Finally, we have described how these general principles have been used in the first international contest on symbol recognition, held during GREC'03. Currently, we are working on the extension of the framework for the next editions of the contest. In it, we plan to add real images with non-segmented symbols and, therefore, we will need to include the new metrics for symbol localization, as discussed in this paper.

Acknowledgments The contest organizers would like to acknowledge all participants of the first contest of performance evaluation of symbol recognition methods, as well as the organizers of the GREC workshop for the promotion and the opportunity given in these contests. The work of Luo Yan and Liu Wenyin was fully supported by grants from the City University of Hong Kong (Project No. 7001771 and 7001842) The work of E. Valveny was partially supported by CICYT TIC2003-09291, Spain.

References

1. Aksoy, S., Ye, M., Schauf, M., Song, M., Wang, Y., Haralick, R., Parker, J., Pivovarov, J., Royko, D., Sun, C., Farneboock, G.: Algorithm performance contest. In: Proceedings of 15th International Conference on Pattern Recognition, vol. 4, pp. 870–876, Barcelona, Spain (2000)
2. Antonacopoulos, A., Gatos, B., Karatzas, D.: ICDAR 2003 page segmentation competition. In: Proceedings of 7th International Conference on Document Analysis and Recognition, Edinburgh (Scotland, UK), pp. 688–689 (2003)
3. Baird, H.S.: The state of the art of document image degradation modeling. In: Proceedings of 4th IAPR International Workshop on Document Analysis Systems, Rio de Janeiro (Brazil) (2000)
4. Chhabra, A., Phillips, I.T.: The 2nd international graphics recognition contest—raster to vector conversion: a report. In: Tombre, K., Chhabra, A.K. (eds.): Graphics Recognition—Algorithms and Systems. Lecture Notes in Computer Science, vol. 1389, pp. 390–410. Springer, Berlin Heidelberg New York (1998)
5. Chhabra, A.K.: Graphic symbol recognition: an overview. In: Tombre, K., Chhabra, A.K. (eds.): Graphics Recognition—Algorithms and Systems. Lecture Notes in Computer Science, vol. 1389, pp. 68–79. Springer, Berlin Heidelberg New York (1998)
6. Clark, A.F., Courtney, P.: Databases for performance characterization. In: Stiehl, H.H., Viergever, M.A., Vincken, K.L. (eds.) Performance Characterization in Computer Vision. Kluwer, Dordrecht (2000)
7. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models: Their training and application. Comput. Vis. Image Underst. **61**(1), 38–59 (1995)
8. Cordella, L.P., Vento, M.: Symbol recognition in documents: a collection of techniques? Int. J. Doc. Anal. Recognit. **3**(2), 73–88 (2000)
9. Courtney, P., Thacker, N.A.: Performance characterization in computer vision: the role of statistics in testing and design. In: Blanc-Talon, J., Popescu, D.C. (eds.) Imaging and Vision Systems: Theory, Assessment and Applications. NOVA Science, Hungtington, NY (2003)
10. Delalandre, M., Trupin, E., Ogier, J., Labiche, J.: Contextual system of symbol structural recognition based on an object-process methodology. Electron. Lett. Comput. Vis. Image Anal. **5**(2), 16–29 (2005)
11. Ghosh, D., Shivaprasad, A.P.: An analytic approach for generation of artificial hand-printed character database from given generative models. Pattern Recognit. **32**, 907–920 (1999)
12. Guyon, I., Haralick, R.M., Hull, J.J., Phipliops, I.T.: Data sets for OCR and document image understanding research. In: Bunke, H., Wang, P.S.P. (eds.) Handbook of Character Recognition and Document Image Analysis, pp. 779–800. World Scientific, Singapore (1997)
13. Haralick, R.: Performance characterization in image analysis: thinning, a case in point. Pattern Recognit. Lett. **13**, 5–12 (1992)
14. Hilaire, X.: A matching scheme to enhance performance evaluation of raster-to-vector conversion algorithms. In: Proceedings of 7th International Conference on Document Analysis and Recognition, vol. 1, pp. 629–633. Edinburgh, Scotland (2003)
15. Kanungo, T., Haralick, R.M., Baird, H.S., Stuetzle, W., Madigan, D.: Document degradation models: parameter estimation and model validation. In: Proceedings of IAPR Workshop on Machine Vision Applications, Kawasaki (Japan), pp. 552–557 (1994)
16. Kanungo, T., Haralick, R.M., Baird, H.S., Stuezle, W., Madigan, D.: A statistical, nonparametric methodology for document degradation model validation. IEEE Trans. Pattern Anal. Mach. Intell. **22**(11), 1209–1223 (2000)
17. Lladós, J., Valveny, E., Sánchez, G., Martí, E.: Symbol recognition: current advances and perspectives. In: Blostein, D., Kwon, Y.-B. (eds.) Graphics Recognition—Algorithms and Applications. Lecture Notes in Computer Science, vol. 2390, pp. 104–127. Springer, Berlin Heidelberg New York (2002)
18. Lopresti, D., Nagy, G.: Issues in ground-truthing graphic documents. In: Blostein, D., Kwon, Y.-B. (eds.) Graphics Recognition—Algorithms and Applications. Lecture Notes in Computer Science, vol. 2390, pp. 46–66. Springer, Berlin Heidelberg New York (2002)
19. Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R., Ashida, K., Nagai, H., Okamoto, M., Yamamoto, H., Miyao, H., Zhu, J., Ou, W., Wolf, C., Jolion, J.M., Todoran, L., Worring, M., Lin, X.: ICDAR 2003 robust reading competitions: entries, results, and future directions. Int. J. Doc. Anal. Recognit. **7**(2-3), 105–122 (2005)
20. Mariano, V.Y., Min, J., Park, J.-H., Kasturi, R., Mihalcik, D., Li, H., Doermann, D., Drayer, T.: Performance evaluation of object detection algorithms. In: Proceedings of the 16th International Conference on Pattern Recognition, Quebec (Canada), vol. 3, pp. 965–969 (2002)
21. Philips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The feret evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **22**(10), 1090–1104 (2000)
22. Phillips, I.T., Chhabra, A.K.: Empirical performance evaluation of graphics recognition systems. IEEE Trans. Pattern Anal. Mach. Intell. **21**(9), 849–870 (1999)
23. Tombre, K., Chhabra, A.K. (eds.): Graphics Recognition—Algorithms and Systems. Lecture Notes in Computer Science, vol. 1389. Springer, Berlin Heidelberg New York (1998)
24. Valveny, E., Dosch, Ph.: Performance evaluation of symbol recognition. In: Marinai, S., Dengel, A. (eds.) Document Analysis Systems VI – Proceedings of 6th IAPR International Workshop on Document Analysis Systems, Florence (Italy). Lecture Notes in Computer Science, vol. 3163, pp. 354–365. Springer, Berlin Heidelberg New York (2004)
25. Valveny, E., Dosch, Ph.: Symbol recognition contest: a synthesis. In: Selected Papers from 5th International Workshop on Graphics Recognition, GREC'03. Lecture Notes in Computer Science, vol. 3088, pp. 368–385. Springer, Berlin Heidelberg New York (2004)
26. Wenyin, L., Dori, D.: A protocol for performance evaluation of line detection algorithms. Mach. Vis. Appl. **9**, 240–250 (1997)
27. 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, Selected Papers from 4th International Workshop on Graphics Recognition, GREC'01. Lecture Notes in Computer Science, vol. 2390, pp. 343–349. Springer, Berlin Heidelberg New York (2002)
28. Wilson, C.L., Geist, J., Garris, M.D., Chellappa, R.: Design, integration and evaluation of form-based handprint and OCR systems. Technical report, National Institute of Standards and Technology, Technical Report NISTIR 5932 (1996)
29. Zhang, Y.J.: A survey on evaluation methods for image segmentation. Pattern Recognit. **29**(8), 1335–1346 (1996)

Annexe C

Réference CV : 4

G. Dupont, S. Adam, Y. Lecourtier, and B. Grilhère. Multi objective particle swarm optimization using enhanced dominance and guide selection. International Journal of Computational Intelligence Research (IJCIR), 4(2) :145-158, 2008.

International Journal of Computational Intelligence Research.
ISSN 0973-1873 Vol.4, No.2 (2008), pp.145–158 doi:10.5019/j.ijcir.2008.134

Multi objective particle swarm optimization using enhanced dominance and guide selection

Gérard Dupont[1 2], Sébastien Adam[1], Yves Lecourtier[1] and Bruno Grilheres[1 2]

[1]Laboratoire d'Informatique de Traitement de l'Information et des Systmes (LITIS), Universit de Rouen, Saint- ´ Etienne-du-Rouvray, France
[2]EADS Defense and Systems, Information Processing and Competence Center, Val de Reuil, France

Abstract: **Nowadays, the core of the Particle Swarm Optimization (PSO) algorithm has proved to be reliable. However, faced with multi-objective problems, adaptations are needed. Deeper researches must be conducted on its key steps, such as solution set management and guide selection, in order to improve its efficiency in this context. Indeed, numerous parameters and implementation strategies can impact on the optimization performance in a particle swarm optimizer. In this paper, our recent works on those topics are presented. We introduce an "dominance variation which enables a finer neighborhood handling in criterion space. Then we propose some ideas concerning the guide selection and memorization for each particle. These methods are compared against a standard MOPSO implementation on benchmark problems and against an evolutionary approach (NSGAII) for a real world problem: SVM classifier optimization (or model selection) for a handwritten digits/outliers discrimination problem.**

Keywords: **Optimization, particle swarm, SVM model selection, multi objective optimizer, epsilon-dominance.**

I. Introduction

In several technical fields, engineers are dealing with complex optimization problems which involve contradictory objectives. Such multi-objective optimization problems have been extensively studied during the last decades. Existing approaches can be classified with respect to the hypotheses which are required for the computation. A common hypothesis is the derivability or continuity of the functions to be optimized. Unfortunately, such hypotheses are not verified for problems with complex models. Thus other ways have been found through meta-heuristic algorithms. Genetic algorithms are famous techniques in that domain and they have shown to be efficient on many optimization problems (see [13]). Recently, some researchers also tackle those problems with multi-objective particle swarm optimizer (see [10]).

Based on the work of James Kennedy and Russel Eberhart presented in [15], the particle swarm optimizers try to find solutions of optimization problems by using techniques inspired by the nature, as the genetic algorithms mimic evolution in species. In the last few years, PSO has been extensively studied and some results have shown that it can compete with other evolutionary algorithms such as genetic algorithms (see [16, 21, 31]). Multi-Objective PSO algorithms (refered as MOPSO in the paper) have also been implemented and have opened a large new field of interest (see [28]).

The aim of this paper is to propose some improvements of particle swarm optimizer dealing with multi-objective problems. These improvements concern the introduction of a new dominance and an original strategy for guide selection.

The paper is organized as follows: section II gives a brief overview on basic definitions involved in multi-objective optimization problems and in particle swarm optimization. In section III, our contributions concerning the dominance and the guide selection strategy are described. In section IV, these contributions are discussed through experimental results on benchmark problems. Finally, the proposed variant of the MOPSO algorithm is applied on a real world problem which concerns SVM multi-model selection for handwritten digit identification.

II. Basic definitions

This section presents the basic formalization of multi objective optimization problems. Then it describes the particle swarm core algorithm and its classical multi-objective implementation (see [10]).

A. Multi-objective optimization problems

Many definitions can be found for multi-objective optimization problems (see [9] for a precise definition of all the following equations). Such problems seek to minimize simultaneously N objective functions fk depending on n parameters in the form:

$$\begin{aligned} f_k: \quad & \mathbb{R}^n \longrightarrow \mathbb{R} \\ & \vec{x} \longrightarrow f_k(\vec{x}) \\ & with \quad k \in [1;N] \end{aligned} \tag{1}$$

In order to express parameter limitations that can be met in real world problems (such as material characteristics in engineering applications), some constraints must be introduced. They reduce the feasible region of $\mathbb{R}^n$ to a smaller one noted *S*. Usually, these constraints are modeled as M equations expressed as inequalities or equalities:

$$g_k(\vec{x}) \geq 0 \quad \text{with} \quad k \in [1;M] \tag{2}$$

$$h_k(\vec{x}) = 0 \quad \text{with} \quad k \in [1;M] \tag{3}$$

The global multi-objective problem can thus be defined as the minimization of:

$$\vec{f}\ (\vec{x}) = \{f_1(\vec{x}), \ldots, f_k(\vec{x}), \ldots, f_N(\vec{x})\} \tag{4}$$

$$\text{given } \vec{x} \in S \quad \leftrightarrow \begin{cases} \vec{g}\ (\vec{x}) \geq 0 \\ \vec{h}\ (\vec{x}) = 0 \end{cases} \tag{5}$$

B. Multi-objective solutions

In most case, multi-objective problems do not have a single global optimal solution according to equation 4 and a new definition of minimizing $\vec{f}\ (\vec{x})$ has to be used. The concept of optimum changes, because in multi-objective optimization problems the purpose is to find trade-off solutions rather than a single solution. Thus to compare those solutions and determine which are useful, the well-known Pareto dominance is commonly used. Based on the work of Vilfredo Pareto (see [25]), it can be expressed as follows:

$$\vec{x}_i \succ \vec{x}_j \leftrightarrow \begin{cases} \forall k \in [1,N], f_k(\vec{x_i}) \leq f_k(\vec{x_j}) \\ \exists k' \in [1,N] \mid f_{k'}(\vec{x_i}) < f_{k'}(\vec{x_j}) \end{cases} \tag{6}$$

In accordance with [9], this expression means that a given decision vector $\vec{x}_i$ dominates another one $\vec{x}_j$ if, and only if none of the corresponding objective function values $f_k(\vec{x_i})$ is worse than $f_k(\vec{x_j})$ and if there is a dimension in the criterion space where it is strictly better. Using such a definition, the Pareto optimal set P_ can be defined as the set of all non dominated vectors (see [29]).

$$P^* \subset \mathbb{R}^n, \vec{x}_i \in P^* \leftrightarrow \nexists\, \vec{x}_j \in \mathbb{R}^n \mid \vec{x}_j \succ \vec{x}_i \tag{7}$$

The set of corresponding objective values in the criterion space constitutes the so-called Pareto front.

The aim of a multi-objective optimization algorithm is to find a good estimation of *P** noted *P* in accordance to some other concepts which can be linked to the problem. As stated in Deb's book [12], the quality of this estimation must be at least measured in terms of diversity of the distribution and spread along the front.

C. PSO core

The PSO is a population based algorithm which deals with swarm intelligence. Each particle in this swarm has a *n* dimensional vector used as a position in the parameter space. At each iteration, particles are moving using some core equations to compute their velocity and decide their movements. The main advantage of PSO is its simple implementation as it can be reduced to the two following equations (see [29]):

$$\begin{aligned} v_{i,t+1} = \quad & \omega.r_0.v_{i,t} + \\ & c_1.r_1.(p_{i,best} - x_{i,t}) + \\ & c_2.r_2.(p_{i,guide} - x_{i,t}) \end{aligned} \tag{8}$$

$$x_{i,t+1} = \quad x_{i,t} + \chi(v_{i,t+1}) \tag{9}$$

$x_{i,t}$ is the position of the i^{th} particle at time *t*, and $v^{i,t}$ its velocity. $p_{i,best}$ and $p_{i,guide}$ are respectively the best position (in term of optimization) that the current particle has found in its path and the position of a particle that has been chosen as a guide. The weights applied to those positions are called the individual and social factors because they respectively depend on the current particle memory of its own best position and on another particle position from the swarm. They are both weighted independently by a coefficient c_x and a random value r_x in [0, 1]. The particles will either tend to explore the parameter space or to further investigate around a previously found solution according to their variations. Thus they have a significant impact on the convergence. ω is the inertia weight which can be constant or time-dependant like in [36]. Large values of this parameter tend to make the particle following its last direction with a turbulence factor r_0 whose value is chosen in [0, 1]. A last part is modeled by the function *X*(). It is generally implemented as a simple factor known as the turbulence factor like in [20] and thus replacing the random part of the inertia weight. However some implementations use it as a velocity normalization function or a constriction factor, keeping the direction but avoiding speed divergence (see [23]).

D. From PSO to MOPSO

Only few modifications need to be made on the core algorithm to adapt it to multi-objective problems. These modifications are presented in algorithm 1. The global PSO algorithm is kept : a loop where particles criteria values are computed then guides selected for each particle and positions updated. The end of the loop relies on stopping criteria which can be simply the number of iteration, the size of archive or based on specific metrics. The main changes are to consider a criterion space of dimensions *N* and to compare the solutions offered by each particle. It increases the algorithm computation cost, but does not change its core. An elitist strategy should be engaged in order to remember only the good parameter combinations and therefore an archive has to be built. It retains only the particle position that can be included in ˆ P, the current Pareto set estimation. In accordance to the cooperative approach in PSO, this system is called the collaborative memory.

Input: S_O a swarm of particles with initilized positions
Output: $\hat{P}_{t_{end}}$ an archive of non-dominated particles
t = 0;
while *stopping criterion not reached* **do**
  Compute objective functions on (S_t);
  foreach *particle* $p_{i,t}$ *in* S_t **do**
    ∗ **if** *current position* $x_{i,t}$ *is best* **then**
      $p_{i,best} = x_{i,t}$;
    end
    ∗ Select new guide $p_{i,guide}$ in archive $(p_{i,t}, \hat{P}_t)$;
    Compute velocity $(p_{i,t})$;
    Test constraints $(p_{i,t})$;
  end
  Update particle position in swarm (S_t);
  ∗ Update Archive $(S_t, \hat{P}_t)$;
  t = t + 1;
end

Algorithm 1: MOPSO pseudo-code with ∗ on features enhanced by our contributions (based on [28]).

Reyes-Sierra proposed a review of state-of-the-art MOPSO variants in [28]. A categorization of the various approaches is presented. It allows to point out that despite the youth of this research field, the variants of MOPSO proposed are very diversified. The most discriminative aspect is the strategy used to manage the multidimensionality of the solution. The simplest technique is to refine the problem through a single objective using aggregation methods (such as a weighted summarization) or to apply an ordering strategy on the different objectives. Sub-population approaches use multiple swarms, optimizing separately each objective but sharing information to propose a global set of solutions. However, as presented in the bibliography, a consensus seems to be established on Pareto dominance based approaches (or combination of approaches) which appear to have better performance (see [28] for a complete description of the MOPSO variants and references).

The study of existing MOPSO variants also allows to point out that dominance and guide selection strategy have a significant impact on the algorithm performance. Thus, our contributions, described in the next sections, are mainly focused on them.

III. An enhanced epsilon dominance and guide selection

In accordance to [28], the major difficulties in the adaptation of PSO to multi objectives problems are : (i) the guide selection (called the leader in the paper), (ii) the maintenance of the non-dominated solutions and (iii) the diversity of the swarm. Our contributions, described in the next sections, are mainly focused on the two first of them. Our proposal can be described as a Pareto dominance based one, using an external archive of non-dominated solutions and a density estimator to select the guide. Indeed, we propose a new guide selection strategy and a variation of the domination concept to ease the archive maintenance. The steps of theMOPSO algorithm impacted by such contributions are highlighted with stars in the algorithm 1.

A. Building the archive

As mentioned before, an archive of solutions eligible for the Pareto set has to be maintained. In order to determine if a particle should be included in the archive, the most common method has been to retain all non-dominated solutions in accordance to the Pareto dominance. The drawback of such an approach is the control of the archive size, which can quickly become very large and hard to maintain, whereas only some key values are needed to obtain a good Pareto Set description. Thus other strategies have to be found to limit the archive size while preserving its diversity and spread along the front.

The $\in$ dominance introduced in [17] and evaluated in [19] presents good capabilities to tackle this problem. Two definitions exist based on the deviation type: absolute (additive $\in$ see equation 10 from [17]) or relative (multiplicative $\in$ see equation 11 from [18]). According to previous studies, the relative definition is commonly chosen as it permits to easily define the $\in$ value and provides more results for smaller objective values.

$$\vec{x}_i \succ \vec{x}_j \quad \leftrightarrow \quad \begin{cases} \forall k \in [1,N], f_k(\vec{x_i}) + \varepsilon \leq f_k(\vec{x_j}) \\ \exists k' \in [1,N] \mid f_{k'}(\vec{x_i}) + \varepsilon < f_{k'}(\vec{x_j}) \end{cases} \tag{10}$$

$$\vec{x}_i \succ \vec{x}_j \quad \leftrightarrow \quad \begin{cases} \forall k \in [1,N], \frac{f_k(\vec{x_i})}{1+\varepsilon} \leq f_k(\vec{x_j}) \\ \exists k' \in [1,N] \mid \frac{f_{k'}(\vec{x_i})}{1+\varepsilon} < f_{k'}(\vec{x_j}) \end{cases} \tag{11}$$

The difference with the classic Pareto dominance can clearly be focused on the figure 1. The first illustration (a) shows the domination area induced by the Pareto dominance for the current particle (in white) on a problem limited to 2 criteria. Other particles are respectively in black, gray or hatched when their are dominated, equivalent or when they dominate the current particle. The illustration (b) shows the $\in$ domination area. It is bigger and allows to dominate elements too much near from the current particle (illustration (c) will be described later). As noticed in [18], this definition allows to quickly achieve an estimation of the Pareto front by modifying the domination area of a particle proportionally to its criterion values. It is one way to manage simultaneously the dominance between particles and the neighborhood in the criterion space and will yield a better diversity along the Pareto front.

However, with such a definition, the difference with the Pareto domination area is larger for particle with bigger objectives values. This could induce a drawback as shown in figure 2 on a benchmark problem, where the domination area of the considered element ($p1$) limits the front description. Particular shapes of the Pareto front estimation (for instance areas with only minor variations on one objective and large variations on another) can thus be

mistaken. This is a consequence of the ∈ dominance definition, which limits the number of particles used to describe the extremes or the parts of the front where one of the criteria is almost constant.

Such a problem was noticed in [18], but surprisingly, no work exists in the literrature about the study of the effects on Pareto front results and no solution has been proposed to avoid this. In order to tackle this problem without involving a CPU greedy clustering method, we introduce an " dominance variant. It limits the domination area introduced by the standard ∈ dominance to local neighborhood in order to avoid the limitations on large criteria value. The figure 1 (c) presents a schematic illustration of this variant in comparison with Pareto and ∈ dominance. One can see that the classic ∈ dominance allows to handle the neighborhood of the considered particle (white one) in the objective space by extending the domination area. Thus closest solutions, which reduce the diversity of the solutions set, are removed. However, it also removes some other particles not present in the local neighborhood because of the global extension of the domination area. Using the ∈ dominance variant allows to limit such extension, keeping its benefits and avoiding the highlighted drawbacks.

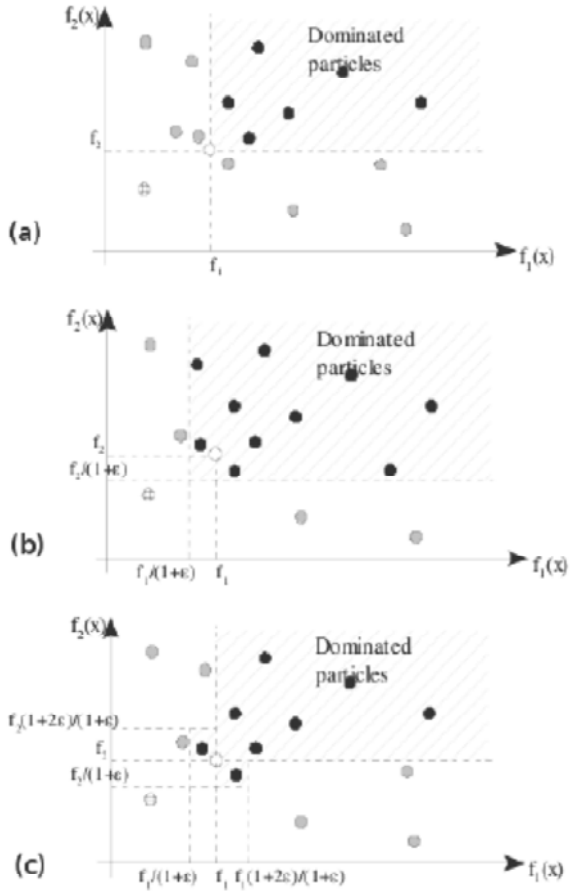


Figure 1 : Illustration of Pareto dominance (a), ∈ dominance (b) and our ∈ dominance variant (c).

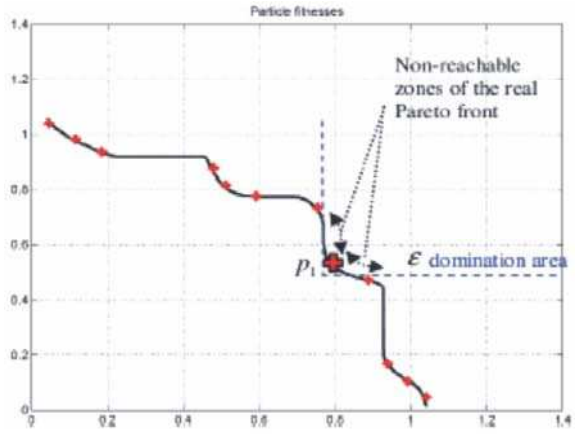


Figure 2 : Example of limitations introduced by ∈ dominance against an estimation of the Pareto front (black line) on TNK problem. The highlighted zone will never be covered by new elements as they are under the ∈ domination area of already present elements (red crosses).

The principle of this variant is to use the implicit neighborhood management introduced by the ∈ dominance. The dominated neighborhood is proportional to ∈ (i.e. multiplicative∈) which is easy to implement and define. The mathematical formalization of such a variant is expressed in equation 12. The first part is simply the Pareto dominance whereas the second part defines the local domination areas in the neighborhood.

$$\vec{x}_i \succ \vec{x}_j \leftrightarrow \begin{cases} \begin{cases} \forall k \in [1,N], f_k(\vec{x_i}) \leq f_k(\vec{x_j}) \\ \exists k' \in [1,N] \mid f_{k'}(\vec{x_i}) < f_{k'}(\vec{x_j}) \end{cases} \\ OR \\ \begin{cases} \exists k' \in [1,N] \mid \\ f_{k'}(\vec{x_j}) < f_{k'}(\vec{x_i}) < \frac{1+2\varepsilon}{1+\varepsilon} f_{k'}(\vec{x_j}) \\ \forall k \in [1,N], \frac{f_k(\vec{x_i})}{1+\varepsilon} \leq f_k(\vec{x_j}) \end{cases} \end{cases} \quad (12)$$

This variant of ∈ dominance allows to overcome the problem mentionned above while maintaining the benefits of classical ∈ dominance. It keeps a good diversity while avoiding the maintenance of a complex data structure for the non-dominated particles induced by methods based on clustering. Such criterion space clustering approaches have been largely tested in [10] with the hypercube strategy, in [21] with the sigma method or in [14] with the dominated trees. The advantages of our variant will be highlighted in the experimentations presented in section IV.

As it is presented in the papers mentioned above, the maintenance of the archive of the non-dominated particles is strongly linked to the guide selection which is one of the core step of the MOPSO. Thus we also contribute on the guide selection behavior.

B. Guide selection behavior

Performance of PSO algorithm depends on the factors

which will influence each particles movement through the core equation 8. The particle will be influenced by its previous position, which is regulated through the inertia factor, its personal memory $p_{i,best}$ and a guide $p_{i,guide}$. Between the numerous possible implementations of personal memory influence, we choose to select the last non-dominated position of the particle to be the individual memorization of its best position. [5] has shown that more complex strategies can provide small improvements, but this approach (called newest strategy in [5]) allows good performance with a very small computational cost.

Then the most important factor is the global guide who will try to help the particle to find to the Pareto front by modifying its trajectory. According to [28], the guide has to be selected in the archive of non-dominated solutions. Nevertheless the selection heuristic can drastically change the swarm convergence behavior.

Our approach is based on the use of a probabilistic framework since it has shown to have better performance in [1]. The idea is to select each particle guide through a roulette wheel selection where each non-dominated solution will have a different selection probability evaluated at each iteration. However, instead of using a computation based on the Pareto domination to determine the probability, we use a local density evaluation in order to tend the swarm to fill the holes of the current Pareto front estimation. Thus for each archive member, the probability is computed as an inverted density measure on its local neighborhood in the criterion space. Such an approach has also been tested in [5] for local best selection with quite good results. A similar approach can also be found in [2] but unfortunately without any further detail on the chosen estimator. However, the choice of the density measurement is not trivial because some particular shapes of the Pareto front or specific constraints can introduce discontinuities. A classic density measure, based on the counting of particles in a fixed area around the current archive element, will be biased by configurations similar to figure 2 : the area could be almost empty because of the front discontinuity. We propose a simple and intuitive solution which provides density estimation on an adaptive local neighborhood. It computes the sum of the inverted distances between the current particle and its K nearest neighbors. Then the selection probability is computed by inverting this estimation and normalizing it as a probability as shown hereafter (where Ψ is the set of the K nearest neighbors of the current particle in the criterion space). This probability needs indeed to be computed again at each archive update.

$$D(x_i) = \sum_{x_j \in \psi} (x_i - x_j)$$

$$then \qquad P(x_i) = \frac{D(x_i)}{\sum_{x_k \in \psi} D(x_i)} \tag{13}$$

According to equation 13, a particle with closest neighbors will have an important local density evaluation and thus a small selection probability.

The last problem to solve is the choice of a decision rule for changing the guide of a particle. Indeed the guide selection strategy has a computational cost. Moreover if the particles change their guide too often (at each iteration) their movements cannot be really influenced by their guide and the social effect can be lost. In mono-objective optimization, this behavior is not a problem because the new guide should always be better than the previous one. However in MOPSO, guides are equivalent since they're all included in *P*. This problem is partly solved by using complex swarm clustering (for example by sub-swarming on each criterion, see [28]), but we propose a more simple technique: enabling a particle guide memorization. Indeed, we did not find any studies on a guide memorization influence. Thereby the guide selection step, highlighted by a star in Algorithm 1, is modified. This is described in Algorithm 2.

Input: a particle p and
$\hat{P}_t$ the current archived of non-dominated particles
Output: a particle p_{guide} out of $\hat{P}_t$
if *p current's position is best position* **then**
  no guide selection;
else
  r = random([0,1]);
  if $r >$ *p's memory threshold*
  AND current guide in $\hat{P}_t$ **then**
    keep current guide;
  else
    select guide($\hat{P}_t$);
  end
end

Algorithm 2: Enhanced guide selection pseudo-code with memory threshold.

The idea is to allow a particle p to keep its previous guide in particular case. To avoid the swarm to only explore locally the front because of the stronger influence of guides, a particle which has been recently added in the archive (which means, when it reaches a non-dominated position) does not select any guide. This is the reason of the first test in the algorithm. In such case, p can be considered to be a pioneer and it is assumed that it does not need any guide. It is completely free to explore any part of the parameter space using only its personal best position and its inertia. In the other case, the particle uses a new characteristic added to the swarm: a guide memory threshold which will define a global behavior of guide memorization. A new guide will be selected for this particle only if its threshold is exceeded as shown hereafter (i.e. the particle remembers its guide) and if its previous guide has not been deleted from archive.

The main advantage of this implementation is that the memorization is under control with the threshold. Experimentations have been conducted on the standard problems in order to select a good trade-off for this new

parameter. The obtained results are presented in the following section.

IV. Evaluation on standards problems

In this section, benchmark problems are used in order to validate our approach against a baseline MOPSO with basic implementation.

A. Evaluation strategies

1) Algorithm setting

As explained in [26] and theoretically studied in [35], the numerous parameters of a PSO algorithm can be adapted to maximize the convergence on each problem. However our experimental approach was to select values which present a good trade-off in order to have a problem-free implementation. As the aim was to study the performance of our contributions concerning dominance and guide selection, there was no need for fine tuning of these parameters. Thus they have been uniformly chosen in controled domains which best fit the state of the art advices (see [26] and [28]):

- Inertia weight !r0 in [0.8; 1.0]
- Individual cognitive factor c1r1 in [1.6; 1.8]
- Social cognitive factor c2r2 in [1.4; 1.6]
- The constriction function _() implemented as a velocity threshold: when a dimension of the velocity vector exceeds the threshold, the whole vector is normalized such as the global direction is kept. Thus it constricts the velocity when it has a dimension greater than 0.1 (with criteria values normalized in [0; 1]).

This approach can be linked to [27]. However we limit the scales for the social and individual cognitive factors to different values since it has shown a statistically significant improvement in mono-objective PSO (see [35]) and in our multi-objective studies. We chose to introduce the uniform randomization through the specified domain instead of using secondary random factor r_x in order to control their variability. The swarm size was limited to 40 elements in order to offer a good trade-off between the number of potential solutions at each iteration and the update rate of the swarm. The number of iterations is not fixed and depends on the problems. For performance comparisons on the experiments, our stopping criteria was a limitation on the number of objective function evaluations, empirically fixed in order to obtain an acceptable estimation of the Pareto front.

2) Benchmark problems

Four problems from the literature have been chosen for the experiments. The first one is BNH, or also called MOPC1 (see [3]). It is considered to be simple because constraints do not introduce serious difficulties in finding the Pareto set and the front does not have any discontinuity or complex convexity. The MOP5, proposed by Viennete, and MOP6

Table 1 : Benchmark functions (f()) and constraints (g()). Name Criteria/constraints

Name	Criteria/constraints
BNH	$f_1(\vec{x}) = 4x_1^2 + 4x_2^2$ $f_2(\vec{x}) = (x_1 - 5)^2 + (x_2 - 5)^2$ $g_1(\vec{x}) \equiv (x_1 - 5)^2 + x_2^2 \leq 25$ $g_2(\vec{x}) \equiv (x_1 - 8)^2 + (x_2 + 3)^2 \geq 7$ $x_1 \in [0, 5] \quad x_2 \in [0, 3]$
MOP5	$f_1(\vec{x}) = \frac{x_1^2 + x_2^2}{2} + \sin\left(x_1^2 + x_2^2\right)$ $f_2(\vec{x}) = \frac{(3x_1 - 2x_2 + 4)^2}{8} + \frac{(x_1 - x_2 + 1)^2}{27} + 15$ $f_3(\vec{x}) = \frac{1}{x_1^2 + x_2^2 + 1} - 1.1e^{-x_1^2 - x_2^2}$ $x_1 \in [-30, 30] \quad x_2 \in [-30, 30]$
MOP6	$f_1(\vec{x}) = x_1$ $f_2(\vec{x}) = (1 + 10x_2)\left[1 - \left(\frac{x_1}{1+10x_2}\right)^2 - \frac{x_1 \sin(8\pi x_1)}{1+10x_2}\right]$ $x_1 \in [0, 1] \quad x_2 \in [0, 1]$
TNK	$f_1(\vec{x}) = x_1$ $f_2(\vec{x}) = x_2$ $g_1(\vec{x}) \equiv x_1^2 + x_2^2 - 1 - 0.1\cos\left(16\arctan\left(\frac{x_1}{x_2}\right)\right) \geq 0$ $g_2(\vec{x}) \equiv (x_1 - 0.5)^2 + (x_2 - 0.5)^2 \geq 0.5$ $x_1 \in [0, \pi] \quad x_2 \in [0, \pi]$

(see [6] for complete references) are two unconstrained problems used to test optimization algorithms against two major difficulties: an increase of the criterion number and a discontinued Pareto front. Then the last problem, called TNK by Tanaka [33], is considered to be quite difficult because of the restriction of the solution space introduced by the constraints. The descriptions of the mathematical functions, as they have been implemented, are shown in table 1.

3) Metrics

Comparing different executions of two multi-objective algorithms is a very complicated task. However, in our case, we only need to compare different variants of the same algorithm. Thus we use only simple metrics to compare the spread and diversity of the front obtained by each implementation.

The spacing metric S (see [30]) measures the homogeneity of the front description by computing the mean distance between each element of the Pareto set estimation. Thus small values are better than large ones. A null value means that the elements are equidistant. This limit cannot be reached with the relative implementation of the $\in$ dominance because of its intrinsic definition which introduces a neighborhood limitation relative to the criterion value. The maximal extension *D* simply measures the diagonal between the extremes elements on each criterion and must be maximized in order to cover the entire front. Then the set coverage *SC* proposed in [37] tries to evaluate the domination of a Pareto front estimation P_A against another one, P_B, by counting the number of elements of P_B which are dominated by a least one element of P_A. By definition if $SC(P_A, P_B) = 1$ and $SC(P_B, P_A) = 0$ we can say that the estimation P_A is better than P_B. They were

respectively computed as presented in equations 14, 15 and 16

Table 2 : Metrics for dominance comparison (left columns results for MOPSO baseline with $\in$ dominance and right with enhanced $\in$ dominance).

	BNH		MOP5		MOP6		TNK	
Objectives evaluations	4000		2000		4000		4000	
Archive size	17.6	**58.4**	66.1	47.3	9.68	**23.9**	10.4	**26.8**
Spacing metric	3.55	**3.2**	0.04	43.4	0.19	**0.12**	0.09	**0.03**
Maximal extension	99.7	**105**	1.55	**65.8**	0.97	**1.05**	1.30	1.30
Set coverage	0.98	0.96	0.12	1	0.92	0.9	0.97	0.95

with normalized objective values.

$$S = \sqrt{\frac{1}{|\hat{P}|}\sum_{i=A}^{|\hat{P}|}\left(d_i - \overline{d}\right)^2} \tag{14}$$

$$with \quad d_i = \min_{\vec{x_i}\in\hat{P}\wedge k\neq i}\sum_{n=1}^{N} \mid f_n(\vec{x_i}) - f_n(\vec{x_k}) \mid$$

$$D = \sqrt{\sum_{n=1}^{N}\left(\max_{\vec{x_i}\in\hat{P}} f_n(\vec{x_i}) - \min_{\vec{x_k}\in\hat{P}} f_n(\vec{x_k})\right)^2} \tag{15}$$

$$SC(\hat{P}_A, \hat{P}_B) = \frac{\mid \left\{x \in \hat{P}_B \mid \exists y \in \hat{P}_A : y \prec x\right\} \mid}{\mid \hat{P}_B \mid} \tag{16}$$

As the algorithm involves random values in its execution, many differences can appear in two different runs. Thus in our experimental protocol, the different configurations of MOPSO used the same initial swarm with random position vectors assigned in the parameter space. Then we repeat 100 times the execution (with different initial swarms) of each implementation of the algorithm. Our aims were to obtain a good estimation of the general algorithm behavior and to enable statistical estimators computation for each metric at each iteration.

The computional cost involved by the enhancement of neighborhood and guide selection was evaluated both on benchmark and real life problems. It appears that the most critical point was the objective computations and that the computational overload in comparison to the baseline was not significative. Thus it has not been studied in the following results.

B. Results and discussion

1) Dominance

We compare the $\in$ dominance variant to the $\in$ dominance classically used in MOPSO on the benchmark problems. Table 2 presents the metric mean values over all executions of our approach (in the right columns and bolded when there is some improvement) against standard $\in$ dominance approach (in the left columns). As the set coverage is a non-symmetric binary measure, we present both the results of our approach against the standard and the standard against our variant.

The results summarized in table 2 must be carefully interpreted. First of all we can see that MOP5 is a problem that highlights the standard $\in$ dominance drawbacks. Since one of the objectives has small variability, the front is extended on very high values. The limitation introduced by the standard $\in$ dominance does not allow to describe those parts and thus the final estimation is very different (and worst) than the one obtained with our variant. Closely considering the set coverage allows a better understanding of the situation: the dissymmetry on the metric implies that all the elements from the Pareto front estimated with our dominance variant dominates the ones from the other approach estimation.

The consequence of this is the large differences on the other metrics: the maximal extension is clearly improved and the spacing metric values are not comparable since the objective values are too different. So on this particular problem, our variant allows to perform a better (or faster) estimation of the Pareto front.

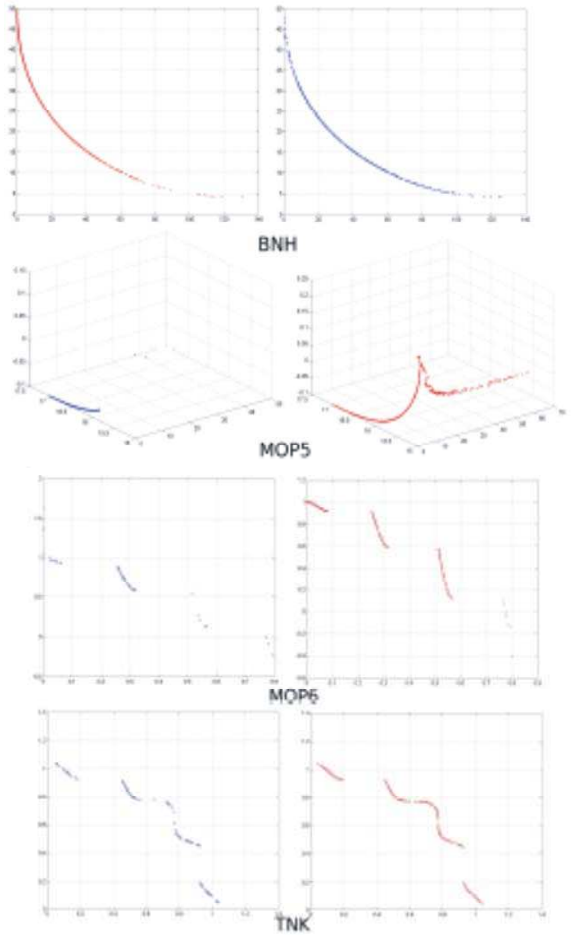


Figure 3 : Dominance comparison on the benchmark problems (the left blue front is for the standard " dominance and the right red one for our variant).

For the other problems, one can observe that the set coverage metrics of both approaches are quite similar and thus we can conclude that the Pareto front estimations are both near the real Pareto front (or near the limit of the algorithm capacities for the number of iterations). As the archive size is always significantly improved by our approach, we can argue that it generally permits to obtain a finer description of the front. This is confirmed by the spacing metric which is also improved and proves that the results are well distributed along the front. Finally, we provide the maximal extension in a specific way in order to allow a better interpretation. The evaluation has been made not on the final front estimation on each runs but on the filtered front. It means that the archive obtained with one approach is reduced by removing all the elements that are dominated by at least one element from the other approach archive. We choose this method because some front estimations contain incorrect elements which corrupt the maximal extension value. The results show that if our approach appears to yield less satisfactory results at first, it is only due to the presence of dominated solutions in the other estimation. Thus its maximal extension artificially grows because of such false Pareto front estimation. This particular difficulty on the metric interpretation highlights the difficulty of quantitative comparison.

Table 3 : Metrics for guide selection behavior comparison (left columns results for MOPSO baseline with random guide selection and right with enhanced guide selection).

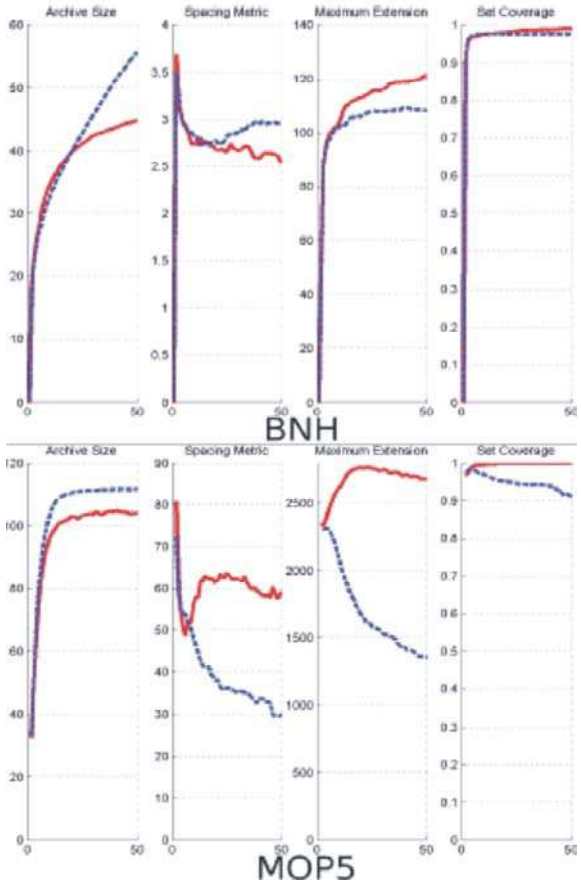
	BNH		MOP5		MOP6		TNK	
Objectives evaluations	4000		2000		4000		4000	
Archive size	55.6	44.8	111	104	18.7	**21.5**	27.5	**27.8**
Spacing metric	2.95	**2.55**	29.9	58.9	0.24	**0.13**	0.03	**0.02**
Maximal extension	108	**121**	1357	**1681**	2.59	2.08	1.29	**1.30**
Set coverage	0.97	**0.99**	0.91	**0.99**	0.95	**0.96**	0.97	0.97

A more thorough comparison requires a qualitative observation of the estimated Pareto front. As seen in Figure 3, the quality of the front is clearly enhanced with our variant: the extremes are better described and the description of parts where a criterion is almost invariant is also enhanced. This is highlighted on MOP5, where the classic $\in$ dominance does not allow describing the right part of the front because of the particular shape of the Pareto front.

It is obvious that the classic $\in$ dominance can also tackle those problems by reducing the epsilon value and allow more elements to be included in the archive. But other parts of the front which are well described will also suffer from this by more and more elements inclusion and thus the archive size bounds can be quickly broken. Moreover, it will not resolve the problem involved in 'flat' parts of the front as our approach can do.

2) Guide selection strategy

Both configurations in this study use the proposed enhanced $\in$ dominance. Their differences are only on the guide management: the first uses a full random selection and no guide memorization whereas the other involves the density based probability to select the guide that can be kept through the next iteration. The number of neighbors was experimentally limited to 4 and the memory factor to 0.6 as it appears to be the most effective values in our experiments (not presented here). Figure 4 shows the evolution of the different metrics through the iterations on each problem. Table 3 presents the mean improvements over all executions of our approach (right columns) against random selection (left columns). BNH: The improvement is not obvious on BNH tests. Such a result is quite logical since the objective functions are quite simple and do not need a strong strategy to allow a good estimation of the Pareto set. Improvements of the front diversity can be seen but through a reduction of archive size.



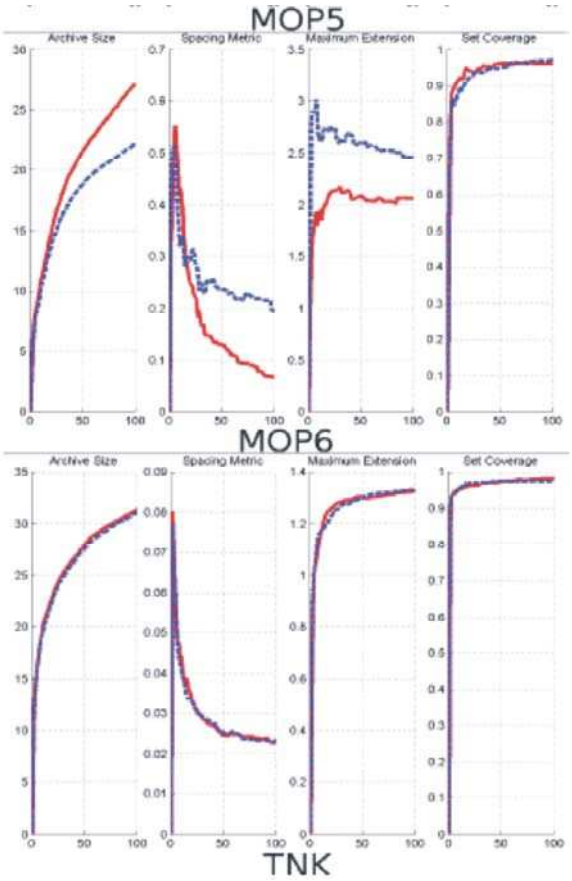


Figure 4 : Evolution of metrics through iteration on different problems (means values for standard guide selection in blue dashed lines and our variant in red lines).

MOP5: The performance of our approach must be well interpreted for this problem. As shown by the dynamic evolution of the metric in figure 4, the results are biased. Indeed after about 20 iterations the values of metrics fall drastically for the random selection. The reason is that the front of this problem is particularly difficult to find as it has a lot of local optimal solutions as explained previously. This is confirmed by the evolution of the set coverage and maximal extension which allow concluding that the front estimated by the probabilistic approach is quite better.

MOP6: The solution is significantly improved by our approach on MOP6 tests. It is quite obvious that this particular problem, which contains much discontinuities on its Pareto front, is better solved by our enhanced guide selection behavior. The only exception is the maximal extension. The reasons are the same as in the precedent study on dominance. TNK: The problem involves a lot of hard constraints which strongly limit the parameter space. Thus our approach based on a density estimator evaluated in the criterion space does not improve the global results since it does not permit to tackle the specific difficulties introduced in this problem.

Such results can be difficult to analyze since some behavioral particularities are kept undetected even when using several metrics. Thus, we interpret the values as relative improvement in order to facilitate the analysis on each problem. The classical qualitative evaluation of the Pareto front has also led us these interpretations. With respect to all the measures, we can conclude that our approach obtained a significant improvement in most cases. As we saw, the higher improvement is reached with difficult problems (i.e. with discontinued front) without strong constraints. However such results are limited to the context of our experiments, which is the comparison between different MOPSO approaches on standard problems. Thus we have also tested our MOPSO in a real world environment against an evolutionary algorithm.

V. SVM model selection using the proposed MOPSO

This section proposes an original application of the proposed MOPSO for tuning the hyper parameters of a classifier. Such a problem is a critical step for building an efficient classification system as this crucial aspect of model selection strongly impacts the performance of a classification system. For a long time, this problem has been tackled using a mono objective optimization process, with the predictive accuracy or error rate as objective. Now, it is well-known that a single criterion is not always a good performance indicator. Indeed, in many real-world problems (medical domain, road safety, biometry, etc...), the miss classification costs are (i) asymmetric as error consequences are class-dependant ; (ii) difficult to estimate, for instance when the classification process is embedded in a more complex system. In such cases, a single criterion might be a poor indicator. Since the works of Bradley [4] concerning the Receiver Operating Characteristics (ROC) curve, classifier model selection has been implicitly considered to be a multi-objective optimization problem, particularly in the context of a two-class classification problem. Indeed, a classifier ROC curve represents the set of trade-offs between False Rejection (FR) and False Acceptance (FA) rates (also known as sensitivity vs. specificity trade-off). As a consequence, some approaches have been proposed in order to choose the classifier hyper parameters using the ROC curve as a performance indicator. Unfortunately, these approaches are always based on a reduction of the FR and FA rates into a single criterion such as the Area Under Curve (AUC) or the FMeasure (FM).

In this section, classifier hyper parameters tuning is explicitly considered to be a multi-objective optimization problem aiming at optimizing simultaneously FA and FR. It is tackled using the proposed MOPSO optimizer. Consequently, the aim is to use the proposed MOPSO to find a set of classifiers in order to select the best set of FA/FR trade-offs. Such a strategy is evaluated on data extracted from a real-world application which takes place in the context of a handwritten digit/outlier discrimination problem.

On can note that some other combinations of SVM classifier and particle swarm optimization (limited to mono-objective optimization) can be found in the literature with different approaches. Two examples can be found in [32] and [24]. In the first one, the PSO is used to select the characteristics (genes in a tumor classification problem) exploited by the SVM classifier and thus appears as a very efficient preprocessing module in the overall classification system. And in the second one, a Modified PSO called the Converging Linear Particle Swarm Optimizer is proposed to replace the traditional learning algorithm. Tested against baseline algorithms on the handwritten characters database from MNIST, it has shown to have similar capabilities. In both studies, an original combination is proposed and promising results are presented. The following sections will describe our own proposal.

The application is quickly described in subsection V-A, in order to justify our choices. The SVM classifier used and its optimization strategy are described in subsection V-B. Finally, obtained results are presented and discussed in V-C.

A. Digits/outliers discrimination

The work described in this section is part of the design of a more complex system which aims at extracting numerical fields (phone number, zip code, customer code, etc.) from incoming handwritten mail document images. The proposed approach is applied to a particular stage of this numerical field extraction system [7]. More precisely, the classifier to be optimized is used as a fast two-class classifier which has to identify the digits among a huge number of irrelevant shapes (words, letters, fragments of words, etc). Consequently, the classifier objective is to reject as many outliers as possible, while accepting as many digits as possible. However, rejecting a digit has a much more serious consequence than accepting an outlier. The rejected data will never be processed and thus a numerical field can be lost. If a non-digit is accepted, it will increase the computation cost on non-relevant data. This problem is a good example of a classification task with asymmetric and unknown misclassification costs since the influence of a FA or a FR rate on the whole system results is unknown a priori. Concerning the classifier to be optimized, the Support Vector Machines classifier has been chosen for its well-known efficiency in a two-class context.

B. SVM classifier and optimization strategy

Support Vector Machines are a well-founded and largely used learning machine algorithm which have been proved to be very effective on several real-world problems. In order to take into account asymmetric misclassification costs, we adopt the strategy proposed in [22] that consists in the introduction of two distinct penalty parameters C^- and C^+ (also called positive and negative margins).

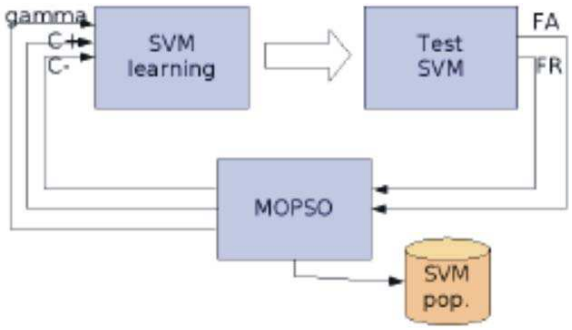


Figure 5 : Schematic view of the SVM optimization strategy through MOPSO.

In such a case, given a set of m training examples x_i belonging to the class y_i, the classical maximization of the dual Lagrangian with respect to the αi becomes: max

$$\max_{\alpha}\left(\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j \kappa(x_i,x_j)\right) \quad (17)$$

subject to the constraints :

$$\begin{cases} 0\le\alpha_i\le C^+, for\ y_i=1 \\ 0\le\alpha_i\le C^+, for\ y_i=1 \\ \sum_{i=1}^{m}\alpha_i y_i=0 \end{cases}$$

Where α_i denotes the Lagrange multipliers, C^- and C^+ are respectively the cost factors for the two classes (−1) and (+1), and k (x_i, x_j) denotes the kernel transformation. In the classical case of a Gaussian (RBF) kernel, k (x_i, x_j) is defined as:

$$k\ (x_i, x_j) = e^{-\gamma\times\|xi-xj\|2} \quad (18)$$

In accordance with [8], we choose to keep the intrinsic optimization of support vector in SVM using the Lagrangian maximization and we apply the optimization process to the classifier hyper-parameters. Hence, our optimization parameters are:

- the kernel parameter of the SVM-rbf : γ
- the penalty parameters introduced above: C^- and C^+.

As explained before, the criteria to be optimized are both the FA rate and the FR rate which are obtained by testing the hyperparameters set on a test database. The proposed strategy is illustrated on figure 5.

C. MOPSO on SVM experimentation and comparison

In this section, the experimental results obtained using the approach shown on figure 5 are presented and discussed. Two kinds of tests are presented. The first one aims at showing the interests of our MOPSO improvements. The second one consists in a comparison of the proposed MOPSO with respectively a state of the art multi-objective algorithm (NSGA-II [11]) and a classic SVM model selection approach.

Our first comparison has been made against a baseline

MOPSO (standard $\in$ dominance and random guide selection) in order to ensure that our contributions concerning MOPSO are efficient on a real world problem. The comparative results are presented on figure 6. As one can see, the problem does not appear to be difficult. The Pareto front estimation does not contain any discontinuity. However the gain of our contributions can be clearly observed. The standard MOPSO mainly focusses its search on the middle part of the front and has a poor description of the extremes. The results obtained using our approach are quite better. One can be observed a better homogeneity of the description and well defined extremes parts.

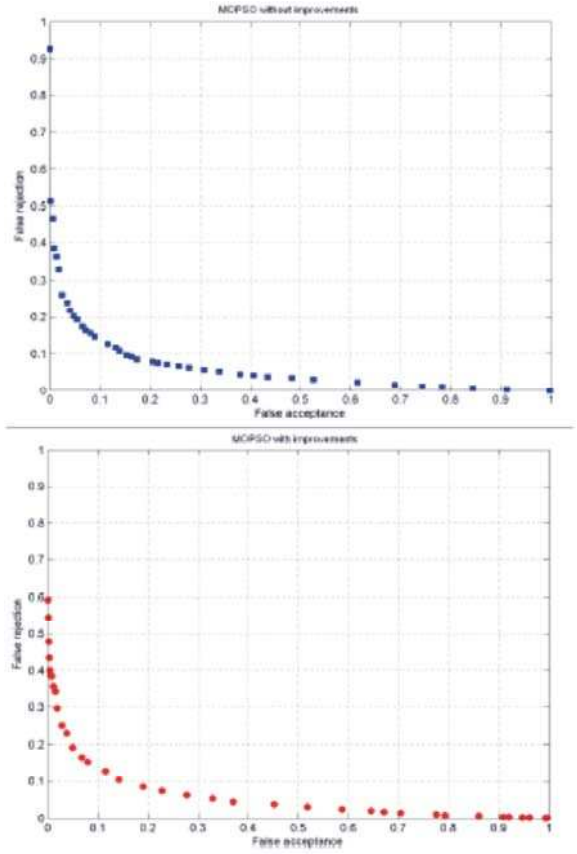


Figure 6 : Final Pareto Front estimation for both baseline MOPSO (up) and enhanced (down) MOPSO.

The second test concerns a comparison between the proposed MOPSO and a state-of-the-art MOEA: the NSGA-II (report to [11] for a complete description). As the approach differs from ours, some adaptations have been needed to offer a fair comparison. The most important parameter is the archive size which is limited to the initial population size in NSGA-II. Thus our MOPSO implementation was modified in order to limit its archive size. Using such a limitation, $\in$ value was dynamically computed with a specific heuristic in order to rebuild the archive. Both algorithms were ran using the same population size (40) for a limited number of objective evaluations (1000). Such values appear as good trade-offs between the running time and the quality of the final Pareto set estimation. The results obtained are shown on figure 7 for the Pareto front estimation and on figure 8 for the metrics previously introduced.

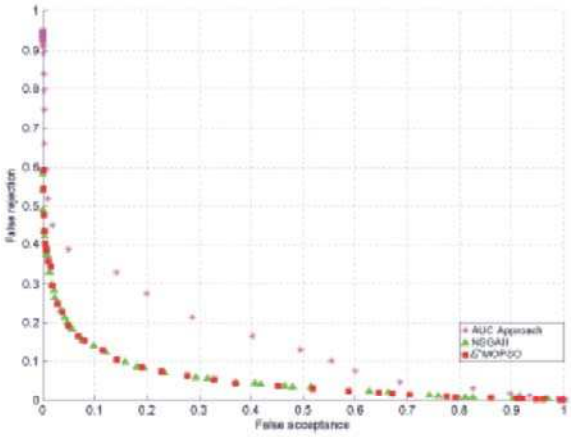


Figure 7 : Final Pareto Front estimation for both approaches (NSAGII in green and enhanced MOPSO in red).

One can note that we also introduce on figure 7 the results obtained using a classical SVM model selection called SVM-perf [34]. This approach has been configured to use the Area Under the ROC curve (AUC) as a single criterion during the classifier learning.

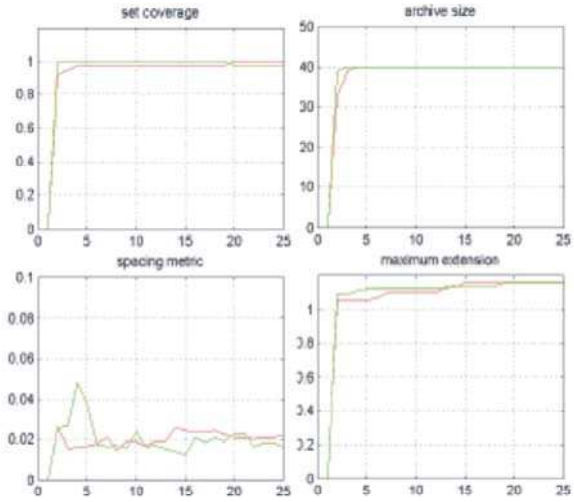


Figure 8 : Comparative values of metrics (NSAGII in green and enhanced MOPSO in red).

One can observe on figure 7 that both MO approaches allow a major improvement of the classic optimization w.r.t. SVMperf approach. Of course, such a comparison is not fair from a theoretical point of view since we compare a ROC curve obtained using a single parameterized classifier (using

AUC as building criterion) with an approach that considers a set of classifiers. Nevertheless, from a practitioner point of view, these results aim at justifying the use of a multi-objective optimization framework in the context of SVM model selection. Indeed, for a chosen FA/FR trade-off, our framework provides a solution to the practitioner which is better than the solution obtained using a single classifier with a given output threshold.

Concerning the comparison of our approach with NSGA-II, the qualitative analysis proposed on figure 7 does not conclude to any dominance between the two multi objective optimizers. The quantitative comparison of metric values confirms this idea. The Figure 8 presents their variations per iteration and shows that both approaches obtain similar values very quickly. Thus the two approaches are quite competitive and perform both well on this problem. Such a result is quite interesting as it shows that our MOPSO implementation can compete with the state-of-the-art MOEA.

VI. Conclusion and further works

This paper introduces two contributions on two intrinsic difficulties faced when adapting the PSO to multi objective optimization: the archive and social guide management. Our variant on $\in$ dominance enables a fast neighborhood management in criterion space and has proved to well maintain the diversity in the archive. Then our guide selection strategy and guide memorization have shown to allow the Pareto front estimation to be enhanced in its difficult parts. The validation of such methods has been made both on standard and real world problems and against a state-of-the-art multi objective optimizer. Our approach appears to be competitive and reliable.

Managing neighborhood, in order to avoid premature convergence and to promote a good spreading of solutions on the Pareto front estimation, is an open problem and several authors have proposed ideas to tackle this problem. This paper proposes an approach which has proven its low computational cost and its performance on a set of problems. A comparison with other proposal remains to be made in a near future.

However, what we tried to prove here was that our implementation allows obtaining a better Pareto set estimation than others using the classic $\in$ dominance. Our proposition on the guide selection allows studying the guide memorization, a topic rarely discussed in other studies. It has shown to allow a significant improvement while keeping the MOPSO performance at the state-of-the-art level on a real world problem. Thus our approach appears as a good improvement to easily handle neighborhood in criterion space.

Much more experiments can then be conducted in order to compare to more MOPSO implementations. But before this, other improvements can be studied to go beyond the ones proposed in this paper. In particular, after proposing a new guide selection strategy, we are looking on the personal best management and selection which is the most natural continuation of our researches. The problem of the extremes handling, which has been partly solved by the neighborhood management, is always present because of the bias introduced by the relative $\in$ dominance. This will also be one of the next big steps of our future work. The management of algorithms parameters also needs to be finer studied and our aim is to reduce the number of algorithm parameters (some successful tests have been conducted on an auto adaptive $\in$). Then, the neighborhood has to be enlarged to the parameter space. It will avoid a guide to be selected when it will add to much turbulence to its movements because its parameters combinaison is too different from the guided particle.

We also want to adapt our experimental approach to a more realistic environment in order to ensure the usability of our particle swarm optimizer. Some experiments will be conducted by considering the kernel choice as a new parameters in the optimization process for SVM model selection. This induces heterogeneity in the parameters but it can be tackled by MOPSO without too many difficulties. This research path is particularly valuable since it really helps the engineers to design their systems which have several heterogeneous parameters. Finally, we plan to enlarge our set of applications in terms of system complexity and domains. Information retrieval systems will be our most promising research paths especially for information extraction tasks through linguistic patterns which involve many parameters.

References

[1] J.E. Alvarez-Benitez, R.M. Everson, and J.E. Fieldsend. Mopso algorithm based exclusively on pareto dominance concepts. Third International Conference on Evolutionary Mutli-Criterion Optimization, pages 726–732, 2005.

[2] Alexandre M. Baltar and Darrell G. Fontane. A generalized multi objective particle swarm optimization solver for spreadsheet models: application to water quality. In AGU Hydrology Days 2006, March 2006.

[3] To Thanh Binh and Urlich Korn. MOBES: A multi objective evolution strategy for constrained optimization problems. In The Third International Conference on Genetic Algorithms (Mendel 97), pages 176–182, Brno, Czech Republic, 1997.

[4] Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern- Recognition, 30:11451159, 1997.

[5] J¨urgen Branke and Sanaz Mostaghim. About selecting the personal best in multi-objective particle swarm optimization. In Parallel Problem Solving from Nature, volume 4193 of Lecture Notes in Computer Science, pages 523–532. Springer, September 2006. ISBN=3- 540-38990-3.

[6] Leticia Cagnina, Susana Esquivel, and Carlos A. Coello Coello. A particle swarm optimizer for multi-objective optimization. Journal of Computer Science & Technology, 5(4), 2005.

[7] Chatelain Cl´ement. Extraction de squences numriques dans des documents manuscrits quelconques. Phd thesis, University of Rouen, December 2006.

[8] Chatelain Cl´ement, Adam S´ebastien, Lecourtier Yves, Heutte Laurent, and Paquet Thierry. Multi-objective optimization for svm model selection. In ICDAR07 – to be published, 2007.

[9] Carlos A. Coello Coello. Evolutionary Multi-Criterion Optimization: First International Conference, volume 1993/2001 of Lecture Notes in Computer Science, chapter A Short Tutorial on Evolutionary Multiobjective Optimization, page 21. Springer Berlin / Heidelberg, emo edition, 2001.

[10] Carlos A. Coello Coello and Maximino Salazar Lechuga. A proposal for multiple objective particle swarm optimization. Computational Intelligence, pages 12–17, May 2002.

[11] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm : Nsgaii. IEEETransactions on Evolutionary Computation, 6:182197, 2002.

[12] Kalyanmony Deb. Multi-Objective Optimization Using Evolutionary Algorithms. John Wiley and Sons, 2001. ISBN 047187339X.

[13] David E. (edward) Goldberg. Genetic algorithms in search, optimization & machine learning. Addison-Wesley Publishing Co. - Reading, Mass, 1989.

[14] J. Fieldsend and S. Singh. A multi-objective algorithm based upon particle swarm optimisation. In The 00 U.K. Workshop on Computational Intelligence, pages 34–44, 2002.

[15] J. Kennedy and R. Eberhart. Particle swarm optimization. Neural Networks, 1995. Proceedings., IEEE International Conference on, 4:1942–1948, 1995.

[16] N. M. Kwok, D. K. Liu, and G. Dissanayake. Evolutionary computing based mobile robot localization. Engineering Applications of Artificial Intelligence, 19(8):857–868, December 2006.

[17] Marco Laumanns, Lothar Thiele, Kalyanmoy Deb, and Eckart Zitzler. Combining convergence and diversity in evolutionary multiobjective optimization. MIT Press in Evolutionary Computation, 10, n3:263–282, 2002.

[18] Sanaz Mostaghim and J¨urgen Teich. The role of "-dominance in multi-objective particle swarm optimization. In Proc. CEC´ 03, the Congress on Evolutionary Computation, volume 3, pages 1764–1771, Canberra, Australia, December 2003.

[19] Sanaz Mostaghim and J¨urgen Teich. Strategies for finding good local guides in multi-objective particle swarm optimization. In Swarm Intelligence Symposium, Indianapolis, USA, April 2003. IEEE service center.

[20] Sanaz Mostaghim and J¨urgen Teich. Covering pareto optimal fronts by subswarms in multi-objective particle swarm optimization. In IEEE Proceedings, World Congress on Computational Intelligence (CEC'04), volume 2, pages 1404–1411, Portland, USA, June 2004.

[21] C. R. Mouser and S. A. Dunn. Comparing genetic algorithms and particle swarm optimisation for an inverse problem exercise. In Rob May and A. J. Roberts, editors, Proc. of 12th Computational Techniques and Applications Conference CTAC-2004, volume 46, pages C89–C101, March 2005.

[22] Osuna, Freund R., and Girosi F. Support vector machines: Training and applications. 1997.

[23] Elpiniki Papageorgiou, Konstantinos Parsopoulos, Chrysostomos Stylios, Petros Groumpos, and Michael Vrahatis. Fuzzy cognitive maps learning using particle swarm optimization. Journal of Intelligent Information Systems, 25(1):95–121, July 2005.

[24] A.P. Paquet, U.; Engelbrecht. Training support vector machines with particle swarms. In Neural Networks, 2003. Proceedings of the International Joint Conference on, volume 2, pages 1593 – 1598, 2003.

[25] Vilfredo Pareto. Cours d'Economie Politique. 1897.

[26] K. E. Parsopoulos and M. N. Vrahatis. Recent approaches to global optimization problems through particle swarm optimization. Natural Computing, 1(2):235–306, June 2002.

[27] Margarita Reyes-Sierra and Carlos A. Coello Coello. Improving pso-based multi-objective optimization using crowding, mutation and epsilon-dominance. In Evolutionary Multi-Criterion Optimization. Third International Conference, volume 3410 of Lecture Notes in Computer Science, pages 505–519. Springer, 2005.

[28] Margarita Reyes-Sierra and Carlos A. Coello Coello. Multi-objective particle swarm optimizers: A survey of the state-of-the-art. International Journal of Computational Intelligence Research (IJCIR), 2:287–308, 2006.

[29] Mara Margarita Reyes-Sierra. Use of Coevolution and Fitness Inheritance for Multi-Objective Particle Swarm Optimization. PhD thesis, Center of Research and Advanced Studies of the National Polytechnic Institute, Mexico City, Mexico, August 25th 2006.

[30] J. R. Schott. Fault tolerant design using single and multi-criteria genetic algorithms. Master's thesis, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, 1995.

[31] Matthew Settles, Brandon Rodebaugh, and Terence

Soule. Comparison of genetic algorithm and particle swarm optimizer when evolving a recurrent neural network. In Springer Berlin / Heidelberg, editor, Genetic and Evolutionary Computation GECCO 2003, volume 2723/2003 of Lecture Notes in Computer Science, pages 148–149, 2003.

[32] Qi Shen, Wei-Min Shi, Wei Kong, and Bao-Xian Ye. A combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification. Talanta, In Press, Corrected Proof, 2006.

[33] M. Tanaka, H. Watanabe, Y. Furukawa, and T. Tanino. GA-based decision support system for multicriteria optimization. In 1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century (Cat. No. 95CH3576-7), volume 2, pages 1556–61, New York, NY, USA, 1995. IEEE.

[34] Joachims Thorsten. A support vector method for multivariate performance measures. In Conference on Machine Learning (ICML), 2005.

[35] F. van den Bergh and A. P. Engelbrecht. A study of particle swarm optimization particle trajectories. Information Sciences, 176(8):937–971, April 2006.

[36] Hong Zhang, C. M. Tam, and Heng Li. Multimode project scheduling based on particle swarm optimization. Computer Aided Civil and Infrastructure Engineering, 21(2):93–103, February 2006.

[37] Eckart Zitzler, Kalyanmoy Deb, and Lothar Thiele. Comparison of multiobjective evolutionary algorithms: Empirical results. Evolutionary Computation, 8(2):173–195, 2000.

Author Biographies

G´erard DUPONT was born in 1982 in Poitiers, France. He received two M.S. degrees in computer engineering and computer science at Rouen University in 2006. Since then, he began a Ph.D. degree in computer science at EADS-DS in Val de Reuil (France) and with the LITIS Laboratory of computer science in Rouen University on implicit feedback learning for semantic information retrieval. His research interests include evolutionary multi objective optimization, swarm intelligence, learning algorithm, information retrieval and semantic.

Sébastien ADAM was born in 1975 in Dieppe, France. He received a PhD in graphical document analysis from the University of Rouen in 2001. This PhD has been led for France Telecom, the historical French telecommunication operator and tackles the problem of multi-oriented and multi-scaled pattern recognition. Then he joined the LITIS labs in Rouen, France. His domains of interest are at the merging of document analysis and multi-objective optimization.

Yves LECOURTIER was born in Marseilles in 1950. After a thesis in signal processing in 1978, and a second thesis in physics (Automatic Control) in 1985 from the University of Paris-Sud, Orsay, France, he joined the University of Rouen as a Professor in 1987. His research domain is in pattern recognition and optimisation, especially for document analysis and text recognition. Pr. Lecourtier is a member of AFRIF, ASTI, IAPR. From 1994 to 2000, he was the chairman of the GRCE, a french society which gather most of the french researchers working in document analysis and text recognition fields.

Bruno GRILHERES joined EADS Information Processing Competence Center in 2002. He has been working on E-democracy and Text Mining. He led the technical architecture activity on IST CyberVote (IST Prize 2006) and Trade Chamber Elections. He has acted as information technology consultant for EADS Defense and Security Global Security and Mission Systems, Airbus. He is currently completing a PhD (to be presented in 2007) on statistical learning for information extraction.

Annexe D

Réference CV : 2

C. Chatelain, S. Adam, Y. Lecourtier, L. Heutte, and T. Paquet. A multi-model selection framework for unknown and/or evolutive misclassification cost problems. Pattern Recognition (PR), 43(3) :815-823, 2010.

Pattern Recognition 43 (2010) 815–823

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A multi-model selection framework for unknown and/or evolutive misclassification cost problems

Clément Chatelain, Sébastien Adam, Yves Lecourtier, Laurent Heutte*, Thierry Paquet

Université de Rouen, LITIS EA 4108, BP12, 76801 Saint Etienne du Rouvray, France

ARTICLE INFO

Article history:
Received 11 January 2008
Received in revised form 24 February 2009
Accepted 5 July 2009

Keywords:
ROC front
Multi-model selection
Multi-objective optimization
ROC curve
Handwritten digit/outlier discrimination

ABSTRACT

In this paper, we tackle the problem of model selection when misclassification costs are unknown and/or may evolve. Unlike traditional approaches based on a scalar optimization, we propose a generic multi-model selection framework based on a multi-objective approach. The idea is to automatically train a pool of classifiers instead of one single classifier, each classifier in the pool optimizing a particular trade-off between the objectives. Within the context of two-class classification problems, we introduce the "ROC front concept" as an alternative to the ROC curve representation. This strategy is applied to the multi-model selection of SVM classifiers using an evolutionary multi-objective optimization algorithm. The comparison with a traditional scalar optimization technique based on an AUC criterion shows promising results on UCI datasets as well as on a real-world classification problem.

1. Introduction

Tuning the hyperparameters of a classifier is a critical step for building an efficient pattern recognition system as this crucial aspect of model selection strongly impacts the generalization performance. In the literature, many contributions in this field have focused on the computation of the model selection criterion, i.e. the value which is optimized with respect to the hyperparameters. These contributions have led to efficient scalar criteria and strategies used to estimate the expected generalization error. One can cite Xi-Alpha bound of [24], the generalized approximate cross-validation of [33], the empirical error estimate of [3], the radius-margin bound of [9] or the maximal-discrepancy of [2]. Based on these criteria, hyperparameters are usually chosen using a grid search, coupled with a cross-validation procedure. In order to decrease the computational cost of grid search, some authors suggest to use gradient-based techniques (e.g. [4,25]). In these works, the performance validation function is adapted in order to be differentiable with respect to the parameters to be optimized.

All the approaches mentioned above, though efficient, use a single criterion as the objective during the optimization process. Now, it is well known that a single criterion is not always a good performance indicator. Indeed, in many real-world pattern recognition problems (medical domain, road safety, biometry, etc.), the misclassification costs are (i) asymmetric as error consequences are class-dependant; (ii) difficult to estimate (for example when the classification process is embedded in a more complex system) or subject to change (for example in the field of fraud detection where the amount of fraud changes monthly). In such cases, a single criterion might be a poor performance indicator.

One solution to tackle this problem is to use as performance indicator the *receiver operating characteristics* (ROC) curve proposed in [6]. Such a curve offers a synthetic representation of the trade-off between the *true positive* (TP) rate and the *false positive* (FP) rate, also known as *sensitivity* vs. *specificity* trade-off. One way to take into account both FP and TP in the model selection process is to resume the ROC curve into a single criterion, such as the F-measure (FM), the break-even point (BEP) or the area under ROC curve (AUC). However, we will show in the following that we can get more advantages in formulating the model selection problem as a true 2-D objective optimization task.

In this paper, our key idea is to turn the problem of the search for a global optimal classifier (i.e. the best set of hyperparameters) using a single criterion or a resume of the ROC curve, into the search for a pool of locally optimal classifiers (i.e. the pool of the best sets of hyperparameters) w.r.t. FP/TP rates. The best classifier among the pool can then be selected according to the needs of some practitioner. Consequently, the proposed framework can be viewed as a multiple model selection approach (rather than a model selection problem) and can naturally be expressed in a multi-objective optimization (MOO) framework. Under particular conditions, we assume that such an approach leads to very interesting results since it enables

* Corresponding author.
E-mail address: Laurent.Heutte@univ-rouen.fr (L. Heutte).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.07.006

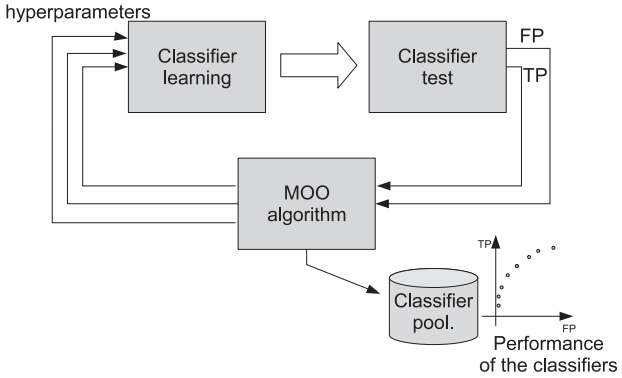


Fig. 1. Multi-model selection framework.

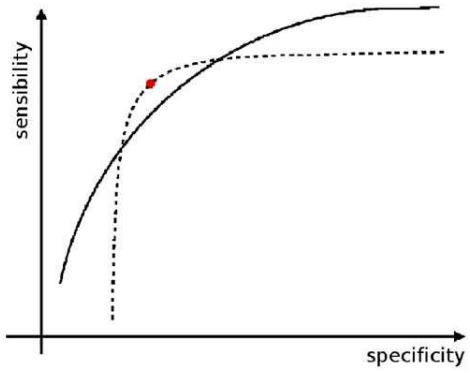


Fig. 2. Comparing ROC curves: the solid ROC curve provides a better AUC than the dashed ROC curve, but is not locally optimal for a given range of specificity (false positive rate).

a practitioner to (i) postpone the choice of the final classifier as late as possible and (ii) to change the classifier without a computationally expensive new learning stage when target conditions change.

Fig. 1 depicts our overall multi-model selection process. The resulting output of such a process is a pool of classifiers, each one optimizing some FP/TP rate trade-off. The set of trade-off values constitutes an optimal front we call "ROC front" by analogy with MOO field.

The remainder of the paper is organized as follows. In Section 2, we detail the rationale behind the ROC front concept and illustrate how our multi-model selection approach may provide solutions that outperform traditional approaches in a MOO framework. Section 3 gives an overview of multi-objective optimization strategies and details the algorithm used in the proposed framework to compute the "ROC front". Section 4 presents a particular application of our approach to the problem of SVM hyperparameter selection and shows that our method enables to reach more interesting trade-offs than traditional model selection techniques on standard benchmarks (UCI datasets). In Section 5, we discuss ways of selecting the best model from the pool of locally optimal models. Then, in order to assess the usefulness of our approach, we present in Section 6 its application on a real world classification problem which consists in a digit/outlier discrimination task embedded in a numerical field extraction system for handwritten incoming mail documents. Finally, a conclusion and future works are drawn in Section 7.

2. The "ROC front" concept

As stated in the Introduction, a model selection problem may be seen from a multi-objective point of view, turning thus into a multi-model selection approach. In the literature, some multi-model selection approaches have been proposed. However, these approaches aim at designing a single classifier and thus cannot be considered as real multi-model selection approaches. Caruana for example proposed in [8] an approach for constructing ensembles of classifiers, but this method aims at combining these classifiers in order to optimize a scalar criterion (accuracy, cross-entropy, mean precision, AUC). Bagging, boosting or error-correcting-output-codes (ECOC) [17] are also classifier ensemble methods that can be viewed as producing single classifiers efficient with respect to a scalar performance metric. In [27], an evolutionary algorithm (EA) based approach is applied to find the best hyperparameters of a set of binary SVM classifiers combined to produce a multi-class classifier.

The approach which is proposed in this paper is different since our aim is not to build a single classifier but a pool of classifiers, each one optimizing both FP and TP rates in the ROC space. In such a context, let us recall that a problem arising when ROC space is used to quantify classifier performance is their comparison in a 2-D objective space: a classifier may be better for one of the objectives (e.g. FP) and worse for the other one (e.g. TP). Consequently, the strict order relation that can be used to compare classifiers when a single objective is only considered becomes unusable and classical mono-objective optimization strategies cannot be applied.

Usually, in ROC space, this problem is tackled using a reduction of the FP and TP rates into a single criterion such as the area under ROC curve (AUC) [30]. However, such performance indicators are a resume of the ROC curve taken as a whole and do not consider the curve from a local point of view. The didactic example proposed in Fig. 2 illustrates this statement. One can see on this figure two synthetic ROC curves. The curve plotted as solid line has a better AUC value, but the corresponding classifier is not better for any specific desired value of FP rate (resp. TP). Consequently, optimizing such a scalar criterion to find the best hyperparameters could lead to solutions that do not fit the practitioner needs in certain context. A better idea could be to optimize simultaneously FP and TP rates using a MOO framework and a dominance relation to compare classifier performance.

Let us recall that the dominance concept has been proposed by Vilfredo Pareto in the 19th century. A decision vector $\vec{u}$ is said to dominate another decision vector $\vec{v}$ if $\vec{u}$ is not worse than $\vec{v}$ for any objective function and if $\vec{u}$ is better than $\vec{v}$ for at least one objective function. This is denoted by $\vec{u} \prec \vec{v}$. More formally, in the case of the minimization of all the objectives, a vector $\vec{u} = (u_1, u_2, \ldots, u_k)$ dominates a vector $\vec{v} = (v_1, v_2, \ldots, v_k)$ if and only if:

$$\forall i \in \{1, \ldots, k\}, \quad u_i \leq v_i \wedge \exists j \in \{1, \ldots, k\} : u_j < v_j$$

Using such a dominance concept, the objective of a multi-objective optimization algorithm is to search for the Pareto optimal set (*POS*), defined as the set of all non-dominated solutions of the problem. Such a set is formally defined as the set:

$$POS = \{\vec{u} \in \vartheta / \neg \exists \vec{v} \in \vartheta, \overrightarrow{f(v)} \prec \overrightarrow{f(u)}\}$$

where ϑ denotes the feasible region (i.e. the parameter space regions where the constraints are satisfied) and $\vec{f}$ denotes the objective function vector. The corresponding values in the objective space constitute the so-called Pareto front.

From our model selection point of view, the *POS* corresponds to the pool of non-dominated classifiers (the pool of the best sets of hyperparameters). In this pool, each classifier optimizes a particular FP/TP trade-off. The resulting set of FP/TP points constitutes an optimal front we call "ROC front". This concept is illustrated with a didactic example as shown in Fig. 3: let us assume that ROC curves have been obtained from three distinct hyperparameter sets. This could lead to the three synthetic curves plotted as dashed lines. One

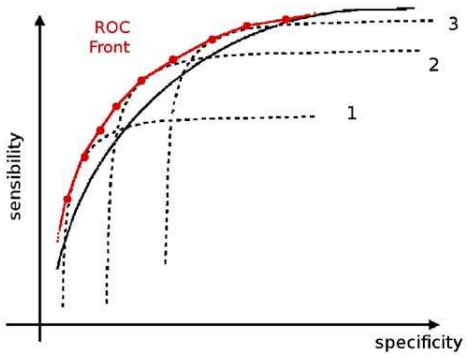


Fig. 3. Illustration of the ROC front concept: the ROC front depicts the FP/TP performance corresponding to the pool of non-dominated operating points.

can see on this example that none of the classifiers dominates the others on the whole range of FP/TP rates. An interesting solution for a practitioner is the "ROC front" (the dotted solid curve), which is made of some non-dominated parts of each classifier ROC curves. The method proposed in this paper aims at finding this "ROC front" (and the corresponding *POS*), using an evolutionary multi-objective optimization (EMOO) algorithm. This class of optimization algorithm has been chosen since evolutionary algorithms (EA) are known to be well-suited to search for multiple Pareto optimal solutions concurrently in a single run, through their implicit parallelism.

In the following section, a brief review of existing EMOO algorithms is proposed and the chosen algorithm is described.

3. Evolutionary multi-objective optimization

As stated earlier, our objective in this paper is to search for a pool of parametrized classifiers corresponding to the optimal set of FP/TP trade-offs. From a multi-objective optimization point of view, this set can naturally be seen as the Pareto optimal set and the set of corresponding FP/TP trade-offs is the ROC front. To tackle such a problem of searching a set of solutions describing the Pareto front, EA are known to be well-suited. This is why we do not consider in our review the approaches that optimize a single objective using the aggregation of different objectives into a single one (e.g. the use of the AUC) or the transformation of some objectives into constraints. For more details concerning these methods, see for example [16].

3.1. Short review of existing approaches

Since the pioneering work of [31] in the mid eighties, a considerable amount of EMOO approaches have been proposed (MOGA from [21], NSGA from [32], NPGA from [23], SPEA from [37], NSGA II from [15], PESA from [12], SPEA2 [36]). In a study reported in [26] the performance of the three most popular algorithms (SPEA2, PESA and NSGA-II) are compared. These three approaches are elitist, i.e. they all use a history archive that records all the non-dominated solutions previously found in order to ensure the preservation of good solutions. This comparative study has been performed on different test problems using as quality measurement the two important criteria of an EMOO, i.e. the closeness to the Pareto front and the solution distribution in the objective space. Indeed, achieving a good spread and a good diversity of solutions on the obtained front is important to give the user as many choices as possible. The results obtained in [26] (which are corroborated in [36,7]) showed that none of the proposed algorithms "dominate" the others in the Pareto sense. SPEA2 and NSGA-II perform equally well in convergence and diversity maintenance. Their convergence through the real Pareto optimal set is inferior to that of PESA but diversity among solutions is better maintained. The study also showed that NSGA-II is faster than SPEA2, because of the expensive clustering of solutions in SPEA2.

In the context of multi-model selection, computation of the objective values is often very time consuming since it involves learning and testing the classifier for each hyperparameter set. Moreover, a good diversity of solutions is necessary since there is no *a priori* information concerning the adequate operating point on the Pareto front. That is why we have chosen to use NSGA-II in the context of our study. We give in the next subsection a concise description of this algorithm. For more details, we refer to [15].

3.2. NSGA-II

NSGA II is a modified version of a previously proposed algorithm called NSGA [32]. It is a population-based, fast, elitist and parameter free approach that uses an explicit diversity preserving mechanism.

Algorithm 1. NSGA-II algorithm.

$P_0 \leftarrow$ pop-init()
$Q_0 \leftarrow$ make-new-pop (P_0)
$t \leftarrow 0$
while $t < M$ **do**
  $R_t \leftarrow P_t \cup Q_t$
  $\mathscr{F} \leftarrow$ non-dominated-sort(R_t)
  $P_{t+1} \leftarrow \emptyset$
  $i \leftarrow 0$
  while $|P_{t+1}| + |\mathscr{F}_i| \leq N$ **do**
    $P_{t+1} \leftarrow P_{t+1} \cup \mathscr{F}_i$
    crowding-distance-assignment($\mathscr{F}_i$)
    $i \leftarrow i + 1$
  end while
  Sort ($\mathscr{F}_i, \prec_n$)
  $P_{t+1} \leftarrow P_{t+1} \cup \mathscr{F}_i[1 : (N - |P_{t+1}|)]$
  $Q_{t+1} \leftarrow$ make-new-pop (P_{t+1})
  $t \leftarrow t + 1$
end while

As one can see in Algorithm 1, the approach starts with the random creation of a parent population P_0 of N solutions (individuals). This population is used to create an offspring population Q_0. For this step, P_0 is first sorted using a non-domination criterion. This sorting assigns to each individual a domination rank. The non-dominated individuals have rank 1, they constitute the front $\mathscr{F}_1$. Then, the others front $\mathscr{F}_i$ are defined recursively by ignoring the lower ranked solutions. This ranking is illustrated on the left of Fig. 4 in the case of a two-objective problem (f_1,f_2). Using the results of the sorting procedure, each individual is assigned a fitness equal to its non-domination level. Then, binary tournament selection, recombination and mutation operators (see [22,15]) are used to create a child population Q_0 with the same size as P_0.

After these first steps, the main loop is applied for M generations. In each loop of this algorithm, t denotes the current generation, $\mathscr{F}$ denotes the result of the non-domination sorting procedure, i.e. $\mathscr{F} = \{\mathscr{F}_i\}$ where $\mathscr{F}_i$ denotes the ith front. P_t and Q_t denote the population and the offspring at generation t, respectively, and R_t is a temporary population.

As one can see, the main loop of the algorithm starts with a merging of the current P_t and Q_t to build R_t. This population of $2N$ solutions is sorted using the non-domination sorting procedure in order to build the population P_{t+1}. In this step, a second sorting criterion is used to keep P_{t+1} to a constant size N during the integration of the successive $\mathscr{F}_i$. Its aim is to take into account the contribution of the solutions to the spread and the diversity of objective function

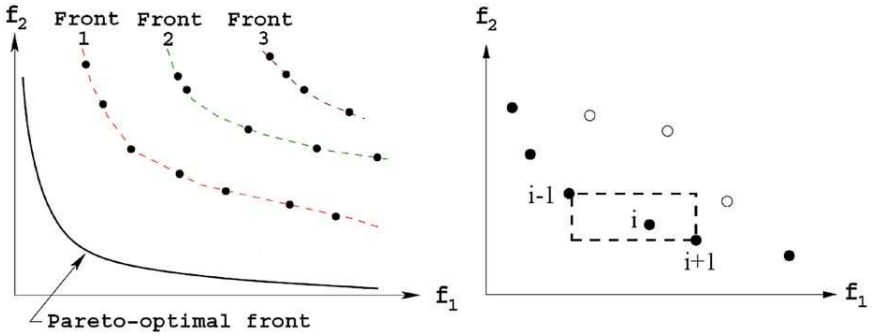


Fig. 4. Illustration of the $\mathscr{F}_i$ concept (left). Illustration of the crowding distance concept (right). The black points stand for the dominant vectors, whereas white ones are dominated.

values in the population. This sorting is based on a measure called *crowding_distance*. This measure which is precisely described in [15] is based on the average distance of the two points on both sides of this point along each of the objectives. This measure is illustrated on the right of Fig. 4. The larger the surface around the considered point, the better the solution from the diversity point of view. Using such values, the solutions in R_t that most contribute to the diversity are preferred in the construction of P_{t+1}. This step is illustrated in Algorithm 1 through the use of Sort($\mathscr{F}_i, \prec_n$), where $\prec_n$ denotes a partial order relation based on both domination and crowding distance. According to this relation, a solution i is better than a solution j if $i_{rank} < j_{rank}$ or if ($i_{rank} = j_{rank}$) and ($i_{distance} > j_{distance}$). One can note that $\prec_n$ is also used in the tournament operator.

Using this algorithm, the population P_t necessarily converges through a set of points of the Pareto front of the problem since non-dominated solutions are preserved along generations. Furthermore, the use of the crowding-distance as a sorting criterion guarantees a good diversity in the population [15]. In the following section, NSGA-II is used in the proposed framework for SVM multi-model selection.

4. Application to SVM multi-model selection

As explained in the previous sections, the proposed framework aims at finding a pool of classifiers, optimizing simultaneously FP and TP rates. The approach can be used for any classifier that uses at least one hyperparameter. In this section, we have chosen to consider support vector machines (SVM) since it is well known that the choice of SVM model parameters can dramatically affect the quality of their solution. Moreover, the problem of SVM model selection is known to be a difficult problem.

4.1. SVM classifiers and their hyperparameters for model selection

As stated in [28], classification problems with asymmetric and unknown misclassification costs can be tackled using SVM through the introduction of two distinct penalty parameters C_- and C_+. In such a case, given a set of m training samples x_i in $\Re^n$ belonging to class y_i:

$$(x_1, y_1)...(x_m, y_m), \quad x_i \in \Re^n,\ y_i \in \{-1, +1\}$$

the maximization of the dual Lagrangian with respect to the α_i becomes

$$\mathrm{Max}_\alpha \left\{ \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right\}$$

$$\text{s.t. the constraints : } \begin{cases} 0 \le \alpha_i \le C_+ & \text{for } y_i = -1 \\ 0 \le \alpha_i \le C_- & \text{for } y_i = +1 \\ \sum_{i=1}^{m} \alpha_i y_i = 0 \end{cases}$$

where α_i denote the Lagrange multipliers and $K(\cdot)$ denotes the kernel. In the case of a Gaussian (RBF) kernel, $K(\cdot)$ is defined as

$$K(x_i, x_j) = \exp(-\gamma \times \|x_i - x_j\|^2)$$

Hence, in the case of asymmetric misclassification costs, three parameters have to be determined to perform an optimal learning of the SVM classifier:

- The kernel parameter of the SVM-rbf: γ.
- The penalty parameters introduced above: C_- and C_+.

In the following, the proposed framework is used in order to select the value of these three hyperparameters.

4.2. Application of NSGA-II for SVM model selection

Two particular points have to be specified for the application of NSGA-II to SVM multi-model selection:

- the solution coding: as said before, three parameters are involved in the learning of SVM for classification problems with asymmetric misclassification costs: C_+, C_- and γ. These three parameters constitute the parameter space of our optimization problem. Consequently, each individual in NSGA-II has to encode these three real values. We have chosen to use a real encoding of these parameters in order to be as precise as possible.
- the evaluation procedure: each individual in the population corresponds to some given values of hyperparameters. In order to compute the performance associated to this individual, a classical SVM learning is performed using the encoded parameter values on a learning dataset. Then, this classifier is evaluated on a test dataset with the classical FP and TP rates as performance criteria.

One can see in Fig. 5 a synthetic scheme of our multi-model selection method.

4.3. Experimental results on UCI datasets

In this subsection, the proposed multi-model selection approach based on the ROC front concept is evaluated and compared with other approaches on publicly available benchmark datasets [1]. First, the experimental protocol of our tests is described. Then, the results

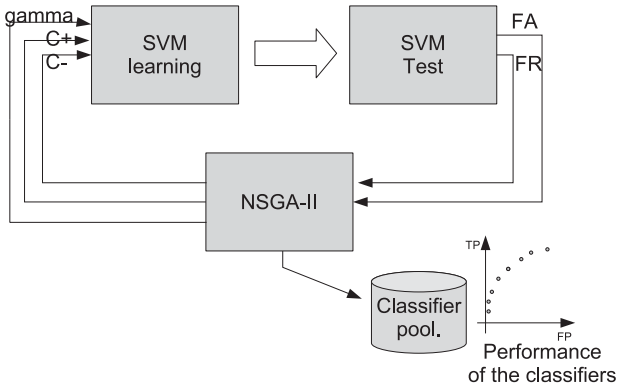


Fig. 5. SVM multi-model selection framework.

Table 1
Number of samples and number of attributes of the considered 2-class UCI problems.

Problem	# samples	# attributes
Australian	690	14
wdbc	569	30
Breast cancer	699	10
Ionosphere	351	34
Heart	270	13
Pima	768	8

are shown and compared with some reference works, and finally several comments on these results are proposed.

Our approach has been applied on several 2-class benchmark datasets publicly available in the UCI machine learning repository on which state-of-the-art results have been published. The number of samples and the number of attributes for each problem are reported in Table 1.

As we propose a real multi-objective approach, the result of our experiments is a pool of classifiers describing the ROC front. Thus, the evaluation of our approach and more precisely its comparison with other approaches of the literature is not easy since as mentioned in the Introduction, comparing some results in a multi-dimensional space is a difficult task. Note that there exist some dedicated measures such as the set coverage metric proposed in [35]. However, to the best of our knowledge, the other referred methods in the literature always consider a single classifier as a solution for a classification problem, which makes it difficult to compare our results with those found in the literature.

Based on this statement, we have therefore chosen to average all the local performance of the ROC front to produce a way to compare our approach to existing ones based on AUC. For that, an area under the ROC front (AUF) is calculated and compared with the area under the ROC curve (AUC) of the other approaches. We do know that this comparison is not theoretically correct since the best results of a pool of classifiers are compared with a curve obtained by varying the threshold of a single classifier. However, the aim of this comparison is not to show that our approach gives better performance but only to highlight the fact that more interesting trade-offs may be locally reached through the ROC front approach. This comparison may also be justified by the fact that finally, in both cases, only one classifier with a unique threshold will be retained for a given problem. We discuss in Section 5 how to select the best model among the pool of classifiers and offer a solution to this problem.

The result of our approach is compared with several works based on the optimization of a scalar criterion for various classifiers: [5] (decision lists and rules sets), [13] (rankboost), [19] (decision trees), [30] (SVMs) and [34] (five models: naive Bayes, logistic, decision tree, kstar, and voting feature interval). We refer to these papers for more explanation of the criterion and the model used.

Table 2
Comparison of the area under the ROC curve (AUC) in the literature with the area under the ROC front (AUF).

Problem	AUC literature	Ref.	AUF
Australian	90.25 ± 0.6	[34]	96.22 ± 1.7
wdbc	94.7 ± 4.6	[19]	99.59 ± 0.4
Breast cancer	99.13	[5]	99.78 ± 0.2
Ionosphere	98.7 ± 3.3	[30]	99.00 ± 1.4
Heart	92.60 ± 0.7	[34]	94.74 ± 1.9
Pima	84.80 ± 6.5	[13]	87.42 ± 1.2

Concerning the application of our multi-objective strategy, a cross-validation procedure has been performed with five folds for each dataset. The results are presented in Table 2, where the first column is the best AUC found until now among the predicted works based on the optimization of a scalar criterion, and the second one is the AUF of our approach.

As expected, one can see that for every dataset the ROC front yielded by the pool of classifiers leads to a higher area than the area under the ROC curve of the other single classifiers. As said before, it is important to emphasize that the AUF cannot theoretically be compared with AUC since the various operating points of the ROC front cannot be reached by a single classifier. However, this comparison with methods which directly optimize AUC clearly shows that our approach enables to reach very interesting local operating points which cannot be reached at the same time by the AUC-based classifiers. Hence, we claim that if the good model can be selected among the pool of classifiers, our approach can lead to better results than AUC-based methods. Despite these interesting results, the model selection problem still remains partly open since the choice of the retained classifier among the set of locally optimal classifiers has to be performed. This crucial final model selection step is discussed in the following section.

5. How to select the best model?

The problem of choosing an operating point in the ROC space is not specific to the proposed approach. For example, when training a single classifier with an AUC criterion, the practitioner still has to choose the appropriate threshold value, i.e. the operating point in the ROC space.

Theoretically, the best operating point must be determined according to Bayes theory by minimizing the following decision function, known as the expected cost and defined as

$$\text{expected cost}(FP, TP) = p(\text{p}).(1 - TP).c(\text{N}, \text{p}) + p(\text{n}).FP.c(\text{Y}, \text{n})$$

where $p(\text{p})$ and $p(\text{n})$ are, respectively, the prior probabilities of (p)ositive samples and (n)egative samples (class distribution), $c(\text{N}, \text{p})$ is the cost of a false negative error and $c(\text{Y}, \text{n})$ is the cost of a false positive error.

Obviously, target conditions ($p(\text{p})$, $p(\text{n})$, $c(\text{N}, \text{p})$, $c(\text{Y}, \text{n})$) are rarely all known at runtime. Consequently, two runtime conditions may be distinguished to select the best model on the ROC front, depending on whether the misclassification costs and the class distributions are known with an acceptable precision or not.

- If the target conditions are known, then *iso-performance lines* proposed in [18] can be used to select the best model. It is based on the projection of the Bayes decision function onto the ROC space. An *iso-performance line* is defined as the set of points providing the same expected cost. The slope of an *iso-performance line* is given by

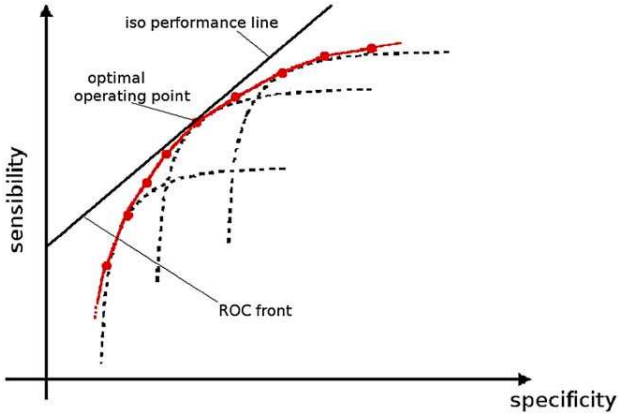


Fig. 6. When the target conditions of a given problem are known, representing the *iso-performance line* allows to select the appropriate operating point.

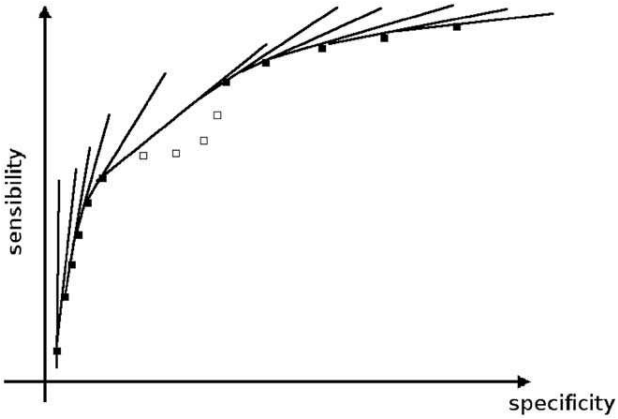


Fig. 7. Browsing all possible *iso-performance lines* on a non-convex ROC front allows to filter the non-filled squares the performance of which can be outperformed.

$$\text{slope} = \frac{p(\text{n}).c(\text{Y},\text{n})}{p(\text{p}).c(\text{N},\text{p})}$$

Using this *iso-performance line* on the ROC space, the optimal operating point can be found by starting from the upper left corner and moving the *iso-performance line* towards the lower right corner. The optimal operating point is the first intersection between the line and the ROC front. This method is illustrated in Fig. 6. We can notice on this figure that the best classifier can be easily selected. Note that in this case, as the accuracy can be computed from the target conditions, a less computational classical scalar-based optimization may be performed, thus avoiding the whole ROC front to be generated. However, if the target conditions are subject to change, generating the whole ROC front is a suitable solution since the adapted operating point can be easily changed using the iso-performance line method, without any additional training stage.

- If the target conditions are unknown at runtime, the expected cost cannot be evaluated. Consequently, the slope of the appropriate *iso-performance line* cannot be determined. Then, the only way for choosing the best classifier is to perform a testing stage in context, i.e. testing each classifier of the ROC front, and choosing the one that best fits the application constraints. We present in Section 6 a real world problem with this kind of scenario.

One can note that, in the second case, browsing all possible *iso-performance lines* could be used in order to "filter" the ROC-front by removing concavities. Indeed, classifiers lying on the concavities of the ROC front cannot be theoretically optimal since any performance on a line segment connecting two ROC points can be achieved by randomly choosing between them [20]. This is illustrated in Fig. 7. Such an idea has been proposed in [29] to generate the ROC convex hull of a set of classifiers. Consequently, one can consider that our proposed method enables to find the optimal ROC-CH.

6. Application to a real-world pattern recognition problem

In this section, an interesting example of real-world problem for which our approach suits better than an AUC-based method is presented.

6.1. Digit/outlier discrimination

The work described in this paper has been motivated by the design of a more complex system that aims at extracting numerical fields (phone number, zip code, customer code, etc.) from incoming handwritten mail document images [10,11] (see Fig. 8). The main difficulty of such a task comes from the fact that handwritten digits may touch each other in the image while some textual parts sometimes are made of separated or touching characters. Fig. 9 gives some examples of segmented components to deal with. In such a variable context, segmentation, detection and recognition of a digit and rejection of textual components must be performed simultaneously.

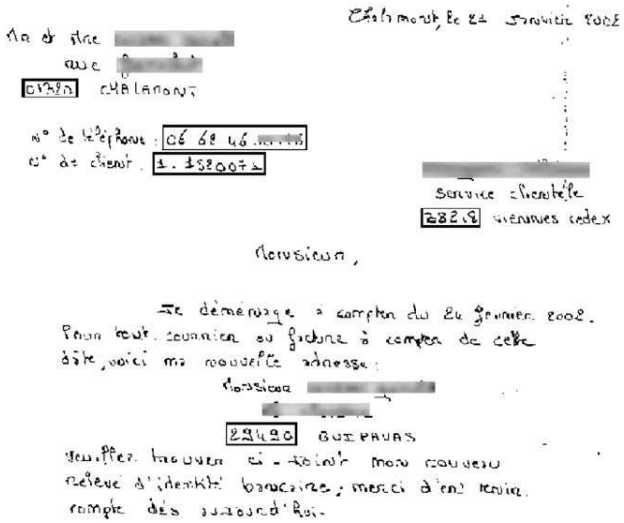


Fig. 8. Example of an incoming mail document. Numerical fields to extract are highlighted.

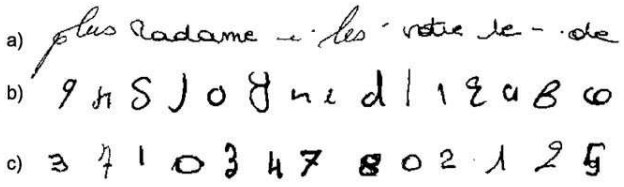


Fig. 9. Examples of digits and outliers. The first line (a) contains shapes which can be considered as "obvious" outliers. The last line (c) contains digits that should be accepted as they are, whereas the middle line (b) contains "ambiguous outliers" (i.e. shaped as digits) that should be rejected by the proposed approach.

In this paper, the proposed approach is applied to a particular stage of the numerical field extraction system. More precisely, the SVM to be optimized is used as a fast two-class classifier prior to the digit recognizer itself, aiming at filtering the "obvious outliers" (see Fig. 9a) from all the other shapes (see Fig. 9b and c) in order to avoid a costly digit recognition stage when it is not necessary. The choice of the SVM classifier has been motivated by its efficiency in a two-class context. Its objective is to reject as many outliers as possible, while accepting as many digits as possible. Further stages of the system deal with digit recognition and ambiguous outlier rejection. This context is a good example of a classification task with asymmetric and unknown misclassification costs since the influence of a FP or a FN on the whole system results is unknown at runtime. In the next subsection, the performance of the proposed system are assessed.

6.2. Experimental results and discussion

In this section, the experimental results obtained using the proposed approach are analysed. These results are compared with those obtained using a state-of-the-art algorithm [30], where a SVM classifier is trained with respect to an AUC criterion. Both NSGA-II and AUC-based approaches have been applied on a learning database of 7129 patterns ($\frac{1}{3}$ digit, $\frac{2}{3}$ outliers), tested and evaluated on a test and a validation database of resp. 7149 and 5000 patterns with the same proportions of digits and outliers. In the case of NSGA-II, the range values for SVM hyperparameters are given in Table 3. Concerning the NSGA-II parameters, we have used some classical values, proposed in [15]. Among them, one can note that the size of the population has been set to 40 in order to have enough points on the Pareto front. The resulting curves are presented in Fig. 10.

Several comments can be made from the obtained results. First, one can remark that each point of the ROC curve obtained for a single classifier trained with AUC criterion is dominated by at least one of the point of the ROC front. Such a result stems from the fact that using an EMOO approach, FP and TP rates are minimized simultaneously through the variation of the three involved SVM hyperparameters, whereas in the case of an AUC approach, a single parametrized classifier is trained to optimize every possible FP/TP trade-offs. Fig. 11 is another illustration of the interest of the ROC front concept. It shows the ROC curves computed from four classifiers which have been selected using the proposed framework. This figure clearly shows that the ROC front corresponds to a set of classifiers which are specialized on some specific ranges of FP/TP trade-offs.

A second remark concerns the possibility when using an EMOO to apply some constraints on the objective values (as in the parameter space). Such a possibility is very useful in the context of our application since it enables to focus on a small part of the ROC front. Indeed, we are particularly interested by a small part of the ROC front since we want the rejection of a digit be as rare as possible

Table 3
Range values for SVM hyperparameters.

Hyperparameter	γ	C_-	C_+
Range	0–1	0–5000	0–5000

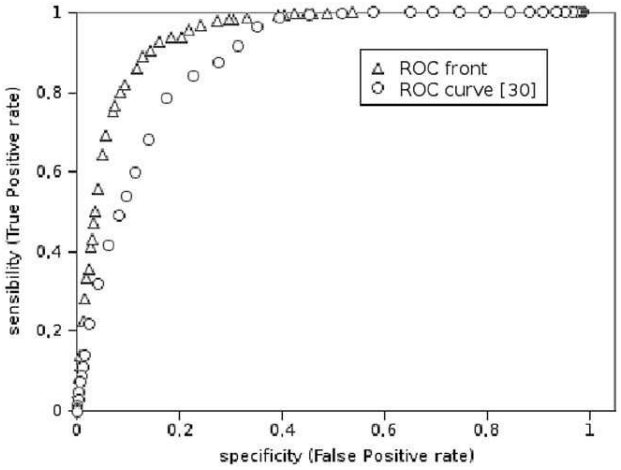


Fig. 10. FP/TP curves obtained using the two approaches: a set of SVM classifiers obtained with NSGA-II (ROC front), and a single SVM classifier trained with AUC criterion (ROC curve).

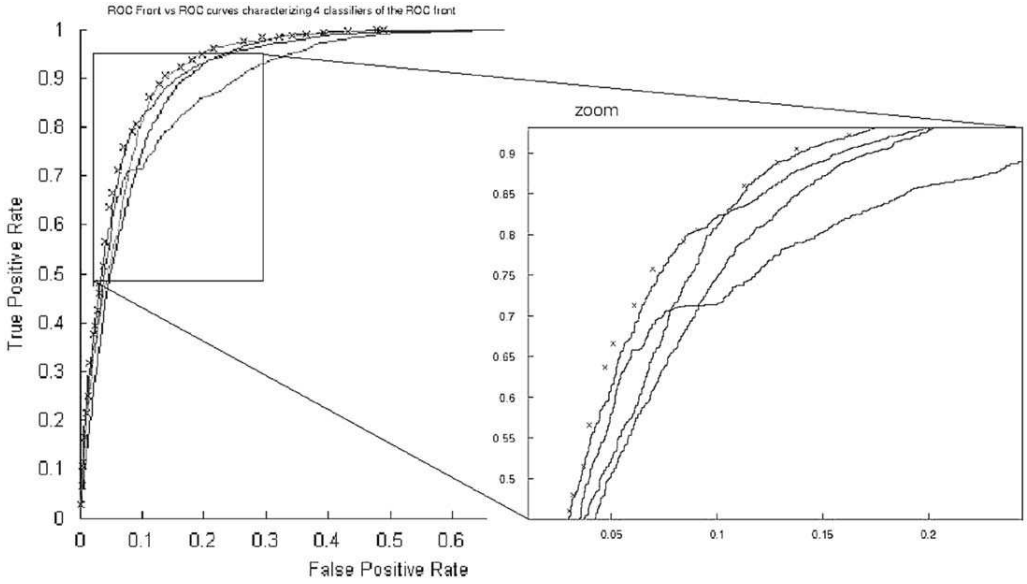


Fig. 11. Illustration of the ROC front concept on a classification dataset. The solid lines are the ROC curves computed from 4 of the 40 classifiers selected using the proposed framework. The performance of the classifiers of the ROC front appear as '×'.

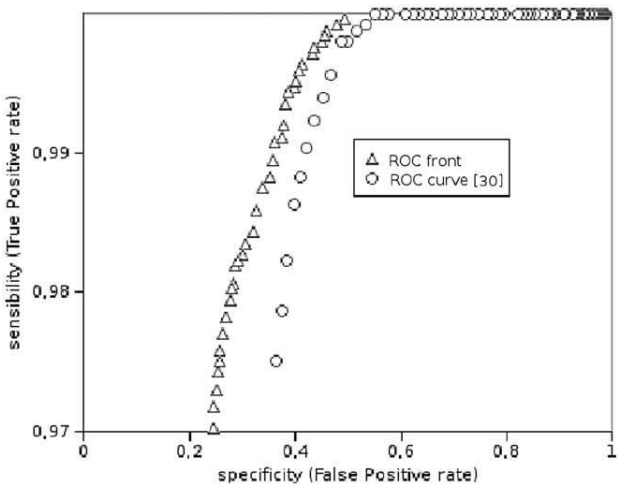


Fig. 12. ROC curve obtained for a true positive rate between 97% and 100%.

Table 4
Recall/precision values of the whole numerical field extraction system for several digit/outlier classifiers, represented here by their TP rate.

Classifier TP rate in %	98.8	99.04	99.26	99.48	99.76	99.96	100
Recall	0.370	0.410	0.440	0.458	0.462	0.481	0.488
Precision	0.110	0.130	0.150	0.176	0.246	0.223	0.152
F1-Measure	0.170	0.197	0.224	0.254	0.321	0.305	0.232

to prevent errors in the whole recognition process, this would imply a null false negative rate (i.e. a 100% TP rate). But on the other hand, Fig. 10 shows that a 100% TP rate leads to a FP higher than 50%. Such a result involves a very time consuming recognition stage, that cannot be accepted regarding our processing time constraints during the decision stage. Thus, we have applied an lower bound of 97% to the TP rate in order to obtain an acceptable trade-off between the recognition quality of the system and the computational constraints. Fig. 12 shows the results obtained with this additional constraint. One can see that such a setting enables to obtain more diversity among the FP/TP trade-offs in the chosen TP range.

6.3. How to select the best model?

Once the ROC front has been built for our application, the final best model among the classifiers has to be selected. As discussed in Section 5, two scenarios may occur at runtime, whether the expected cost can be computed or not. In our digit/outlier discrimination problem, this expected cost cannot be computed since the classification task is embedded in the whole numerical field extraction application and is evaluated by recall/precision measures. Hence, a test stage in context has to be performed by successively embedding each classifier of the front in the whole system. Table 4 presents the results obtained by the whole numerical field extraction system for several digit/outlier classifiers of the ROC front, i.e. for several FP/TP trade-offs.

As one can expected the true positive rate has to be very high to provide good recall and precision values since rejecting a digit may imply to miss a numerical field. We do not show the results for the classifiers the TP rate of which is lower than 98.8% since both recall and precision are lower than those presented in Table 4. Finally, given the final application constraints, the system designer is able to choose the model that best fits the industrial needs. As an example, if one choose to maximize the F1-measure, the classifier providing TPR = 99.76% will be selected. The results of this real-world application corroborate the idea that model selection must be considered as long as possible as a multi-objective optimization task in a pattern recognition system.

7. Conclusion

In this paper, we have presented a framework to tackle the problem of classifier model selection with unknown and/or evolutive misclassification costs. The approach is based on a multi-model selection strategy in which a pool of classifiers is trained in order to depict an optimal ROC front. Using such a front, it is possible to choose the FP/TP trade-off that best fits the application constraints. An application of this strategy with evolutionary multi-objective optimization for the training of a set of SVM classifiers has been proposed, with a validation on both UCI datasets and a real-world application on the discrimination of handwritten digits from outliers. Obtained results have shown that our approach enables to reach better local operating points that state-of-the-art approaches based on the area under ROC curve criterion. As a conclusion, one can say that an AUC-based approach suits pattern recognition problems where the operating point may vary, whereas our approach better suit problems where the operating point is supposed to be static.

The proposed approach is simple and generic and can thus be of great interest for the practitioner who has to optimize a classifier in the context of unknown and/or evolutive misclassification costs. It can be applied to other parametric classifiers (KNN, Neural network, etc.) with other optimization methods [14]. Moreover, it can be easily extended through the introduction of other parameters (kernel type) or objectives (number of support vectors, decision time).

In our future works, we plan to extend the approach to the multi-class problem. We also plan to apply a multi-objective optimization strategy to the whole numerical field extraction system, using recall and precision as criteria.

References

[1] D.J. Newman A. Asuncion, UCI machine learning repository, 2007.
[2] D. Anguita, S. Ridella, F. Rivieccio, R. Zunino, Hyperparameter design criteria for support vector classifiers, Neurocomputing 55 (1–2) (2003) 109–134.
[3] N.E. Ayat, M. Cheriet, C.Y. Suen, Automatic model selection for the optimization of SVM kernels, Pattern Recognition 30 (2004) 1733–1745.
[4] Y. Bengio, Gradient-based optimization of hyperparameters, Neural Computation 12 (2000) 1889–1900.
[5] H. Boström, Maximizing the area under the ROC curve using incremental reduced error pruning, in: Proceedings of ROCML, 2005.
[6] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition 30 (1997) 1145–1159.
[7] L.T. Bui, D. Essam, H.A. Abbass, D. Green, Performance analysis of multiobjective evolutionary methods in noisy environments, in: Proceedings of APS 2004, pp. 29–39.
[8] R. Caruana, A. Niculescu-Mizil, G. Crew, A. Ksikes, Ensemble selection from libraries of models, in: Proceedings of ICML, 2004.
[9] O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee, Choosing multiple parameters for support vector machines, Machine Learning 46 (1) (2002) 131–159.
[10] C. Chatelain, L. Heutte, T. Paquet, Segmentation-driven recognition applied to numerical field extraction from handwritten incoming mail documents, in: Document Analysis System, Lecture Notes in Computer Sciences, vol. 3872, 2006, pp. 564–575.
[11] C. Chatelain, L. Heutte, T. Paquet, A two-stage outlier rejection strategy for numerical field extraction in handwritten documents, in: Proceedings of ICPR, 2006, pp. 224–227.
[12] D.W. Corne, J.D. Knowles, M.J. Oates, The Pareto envelope-based selection algorithm for multiobjective optimization, in: Parallel Problem Solving from Nature, 2000, pp. 839–848.
[13] C. Cortes, M. Mohri, AUC optimization vs. error rate minimization, in: Advances in NIPS, MIT Press, Cambridge, MA, 2004.
[14] B.F. de Souza, A.C.P.L.F. de Carvalho, R. Calvo, R.P. Ishii, Multiclass SVM model selection using particle swarm optimization, in: Proceedings of HIS, 2006, p. 31.
[15] K. Deb, S. Agrawal, A. Pratap, T. Meyarivan, A fast elitist nondominated sorting genetic algorithm for multiobjective optimization: NSGA-II, IEEE Transactions on Evolutionary Computation (2002) 182–197.
[16] K. Deb, Multi-Objective Optimization Using Evolutionary Algorithms, Wiley, New York, NY, USA, 2001.
[17] T.G. Dietterich, G. Bakiri, Solving multiclass learning problems via error-correcting output codes, Journal of Artificial Intelligence Research 2 (1995) 263–286.

[18] T. Fawcett, ROC graphs: notes and practical considerations for researchers, Technical Report, HP Laboratories, 2004.
[19] C. Ferri, P. Flach, J. Hernandez-Orallo, Learning decision trees using the area under the ROC curve, in: Proceedings of ICML, 2002, pp. 139–146.
[20] P.A. Flach, S. Wu, Repairing concavities in ROC curves, in: Proceedings of the 2003 UK Workshop on Computational Intelligence, University of Bristol, August 2003, pp. 38–44.
[21] C.M. Fonseca, P.J. Flemming, Genetic algorithm for multiobjective optimization: formulation, discussion and generalization, in: Proceedings of ICGA, 1993, pp. 416–423.
[22] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Longman Publishing Co., Inc., Boston, MA, USA, 1989.
[23] J. Horn, N. Nafpliotis, D.E. Goldberg, A niched Pareto genetic algorithm for multiobjective optimization, in: Proceedings of IEEE-WCCC, 1994, pp. 82–87.
[24] T. Joachims, Making large-scale support vector machine learning practical, in: A. Smola, B. Scholkopf, C. Burges (Eds.), Advances in Kernel Methods, MIT Press, Cambridge, MA, 1998.
[25] S. Keerthi, V. Sindhwani, O. Chapelle, An efficient method for gradient-based adaptation of hyperparameters in SVM models, in: B. Schölkopf, J. Platt, T. Hoffman (Eds.), Advances in Neural Information Processing Systems, vol. 19, MIT Press, Cambridge, MA, 2007, pp. 673–680.
[26] V. Khare, X. Yao, K. Deb, Performance scaling of multiobjective evolutionary algorithm, Technical Report—SCS, University of Birmingham, 2002, pp. 1–70.
[27] G. Lebrun, O. Lezoray, C. Charrier, H. Cardot, An EA multi-model selection for SVM multiclass schemes, in: Proceedings of IWANN, 2007, pp. 257–264.
[28] E. Osuna, R. Freund, F. Girosi, Support vector machines: training and applications, Technical Report, 1997.
[29] F. Provost, T. Fawcett, Robust classification for imprecise environments, Machine Learning 42 (3) (2001) 203–231.
[30] A. Rakotomamonjy, Optimizing AUC with support vector machine, in: Proceedings of ECAI Workshop on ROC Curve and AI, 2004, pp. 469–478.
[31] J.D. Schaffer, J.J. Grefenstette, Multiobjective learning via genetic algorithms, in: Proceedings of IJCAI 1985, 1985, pp. 593–595.
[32] N. Srinivas, K. Deb, Multiobjective optimization using nondominated sorting in genetic algorithms, Evolutionary Computation 2 (3) (1994) 221–248.
[33] G. Wahba, X. Lin, F. Gao, D. Xiang, R. Klein, B. Klein, The bias-variance tradeoff and the randomized GACV, in: Proceedings of NIPS, 1999, pp. 620–626.
[34] S. Wu, A scored AUC metric for classifier evaluation and selection, in: Proceedings of ROCML, 2005.
[35] E. Zitzler, K. Deb, L. Thiele, Comparison of multiobjective evolutionary algorithms: empirical results, IEEE Transactions on Evolutionary Computation 2 (8) (1999) 173–195.
[36] E. Zitzler, M. Laumanns, L. Thiele, SPEA2: improving the strength Pareto evolutionary algorithm, Technical Report, Computer Engineering and Networks Laboratory (TIK), ETH Zurich, 2001.
[37] E. Zitzler, L. Thiele, Multiobjective evolutionary algorithms: a comparison case study and the strength Pareto approach, IEEE Transactions on Evolutionary Computation 3 (4) (1999) 257–271.

About the Author—CLÉMENT CHATELAIN is an Assistant Professor in the Department of Information Systems Engineering at the INSA of Rouen, France. His research interests include document analysis, handwriting recognition and machine learning. His teaching interests include signal processing, automatic and pattern recognition. Dr. Chatelain received his PhD "Numerical sequences extraction from weakly constrained handwritten documents" from the University of Rouen in 2006.

About the Author—SÉBASTIEN ADAM was born in 1975 in Dieppe, France. He received a PhD in graphical document analysis from the University of Rouen in 2001. This PhD has been led for France Telecom, the historical French telecommunication operator and tackles the problem of multi-oriented and multi-scaled pattern recognition. Then he joined the LITIS labs in Rouen, France. His domains of interest are at the merging of document analysis and multi-objective optimization.

About the Author—YVES LECOURTIER was born in Marseille in 1950. After a thesis in signal processing in 1978, and a second thesis in Physics (automatic control) in 1985 from the University of Paris-Sud, Orsay, France, he joined the University of Rouen as a Professor in 1987. His research domain is in pattern recognition and optimization, especially for document analysis and text recognition. Pr. Lecourtier is a member of AFRIF, ASTI, IAPR. From 1994 to 2000, he was the chairman of the GRCE, a French society which gathers most of the French researchers working in document analysis and text recognition fields.

About the Author—LAURENT HEUTTE (30/05/1964) received his PhD degree in Computer Engineering from the University of Rouen, France, in 1994. From 1996 to 2004, he was a Senior Lecturer in Computer Engineering and Control System at the University of Rouen. Since 2004, he has been a Professor in the same university. Professor Heutte's present research interests are multiple classifier systems, off-line cursive handwriting analysis and recognition, handwritten document layout analysis and information extraction from handwritten documents. Since 2003, he is an Associate Editor of Pattern Recognition journal and the representative member of the French association for pattern recognition (AFRIF) in the Governing Board of the IAPR. He is currently the Head of the "Document and Learning" group in LITIS lab, University of Rouen.

About the Author—THIERRY PAQUET received the PhD degree from the University de Rouen in 1992 in the field of Pattern Recognition. From 1992 to 2002 he has been appointed as a Senior Lecturer at the University of Rouen where he taught Signal and Image Processing. From 1992 to 1996 he was involved in an industrial collaboration with Matra MCS and the French Postal Research Center (SRTP) for the automatization of mail sorting and bank checks reading. Thierry PAQUET was appointed as a full professor in 2002 at the University of Rouen. His current research area concern statistical Pattern Recognition and Image Processing for Document Image Processing including Handwriting Analysis and Recognition. Thierry Paquet is Vice Director of the LITIS laboratory at the University of Rouen since 2007. He is also President of the French association Research Group on Document and Written Communication.

Annexe E

Réference CV : 1

R. Raveaux, S. Adam, P. Héroux, and É. Trupin. Learning graph prototypes for shape recognition. Computer Vision and Image Understanding (CVIU), 115(7) :905-918, 2011.

Computer Vision and Image Understanding 115 (2011) 905–918

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Learning graph prototypes for shape recognition

Romain Raveaux [a], Sébastien Adam [b,*], Pierre Héroux [b], Éric Trupin [b]

[a] Université de la Rochelle – L3I EA 2128, BP 12, 17042 La Rochelle cedex 01, France
[b] Université de Rouen – LITIS EA 4108, BP 12, 76801 Saint-Etienne du Rouvray, France

ARTICLE INFO

Article history:
Received 26 November 2009
Accepted 1 December 2010
Available online 12 March 2011

Keywords:
Graph classification
Graph prototypes
Median graphs
Discriminative graphs
Genetic algorithm
Symbol recognition

ABSTRACT

This paper presents some new approaches for computing graph prototypes in the context of the design of a structural nearest prototype classifier. Four kinds of prototypes are investigated and compared: *set median graphs*, *generalized median graphs*, *set discriminative graphs* and *generalized discriminative graphs*. They differ according to (i) the graph space where they are searched for and (ii) the objective function which is used for their computation. The first criterion allows to distinguish *set prototypes* which are selected in the initial graph training set from *generalized prototypes* which are generated in an infinite set of graphs. The second criterion allows to distinguish *median graphs* which minimize the sum of distances to all input graphs of a given class from *discriminative graphs*, which are computed using classification performance as criterion, taking into account the inter-class distribution. For each kind of prototype, the proposed approach allows to identify one or many prototypes per class, in order to manage the trade-off between the classification accuracy and the classification time.

Each graph prototype generation/selection is performed through a genetic algorithm which can be specialized to each case by setting the appropriate encoding scheme, fitness and genetic operators.

An experimental study performed on several graph databases shows the superiority of the generation approach over the selection one. On the other hand, discriminative prototypes outperform the generative ones. Moreover, we show that the classification rates are improved while the number of prototypes increases. Finally, we show that discriminative prototypes give better results than the median graph based classifier.

1. Introduction

Labeled graphs are powerful data structures for the representation of complex entities. In a graph-based representation, vertices and their labels describe objects (or part of objects) while labeled edges represent interrelationships between the objects. Due to the inherent genericity of graph-based representations, and thanks to the improvement of computer capacities, structural representations have become more and more popular in many application domains such as computer vision, image understanding, biology, chemistry, text processing or pattern recognition. As a consequence of the emergence of graph-based representations, new computing issues such as graph mining [1,2], graph clustering [3,4] or supervised graph classification [5–7] provoked a growing interest.

This paper deals with the supervised graph classification problem. In the literature, this problem is generally tackled using two kinds of approaches. The first one consists in using kernel based algorithms such as Support Vector Machines (SVM) or Kernel Principal Component Analysis (KPCA) [8–13]. Using such methods, the graph is embedded in a feature space composed of label sequences which are obtained through a graph traversal. The kernel values are then computed by measuring the similarity between label sequences. Such approaches have proven to achieve high performance but they are computationally expensive when the dataset is large. The second family consists in using a *k*-Nearest Neighbors (*k*-NN) rule in a dissimilarity space, using a given dissimilarity measure. This kind of approach is the most frequently chosen for its simplicity to implement and its good asymptotic behavior. However, it suffers from three major drawbacks: its combinatorial complexity, its large storage requirements and its sensitivity to noisy examples. A classical solution to overcome these problems consists in reducing the learning dataset through an object prototype learning procedure and to use a Nearest Prototype Classifier (NPC). Such a prototype-based strategy is not inherent to the graph classification problem. It has already been tackled for comparing shapes in computer vision application, e.g. in the approach described in [14] that learns some contour prototypes. It has also been studied for a long time in the context of statistical pattern

* Corresponding author. Fax: +33 2 32 95 52 10.
E-mail addresses: Romain.Raveaux@univ-lr.fr (R. Raveaux), Sebastien.Adam@univ-rouen.fr (S. Adam), Pierre.Heroux@univ-rouen.fr (P. Héroux), Eric.Trupin@univ-rouen.fr (É. Trupin).

1077-3142/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2010.12.015

recognition, using either prototype selection methods (see e.g [15,16]) or prototype generation methods (see e.g. [17,18]).

In the field of structural pattern recognition, there also has been some recent efforts dedicated to the learning of prototypes. Among them, one can cite the pioneering approach proposed in [19] which builds prototypes by detecting subgraphs that occur in most graphs. Another approach concerning trees is proposed in [20]. It consists in learning some kinds of tree prototypes through the definition of a superstructure called tree-union that captures the information about the tree training set. In the domain of graphs, the approaches proposed in [21,22] aim at creating super-graph representations from the available samples. One can also cite the interesting work of Marini proposed in [23] that generates some *creative prototype* by applying to a seed model a well selected set of editing operation. A last approach which is probably the most frequently used concerns median graphs [24–28]. In a classification context, median graphs are computed independently in each class through a minimization process of the sum of distances to all input graphs. Two kinds of median graphs are proposed in the literature: the set median graphs (*smg*) and the generalized median graphs (*gmg*). The only difference between them lies in the space where the medians are searched for. In the first case, the search space is limited to the initial set of graphs (the problem is thus a graph prototype selection problem) whereas in the second case, medians are searched among an infinite set of graphs built using the labels of the initial set (the problem is thus a graph prototype generation problem). Generalized median graphs approaches have proven to keep the most important information in the classes and reject noisy examples [25]. However, a drawback of median graphs when they are used as learning samples of a classification process, as for the all the approaches mentionned before, is that they do not take into account the inter-classes data distribution. In other words, median graphs are rather generative prototypes than discriminative ones.

In this paper, we overcome this drawback by using a discriminative approach while searching an optimal set of prototypes. Thus, it is the classification performance obtained on a validation dataset which is used as criterion in the prototype optimization process. Hence, we propose to use a graph based genetic algorithm in order to learn a set of graph prototypes, called discriminative graphs (*dg*), which minimize the error rate of a classification system. Two configurations are successively considered for extracting the discriminative graphs. In the first one, a single prototype is generated for each class of the classification problem, as in the case of median graphs. Then, this concept is extended to the extraction of multiple prototypes for each class in order to obtain a better description of the data. This extension is also considered in the case of median graphs in order to provide a suitable comparison. In both configurations, we show that discriminative graphs, and particularly multiple discriminative graphs, enable to obtain very good classification results while considerably reducing the number of dissimilarity computations in the decision stage.

Four datasets are used in the experimental protocol. The first is a huge synthetic dataset. The others are real-world datasets consisting of graphs built from a graphical symbol recognition benchmark [29] for the second and the third and from character recognition for the fourth. The classification performance obtained using discriminative graphs and median graphs are compared on these four datasets.

The paper is organized as follows. In section 2, the most important concepts and notations concerning median graphs and discriminative graphs are defined. In section 3, the proposed approach for graph prototypes extraction is detailed. Section 4 describes the experimental evaluation of the algorithm and discusses results. Finally, Section 5 offers some conclusions and suggests directions for future works.

2. Definitions and notations

In this work, the problem which is considered concerns the supervised classification of directed labeled graphs. Such graphs can be defined as follows:

Definition 1. A directed labeled graph G is a *4-tuple* $G = (V, E, \mu, \xi)$ where:

- V is the set of vertices,
- $E \subseteq V \times V$ is the set of edges,
- $\mu: V \rightarrow L_V$ is a function assigning a label to a vertex,
- $\xi: E \rightarrow L_E$ is a function assigning a label to an edge.

A graph classification algorithm aims at assigning a class to an unknown graph using a mapping function f. This function is usually induced from a learning stage which can be defined as follows:

Definition 2. Let χ be the set of the labeled graphs. Given a graph learning dataset $L = \{\langle g_i, c_i\rangle\}_{i=1}^{M}$, where $g_i \in \chi$ is a labeled graph and $c_i \in C$ is the class of the graph among the N classes. The learning of a graph classifier consists in inducing from L a mapping function $f(g)$: $\chi \rightarrow C$ which assigns a class to an unknown graph.

In this paper, graph classification is tackled with a Nearest Prototype Classifier (NPC), *i.e.* with a NN rule applied on a reduced set of representative graph prototypes. Hence, the learning stage of the classifier consists in generating these prototypes. The objectives are (i) to overcome the well-known disadvantages of a k-NN procedure, *i.e.* the large storage requirements, the large computational effort and the sensitivity to noisy examples and (ii) to keep classification performance as high as possible.

As mentioned before, median graphs are frequently used as representative in a graph classification context. Two kinds of median graphs may be distinguished: the set median graph *smg* and the generalized median graph *gmg*. Both are based on the minimization of the sum of distances (SOD) to all input graphs. Formally, they are defined as follows:

Definition 3. Let $d(.,.)$ be a distance or a dissimilarity function that measures the dissimilarity between two graphs. Let $S = \{g_1, g_2, \ldots, g_n\}$ be a set of graphs. The set median graph (*smg*) of S is defined by:

$$smg = \arg\min_{g \in S} \sum_{i=1}^{n} d(g, g_i) \tag{1}$$

According to this definition, *smg* necessarily belongs to the set S. This definition has been extended in [25] to the generalized median graph (*gmg*) which does not necessarily belong to S:

Definition 4. Let $d(.,.)$ be a distance or a dissimilarity function that measures the dissimilarity between two graphs. Let $S = \{g_1, g_2, \ldots, g_n\}$ be a set of graphs. Let U be the infinite set of graphs that can be built using the labels of S. The generalized median graph (*gmg*) of the subset S is defined by:

$$gmg = \arg\min_{g \in U} \sum_{i=1}^{n} d(g, g_i) \tag{2}$$

Median graphs, generalized or not, have already been used as class representatives in a classification process, e.g. in [25–27]. In this case, if N is the number of classes in the learning dataset L, N *smg* (resp. *gmg*) are computed independently (one for each class) and the resulting graph set constitutes the learning dataset $SMG = \{smg_i\}_{i=1}^{N}$ (resp. $GMG = \{gmg_i\}_{i=1}^{N}$) of the nearest prototype classi-

fier. It has been shown in [25] that generalized median graphs capture the essential information of a given class. However, such prototypes do not take into account the inter-class distribution of learning samples.

In order to overcome this problem, we propose to use discriminative graphs (*dg*) as prototypes for graph classification. The main difference between median graphs and discriminative graphs lies in the criterion which is used to generate the prototypes. In the case of *dg*, rather than optimizing a sum of intra-class distances, prototypes are generated in order to minimize the classification error rate obtained on a validation dataset. Obviously, as in the case of median graphs, these prototypes can be computed in the initial set of graphs, leading to set discriminative graphs (*sdg*), or in the whole set of graphs, leading to generalized discriminative graphs (*gdg*). As a consequence, the *dg* for each class are related to each other and can not be expressed independently. The set *SDG* of sdg_i can be defined as follows:

Definition 5. Let N be the number of classes in the learning dataset L. Let T be a validation dataset and let $\Delta(T, \{g_i\}_{i=1}^N)$ be a function computing the error rate obtained by a 1-NN classifier on T using the graph prototypes $\{g_i\}_{i=1}^N \in L$ as learning samples. Then the set *SDG* composed of the sdg_i of each class is given by:

$$SDG = \{sdg_1, sdg_2, \ldots, sdg_N\} = \arg\min_{\{g_i\}_{i=1}^N \subset L} \Delta\left(T, \{g_i\}_{i=1}^N\right) \quad (3)$$

In the same way, the set *GDG* of *gdg* is defined as follows:

Definition 6. Let N be the number of classes in the learning dataset L. Let U be the infinite set of graphs that can be built using labels from L. Let T be a validation dataset and let $\Delta(T, \{g_i\}_{i=1}^N)$ be the error rate obtained by a 1-NN classifier on T using the graph prototypes $\{g_i\}_{i=1}^N \in U$ as learning samples. Then the set *GDG* composed of the *gdg* of each class is given by:

$$GDG = \{gdg_1, gdg_2, \ldots, gdg_N\} = \arg\min_{\{g_i\}_{i=1}^N \subset U} \Delta\left(T, \{g_i\}_{i=1}^N\right) \quad (4)$$

The concepts presented above involve the generation of a single prototype for each class. In some particular applications, it may be interesting to generate m prototypes for each class in order to obtain a better description of the data. In the following, we give the definition of such prototypes called *m-gdg*.[1]

Definition 7. Let N be the number of classes in the learning dataset L. Let U be the infinite set of graphs that can be built using labels from L. Let m be the number of prototypes to be computed in each class. Let T be a validation dataset and let $\Delta(T, \{g_{ik}\}_{i=1,k=1}^{N,m})$ be the error rate obtained by a 1-NN classifier[2] on T using the graph prototypes $\{g_{ik}\}_{i=1,k=1}^{N,m} \in U$ as learning samples. Then the set *mGDG* composed of the *m-gdg* of each class is given by:

$$\begin{aligned} mGDG &= \{gdg_{11}, \ldots, gdg_{1m}, \ldots, gdg_{N1}, \ldots, gdg_{Nm}\} \\ &= \arg\min_{\{g_{ik}\}_{i=1,k=1}^{N,m} \subset U} \Delta\left(T, \{g_{ik}\}_{i=1,k=1}^{N,m}\right) \end{aligned} \quad (5)$$

In order to provide some fair comparisons in the experimental protocol, we also extend the median graph concept to multiple prototypes. In this case, the *m-gmg* (as well the *m-smg*) are defined independently for each class:

[1] The definition of *m-sdg* is easily obtained through the change of the search space from U to S.

[2] In this case, a k-NN procedure with $k > 1$ will be considered in future works, for example to allow the system to reject some patterns.

Definition 8. Let $d(.,.)$ be a distance or a dissimilarity function that measures the dissimilarity between two graphs. Let n be the number of samples in the considered class. Let m be the number of prototypes, gp_k be the prototypes and g_i be the graphs of the considered class. Then, the set *mGMG* composed of the *m-gmg* for the considered class is given by:

$$mGMG = \{gmg_1, \ldots, gmg_m\} = \arg\min_{\{gp_k\}_{k=1}^m \subset U} \sum_{i=1}^{n} \min_{k \in \{1,m\}} d(gp_k, g_i) \quad (6)$$

The algorithms involved in the computation of the different kinds of representative prototypes are presented in the following section.

3. Genetic algorithms for graph prototypes generation

In Section 2, the graph prototype search problem has been defined as an optimization process. Two kinds of prototypes have been distinguished: (i) set prototypes and (ii) generalized prototypes.

(i) The set prototype search problem consists in selecting the m prototypes per class which optimize an objective function. A combinatorial exploration of the solution space would result in evaluating the criterion for each of the potential solutions. If we consider that each of the N classes contains n_i elements, there are

$$\binom{m}{n_1} \times \binom{m}{n_2} \times \cdots \times \binom{m}{n_N} \quad (7)$$

combinations for selecting m prototypes to represent each class. For a quite simple problem with two classes and 100 graphs in each class, the search for five prototypes per class would result in more than 75×10^6 evaluations of the criterion. Hence, a complete exploration of the solution space rapidly becomes intractable. Many heuristic methods such as multistart, genetic algorithms or tabu search [18] have been used to tackle the problem of set prototype search when dealing with vectorial data. Among them, genetic based methods have shown good performance [30,18].

(ii) The generalized prototype search problem can also be stated as an optimization problem. However, it cannot be solved with a combinatorial approach since the set U in which the solutions are searched for is unbounded (only a subset S of U is known). In [24], the authors use genetic algorithms to approximate the generalized median graph of a set of graphs. In the context of computing a single generative prototype, they report that the solution reached by a genetic approach is often the optimal solution. In this paper, we also propose to use genetic algorithms but to solve both the set/generalized median/discriminative prototype extraction problem. The next subsections precisely describe our approach.

3.1. Genetic algorithm

Genetic Algorithms (GA) are evolutionary optimization techniques with a wide scope of applications [31]. They have been used to solve many combinatorial problems [32]. An individual of a GA corresponds to a possible solution of an optimization problem. The relationship between this individual and the corresponding solution is given by an appropriate encoding. The quality of each individual is evaluated thanks to a score function which enables to quantify the quality of the corresponding solution. In order to converge to the optimal solution, individuals from a size-limited population are randomly selected at each generation according to a fitness value which is computed using the scores of all the indi-

viduals of the population. New individuals are then generated from those selected individuals thanks to genetic operators such as crossover or mutation. From a general point of view, the crossover operator aims at promoting the exchange of *good* genetic material between individuals of the previous generation. The mutation operator is used to promote genetic diversity and to explore the solution space. Given these general principles, solving a specific optimization problem using GA requires the definition of:

- an appropriate encoding of the solutions;
- a function which evaluates the score of the individual;
- a selection strategy;
- some dedicated genetic operators (mutation and crossover operators).

The following paragraphs tackle each of these points for both graph prototype selection and generation, and describe the proposed genetic algorithm.

3.2. Individual encoding

The encoding aims at giving a one-to-one relationship between the individuals manipulated by the GA and the solutions of the optimization problem. As defined before, the prototype selection/generation problem aims at providing m prototypes for each of the N classes. So, we adopt a general scheme where an individual contains $m \times N$ genes, and each gene encode a graph prototype. An example is given in Fig. 1. In this example, the individual encodes two prototypes for each class in a 3 classes problem and $g_{i,j}$ is the ith graph prototype describing class j. Obviously, this encoding is specialized for each problem.

3.2.1. Set prototype problem encoding

As stated in Section 2, the possible solutions of a set prototype problem are the combinations of m elements selected from each class in the initial graph set. For this kind of problem, an individual can be defined by a list of $N \times m$ integers which is structured as a sequence of N m-sets. Each m-set describes one of the N classes and contains the m indices of the elements from the initial set which are selected as prototype. The exemple in Fig. 2 presents the encoding of an individual for a 3-class problem where 2 prototypes are selected to describe each class. This individual indicates that class 1 is described with elements 1 and 3 of a learning subset composed of the graphs of the first class, that class 2 is described with elements 5 and 2 of the class, and that class 3 is described with graphs the indices of which are 7 and 3 in the third class subset.

3.2.2. Generalized prototype problem encoding

The index model used in the set prototype problem can not be used for the solution encoding of the generalized prototype problem since the definition of generalized (median and discriminative) graphs implies that prototypes may be outside of the initial set of graphs. As a consequence, each gene of an individual can not be a *simple* index and has to encode all the information contained in the corresponding graph. We have chosen to represent each graph with its adjacency matrix. Hence, an individual can be defined by a list of $N \times m$ adjacency matrices, structured as a sequence of N m-sets. Fig. 3 illustrates such an encoding where only one of the six genes is represented.

$g_{1,1}$	$g_{2,1}$	$g_{1,2}$	$g_{2,2}$	$g_{1,3}$	$g_{2,3}$

Fig. 1. General encoding scheme for the m prototypes problem. Each individual contains $m \times N$ genes. Each one corresponds to a graph prototype.

1	3	5	2	7	3

Fig. 2. Set prototype encoding scheme for the m prototypes problem. Each individual contains $m \times N$ genes. Each gene is the index of the graph in the considered class of the learning dataset.

3.3. Fitness function

A fitness function aims at evaluating how the solution encoded by an individual is good for the optimization problem with respect to the entire population. The computation of a fitness value relies on two steps. First, the score of the individual has to be evaluated. It corresponds to the value of the objective function to be optimized. Then, this value is normalized with respect to the scores of all the individuals of the population. As mentioned in Section 2, objectives are different for the median prototype problem and for the discriminative prototype problem. As a consequence, score functions differ for each problem.

3.3.1. Score function for median prototypes

As defined in Section 2, the score function in the median prototype problem is given by:

$$S_\alpha = \sum_{i=1}^{N} \left(\sum_{j=1}^{n_i} \min_{k \in [1,m]} d(L_{ij}, smg_{ik}) \right) \tag{8}$$

where N is the number of classes, n_i is the number of elements of class i in the learning dataset, m is the number of prototypes per class, L_{ij} is the jth sample of class i, and smg_{ik} is the kth prototype of class i in the individual α.

3.3.2. Score function for discriminative prototypes

The score value of an individual in the discriminative prototype problem is a function which is directly linked to the error rate of the Nearest Prototype Classifier evaluated on a validation dataset T using the prototypes encoded in the individual. It is given by:

$$S_\alpha = \Delta\left(T, \{g_{ik}\}_{i=1,k=1}^{N,m}\right) \tag{9}$$

where T is the validation dataset, N is the number of classes, m is the number of prototypes per class, g_{ik} is the kth prototype of class i in the individual and $\Delta(T, \{g_{ik}\}_{i=1,k=1}^{N,m})$ is the error rate obtained by a 1-NN classifier on T using the graph prototypes $\{g_{ik}\}_{i=1,k=1}^{N,m}$ as learning samples.

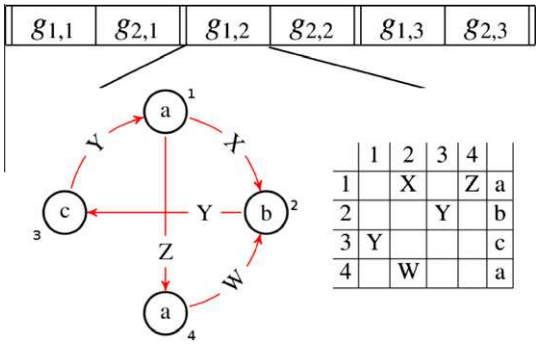


Fig. 3. Generalized prototype encoding scheme for the m prototypes problem. Each individual contains $m \times N$ genes. Each gene is an adjacency matrix describing the corresponding graph. Only $g_{1,2}$ is represented here. In the adjacency matrix, the digits state for vertex identifiers. a, b, and c are vertices labels, they appear in the last column of the matrix. W, X and Y are edge labels, they appear in the adjacency matrix at the line (resp. column) corresponding to the source (resp. target) vertex.

The computation of both the Δ value of Eq. (9) and the S_α value of Eq. (8) makes use of graph distance computation. The following paragraph discusses our choice for this distance definition.

3.3.3. Distance computation

Any kind of distance can be used in the proposed framework (graph edit distance [33,34] or its approximations [35], distance based on the maximum common subgraph [36], distance based on graph union [37], etc.). In the experiments proposed in section 4, the graph comparison computation is performed using a dissimilarity measure proposed by Lopresti and Wilfong [38]. This measure is based on graph probing which has been proved to be a lower bound for the reference graph edit distance within a factor of 4.

Let g be a directed attributed graph with edges labeled from a finite set $L_E = \{l_1,\ldots,l_a\}$. A given vertex of g can be represented with its edge structure as a $2a$-tuple of non-negative integers $\{x_1,\ldots,x_a, y_1,\ldots,y_a\}$ such that the vertex has exactly x_i incoming edges labeled l_i and y_j outgoing edges labeled l_j.

In this context, two types of probes are defined in [38]:

- $P_1(g)$: a vector which gathers the counts of vertices sharing the same edge structure for all encountered edge structures;
- $P_2(g)$: a vector which gathers the number of vertices for each vertex label.

Based on these probes and on the L_1-norm, the graph probing distance between two graphs g_1 and g_2 is given by:

$$gpd(g_1,g_2) = L_1(P_1(g_1),P_1(g_2)) + L_1(P_2(g_1),P_2(g_2)) \quad (10)$$

The graph probing distance respects the non-negativity, symmetry, and triangle inequality properties of a metric, but it does not respect the uniqueness property. In other words, *gpd* is a pseudo-metric and two non-isomorphic graphs can have the same probes.

However, the main advantage of graph probing in this study is its low computational cost (linear function of the vertex number). Due to the intensive use of distance computations during the genetic algorithm, this property makes the graph probing distance a good candidate. Nevertheless, it is important to note that any kind of dissimilarity measure may be used in the proposed framework.

3.3.4. Fitness computation

Once the score value of an individual has been computed, a second step of individual evaluation consists in computing its fitness, through a normalization of the score value with respect to all the individuals of the population. We use the following classical fitness assignment procedure in this scope:

$$F_\alpha = \frac{S_\alpha}{\sum_{i=1}^{\rho} S_i} \quad (11)$$

3.4. Selection strategy

The selection operator aims at selecting a proportion of the existing population to breed a new generation. Individual solutions are selected through a fitness-based process, where fitter solutions (as measured by the fitness function defined in Eq. (11)) are typically more likely to be selected. We use the well-known roulette wheel strategy [31] in which the probability of an individual to be selected is proportional to its fitness value. In the whole reproduction process, an elitism mechanism is coupled with this selection strategy such that the μ best individuals from the previous generation are ensured to be in the next generation.

3.5. Crossover

As mentioned before, the crossover operator is designed to generate offsprings from selected individuals. The exchange of genetic material aims at generating offsprings sharing *good* genes from their parents.

In our case, the crossover is performed by a random exchange of prototypes between the parent for each class. Fig. 4 illustrates the crossover operation. The operation is the same for all the kinds of prototypes. In the case of set prototypes, where the graphs prototypes are designated by indices, only indices are permuted whereas the complete graph descriptions are exchanged when dealing with the generalized prototype problem.

3.6. Mutation

Mutations are used to promote genetic diversity and allow the exploration of regions of the solution space which can not be reached only with crossover. As the solution space is different for set prototype and generalized prototype problems, the mutation operator has to be specialized for each case.

3.6.1. Mutation for set prototype problem

In the set prototype problem, the solution space is defined by the combinations allowing the selection of m prototypes for each class. An elementary modification of an individual would consist in replacing a prototype by an element from the same class that is not already selected in the individual. Hence, considering the index model used to represent graphs, a simple way to perform a mutation is to arbitrarily substitute an index values by a random integer. Fig. 5 illustrates the mutation process. In this example, we can observe that element 3 has been replaced by element 4 in the mutated version of the description of class 1. In the same way, instance 5 has been replaced by instance 6 in the description of class 2. Finally, the mutated version describes class 3 using the element 5 instead of element 3.

3.6.2. Mutation for the generalized prototype problem

In the generalized prototype problem, the solution space is not restricted to the combinations of elements selected in L. Graphs that are not element of L can be generated as prototypes. As a consequence, the mutation operation can not be restricted to an index modification. It must be able to produce new graphs. To do this, a

$g_{1,1}$	$g_{2,1}$	$g_{1,2}$	$g_{2,2}$	$g_{1,3}$	$g_{2,3}$
$g'_{1,1}$	$g'_{2,1}$	$g'_{1,2}$	$g'_{2,2}$	$g'_{1,3}$	$g'_{2,3}$

(a) Pair of individuals selected for the crossover operation : the parents

$\mathbf{g'_{1,1}}$	$g_{2,1}$	$\mathbf{g'_{1,2}}$	$\mathbf{g'_{2,2}}$	$g_{1,3}$	$g_{2,3}$
$\mathbf{g_{1,1}}$	$g'_{2,1}$	$\mathbf{g_{1,2}}$	$\mathbf{g_{2,2}}$	$g'_{1,3}$	$g'_{2,3}$

(b) Pair of children generated by the crossover operation.

Fig. 4. Illustration of the crossover operator: two selected parents (a) generate two offsprings (b). Genes 1, 3 and 4 have been swapped during the operation.

1	3	5	2	7	3

(a) individual selected for mutation

1	**4**	**6**	2	7	**5**

(b) individual resulting from the mutation operation

Fig. 5. Illustration of the mutation operator for set prototypes: genes 2, 3 and 6 have mutated.

random edit operation is applied to the graph prototypes that are included in the individual. For each graph of a given individual, a first random choice according to a mutation probability enables to decide if a mutation is applied or not. Then, one of the six following possible operations illustrated on Fig. 6 is chosen randomly:

- *Vertex deletion*: delete a randomly chosen vertex and all its connected edges. This operation corresponds to the deletion of a row and a column in the adjacency matrix (see Fig. 6b).
- *Edge deletion*: delete a randomly chosen edge. This operation corresponds to the deletion of an edge value in the adjacency matrix (see Fig. 6c).
- *Vertex insertion*: insert a new vertex in the graph with a randomly chosen label among the vertex label dictionary. This operation corresponds to the addition of a new row and a new column in the adjacency matrix. The label column is also updated using the randomly chosen label (see Fig. 6d).
- *Edge insertion*: insert a new edge between two random vertices with a randomly chosen label among the edge label dictionary. This operation corresponds to the addition of a randomly labeled edge in the adjacency matrix (see Fig. 6e).
- *Vertex substitution*: substitute the label of a randomly chosen vertex using the vertex label dictionary. This operation corresponds to the modification of the label column for the randomly chosen vertex(see Fig. 6f).
- *Edge substitution*: substitute the label of a randomly chosen edge using the edge label dictionary. This operation corresponds to the modification of the label for the randomly chosen edge (see Fig. 6g).

3.7. Proposed algorithm

Algorithm 1 gives the generic structure of the GA used for the graph prototype generation/selection problems. This algorithm complies with the principles defined in Section 3.1 and is specialized by setting the adapted encoding, fitness function and genetic operators presented previously.

First, an initialization procedure aims at building the initial population where each individual corresponds to a possible solution of the optimization problem. In the case of set prototypes, distinct indices are randomly chosen for each individual in order to represent the N classes with $N \times m$ graphs. For generalized prototypes, we have chosen to initialize the individuals with randomly chosen graphs from the learning dataset, since it has been shown in [24] that it is a better solution than a complete random procedure.

Then, the GA iterates over the generations, building new size-limited populations from the previous ones. Each new generation is composed of:

- the μ best individuals from the previous one. Such an elitist strategy ensures the convergence of the algorithm.
- mutated or crossed version of individuals that have been selected from the previous generation.

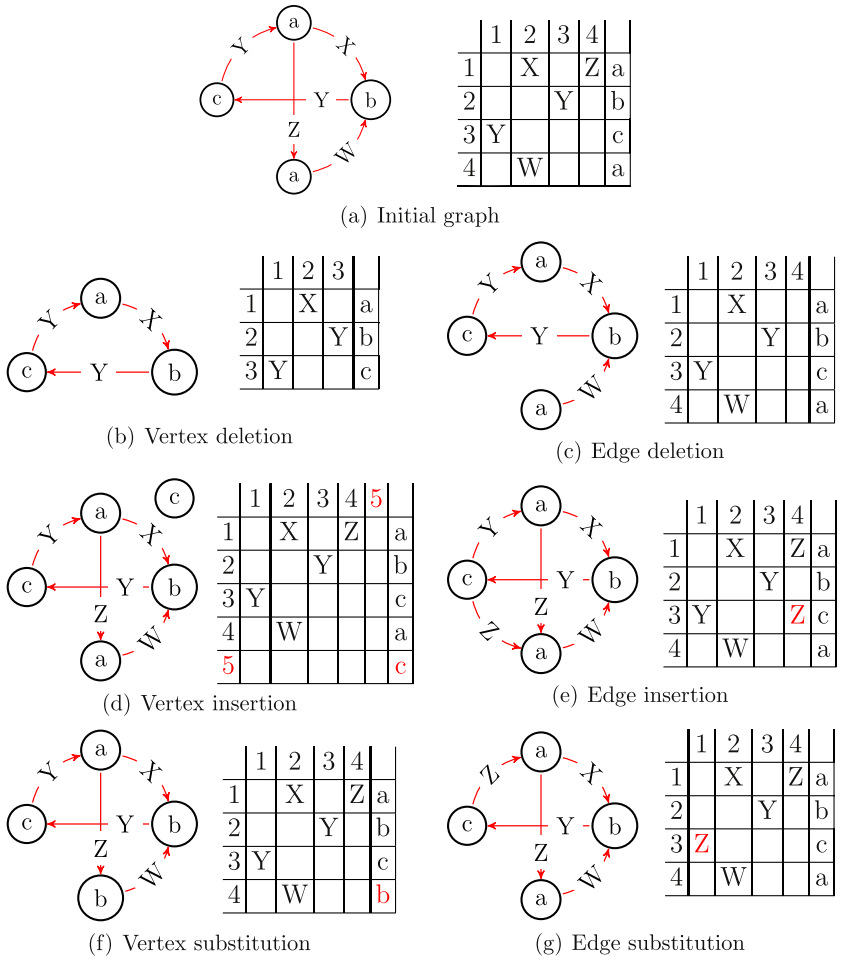


Fig. 6. Illustration of the mutation operators on both generalized graphs and the corresponding adjacency matrices.

Finally, the algorithm provides the best individual from the last generation as the best solution of the optimization procedure.

Algorithm 1. Genetic algorithm

Require: L: the training set
Require: T: the validation set
Require: m: number of prototypes per class
Require: populationSize
Require: generationNumber
Require: mutationRate
Require: μ: elitism value
Ensure: A set of $N \times m$ prototypes
  Pop[0][] ← popInit(L,T,m,populationSize)[1]
  popEval(Pop[0],L,T)
  fitnessEval(Pop[0])
  for i = 1 to generationNumber **do**
    Pop[i][1:μ] ← μ best individuals in Pop[$i-1$]
    $j \leftarrow \mu + 1$
    while $j \leqslant$ *populationSize* **do**
      op← choice between mutation and crossover [2]
      if op = mutation **then**
        ind← select an individual in Pop[$i-1$] [3]
        Pop[i][j]← mutation(ind)
        $j \leftarrow j + 1$
      else
        ind_1← select an individual in Pop[$i-1$] [3]
        ind_2← select an individual in Pop[$i-1$] [3]
        ($newInd_1$,$newInd_2$)← crossover(ind_1,ind_2)
        Pop[i][j]← ind_1
        Pop[i][$j+1$]← ind_2
        $j \leftarrow j + 2$
      end if
      popEval(Pop[i],L,T)
      fitnessEval(Pop[i])
    end while
  end for
  return the best individual from the last generation
[1] T is not used for the initialization in the case of discriminative graphs
[2] This choice is made according to *mutationRate*
[3] Selection is done using a roulette wheel according to fitness values

4. Experimental results and analysis

This section is devoted to the experimental evaluation of the proposed approach. First, both the datasets and the experimental protocol are described before investigating and discussing the merits of the proposed approach.

4.1. Dataset description

The experiments described in this section have been carried out on four databases. The first one is composed of synthetic data allowing (i) an evaluation in a general context on a huge dataset and (ii) an evaluation with respect to the number of classes. The others sets are domain specific, they are related to pattern recognition issues where graphs are meaningful. Each dataset has been split into three subsets respectively called training subset, validation subset and test subset. The content of each database is summarized in Table 1. For each dataset, this table gives: the number of classes (Classes), the total number of data (Samples), the sizes of learning/validation/test datasets and the mean properties of the graphs.

Table 1
Properties of the four datasets (A,B,C,D) used in the experiments: number of graphs, distribution of the graphs in the learning/validation/test subsets and properties of the graphs in the dataset.

	A	B	C	D
$\|Classes\|$ (N)	50	10	32	15
$\|Samples\|$	28,229	200	12,800	6750
$\|Training\|$	10,596	88	7200	3796
$\|Validation\|$	14,101	56	3200	1688
$\|Test\|$	3532	56	2400	1266
$\|vertices\|_{mean}$	12.03	5.56	8.84	4.7
$\|edges\|_{mean}$	9.86	11.71	10.15	3.6
$\|degree\|_{mean}$	1.63	4.21	1.15	1.3

4.1.1. Synthetic dataset: Base A

This dataset contains over 28,000 graphs, roughly identically distributed in 50 classes (about 560 graphs per class). The graphs are directed with edges and vertices labeled from two distinct alphabets. They are built using a modified version of the generic framework used to construct random graphs proposed in [39]. Since this framework does not aim at depicting classes, in the sense of similar graphs, we add a second step to the data generation process in order to create classes of graphs. In the initial step a number N (where N is the desired number of classes) of graphs are constructed using the Erdös-Rényi model [39]. This model takes as input the number of vertices of the graph to be generated, and the probability of generating an edge between two vertices. A low probability for edges leads to sparse graphs, that typically occur in proximity based graph representations found in pattern recognition. In the second step, each of the generated graphs are modified using two processes. In a first stage edges and vertices are randomly deleted or relabeled according to a given probability. Then, a second stage of modifications is applied by selecting a vertex from a graph and replacing it with a random subgraph. The whole process leads to graph classes which have an intra-class similarity greater than the inter-class similarity. Numerical details concerning this dataset are presented in Table 1. The large size of this dataset is a key point to measure up our approach to the scalability problem.

4.1.2. Symbol recognition related dataset: Base B

This second dataset contains graphs which are generated from a corpus of 200 noisy symbol images, corresponding to 10 ideal models (classes) proposed in a symbol recognition contest [29] (GREC workshop). The class distribution is given in Table 2. In a first step, considering the symbol binary image, both black and white connected components are extracted. These connected components are then automatically labeled with a partitional clustering algorithm [40] using Zernike moments as features[41]. Using these labeled items, a graph is built. Each connected component correspond to an attributed vertex in this graph. Then, edges are built using the following rule: two vertices are linked with an undi-

Table 2
Class sizes of the dababase B.

Class	Samples
1	25
2	13
3	17
4	13
5	20
6	39
7	22
8	17
9	17
10	17

rected and unlabeled edge if one of the vertices is a neighbor of the other vertex in the corresponding image. This neighborhood is decided according to the distance between the centroids of each connected components with respect to a predefined threshold (see [42] for more details). An example of the association between two symbol images and the corresponding graphs is illustrated in Fig. 7. Numerical details concerning this dataset are presented in Table 1.

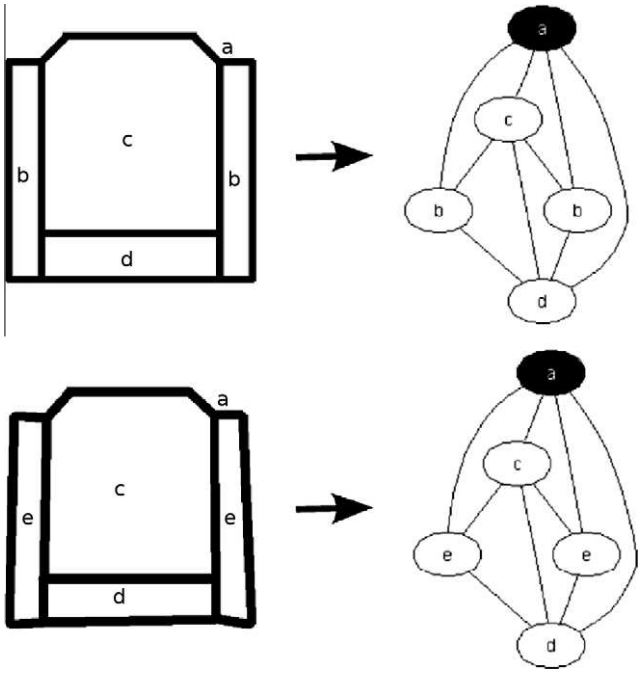


Fig. 7. From symbols to graphs through connected component analysis. At the top: a model symbol. At the bottom: a distorded symbol. In both graphs, the vertex *a* denotes the black connected component whereas the others denote white connected components. In the bottom graph (distorded version), the label *e* has replaced the label *b* of the initial.

4.1.3. Ferrer dataset: Base C

This third dataset is also related to the symbol recognition problem. It is derived from the GREC database [29]. It is composed of 12,800 graphs identically distributed among 32 classes (examples of symbols are given on Fig. 8). These graphs are built using a slightly modified version of the approach proposed in [26]. Using Ferrer's approach, a symbol is represented as an undirected labeled graph which stems from a vectorial representation of the symbol image. In this graph, the vertices correspond to the Terminal Points (TPs) and the Junction Points (JPs) of the vectorial representation and the edges correspond to the segments which connect those points in the image. The information associated to vertices or edges are their cartesian coordinates (x,y). Due to the graph spectral theory limitation, Ferrer's graphs have to be labeled using real positive or null values and can not handle complex objects. This restriction leads to the construction of two graphs for a single symbol: a graph Gx labeled with x coordinates and a graph Gy with y coordinates, as

Table 3
Parameters used for the genetic algorithm in the proposed experiments.

	Acronym	Value
Population Size	ρ	200
Mutation rate	σ	0.3
♯ of generations	G	100
♯ of runs	W	10

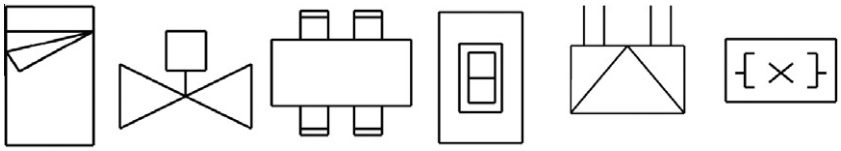


Fig. 8. Examples of symbols used to build the graphs of the Ferrer dataset [29] – base C.

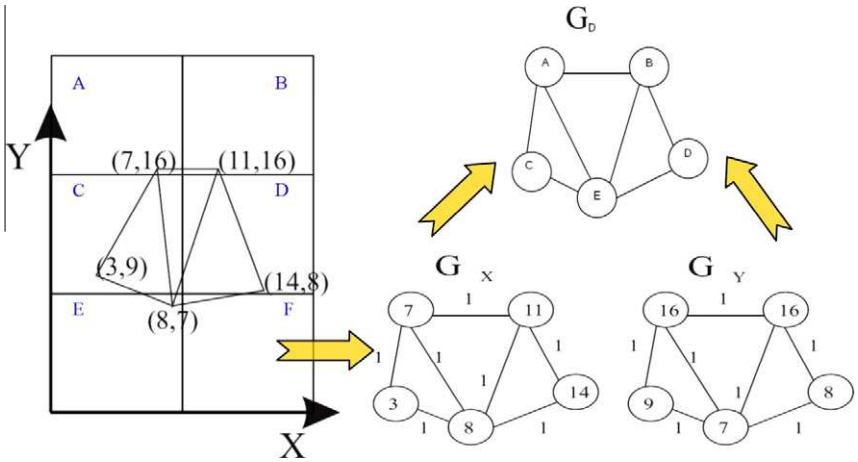


Fig. 9. From symbols to graphs using a 2D mesh. On the left, a vectorized symbol. One the bottom right, the two graphs G_x and G_y obtained using Ferrer's approach. The vertices correspond to the Terminal Points (TPs) and the Junction Points (JPs) of the vectorial representation, labeled with either their x coordinates (on the left) or their y coordinates (on the right). The edges correspond to the segments which connect those points in the image. On the top right, the graphs used to evaluate the proposed approach where the vertices label are obtained through a discretization of $\mathbb{R}^2$.

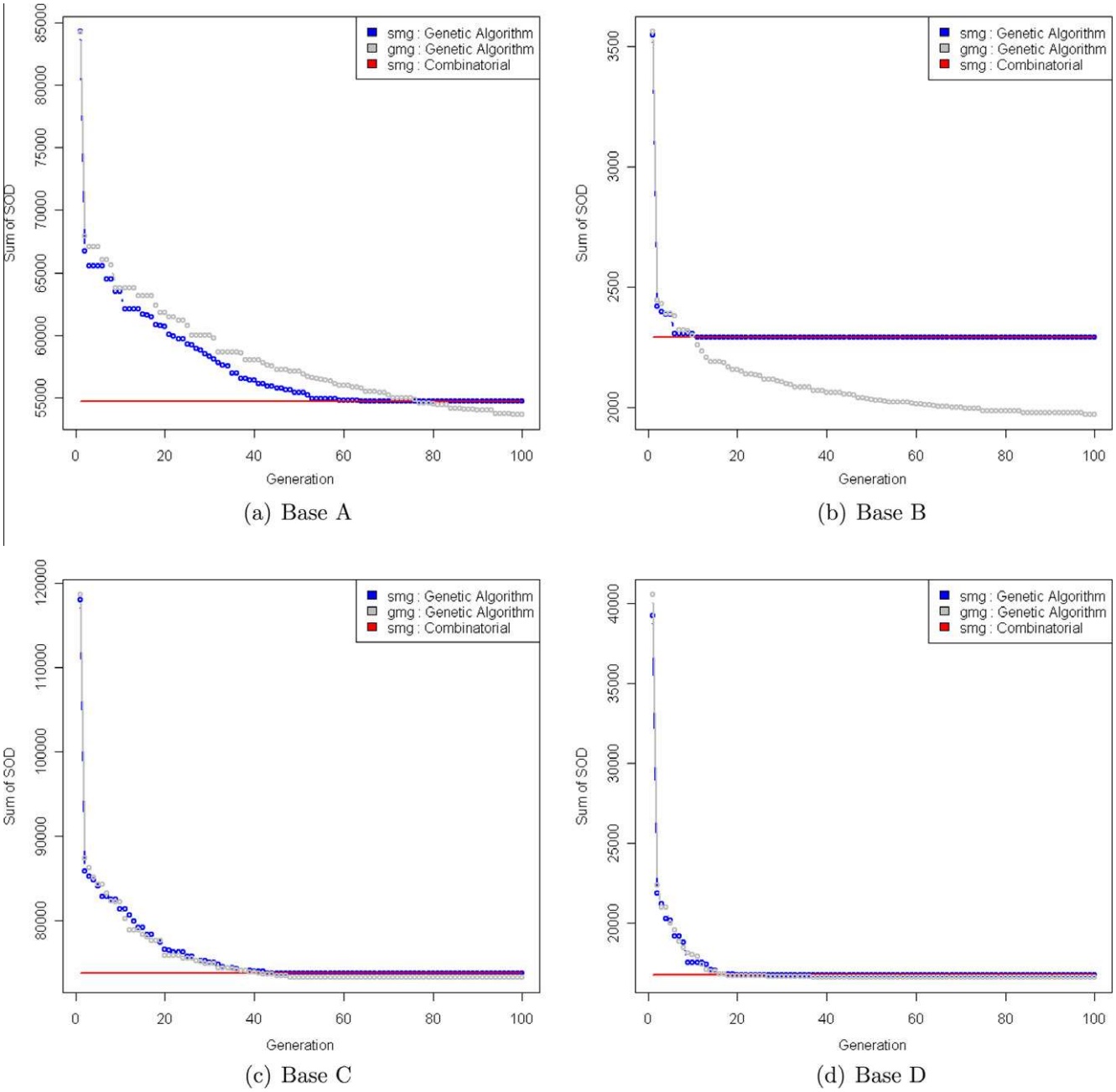


Fig. 10. Evolution of the sum of SOD with respect to the generation number obtained using the proposed genetic algorithm for the computation of *smg* (blue curve) and *gmg* (gray curve) on the four datasets. The red line states for the sum of SOD obtained using a combinatorial approach. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

shown on Fig. 9. In our case, the chosen graph signature imposes the use of nominal labels. Consequently, a 2-Dimensional mesh is applied to achieve the JP and TP discretisation (see the top right of Fig. 9). An experimental study which is not presented in this paper has been used in order to choose mesh granularity.

In order to prove the robustness of such a graph representation against noise, 4 different levels of distortion were introduced in [26]. These distortions are generated by moving each TP or JP randomly within a circle of radius r (given as a parameter for each level) centered at original coordinates of the point. If a JP is randomly moved, all the segments connected to it are also moved. With such distortions, gaps in line segments, missing line segments and wrong line segments are not allowed. Moreover, the number of vertices of each symbol is not changed.

4.1.4. Letter database: Base D

This last database consists of graphs representing distorted letter drawings. It is a slightly modified version of the letter dataset proposed in the IAM graph database repository [43][3] where LOW, HIGH and MED parts of the dataset have been merged. It considers the 15 capital letters of the Roman alphabet that consists of straight lines only (A, E, F, etc.). For each class, a prototype line drawing is manually constructed. To obtain arbitrarily large sample sets of drawings with strong distortions, arbitrarily distortion operators are applied to the prototype line drawings. This results in randomly shifted, removed, and added lines. These drawings are then con-

[3] Available at http://www.greyc.ensicaen.fr/iapr-tc15/.

verted into graphs in a simple manner by representing lines by edges and ending points of lines by vertices. Each vertex is labeled with a two-dimensional attribute giving its position. Since our approach only focuses on nominal attributes, a quantification is performed by the use of a mesh, as in the case of database C. This dataset contains 12,800 graphs, identically distributed among the 15 classes. More information concerning those data are given in Table 1.

4.2. Experimental protocol

The experiments proposed in this section aim at comparing the classification performance which can be reached using the different graph prototypes defined in Section 2. To achieve such a goal, the following protocol has been applied.

First, each dataset has been split into three subsets respectively called training subset (*Tr*), validation subset (*Tv*) and test subset (*Ts*). These subsets are used differently according to the prototypes which are involved.

In the case of using discriminative graphs as prototypes, the training set is used to generate the initial population of the GA, as explained in Section 3.7. Hence, individuals of the first generation are composed of graphs of *Tr*. The validation set *Tv* is involved in the evaluation of the individuals using the 1-NPC classifier during the GA. Finally, the test set is used for evaluating the quality of the best individual (i.e. the best classifier) found at the end of the algorithm. Using such a split, the final performance of the proposed approach is evaluated on a set that has not been considered in the graph prototype learning stage.

In the case of using median graphs as prototypes, the learning process does not involve a classification stage. Consequently, the training and the validation subsets are merged together for medians computation and the test set is used for evaluating the final performance.

Concerning the number m of prototypes to be computed for each class, different values have been tested in the protocol. These values have been chosen with respect to the properties of the dataset.

Furthermore, since GA's are stochastic algorithms, it is necessary to estimate the variability of the results in order to assess the statistical significance of the performance. This was done by running W times the GA and then calculating the conventional couple average and standard deviation $\langle \overline{Rec}, \sigma \rangle$ at the end of the W runs.

Algorithm 2 gives an overview of the whole protocol. The entire experimental session was performed according to the setting described in Table 3, these latter parameters have been chosen experimentaly.

Table 4
A single prototype per class, a comparison.

%	*smg*		*gmg*		*sdg*		*gdg*	
	$\overline{Rec}$	σ	$\overline{Rec}$	σ	$\overline{Rec}$	σ	$\overline{Rec}$	σ
Base A	33.75	0.0	36.00	1.52	66.10	0.981	66.67	1.59
Base B	62.5	0.0	75	0.0	71.42	2.5	83.39	2.5
Base C	86.92	0.0	85.48	2.05	86.58	0.596	90.70	0.59
Base D	69.61	0.0	69.14	0.34	69.67	0.67	71.24	1.47

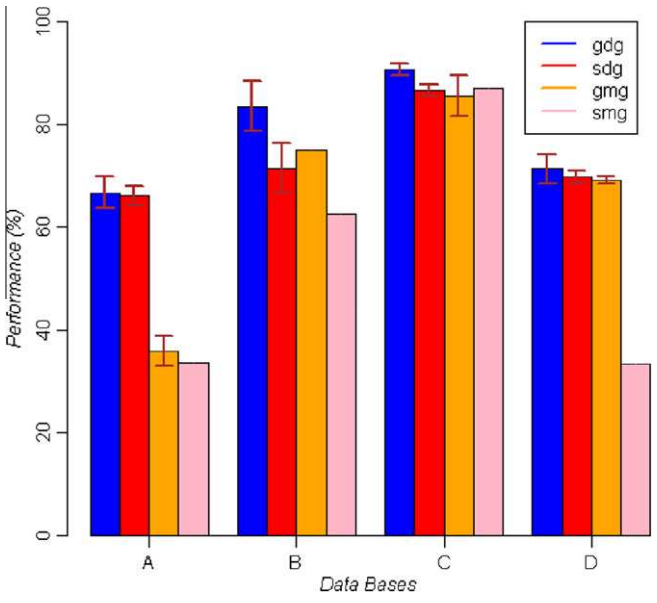


Fig. 11. Recognition rates obtained using a 1-NN rule applied on *Ts* and using *gdg*, *sdg*, *gmg* and *smg* as learning prototypes for the four datasets.

Algorithm 2. Experimental protocol

Require: *Tr*: the training dataset
Require: *Tv*: the validation dataset
Require: *Ts*: the test dataset
Require: *W*: the number of runs
Require: $m[m_{max}]$: the m_{max} values of m to be tested [1]
Require: ga_{param}: GA parameters [2]
ensure: $m_{smg}[m_{max}]$,$\sigma_{smg}[m_{max}]$
ensure: $m_{gmg}[m_{max}]$,$\sigma_{gmg}[m_{max}]$
ensure: $m_{sdg}[m_{max}]$,$\sigma_{sdg}[m_{max}]$
ensure: $m_{gdg}[m_{max}]$,$\sigma_{gdg}[m_{max}]$
  for $j = 1$ to m_{max}
    for $i = 1$ to W
      smg[i][1:j] ← GA(Tr,Tv,$m[j]$,ga_{param}) [3]
      gmg[i][1:j] ← GA(Tr,Tv,$m[j]$,ga_{param}) [3]
      sdg[i][1:j] ← GA(Tr,Tv,$m[j]$,ga_{param}) [3]
      gdg[i][1:j] ← GA(Tr,Tv,$m[j]$,ga_{param}) [3]
      err_{smg}[i] ← err1ppv(Ts,smg[i][1:j])
      err_{gmg}[i] ← err1ppv(Ts,gmg[i][1:j])
      err_{sdg}[i] ← err1ppv(Ts,sdg[i][1:j])
      err_{gdg}[i] ← err1ppv(Ts,gdg[i][1:j])
    end for
    $m_{smg}[j]$ ← mean(err_{smg}[i])
    $\sigma_{smg}[j]$ ← std(err_{smg}[i])
    $m_{gmg}[j]$ ← mean(err_{gmg}[i])
    $\sigma_{gmg}[j]$ ← std(err_{gmg}[i])
    $m_{sdg}[j]$ ← mean(err_{sdg}[i])
    $\sigma_{sdg}[j]$ ← std(err_{sdg}[i])
    $m_{gdg}[j]$ ← mean(err_{gdg}[i])
    $\sigma_{gdg}[j]$ ← std(err_{gdg}[i])
  end for
[1] m values differ according to the considered dataset
[2] Include populationSize, generationNumber, mutationRate and μ
[3] Each GA is specialized to the kind of prototypes to be computed

From this stage, our experiments are organized in a five step methodology. First, a study on set median graph computation is carried out to prove the good convergence of the proposed genetic algorithm. Second, an evaluation of the classification performance that can be reached using *smg*, *gmg*, *sdg* and *gdg* ($m = 1$) as prototypes is performed. Third, we have investigated the influence of m value on the obtained results when multiple prototypes are used for each class. These results are compared to those obtained by a 1-NN classifier trained on the whole learning base ($Tr \cup Tv$), without reduction. Fourth, a closer look is given to the number of classes

impact. Finally, the time complexity is benchmarked though different points of view, the prototype nature and the number of classes.

4.3. Algorithm convergence

In the particular case of computing a single set median graph *smg* for a given class, the problem is computionally feasible and reachable in $O(N^2)$ where N is the number of elements in the given class. Therefore, it is interesting to compare the set median graphs when they are calculated in a computional way and by GA. This test is illustrated in Fig. 10 which reports the sum of the SOD for all classes when the computation is done (i) in a deterministic way (red line) and (ii) when using GA (blue curve for *smg* and gray curve for *gmg*). Results highlight that our algorithm always reaches the global optimum and moreover that few generations are needed to obtain this good performance. In addition, over the four databases, the lowest SODs are achieved by the generalized median graphs. Such a result shows the capacity of our algorithm to build efficient generalized graphs.

4.4. Classification performance with a single prototype

The first classification experiments which have been performed aim at comparing the performance in graph classification obtained on datasets A, B, C, D using an 1-NPC when choosing a single representative per class. The obtained classification rates are reported in Table 4 and illustrated in Fig. 11. Such results lead to several remarks. First of all, regarding all the databases, results obtained by *gmg* are better than those results obtained by *smg*. This latter observation corroborates the idea that *gmg* have a better modeling behavior than *smg*. This observation relies on a straightforward explanation, *gmg* belong to a more complete graph space while

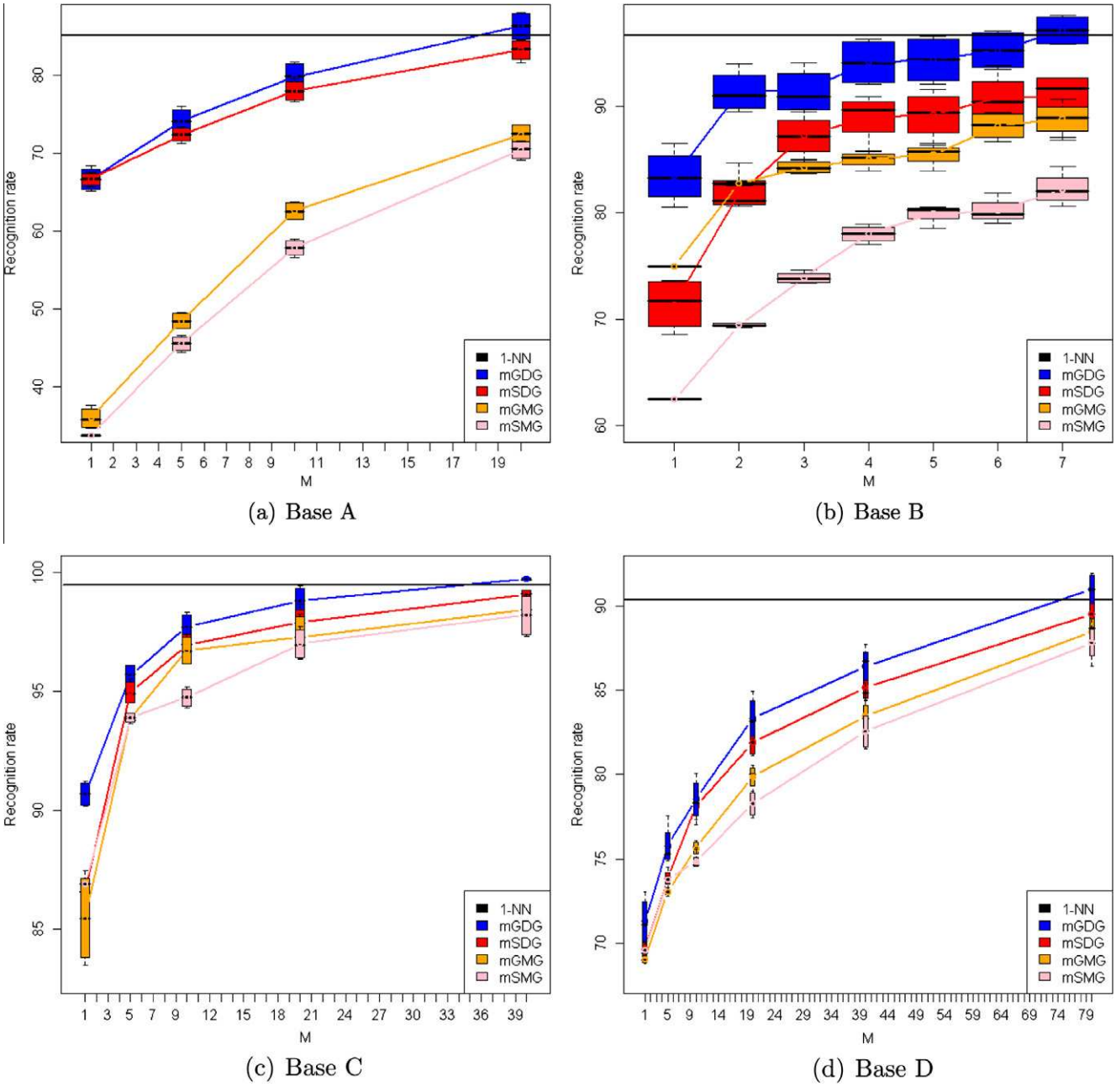


Fig. 12. Recognition rate evolution according to m for each kind of prototypes and on the four datasets.

smg are limited to elements constituting the training dababase. Secondly, another remark states the case that the discriminative approaches outperform the generative ones. This statement relies on the comparisons between (*sdg vs. smg*) and (*gdg vs. gmg*). In both cases, the discriminative graph performance exceed median graph results in a significant way. These important improvements justify to choose *gdg* in order to synthetize a given graph set in a classification context.

4.5. Classification performance with regard to the number of prototypes

This second part of experiments aims at investigating the influence of the number m of prototypes on classification results. The results illustrated in Fig. 12 clearly show that the classification rate is improved when increasing the number of representatives for both median and discriminative graphs. This fact shows that a larger number of prototypes tends to better describe the difficult problems of classification. Also we noticed that the use of a very restricted representative set (i.e. $m = 1$) leads to a lower recognition rate in comparison to the results obtained by a 1-NN classifier trained on the whole learning dataset ($Tr \cup Tv$). However, the time and memory complexities are considerably reduced since there are only N distances to be calculated. Nevertheless, when increasing the number of prototypes, performance match and even outperform the quality of the 1-NN classifier (see Table 5) while maintaining the reduction rate quite high. This trade-off to be made between CPU resources and accuracy gives a solution to tackle the scalability problem and consequently to face large data sets taking fast decisions in the classification stage.

4.6. Impact of the number of classes

Thanks to our synthetic graph generator, the number of classes can be tuned to evaluate the algorithm behavior according to this criterion. In addition, the scalability problem can be addressed reaching a number of classes up to 50. This comparison is presented in Fig. 13. Implicitly, a higher number of classes will lead to a more complicated issue, in such a way that the recognition rate will be deteriorated. When increasing the number of classes, the gap in term of accuracy between modelizing and discriminative graphs is more important. This difference of accuracy starts from 3.68% in the 5-classes problem to reach 21.3% when the number of classes is 50. The higher is the number of classes, the larger is the gap between modelizing and discriminative graphs. This advantage makes discriminative graphs suitable for difficult classification problems. Independently from the number of classes, it is interesting to report the following statements. This test strengthened our prior observations. The *gmg* better modelizes classes than *smg* and *gdg* outperform all the others prototypes over the four subsets.

4.7. Time complexity analysis

As a matter of fact, learning algorithms are performed off-line. In such a configuration, it seems reasonable to mention that time complexity is not a crucial issue. It is much more significant to be fast at the decision stage. However, a way to compare the computational cost of the different types of prototypes was to undertake an empirical study. The algorithm complexity is directly linked to the number of classes, the influence of the dataset size is depicted by the Fig. 14. A first comment illustrates the strong impact of the class number on the computational cost when producing a discriminative graph. Moreover a comparison of the runtime execution according to the kind of prototypes on the largest data-

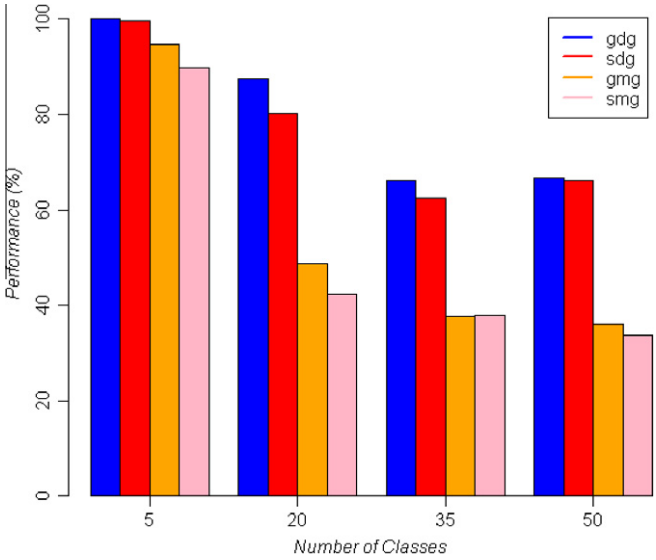


Fig. 13. Performance comparison between the different kinds of prototypes with respect to the number of classes on different subsets of the database A.

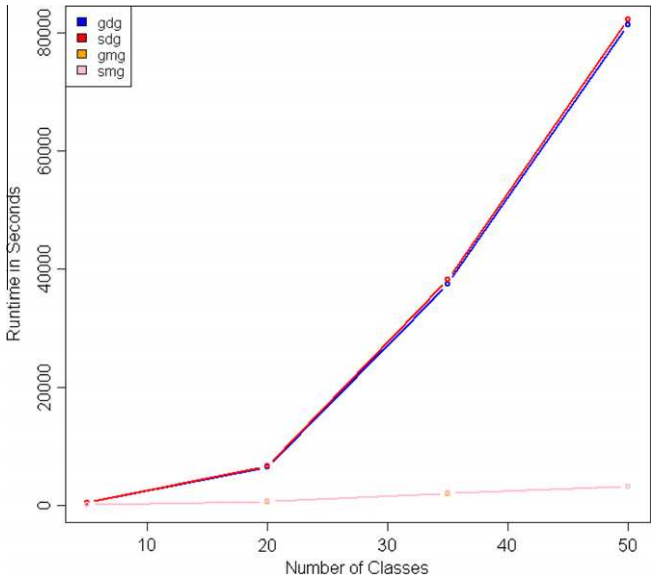


Fig. 14. Run-time evolution with respect to the number of classes on different subsets of the database A.

Table 5
Reduction rate and performance comparisons between *gdg* and a 1-NN classifier using the entire learning set $Tr \cup Tv$. Reduction rate stands for $100 - \frac{m \times N}{|Tr \cup Tv|}$.

	Base A		*Base B*		*Base C*		*Base D*	
	gdg	1-NN	*gdg*	1-NN	*gdg*	1-NN	*gdg*	1-NN
Reduc. rate (%)	92.92	0	50.71	0	86.67	0	76.3	0
Rec (%)	86.34	85.16	97.14	96.43	99.71	99.47	91.04	90.16

base has been led. The complexity of the median graph search came out from this test. The SOD criterion is less demanding in term of distance computation, therefore, it is less time consuming. At worst case, in our experiments on the largest database, the median graph computation was 15 times faster. However, this overload does not discourage the use of discriminative graphs since the gain they imply is really significant. It is a commonplace in machine learning to state the case that training algorithms require much time and many computations to assimilate the data variability.

5. Conclusion and future works

This paper has presented several approaches for the construction of prototype-based structural classifiers. These approaches have been experimentally compared according to several criteria on both synthetic and real databases.

The experimental results first confirm that the generalized median graph approximated using a genetic algorithm has a better modeling ability than the set median graph. Moreover, the results show that prototypes which take into account the whole classification problem (discriminative approach) offer better results than the class centered median graph approach.

Furthermore, the proposed GA framework allows to synthetize m graph prototypes per class. The experimental results illustrate that, when m increases, the classification problem is better described and the performance improves and converges rapidly towards the classification rate of a 1-NN classifier applied on the whole learning dataset.

Finally, the assessments carried out on four datasets expressed that *gdg* and *m-gdg* obtain better or comparable results, in terms of accuracy, than the state-of-the-art prototypes schemes for structural data on multi-class graph classification problem. Our contribution gives the proof for the following key points: (i) genetic algorithms are well suited to deal with graph structures and (ii) the recognition rate on a validation dataset is a better criterion of the optimization process than a classical SOD in a classification context. Also, the scalability to large graph datasets has been assessed on a synthetic database with success. This observation illustrates that a prototype-based classifier is well suited to manage masses of structural data.

Short-term, we intend to investigate the ability of setting a different number of prototypes for each class. This strategy would allow to distribute a global number of prototypes among the classes and then to automatically fit the difficulty of the classification problem. This modification impacts on the algorithm and requires a redefinition of the genetic algorithm (problem coding and genetic operators).

We also intend to investigate the ability to propose several prototype sets for different values for m. These sets would correspond to different trade-offs between the concurrent objectives that are the recognition rate and the reduction of the training set which allows to reduce the classification time and spatial complexity. A multi-objective procedure [44] could be used to optimize these non commensurable criterions. Finally, a human operator would *a posteriori* make the final decision according to the use case.

Finally, the reject of elements which do not belong to any known class is a feature which is often required when classifiers are faced with actual data. When dealing with Nearest Neighbor rule, it is generally implemented throught the definition of threshold values. In the same time, the reject of an element is often prefered to a misclassification. This kind of feature can be undertaken with k nearest neighbors rules with values of k greater than 1. Future works should be dedicated to include reject consideration as an additional criterion to be optimized while maintaining the classification rate as high as possible. In this case again, a multi-objective procedure could be useful.

References

[1] M. Kuramochi, G. Karypis, Finding frequent patterns in a large sparse graph, Data Mining and Knowledge Discovery 11 (3) (2005) 243–271.
[2] A. Inokuchi, T. Washio, H. Motoda, Complete mining of frequent patterns from graphs: mining graph data, Machine Learning 50 (3) (2003) 321–354.
[3] H. Zanghi, C. Ambroise, V. Miele, Fast online graph clustering via erdos renyi mixture, Pattern Recognition 41 (12) (2008) 3592–3599.
[4] H. Qiu, E.R. Hancock, Graph matching and clustering using spectral partitions, Pattern Recognition 39 (1) (2006) 22–34.
[5] S. Auwatanamongkol, Inexact graph matching using a genetic algorithm for image recognition, Pattern Recognition Letters 28 (12) (2007) 1428–1437.
[6] M. Neuhaus, H. Bunke, Inexact graph matching using a genetic algorithm for image recognition, Pattern Recognition 39 (10) (2006) 1852–1863.
[7] M.A. Lozano, F. Escolano, Protein classification by matching and clustering surface graphs, Pattern Recognition 39 (4) (2006) 539–551.
[8] H. Kashima, K. Tsuda, A. Inokuchi, Marginalized kernels between labeled graphs, in: Proceedings of the Twentieth International Conference on Machine Learning, 2003, pp. 321–328.
[9] H. Kashima, K. Tsuda, A. Inokuchi, Kernel for graph, in: Kernel Methods in Computational Biology, 2004, pp. 155–170.
[10] F. Suard, V. Guigue, A. Rakotomamonjy, A. Bensrhair, Pedestrian detection using stereovision and graph kernels, in: Proceedings of the IEEE Intelligent Vehicle Symposium, 2005, pp. 267–272.
[11] P. MahT, N. Ueda, T. Akutsu, J.-L. Perret, J.-P. Vert, Extensions of marginalized graph kernels, in: Proceedings of the Twenty-First International Conference on Machine Learning, 2004, pp. 552–559.
[12] P. MahT, N. Ueda, T. Akutsu, J.-L. Perret, J.-P. Vert, Graph kernels for molecular structure-activity relationship analysis with support vector machines, Journal of Chemical Information and Modeling 45 (4) (2005) 939–951.
[13] S.V.N. Vishwanathan, N.N. Schraudolph, R. Kondor, K. Borgwardt, Graph kernels, Journal of Machine Learning Research 11 (2010) 1201–1242.
[14] W.Y. Chen, W.L. Hwang, T.C. Lin, Planar-shape prototype generation using a tree-based random greedy algorithm, IEEE Transactions on Systems, Man, and Cybernetics, Part B 36 (3) (2006) 649–659.
[15] B.V. Dasarathy, Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques, IEEE Computer Society Press, Los Alamitos, 1990.
[16] P.E. Hart, The condensed nearest neighbour rule, IEEE Transactions on Information Theory 14 (5) (1968) 515–516.
[17] C.-L. Chang, Finding prototypes for nearest neighbor classifiers, IEEE Transactions on Computers 23 (11) (1974) 1179–1184.
[18] J.C. Bezdek, T.R. Reichherzerand, G.S. Lim, Y. Attikiouzel, Multiple-prototype classifier design, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 28 (1) (1998) 67–79.
[19] J. Jia, K. Abe, Automatic generation of prototypes in 3d structural object recognition, in: ICPR '98: Proceedings of the 14th International Conference on Pattern Recognition, vol. 1, 1998, p. 697.
[20] A. Torsello, E.R. Hancock, Learning shape-classes using a mixture of tree-unions, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (6) (2006) 954–967.
[21] B. Bonev, F. Escolano, M.A. Lozano, P. Suau, M. Cazorla, W. Aguilar, Constellations and the unsupervised learning of graphs, in: GbRPR, 2007, pp. 340–350.
[22] H. Bunke, P. Foggia, C. Guidobaldi, M. Vento, Graph clustering using the weighted minimum common supergraph, in: GbRPR, 2003, pp. 235–246.
[23] S. Marini, M. Spagnuolo, B. Falcidieno, Structural shape prototypes for the automatic classification of 3d objects, IEEE Computer Graphics and Applications 27 (4) (2007) 28–37.
[24] X.J.H. Bunke, A. Mnnger, Combinatorial search versus genetic algorithms: a case study based on the generalized median graph problem, Pattern Recognition Letters 20 (11) (1999) 1271–1277.
[25] X. Jiang, A. Mnnger, H. Bunke, On median graphs: Properties, algorithms, and applications, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (10) (2001) 1144–1151.
[26] M. Ferrer, E. Valveny, F. Serratosa, Spectral median graphs applied to graphical symbol recognition, in: CIARP, 2006, pp. 774–783.
[27] M. Ferrer, F. Serratosa, E. Valveny, On the relation between the median and the maximum common subgraph of a set of graphs, in: GbRPR, 2007, pp. 351–360.
[28] A. Hlaoui, S. Wang, Median graph computation for graph clustering, Soft Computing – A Fusion of Foundations, Methodologies and Applications 10 (1) (2005) 47–53.
[29] E. Valveny, P. Dosch, Symbol recognition contest: a synthesis, in: J. Lladós, Y.B. Kwon (Eds.), Selected Papers of the 5th IAPR International Workshop on Graphics Recognition, Lecture Notes in Computer Science, vol. 3088, Springer-Verlag, 2004, pp. 368–385.
[30] L.I. Kuncheva, Editing for the k-nearest neighbors rule by a genetic algorithm, Pattern Recognition Letters 16 (8) (1995) 809–814.
[31] D.E. Goldberg (Ed.), Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, 1989.
[32] C.R. Reeves (Ed.), Modern Heuristic Techniques for Combinatorial Problems, Blackwell Scientific Press, 1993 (Chapter: Genetic Algorithms, pp. 151–196).
[33] H. Bunke, On a relation between graph edit distance and maximum common subgraph, Pattern Recognition Letters 18 (8) (1997) 689–694.

[34] X. Gao, B. Xiao, D. Tao, X. Li, A survey of graph edit distance, Pattern Analysis and Applications 13 (1) (2010) 113–129.
[35] K. Riesen, H. Bunke, Approximate graph edit distance computation by means of bipartite graph matching, Image Vision Computing 27 (7) (2009) 950–959.
[36] H. Bunke, K. Shearer, A graph distance metric based on the maximal common subgraph, Pattern Recognition Letters 19 (3–4) (1998) 255–259.
[37] W.D. Wallis, P. Shoubridge, M. Kraetz, D. Ray, Graph distances using graph union, Pattern Recognition Letters 22 (6–7) (2001) 701–704.
[38] D.P. Lopresti, G.T. Wilfong, A fast technique for comparing graph representations with applications to performance evaluation, IJDAR 6 (4) (2003) 219–229.
[39] P. Erdos, A. RTnyi, On random graphs, Publicationes Mathematicae Debrecen 6 (1959) 290–297.
[40] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, 1990.
[41] A. Khotanzad, Y.H. Hong, Invariant image recognition by zernike moments, IEEE Transactions on PAMI 12 (5) (1990) 489–497.
[42] E. Barbu, P. Heroux, S. Adam, E. Trupin, Clustering document images using a bag of symbols representation, in: Proceedings of the 8th International Conference on Document Analysis and Recognition, 2005, pp. 1216–1220.
[43] K. Riesen, H. Bunke, Iam graph database repository for graph based pattern recognition and machine learning, in: SSPR & SPR '08: Proceedings of the 2008 Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, 2008, pp. 287–297.
[44] K. Deb, Multi-Objective Optimization Using Evolutionary Algorithms, John Wiley & Sons, Inc., New York, NY, USA, 2001.

Annexe F

Réference CV : 25

H. Locteau, R. Raveaux, S. Adam, Y. Lecourtier, P. Héroux, and E. Trupin. Approximation of digital curves using a multi-objective genetic algorithm. In Proceedings of the International Conference on Pattern Recognition (ICPR'06), pages 716-719. 2006.

Approximation of Digital Curves using a Multi-Objective Genetic Algorithm

Hervé Locteau, Romain Raveaux, Sébastien Adam, Yves Lecourtier, Pierre Héroux, Eric Trupin
LITIS Labs – University of Rouen, FRANCE
Herve.Locteau@univ-rouen.fr

Abstract

In this paper, a digital planar curve approximation method based on a multi-objective genetic algorithm is proposed. In this method, the optimization/exploration algorithm locates breakpoints on the digital curve by minimizing simultaneously the number of breakpoints and the approximation error. Using such an approach, the algorithm proposes a set of solutions at its end. The user may choose his own solution according to its objective. The proposed approach is evaluated on curves issued from the literature and compared successfully with many classical approaches.

1. Introduction

Approximation of digital planar curves using vertices and/or circular arcs is an important issue in pattern recognition and image processing. It is a classical way to represent, store and process digital curves. For example, approximation results are frequently used for shape recognition. The problem can be stated as follows: Given a curve $C = \{C_i \equiv (x_i, y_i)\}_{i=1}^{N}$ constituted of N ordered points, the goal is to find a subset $S = \{S_i \equiv (x_i, y_i)\}_{i=1}^{M}$ of M ordered points and the corresponding parameter set $P = \{P_i \equiv (xc_i, yc_i)\}_{i=1}^{M}$. S contains the extremities of the line segments or the circular arcs (sometimes called breakpoints) and P the parameters of the best approximation of the set of points between each couple of breakpoints (a specific value is applied in the case of segment)

Whereas many paradigms have been proposed to solve the problem of polygonal approximation or the problem of approximation with circular arcs, much less papers were proposed concerning the approximation of digital curves with both representations. Among the existing papers [1][2][3][4], an approach recently proposed in [4] consists in using **G**enetic **A**lgorithms (GA) in order to find a near-optimal approximation. In such a case, the approximation of digital curves is considered as an optimization process. The algorithm automatically selects the best points of the curves by minimizing a given criterion. In [4], the number N of breakpoints to be obtained is fixed and the method uses the concept of genetic evolution to obtain a near-optimal approximation.

In this paper, we adopt the same paradigm and we propose a new GA for the approximation of digital curves. The originality of the described approach is the use of a multi-objective optimization process. Such a new viewpoint enables the user of the system to choose a trade-off between different quality criteria after a single run of the GA.

The remainder of the paper is organized as follows. In section 2, an introduction to the multi-objective optimization problem is proposed and our algorithm is presented. In section 3, the application of this algorithm to the approximation problem is shown. Section 4 presents the experimentally obtained results and a comparison with existing approaches. Section 5 gives the concluding remarks.

2. Multi objective optimization GA

When an optimization problem involves more than one objective function, the task of finding one or more optimum solutions is known as multi-objective optimization. Some classical textbooks on this subject have been published, e.g. [5]. We just recall here some essential notions in order to introduce the proposed algorithm. The main difference between single and multi-optimization tasks lies in the requirement of compromises between the various objectives in the multi-optimization case. Even with only two objectives, if they are conflicting, the improvement of one of them leads to a deterioration of the other one. For example, in the context of polygonal approximation, the decrease of the approximation error always leads to an increase of the vertices number. Two main approaches are used to overcome this

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

problem in the literature. The first one consists in the combination of the different objectives into a single one (the simpler way being to use a linear combination of the various objectives), and then to use one of the well-known techniques of single objective optimization (like gradient based methods, simulated annealing or classical genetic algorithm). In such a case, the compromise between the objectives is a priori determined through the choice of the combination rule. The main critic addressed to this approach is the difficulty to choose a priori the compromise. It seems a better idea to postpone this choice after having several candidate solutions at hand. This is the goal of Pareto based method using the notion of dominance between candidate solutions. A solution dominates another one if it is better for all the objectives. This dominance concept is illustrated on figure 1. Two criteria J_1 and J_2 have to be minimized. The set of non-dominated points that constitutes the Pareto-Front appears as 'O' on the figure, while dominated solutions are drawn as 'X'.

Using such a dominance concept, the objective of the optimization algorithm becomes to determine the Pareto front, that is to say the set of non-dominated points. Among the optimization methods that can be used for such a task, genetic algorithms are well-suited because they work on a population of candidate solutions. They have been extensively used in such a context. The most common algorithms are VEGA – Vector Evaluated Genetic Algorithm – [6], MOGA – Multi Objective Genetic Algorithm –approach [7], NSGA – Non-Dominated Sorting Genetic Algorithm – [8], NSGA II [9], PAES – Pareto Archived Evolution Strategy – [10] and SPEA – Strength Pareto Evolutionary Algorithm – [11]. The strategies used in these contributions are different, but the obtained results mainly vary from the convergence speed point of view. A good review can be found in [12].

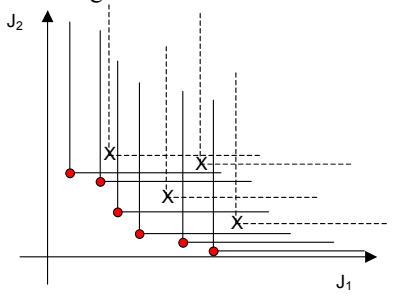


Fig. 1. Illustration of the Pareto Front concept

The proposed genetic algorithm is elitist and steady-state. This means that (i) it manages two populations and (ii) the replacement strategy of individuals in the populations is not made as a whole, but individual per individual. The two populations are a classical population, composed of evolving individuals and an "archive" population composed of the current Pareto Front elements. These two populations are mixed during the iterations of the genetic algorithm. The first population guarantees space exploration while the archive guarantees the exploitation of acquired knowledge and the convergence of the algorithm.

Based on such concepts, our optimization method uses the following algorithm:

Population (I) and Archive (A) Initialization
do
- Random selection of two individuals I_1 and I_2 in (I)
- Crossover between I_1 and I_2 to generate I_3 and I_4
- Mutation applied to the generated children I_3 and I_4
- Evaluation of children I_3 and I_4
- Selection either of the dominant individual I_5 between mutated children (if it exists) or random selection of I_5 between I_3 and I_4
- Random selection of (I_6) a in (A)
- Crossover between I_5 and I_6 to generate I_7 and I_8
- Evaluation of children I_7 and I_8
- Test for the integration of I_7 and I_8 in (A)
- Test for the integration of I_7 and I_8 in (I)

While i < the maximal generation number

This algorithm has been designed in order to be applied to various problems. The design of a new application consists in the choice of a coding scheme for individuals, in the design of the evaluation method and in the choice of both parameters values and of some specific operators. In its current implementation, the coding of an individual is a classical bit string. Crossover is a well-known 2-points crossover whereas initialization and mutation are application-dependent. Concerning the replacement strategy, several choices can be made for the integration of a candidate individual in the archive. The simplest is a dominance test between the candidate and the archive elements. The candidate is inserted within the archive if no archive element dominates it. In the same time, archive elements dominated by the candidate are eliminated from the archive. A problem reported in the literature on evolutionary multi-objective optimization is the possible bad exploration of Pareto front: the archive population elements concentrate on only some parts of the front. This difficulty is overcome in our approach by defining a minimal distance between two points in the objective space. This algorithm has been tested on classical multi-objective problems such as BNH or TNK [13]. The obtained results have shown the quality of the proposed approach since it is able to find a similar approximation of the Pareto Front for the same number of calls to the evaluation function.

3. Application to curve approximation

In order to apply the algorithm presented above to the curve approximation problem, an individual has to represent a possible solution to the approximation problem. That is why an individual is composed of N genes, where N is the number of points in the initial curve. A gene is set to '1' if the point is kept as a breakpoint, '0' if it is not. An example of an individual coding is given in figure 2. Each point C_i of the curve S corresponds to a bit in the chromosome. In the example of figure 2, the individual is a binary string of 45 genes corresponding to the initial C_1-C_{45}.. The approximation is composed of 2 line-segments and 6 circular arcs. The corresponding breakpoints are respectively C_3, C_5, C_{20}, C_{29}, C_{35}, C_{37}, C_{41}, and C_{44}. Such an approximation (the optimal approximation for 8 breakpoints) corresponds to the individual "001010000000000000010000 000010000010100010010".

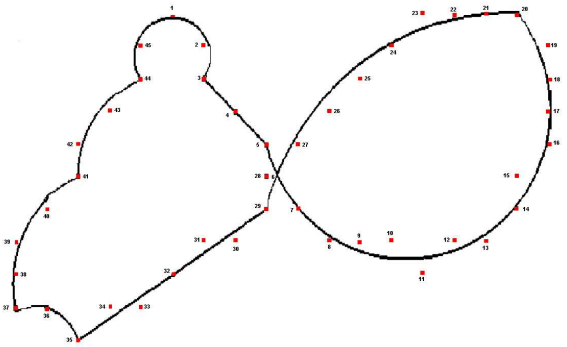


Fig. 2: An example of the coding scheme

Using such a coding scheme, the GA described in section 2 is applied. In order to reduce the number of iteration in the GA, a specific initialization operator is used. It is based on a simple analysis of the curve to be approximated. An histogram of the curvature along the curve is first computed. During initialization, for each point, a probability to be selected is deduced from this histogram. This strategy enables to avoid the selection of collinear points and on the contrary enables to select points with high curvature. A specific mutation operator is also used. It is based on the shift of a selected point to the preceding or the next one.

Concerning the criteria to be optimized, two objectives have been included in the current version. The first one is the Integral Square Error (ISE) and the second one is the number of points. This enables to have a trade-off between the precision of the result and the number of line segments, thanks to elements of the Pareto front. One can note that the use of a discrete objective (vertices number) guarantees itself the diversity on the Pareto front, we do not need to specify any minimal distance between any couples of solutions of the Pareto Front. For the computation of the ISE, the error is computed both in the case of line-segments and circular arcs and the best solution is kept as P_i. Circular arcs are obtained using a LMS approach [4].

4. Experimental results

In order to assess the performances of the proposed algorithm, it has been applied to the four broadly used digital curves presented in [14] and proposed in Fig. 3.

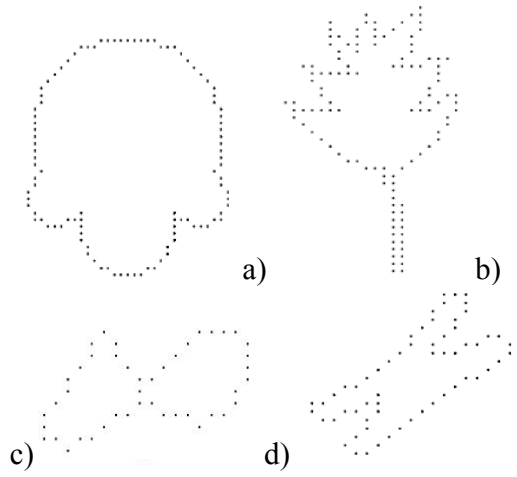


Fig. 3. The four digital test curves

Such tests allow to test the performances of the proposed algorithm versus those of published approaches. For each of these curves, the program has been run for 2000 generations, using a population size of 100 individuals. Such a parameter set involves about 8000 calls to the evaluation method (see the algorithm below). The mutation rate has been fixed to 0.05 and the crossover rate to 0.6. As said before, the output of the presented algorithm is not a single ISE for a number of vertices given a priori. It consists in the whole Pareto front of the optimization problem. That is why the result is a set of couple (ISE – number of vertices). As an example, figure 4 shows the set of couple obtained at the end of the algorithm applied on the "semicircle" curve. Another remark has to be done. Since GA are stochastic, results may be different at independent runs. That is why, in these experiments, we give (table 1) both the best (B) and the worst (W) ISE for each number of vertices obtained after 5 independent runs on each curve. The obtained results can be compared with the results of table 2 issued from an existing comparative study [4]. As one can see on table 1 and 2, results obtained using the GA approach

enables to obtain competitive results. Moreover, theses tables also show the stability of the proposed approach since best (B) and worst (W) results are generally the same for the 5 runs.

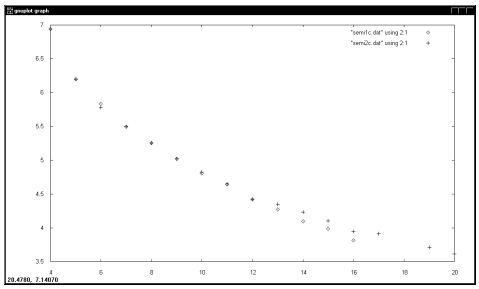


Fig. 4: Two obtained approximations

Table 1 : Results obtained using the GA

Fig 3a			Fig 3b			Fig 3c			Fig 3d		
N	B	W	N	B	W	N	B	W	N	B	W
4	6.9	6.9	12	43.5	43.9	5	5.2	5.9	9	4.6	4.6
5	6.1	6.1	14	22.1	22.7	6	3.0	3.4	10	2.4	2.4
6	5.7	5.8	16	10.7	10.7	7	2.6	2.8	11	1.9	2.0
7	5.4	5.7	18	7.3	7.4	8	2.3	2.3	12	1.6	1.7
8	5.2	5.2	25	3.2	3.3	9	1.9	1.9	13	1.4	1.4
12	4.2	4.4	27	2.9	3.0	10	1.5	1.5	14	1.2	1.2
14	3.8	4.0	29	2.7	2.8	11	1.2	1.3	15	1.0	1.1
22	2.3	2.4	31	2.6	2.8	13	0.7	0.8	16	0.9	1.0

Table 2 : Best results found in the literature for the approximation of the curves of figure 3

Fig 3a		Fig 3b		Fig 3c		Fig 3d	
N°	ISE	N°	ISE	N°	ISE	N°	ISE
4	6.9	16	10.9	6	3.0	10	2.6
6	6.4	18	7.4	8	2.3	11	2.1
12	10.9	27	8.8	9	2.0	15	1.2
14	17.7	29	14.9	13	5.9		
22	20.6	31	1.6				

5. Conclusion and future works

In this paper, we have proposed a new approach for the approximation of curves. This approach is inspired from previous approaches in the way that it considers the polygonal approximation as an optimization process. The fundamental difference with the existing approaches lies in the fact that we use a multi-objective optimization process while other contributions only optimize a unique objective, that is to say the ISE. One can see several interests in such an approach. As many solutions are proposed, the user or the system may choose the optimal solution regarding its constraints. Another interest is to offer the possibility to add a new objective easily. As an example, such an approach may be used for the vectorization of shape contours by adding a parallelism constraint.

7. References

[1] C. Ichoku, B. Deffontaines and J. Chorowicz, "Segmentation of digital plane curves: a dynamic focusing approach", Pattern Recognition Letters, 17, 1996, pp 741–750.

[2] P.L. Rosin and G.A.W. West, "Nonparametric segmentation of curves into various representations", IEEE Trans. Pattern Anal. Machine Intell., 17, 1995, pp 1140-1153.

[3] J-H. Horng and J.T. Li, "A dynamic programming approach for fitting digital planar curves with line segments and circular arcs", Pattern Recognition Letters, 22, 2001, pp 183–197.

[4] B. Sarkar, L.K. Singh and D. Sarkar, "Approximation of digital curves with line segments and circular arcs using genetic algorithms", *Pattern Recognition Lett.* 24, 2003, 2585-2595.

[5] K. Deb, "Multi-Objective optimization using Evolutionary algorithms", Wiley, London, 2001.

[6] J.D. Schaffer and J.J. Grefenstette, "Multiobjective learning via genetic algorithms", *In Proceedings of the 9th international joint conference on artificial intelligence*, Los Angeles, California, pp 593-595, 1985.

[7] C.M. Fonseca, P.J. Fleming, "Genetic algorithm for multi-objective optimization: formulation, discussion and generalization", In Stephanie editor, *Proceedings of the fifth international conference on genetic algorithm*, San Mateo, California, pp 416-423, 1993.

[8] N. Srinivas, K. Deb, "Multiobjective optimization using nondominated sorting in genetic algorithm", *Evolutionary Computation* 2, 1994, pp 221-248.

[9] K. Deb, S. Agrawal, A. Pratab and T. Meyarivan, "A fast and elitist multi-objective genetic algorithm: NSGA-II", *IEEE Transactions on Evolutionary Computation* 6, 2000, pp 182-197.

[10] J.D. Knowles, D.W. Corne, "Approximating the nondominated front using the Pareto archived evolution strategy", *Evolutionary computation* 8, 2000, pp 149-172.

[11] E. Zitzler, L. Thiele, "Multiobjective evolutionary algorithms : a comparative study and the strength pareto approach", *IEEE Transactions on Evolutionary Computation* 3, 1999, pp 257-271.

[12] C. A. Coello Coello, "A short tutorial on Evolutionary Multiobjective Optimisation", In Eckart Zitzler, Kalyanmoy Deb, Lothar Thiele, Carlos A. Coello Coello and David Corne (editors), *First International Conference on Evolutionary Multi-Criterion Optimization,* Lecture Notes in Computer Science, . Springer-Verlag n° 1993, pp 21-40, 2001.

[13] D. Chafekar, J. Xuan, K. Rasheed, "Constrained Multi-objective Optimization Using Steady State Genetic Algorithms", *In Proceedings of Genetic and Evolutionary Computation Conference*, Chicago, Illinois, pp 813-824, 2003.

[14] R.T. Teh and Chin, "On the detection of dominant points on digital curves", *IEEE transaction on Pattern Analysis and Machine Intelligence* 23 , 1989, pp 859-872.