Identification de documents par classification monoclasse

Nicolas Sidère* — **Jean-Yves Ramel**** — **Sabine Barrat**** — **Vincent Poulain D'Andecy***** — **Saddok Kebairi*****

* *INSA Centre Val de Loire, LI EA 6300, 41000 Blois, France*
** *Université François Rabelais Tours, LI EA 6300, 37200 Tours, France*
*** *IteSoft, Parc d'Andron, Le Séquoia, 30470 Aimargues, France*

RÉSUMÉ. *Utilisée dans un contexte industriel, la classification d'images de documents nécessite le respect de certaines contraintes; par exemple, être confronté à une grande variabilité des documents et/ou du nombre de classes. Dans cet article, nous répondons à ce problème en présentant une nouvelle approche basée sur la spécialisation du vecteur de caractéristiques et d'un classificateur pour chaque classe, contrairement à la majorité des méthodes qui traitent l'ensemble des classes. Cette approche permet alors d'introduire de nouvelles classes sans contraindre le système à un nouvel apprentissage. Pour cela, nous calculons un vecteur de caractéristiques générique qui sera ensuite spécialisé en classant les caractéristiques selon un score de stabilité. Finalement, un classificateur monoclasse de type K plus proche voisins est entrainé en utilisant ce vecteur. Les expérimentations menées révèlent de bons taux de classification prouvant une adaptabilité de notre système sur des problèmes complexes.*

ABSTRACT. *Document image classification in an industrial context requires to respect some constraints such as dealing with a large variability of documents and/or number of classes. In this article, we answer this problem by presenting a new methodology focused on an idea of specializing the features and the classifier for each class, whereas most methods deal with all classes at the same time. The benefit of this approach is to enable the industrial system to introduce a new class without re-training the current classifier. We first compute a generalized vector of features in order to specialize it by ranking the features according a stability score. Finally, a one-class K-nn classifier is trained by using the specific features for a chosen class. Conducted experiments reveal good classification rates proving the ability of our system to deal with a large range of classes of documents.*

MOTS-CLÉS : *Classification d'images de document, Classification monoclasse.*

KEYWORDS: *Document image classification, One-class classification.*

CIFED 2014, pp. 277–290, Nancy, 18-21 mars 2014

1. Introduction

Aujourd'hui, beaucoup d'applications viennent assister l'être humain dans des tâches fastidieuses telles que la distribution de courrier entrant, l'archivage de documents ou encore la recherche dans des fonds documentaires.

Au cœur de ces applications, la classification d'images de document, *ie* attribuer une classe – un ensemble prédéfini de documents présentant une homogénéité – à l'image d'un document, est encore un domaine de recherche actif ((Saund, 2011)). Dans un contexte industriel, la conception d'un système complet de classification des documents peut rencontrer des difficultés telles que :

– la diversité des tâches à accomplir (automatisation des tâches bureautiques, la lecture optique des formulaires, les bibliothèques numériques, ...),

– la composition et l'utilisation des données d'apprentissage destinées à construire les modèles de classes (faible nombre d'exemples, mauvaise représentativité de l'espace, ...),

– le choix des caractéristiques (bas et/ou haut niveau sémantique) utilisées pour la comparaison du contenu des images,

– la robustesse des traitements effectués sur l'image pour extraire ces caractéristiques,

– la maintenance du système lorsqu'il s'agit d'introduire une nouvelle classe après plusieurs mois de mise en production (évolution continue).

Partant de ce constat, l'objectif principal est de définir un système de classification adaptatif, ce qui apparaît comme un véritable défi en raison de la grande variété de documents, des divers critères utilisés pour définir les classes de documents et l'ambiguïté pouvant être liée à une mauvaise définition (ou floue) des classes de documents. Cette idée de généricité est souvent opposé à la performance – en terme de taux de classification – en particulier lorsque le nombre d'exemples disponibles pour l'apprentissage est faible. Nous pensons que cette contrainte doit être prise en compte à chaque étape du système, *ie* de l'extraction de caractéristiques qui caractériseront le document à la classification qui sélectionne la classe de documents la plus similaire.

Ces problématiques sont majoritairement résolues par l'utilisation de techniques de classification dites supervisées. L'application de tels algorithmes pour un problème réel dans un contexte industriel présente deux inconvénients majeurs :

– le nombre de classes doit être connu *a priori*. Les algorithmes incrémentaux qui permettent ces évolutions existent ((Bouguelia *et al.*, 2013)) mais sous réserve d'un paramétrage fastidieux ou limité à l'utilisation de vecteurs de caractéristiques de taille fixe ; les classes nécessitant de nouvelles caractéristiques ne peuvent donc pas être intégrées.

– l'ensemble d'apprentissage doit avoir une taille conséquente pour entrainer le classificateur et obtenir des taux de classification intéressant.

Ces deux conditions sont d'autant plus contraignantes que le nombre de documents à classifier est grand. Il est donc obligatoire de pouvoir construire une vérité terrain complète. Cette opération demande souvent beaucoup de temps, ce qui a pour conséquence de limiter l'utilisation des méthodes de classification supervisée classiques dans un contexte industriel.

Dans cet article, nous nous concentrons sur le développement d'un système complet de classification : de la caractérisation d'une image de document, au choix du classificateur ainsi que son paramétrage. Pour respecter la généricité et la capacité du système à évoluer selon la diversité des problèmes du monde réel qui sont rencontrés par les éditeurs de logiciels, notre idée principale est de proposer des approches génériques associées à des méthodes de spécialisation pour pouvoir à la fois s'adapter à toutes les problématiques et conserver robustesse et précision. Ainsi, dans la section 2 nous présentons d'abord le vecteur de caractéristiques que nous proposons et la méthode associée destinée à spécialiser le vecteur selon les données et la classe à traiter. Deuxièmement, dans la section 3, nous décrivons un classificateur K-ppv monoclasse capable de généraliser le modèle de classe à partir d'échantillons d'apprentissage issus d'une unique classe. Ensuite, nous montrons les expérimentations que nous avons menées sur une base de données privée issue d'un problème industriel réel. Enfin, la section 5 conclut le papier et propose quelques perspectives.

2. Le vecteur de caractéristiques

Le choix des caractéristiques étant une étape très importante dans la conception d'un système de classification, une multitude de travaux a été consacrée à ce problème. Ces méthodes peuvent être divisées en deux catégories principales :

– Les méthodes utilisant des attributs issus de l'image, comme les caractéristiques bas-niveau ou les primitives structurelles. Un état de l'art complet est présenté dans (Chen et Blostein, 2007).

– Les méthodes utilisant des caractéristiques dîtes textuelles, comme par exemple la fréquence de mots ou des histogrammes d'entités (lettres, syllabe, ...). Dans ce cas là, il est souvent nécessaire d'utiliser des méthodes de reconnaissance automatique de caractères pour extraire l'information textuelle de l'image.

En raison de la pertinence de l'information véhiculée par les mots, ou d'autres entités de texte, les caractéristiques textuelles semblent être le meilleur choix pour obtenir de bons taux de classification, qu'elles soient utilisées seules ou en association avec des caractéristiques issues de l'image. Malheureusement, cette hypothèse n'est pas vérifiée dans de nombreux cas. En premier lieu, ces procédés reposent sur une étape de transcription qui peut être perfectible. Par exemple, des images de documents produits par un smartphone ou un télécopieur peuvent être détériorées et la faible qualité de l'image affectera automatiquement les résultats de l'étape de reconnaissance de caractères. Toutes ces erreurs vont détériorer les taux de classification. De plus, certains documents sont principalement composés d'éléments graphiques (tels que les

dessins, les schémas ou les partitions) et l'information textuelle qu'ils contiennent est moins importante. Dans ce cas, l'utilisation de méthodes basées texte semble moins pertinente. Notre travail se focalise donc sur une classification d'images de documents en utilisant des caractéristiques basées image.

Dans (Cesarini *et al.*, 2001), les auteurs identifient deux catégories de caractéristiques non textuelles qui sont souvent utilisées en classification d'image de documents.

– **Les caractéristiques bas niveau** qui sont extraites directement de l'image. Leur représentation la plus usuelle est un vecteur caractérisant l'image en utilisant des indices bas niveau tels que les densités de niveau de gris, les projections ou encore des informations texture. Ces caractéristiques peuvent être de nature numérique ou symbolique. Si les caractéristiques sont extraites sur l'image entière, on parle alors de *caractéristiques d'image globales*. Dans le cas contraire, si les caractéristiques sont extraites à partir de régions d'une image, le terme utilisé est *caractéristiques d'image locales*

– **Les caractéristiques structurelles** visent à caractériser les relations entre les objets dans la page. Généralement, ils sont obtenus après un processus de segmentation suivi d'une analyse de la structure, qu'elle soit physique (mise en page) ou logique (décomposition en titres, sous-titres, sections, paragraphes, ...). Ces caractéristiques sont le plus souvent intégrées dans des structures de données telles que des arbres ou des graphes.

2.1. *Liste des caractéristiques*

Notre idée principale est de développer un vecteur générique qui sera spécialisé selon la classe de document en conservant uniquement les caractéristiques pertinentes. Le vecteur de caractéristiques doit alors intégrer un large éventail de caractéristiques pour être en mesure de pouvoir sélectionner les plus pertinentes lors de l'étape de spécialisation.

L'état de l'art se référant à la caractérisation d'images de documents est très riche. Dans (Shin *et al.*, 2001), les auteurs extraient un vecteur caractéristique de grande dimension basé sur des statistiques bas niveau telles que la taille de l'image, sa densité, les périmètres de composantes connexes, le nombre de lignes horizontales et verticales. Dans (Heroux *et al.*, 1998), Héroux *et al.* utilise une technique basée sur une caractérisation visuelle. L'image est découpée en une grille $n * n$ avec plusieurs valeurs de n. Pour chaque niveau, la densité de pixels noirs est calculée. Dans (Kumar *et al.*, 2012), les auteurs présentent une méthode basée sur les caractéristiques de mise en page. D'autres travaux sont basés sur des méthodes qui viennent de la vision par ordinateur tels que l'algorithme de détection d'objets développé par Viola et Jones ((Usilin *et al.*, 2010)) ou l'utilisation de points caractéristiques saillants ((Chen *et al.*, 2012)).

Ainsi, dans ces travaux, nous proposons de construire un vecteur caractéristique unique combinant à la fois des informations sur l'image et de l'information structurelle. Les caractéristiques choisies sont décrites dans les trois paragraphes suivants.

2.1.1. *Caractéristiques fond/forme*

Premièrement, nous proposons d'extraire les caractéristiques intégrant des informations sur la dispersion des pixels contenus dans l'image. Nous considérons alors une image binaire où les pixels noirs représentent le contenu (la forme), et les pixels blancs le fond. Ainsi, nous avons sélectionné les caractéristiques suivantes :

– le ratio largeur/hauteur du contenu de l'image,

– la densité : proportion des pixels *forme* sur les pixels *fond*,

– le centre de gravité de l'image (coordonnées x and y),

– histogramme de projection (horizontale et verticale) des pixels *forme* quantifiée en 10 valeurs.

2.1.2. *Caractéristiques de primitives*

La deuxième catégorie de caractéristiques est basée sur des primitives qui possèdent un niveau de détail supérieur par rapport aux pixels. Pour obtenir ces primitives, nous effectuons un suivi de contour des formes qui nous permet d'extraire la boîte englobante de chaque composant. Pour l'obtenir, nous avons choisi la méthode d'approximation polygonale présentée par Wall et Danielsson ((Wall et Danielsson, 1984)). Ensuite, ces polygones sont caractérisés selon certaines de leurs propriétés pour construire la deuxième partie du vecteur de caractéristiques, qui est décrite par :

– le nombre de composantes connexes, classées selon trois tailles : petite, moyenne et grande,

– le nombre d'occlusions, classées selon trois tailles : petite, moyenne et grande,

– les histogrammes des lignes droites, classées selon trois tailles (petite, moyenne et grande) et trois orientations (verticale, horizontale et inclinée),

– les histogrammes des orientations de contours classées selon huit orientations (0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°).

2.1.3. *Caractéristiques localisées par zone*

Pour inclure des caractéristiques topologiques dans notre vecteur , nous proposons d'inclure également des informations sur les primitives en fonction de leur position dans l'image. Ainsi, l'image est découpée en 9 zones de tailles équivalentes (3 lignes * 3 colonnes). Ensuite, les caractéristiques suivantes sont calculées pour chacune des tuiles :

– nombre de composantes connexes (trois tailles)

– nombre d'occlusions (trois tailles)

– nombre de lignes (trois tailles * trois orientations)

2.2. *Spécialisation du vecteur de caractéristiques*

En amont du processus de classification, la sélection des caractéristiques importantes et l'élimination des caractéristiques non pertinentes est une étape importante pour améliorer la qualité de la classification. Bien que la sélection de caractéristiques dans les problèmes de classification multiclasse a fait l'objet de beaucoup de recherches, peu de méthodes de réduction de dimension sont disponibles pour une utilisation dans les problèmes de classification monoclasse ((Fodor, 2002)). En effet, l'évaluation du pouvoir discriminant des caractéristiques est un problème difficile quand aucun exemple négatif n'est disponible. De ce constat, nous vous proposons d'estimer la stabilité (robustesse) des caractéristiques par le calcul d'un score basé sur deux indices :

– P : pourcentage des valeurs dont la dispersion autour de la moyenne est d'un écart type.

– R : rapport entre l'écart interquartile (différence entre le troisième et le premier quartile) et l'étendue des données (différence entre les valeurs maximales et minimales)

A partir de ces deux valeurs, un score est calculé tel que $S = P + (1 - R)$. Ce score varie entre 0 et 2 (2 signifiant que les valeurs de cette caractéristiques sont identiques pour l'ensemble des données). Ainsi, un classement des caractéristiques en fonction de leur stabilité pour une classe est possible. Cette méthode permettra une réduction de l'ensemble des caractéristiques en sélectionnant les N plus robustes.

3. Classification

La classification est une tâche qui requiert un classificateur, *ie.* une fonction qui assigne une étiquette de classe pour les cas décrits par un ensemble de caractéristiques. L'induction de classificateurs dans les ensembles de données étiquetées (on parle d'apprentissage supervisé) est un problème central.

Toutefois, dans cet article, nous devons faire face à des contraintes spécifiques liées au contexte industriel de ce travail. Le nombre de classes est important et de nouvelles classes peuvent apparaître au cours du temps (et le système doit gérer ces nouvelles classes). Toutes les classes n'ont pas le même nombre d'images d'apprentissage et de nombreuses classes sont très petites. Pour surmonter ces problèmes, nous proposons une méthode consistant à modéliser un classificateur pour chaque classe et permettant ainsi de traiter les vecteurs de caractéristiques spécifiques à chaque classe.

Dans sa thèse ((Tax, 2001)), Tax aborde la problématique de la classification monoclasse et offre un état de l'art des méthodes existantes en identifiant trois catégories :

– les classificateurs monoclasse de la première catégorie sont basés sur des estimations de densité des données d'apprentissage ((Tarassenko *et al.*, 1995)) et supposent ensuite de fixer un seuil sur cette densité. Plusieurs modèles de densité peuvent être appliqués : la loi normale, le mélange de gaussiennes, et la densité de Parzen sont le plus souvent utilisés. Malheureusement, tous ces modèles ont besoin d'un grand nombre d'échantillons d'apprentissage pour surmonter la malédiction de la dimensionnalité ((Duda et Hart, 1973)).

– Les techniques à base de frontières se focalisent sur l'estimation de frontières fermées autour de la classe. Les méthodes des k-centres, des plus proches voisins ou les SVM sont majoritairement utilisés. Du fait qu'elles reposent sur des mesures de distances entre les objets, ces méthodes sont sensibles à la dimension des vecteurs de caractéristiques mais nécessitent un ensemble d'apprentissage réduit par rapport aux méthodes précédentes.

– La dernière catégorie de techniques rassemble les méthodes de reconstruction. En utilisant les connaissances *a priori* sur les données et en établissant des hypothèses sur le processus de génération, un modèle est choisi et adapté aux données. La plupart de ces méthodes fait des hypothèses sur les caractéristiques de clustering des données (k-means, cartes d'auto-organisation, ...) ou leur distribution (ACP ou un mélange d'ACP, ...). Le principal inconvénient de ces méthodes est leur faible robustesse au bruit et leur sensibilité à la présence de valeurs aberrantes dans l'ensemble des données d'apprentissage.

En conséquence, nous avons pensé que, pour répondre à notre problématique, les méthodes basées frontières semblaient être les plus adaptées. Après quelques essais préliminaires concluants avec différents classificateurs, nous avons développé une approche inspirée (Gesu *et al.*, 2009 ; Gesù et Bosco, 2007) où les auteurs définissent un K-ppv monoclasse utilisant d'un seuil pour décider si les données appartiennent à la classe ou non. Pour définir ce seuil, nous proposons d'utiliser un petit ensemble de données d'apprentissage (10 documents) et de calculer la moyenne (M) des distances entre chaque paire de documents comme suit :

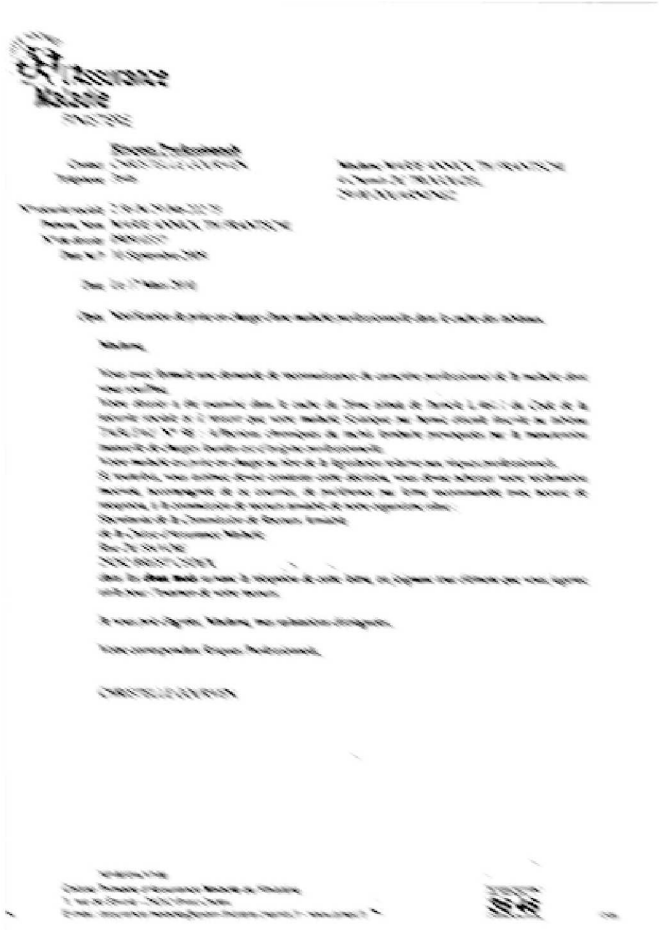
$$\gamma(x, C) = \begin{cases} 1 & if \quad \sum_{k=1}^{K} \delta(x, t_k) > \frac{K}{2} \\ O & sinon. \end{cases}$$

avec $\delta(x,t) = \begin{cases} 1 & si \quad \alpha d(x,t) > M \\ 0 & sinon. \end{cases}$

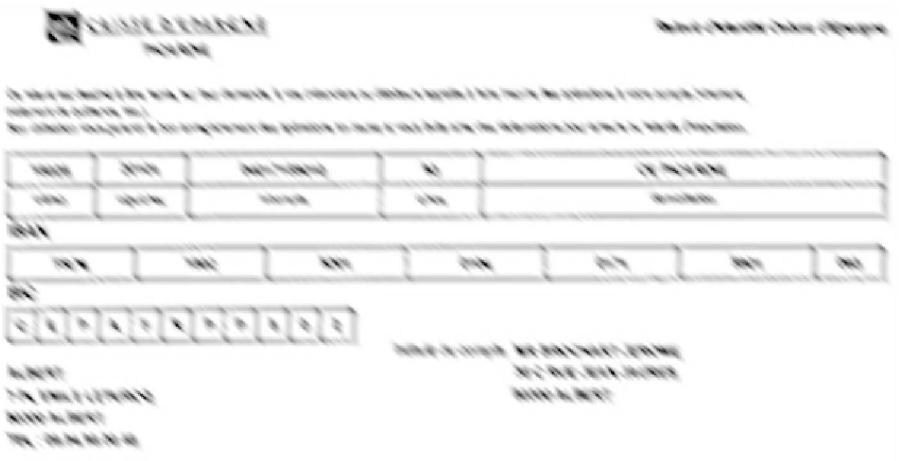
où : M est la distance (euclidienne) moyenne entre tous les exemples de l'ensemble d'apprentissage, $d(x,t)$ est la distance entre x et t, α est le paramètre de tolérance/rejet et K le nombre de plus-proches voisins.

4. Experimentations

Nous avons effectué des expérimentations sur une base de données privée qui est composée de 544 images de documents réparties en 8 classes (voir le tableau 1).

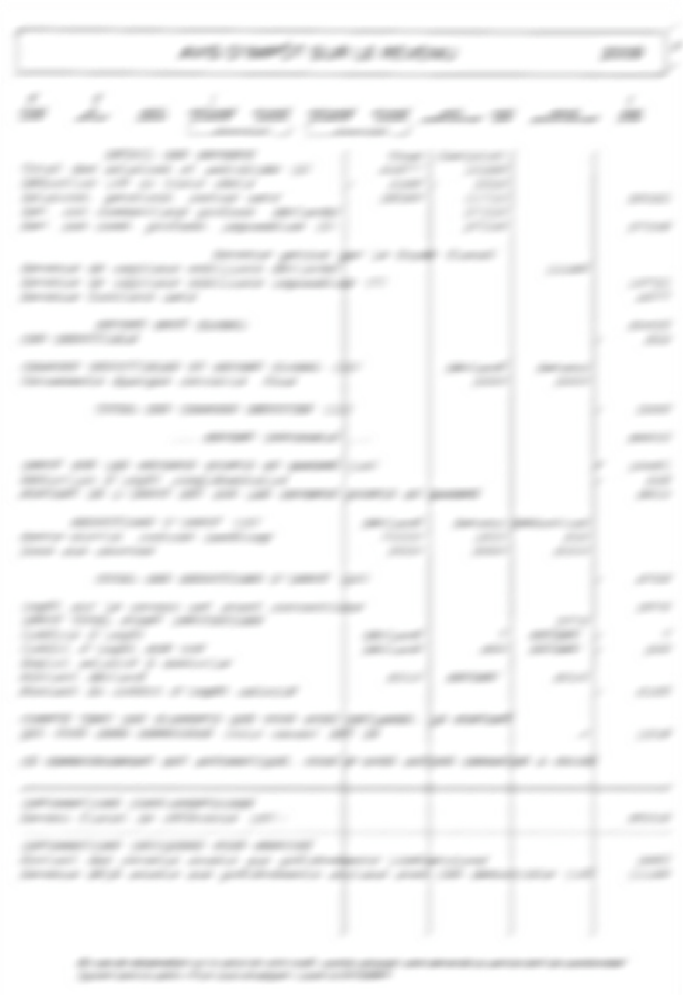


(a) ATT_SS class



(b) RIB_GT class

(c) RIB_GT class



(d) IMPOT_VER class

Figure 1. *Exemples d'images de document issues de la base*

Classe	Acronyme	Nb d'images
RIB Grande taille	RIB_GT	76
RIB Petite taille	RIB_PT	16
RIB Petite et Grande taille	RIB_MIXT	49
Feuille Impot (Verso)	IMPOT_VER	223
Feuille Impot (Recto)	IMPOT_REC	63
Attestation Sécurité Sociale	ATT_SS	60
Livret Famille	LIV_FAM	36
Carte Grise	CG	21

Tableau 1. *Description détaillée des classes*

Cette base de test est constituée de documents numérisés issus d'un problème réel. Les deux difficultés majeures sont :

– la variabilité des documents à l'intérieur d'une même classe. Par exemple la classe "RIB_GT"(Fig. 1(b) et 1(c)) contient des relevés d'identité bancaire pouvant provenir de banque différentes et sont donc différents.

– les différences de contenu entre deux classes. Par exemple, les documents issus de la classe "IMPOT_VER" (Fig. 1(d)) présente une structure forte (séparation de colonnes avec des traits verticaux) contrairement à la classe "ATT_SS" (Fig.1(a)) qui contient majoritairement du texte sans élément structurant.

Ces deux propriétés soulignent la nécessité de construire un grand vecteur de caractéristique générique pour pouvoir le spécialiser ensuite selon les classes.

Pour cette expérimentation, nous avons choisi au hasard dix documents de chaque classe pour former les ensembles d'apprentissage. Ensuite, nous avons entrainé huit classificateurs monoclasse à partir de ces documents (un classificateur K-ppv pour chaque classe) et effectué plusieurs classifications (chaque classificateur est testé sur l'ensemble des données restantes et classe les documents selon leur appartenance à la classe ou non) . Nous avons également optimisé les paramètres N (nombre de caractéristiques après l'étape de spécialisation, varie de 30 à 80) et α (le coefficient tolérance/rejet du classificateur, de 0,7 à 1,3). Nous avons reproduit ce processus cinq fois pour parvenir à des résultats de type validation croisée.

Le tableau 2 illustre les résultats obtenus par notre système. Les deux premières colonnes indiquent les paramètres qui obtiennent les meilleurs résultats. Les quatre colonnes suivantes affichent le nombre de documents classifiés selon selon leur résultats : vrais positifs (V_p), vrais négatifs (V_n), faux positifs (F_p) et de faux négatifs (F_n). Nous avons conduit ces expérimentations avec pour objectif de minimiser le nombre de faux positifs, *ie.* avec un taux élevé de rejet afin d'obtenir la plus faible probabilité de retrouver des documents négatifs, appelée confusion. En fait, nous avons suivi une

contrainte issue d'un problème réel : aucun document ne doit être classé positif à tort. En effet, les confusions ont des conséquences plus lourdes dans toute une chaîne de traitement que de classer un bon document comme négatif.

Classe	N	α	V_p	V_n	F_p	F_n	Taux
ATT_SS	50	1.2	18	414	0	32	93.10%
RIB_PT	30	0.9	6	458	0	0	100.00%
RIB_GT	60	0.8	17	398	0	49	89.43%
RIB_MIXT	70	0.8	10	425	0	29	93.75%
IMPOT_VER	70	0.7	22	411	0	31	93.31%
IMPOT_REC	60	1.0	54	251	0	159	65.73%
LIV_FAM	50	0.7	3	438	0	23	95.04%
CG	60	0.7	6	453	0	5	98.92%

Tableau 2. *Résultats d'un classificateur monoclasse k-ppv appliqué sur les 8 classes de la base. Les paramètres sont optimisés pour obtenir la confusion la plus faible possible.*

Les résultats montrent le bon comportement de notre système. Pour toutes les classes, le classificateur, combiné avec la sélection de caractéristiques stables rejette correctement les documents qui n'appartiennent pas à la classe avec un très bon taux de vrais positifs (V_p). Malheureusement, certains documents positifs sont également rejetés. Mais comme nous l'avons mentionné ci-dessus, les paramètres ont été optimisés pour obtenir ce taux de rejet élevé. Avec une contrainte plus relaxée (autorisant certaines confusions) le taux global de classification est amélioré. Les résultats sont présentés dans le tableau 3.

Pour 5 classes, le taux global de classification a augmenté. Ce résultat montre l'adaptation possible au problème selon le paramétrage du système. Ce paramétrage peut être effectué lors d'une étape de validation sur une base de données dédiée.

Dans la dernière expérimentation, nous proposons d'utiliser un classificateur SVM (Machines à Vecteurs Supports) dédié à la classification monoclasse. Le SVM a été proposé par Schölkopf (Schölkopf *et al.*, 2001) et le SVM monoclasse (ou SVDD) dans (Tax, 2001). Le SVDD à un fonctionnement similaire au SVM. Après la projection de l'espace des vecteurs de caractéristiques via un noyau Φ, le système traite l'origine comme le seul membre de l'ensemble des données négatives et tente de déterminer la marge maximale entre la classe et ce membre. Cette méthode a été utilisée pour un problème de recherche d'images dans (Chen *et al.*, 2001) et pour le classement des documents dans (Manevitz et Yousef, 2002).

Dans le tableau 4, nous pouvons observer que les résultats obtenus après une optimisation des paramètres du SVM, en suivant le même processus que ci-dessus (validation croisée), mais cette fois, les ensembles d'apprentissage représente $4/5^{\text{e}}$ de

Classe	N	α	T_p	T_n	F_p	F_n	Taux
ATT_SS	50	1.2	19	413	1	31	93.10%
RIB_PT	30	0.9	6	458	0	0	100.00%
RIB_GT	60	1.0	44	383	15	22	92.02%
RIB_MIXT	50	0.7	13	424	1	26	94.18%
IMPOT_VER	30	0.9	37	397	14	16	93.53%
IMPOT_REC	70	1.2	141	229	22	72	79.74%
LIV_FAM	50	0.7	14	430	8	12	95.68%
CG	60	0.7	7	452	1	4	98.92%

Tableau 3. *Résultats d'un classificateur monoclasse k-ppv appliqué sur les 8 classes de la base. Les paramètres sont optimisés pour obtenir le meilleur taux de classification global*

Classe	T_p	T_n	F_p	F_n	Taux
ATT_SS	6	79	18	6	77.98%
RIB_PT	0	56	50	3	51.38%
RIB_GT	4	64	30	11	62.39%
RIB_MIXT	2	38	61	8	36.70%
IMPOT_VER	32	57	7	13	79.82%
IMPOT_REC	3	81	15	10	77.06%
LIV_FAM	2	72	30	5	67.89%
CG	2	73	32	2	68.81%

Tableau 4. *Résultats d'un classificateur monoclasse SVM (SVDD) appliqué sur les 8 classes de la base. Les paramètres sont optimisés pour obtenir le meilleur taux de classification global*

l'ensemble des données. Cette distribution est plus adaptée à un SVM. Sauf pour la classe IMPOT_REC pour laquelle les taux obtenus sont équivalents à un k-ppv, les résultats semblent être plus faibles si l'on considère le taux élevé de faux positifs (des documents qui n'appartiennent pas à la classe, mais classés comme positifs). La cause de ce faible score est dû au faible nombre d'exemples d'apprentissage de notre problème. Ici, nous avons utilisé 80% (*i.e.* de 12 à 195 documents selon les classes) des ensembles pour construire le jeu de données d'apprentissage et il semble que ce pourrait être insuffisant pour obtenir une classification de bonne qualité, les SVM nécessitant des bases d'apprentissage de taille importante.

5. Conclusions et perspectives

Cet article présente un processus complet d'un système de classification d'images de document respectant des contraintes issues d'une problématique industrielle. Nous avons d'abord proposé un grand vecteur de caractéristiques capable de recueillir une description statistique et structurelle complète d'une image de document. Ensuite, nous avons proposé de spécialiser ce vecteur de caractéristiques pour chaque classe en utilisant un score de stabilité. Enfin, nous avons développé un classificateur monoclasse utilisant peu d'exemples d'apprentissage. Les expérimentations que nous avons menées montrent que notre système est performant au vu des résultats obtenus par d'autres classificateurs. De plus, notre système peut être optimisé en fonction de l'objectif a atteindre en paramétrant le compromis entre la confusion et le taux global de la classification. Nos futurs travaux seront axés sur le développement de nouvelles méthodes de sélection de caractéristiques et/ou de nouveaux classificateurs monoclasse afin d'améliorer les résultats de la classification.

6. Bibliographie

Bouguelia M.-R., Belaid Y., Belaïd A., « A Stream-Based Semi-Supervised Active Learning Approach for Document Classification », *International Conference on Document Analysis and Recognition*, p. 611-615, 2013.

Cesarini F., Lastri M., Marinai S., Soda G., « Encoding of modified X-Y trees for document classification », *International Conference on Document Analysis and Recognition, 2001.*, p. 1131-1136, 2001.

Chen N., Blostein D., « A survey of document image classification : problem statement, classifier architecture and performance evaluation », *IJDAR*, vol. 10, n° 1, p. 1-16, 2007.

Chen S., He Y., 0004 J. S., Naoi S., « Structured document classification by matching local salient features. », *International Conference on Pattern Recognition*, p. 653-656, 2012.

Chen Y., Zhou X. S., Huang T., « One-class SVM for learning in image retrieval », *International Conference on Image Processing*, vol. 1, p. 34-37 vol.1, 2001.

Duda R. O., Hart P. E., *Pattern Classification and Scene Analysis*, John Willey & Sons, New York, 1973.

Fodor I., A Survey of Dimension Reduction Techniques, Technical report, 2002.

Gesù V. D., Bosco G. L., « Combining One Class Fuzzy KNN s », *in* F. Masulli, S. Mitra, G. Pasi (eds), *International Workshop on Fuzzy Logic and Applications*, vol. 4578 of *Lecture Notes in Computer Science*, Springer, p. 152-160, 2007.

Gesu V. D., Bosco G. L., Pinello L., « A one class KNN for signal identification ; a biological case study », *Int. J. Knowl. Eng. Soft Data Paradigm.*, vol. 1, n° 4, p. 376-389, 2009.

Heroux P., Diana S., Ribert A., Trupin P., « Classification method study for automatic form class identification », *International Conference on Pattern Recognition*, vol. 1, p. 926-928 vol.1, 1998.

Kumar J., Ye P., Doermann D. S., « Learning document structure for retrieval and classification. », *International Conference on Pattern Recognition*, p. 1558-1561, 2012.

Manevitz L. M., Yousef M., « One-class svms for document classification », *J. Mach. Learn. Res.*, 2002.

Saund E., « Scientific challenges underlying production document processing. », *in* G. Agam, C. Viard-Gaudin (eds), *Document Recognition and Retrieval*, 2011.

Schölkopf B., Platt J. C., Shawe-Taylor J. C., Smola A. J., Williamson R. C., « Estimating the Support of a High-Dimensional Distribution », *Neural Comput.*, vol. 13, n° 7, p. 1443-1471, July, 2001.

Shin C., Doermann D., Rosenfeld A., « Classification of document pages using structure-based features », *International Journal on Document Analysis and Recognition*, vol. 3, n° 4, p. 232-247, 2001.

Tarassenko L., Hayton P., Cerneaz N., Brady M., « Novelty detection for the identification of masses in mammograms », *International Conference on Artificial Neural Networks*, p. 442-447, 1995.

Tax D. M. J., One-class classification : Concept learning in the absence of counter-examples, PhD thesis, Technische Universiteit Delft, 2001.

Usilin S., Nikolaev D. P., Postnikov V. V., Schaefer G., « Visual appearance based document image classification. », *International Conference on Image Processing*, p. 2133-2136, 2010.

Wall K., Danielsson P.-E., « A fast sequential method for polygonal approximation of digitized curves », *Computer Vision, Graphics, and Image Processing*, vol. 28, n° 2, p. 220 - 227, 1984.