Une stratégie originale de classification basée sur le calcul de distances avec sélection de caractéristiques : application à la classification d'images de documents

An Original Distance-Based Classification Strategy with Feature Selection : Application to Document Image Classification

Fabien Carmagnac[1,2] Pierre Héroux[2] Éric Trupin[2]

[1] A2IA S.A.
40 bis, rue Fabert
75 007 Paris cedex

[2] Laboratoire Perception Systèmes Information
CNRS FRE 2645 - UFR des Sciences et Techniques
Université de Rouen
76 821 Mont-Saint-Aignan cedex

Fabien.Carmagnac@a2ia.com
Pierre.Heroux@univ-rouen.fr
Eric.Trupin@univ-rouen.fr

Résumé

Cet article aborde la classification supervisée d'images de document. Il présente une stratégie originale basée sur le calcul de distances et la sélection de caractéristiques. On montre comment le calcul de la distance, selon un jeu de caractéristiques et une métrique, entre l'image du document à classer et l'image représentante d'une classe (point de vue) peut être utilisée pour calculer une fonction d'appartenance pour toutes les autres classes. Le choix du meilleur point de vue revient à choisir la meilleure image représentante dans le meilleur espace de caractéristiques. Cette idée est exploitée dans un algorithme itératif qui parcourt un arbre de décision construit de façon dynamique. L'apprentissage consiste à évaluer les distances entre les différentes classes et à choisir une image représentante par classe. Lors de l'étape de classification, on calcule en chaque nœud, la distance D entre l'image à classer et le meilleur point de vue. Les classes dont les distances diffèrent de D sont alors éliminées. Cette stratégie est implantée dans un module de classification d'images de document d'un produit commercial dont le but est la lecture automatique des champs manuscrits sur des documents. Ce module permet de traiter des documents de différentes classes. Des résultats expérimentaux préliminaires sont présentés et comparés avec ceux obtenus grâce à un classifieur k-ppv.

Mots Clef

Classification supervisée d'images de document, calcul de distances

Abstract

This article deals with supervised classification of document images. It presents an original strategy based on distance computation and feature selection. We show how the computation of the distance, according to a feature set and a metric, between an image to classify and the representative image of a class (point of view) can be used to compute a membership function with all other classes. The selection of the best point of view corresponds to select the best representative image in the best feature space. This idea is used in an iterative algorithm which traverses a dynamically built decision tree. The learning step consists in evaluating the distances between the different classes and on choosing a representative image per class. During the classification process, in each node we compute the distance D between the image to be classified and the best point of view. Then we eliminate the classes whose distances with the point of view do not correspond with D. This strategy is implanted as a document image classification module in a comercial product which aims at the automatic reading of handwritten fields on documents. This module allows to process documents from multiple classes. Preliminary experimental results are presented and compared with the results from a kNN classifier.

Keywords

Document image supervised classification, distance computation

1 Introduction

Depuis quelques années, les systèmes de lecture automatique hors ligne de l'écriture manuscrite se sont améliorés de façon significative (codes postaux, champs de formulaires, montants de chèque), donnant naissance à des produits commerciaux. Les bons taux de lecture de ces systèmes permettent d'orienter maintenant les recherches vers des documents moins contraints tels que les bons de commandes, voire les courriers libres. Pour valider les résultats de la lecture automatique sur de tels documents semi-structurés, il est nécessaire d'associer un modèle de lecture à chaque classe de documents. Ce modèle de lecture contenant des informations relatives aux champs du document : position, nature (lettres, chiffres), objets logiques (dates, montants, numéros de client), et des règles de cohérences entre les différents champs.

Certains systèmes ne traitent qu'un type unique de documents. D'autres, au contraire, sont confrontés à un flux de documents hétérogènes. Dans ce cas, le système doit préalablement identifier le type du document, et donc le modèle de lecture adéquat, au moyen d'un module de classification d'images de document. Dans beaucoup de systèmes, les classes de documents à discriminer sont connues au moment de la conception du module de classification, si bien qu'il est alors possible, grâce à une étude préalable, de déterminer quelles sont les caractéristiques les plus discriminantes [10] [5]. Des systèmes sont maintenant conçus sans connaissance *a priori* du nombre et de la nature des classes de documents. Il est alors nécessaire de déterminer automatiquement les meilleures caractéristiques pour un jeu de classes de documents donné.

Cet article présente une stratégie de classification d'images de document qui réalise automatiquement la sélection de caractéristiques pour tout jeu de classes de documents. Ce module est couplé à un système de lecture automatique des champs manuscrits devant disposer d'un modèle de lecture. En associant un modèle de lecture à chaque classe de documents, le module de classification permet au système de pouvoir traiter un flux de documents hétérogènes.

Ce système est vendu à des utilisateurs finaux tels que des services administratifs ou des banques ne disposant pas d'expertise en traitement ou en classification de documents. Le module doit donc, pour chaque problème pouvant être soumis au système, déterminer le jeu de caractéristiques le plus efficient pour discriminer les classes. Si les caractéristiques disponibles ne sont pas assez efficaces pour un problème particulier, de nouvelles caractéristiques doivent être facilement ajoutées et mises à disposition du module de classification.

Notre stratégie réalise la sélection de caractéristiques pour un problème donné simultanément à la classification. Elle est basée sur le calcul de distances entre images de documents. Dans la section 2, nous précisons le contexte dans lequel le module de classification des images de document est utilisé et nous exposons les contraintes qu'il doit satisfaire pour une intégration dans un produit commercial. La section 3 introduit les principales définitions. En particulier, elle précise les notions de classes de documents et introduit le concept de distances, au sens d'un jeu de caractéristiques et d'une métrique, entre deux images de documents. La section 4 présente les principes de notre stratégie de classification. La section 5 détaille les différentes étapes de l'apprentissage supervisé et la section 6 décrit les traitements appliqués à une image de document pour en déterminer la classe. Enfin, les premiers résultats expérimentaux sont exposés dans la section 7. Ces résultats sont comparés avec ceux données par des classifieurs k-ppv utilisant les mêmes jeux de caractéristiques.

2 Contexte et contraintes

Le module de classification d'images de document présenté dans cet article complète un système existant de lecture automatique des champs manuscrits sur des formulaires d'une classe unique dont on dispose du modèle de lecture. Le but de ce module est de déterminer la classe d'une image de document, acquise par un scanner, de telle sorte que le système puisse traiter des documents de classes différentes. Lorsque la classe du document a été déterminée, le modèle de lecture correspondant lui est associé et il est alors traité de la même façon qu'avec la version antérieure du système de lecture. Le modèle de lecture comprend des règles (position, nature, syntaxe, signification, cohérence) qui permettent de guider, de valider, et donc d'améliorer la lecture automatique. Si le champ à lire est, par exemple, une date, les valeurs permises sont différentes pour le jour, le mois et l'année. Si le champ à lire est l'identité d'une personne, il faudra utiliser un module de lecture configuré pour lire uniquement des lettres et sans validation par un dictionnaire.

Cependant, l'utilisation de ce module de classification au sein d'un produit commercial suppose le respect d'un certain nombre de contraintes. Tout d'abord, il doit être capable de déterminer la classe d'un document parmi une centaine de classes candidates. Le temps nécessaire à la classification doit être inférieur à une limite fixée par l'utilisateur (1 à 2 secondes sur un micro-ordinateur cadencé à 1GHz). Il doit en conséquence être une fonction sublinéaire du nombre de classes candidates. En effet, le temps nécessaire pour discriminer la classe d'un document parmi 100 classes candidates ne doit pas être 10 fois supérieur au temps requis pour le classer parmi 10 classes. De plus, les utilisateurs finaux n'ont pas d'expertise en analyse d'images de documents. Ils ne disposent pas des connaissances nécessaires pour déterminer les caractéristiques les plus performantes pour la classification. Le module de classification doit pouvoir adapter automatiquement ses paramètres pour être le plus robuste possible sur chaque jeu de documents, et ce, en évaluant ses performances de classification.

3 Définitions

Cette section introduit la terminologie utilisée dans la description du principe de classification. Les concepts de clas-

se de documents, de caractéristiques, ainsi que différents types de distances y sont détaillés. Ainsi, nous définissons la distance entre deux images de document, la distance entre une image de documents et une classe et la distance entre deux classes de document.
Comme évoqué plus haut, une classe de document est définie comme l'ensemble des documents sur lequel est appliqué le même modèle de lecture. Dans notre cas, le modèle de lecture est défini par l'utilisateur final. Une classe de documents est donc définie par les traitements ultérieurs qui leur seront appliqués (lecture automatique des champs en l'occurence). Si les champs à lire de deux documents sont identiques (position, nature, signification), ces deux documents sont considérés comme appartenant à la même classe, et ce, même si les images sont par ailleurs différentes. D'autre part, deux documents dont les images sont très proches peuvent appartenir à deux classes différentes si, par exemple, les positions de deux champs sont interverties.
Ainsi, selon la nature des traitements ultérieurs, deux documents identifiés peuvent appartenir ou non à la même classe.

Notation 3.1 *bases d'images de document*
L'ensemble des classes du problème est noté $ClassSet$.
L'ensemble des images de document est noté $ImageSet$.
L'ensemble des images appartenant à la classe C *est noté* $ImageSet_C$.
$Card(ImageSet_C)$ *est le nombre d'images de document de la classe* C.

Les caractéristiques utilisées peuvent être numériques, syntaxiques et/ou structurelles. Il est possible d'associer à chaque jeu de caractéristiques plusieurs métriques permettant alors de calculer autant de distances entre deux images de document.

Definition 3.2 *Caractéristiques et métriques*
L'espace de caractéristiques associé au jeu de caractéristiques F *est noté* $Space_F$. M_F *est une métrique associée au jeu de caractéristiques* F.

$$\begin{aligned} F &: ImageSet \rightarrow Space_F \\ M_F &: Space_F \times Space_F \rightarrow [0,1] \end{aligned}$$

L'ensemble des jeux de caractéristiques disponibles est noté $FeatureSet$.
L'ensemble de toutes les métriques est noté $MetricSet$.

Les métriques associées aux jeux de caractéristiques peuvent être calculées grâce aux distances les plus communes (distance euclidienne, distance de Hamming, distance du Max, distance du Min [3], distance d'édition [15], distance entre graphes [4]).

Definition 3.3 *Distance entre images de document*
La distance entre deux images de documents I_1 *et* I_2 *selon le jeu de caractéristiques* F *et la métrique* M_F *est définie par :*

$$\begin{aligned} D_{M_F} &: ImageSet \times ImageSet \rightarrow [0,1] \\ D_{M_F}(I_1,I_2) &= M_F(F(I_1),F(I_2)) \end{aligned}$$

Une distance calculée entre deux images de document provenant de la même classe est appelée distance intra-classe. Une distance calculée entre deux images de document provenant de deux classes différentes est appelée distance inter-classe.
Pour chaque métrique M_F et pour chaque classe, nous définissons une image représentante. Nous détaillons en section 5 comment est effectué le choix de cette image.

Notation 3.4 *L'image de document représentant la classe* C *au sens de la métrique* M_F *est notée* $I^*_{M_F,C}$.

Nous définissons ensuite la distance d'une image I quelconque à une classe C conformément à la métrique M_F comme la distance entre I et l'image représentant de la classe C pour cette métrique M_F.

Definition 3.5

$$\begin{aligned} D_{(M_F,C)} &: ImageSet \rightarrow [0,1] \\ D_{(M_F,C)}(I) &= D_{(M_F,C)}\left(I^*_{M_F,C},I\right) \\ &= M_F(F(I^*_{M_F,C}),F(I)) \end{aligned}$$

Afin d'évaluer l'éloignement entre deux classes C et C', nous définissons les distances inter-classe comme les distances qui séparent le représentant de C des éléments de C'.

Notation 3.6 *Distances intra-classe et inter-classe*
La liste des distances inter-classe entre $I^*_{M_F,C}$ *et les éléments de* C' *est notée* $DX_{M_F,C}(C')$.
La liste des distances intra-classe $DX_{M_F,C}(C)$ *est notée* $DI_{M_F,C}$.

4 Principe de classification

Traditionnellement, les images de documents sont représentées par des points dans différents espaces de caractéristiques [2] [9] [11] [12]. Les classes d'images sont alors représentées par des nuages de points. Notre approche est différente des méthodes de classification couramment utilisées [6], [7], [8] (réseaux de neurones, plus proches voisins, arbres de décision, machines à support vecteur), qui cherchent à établir les frontières entre les classes dans l'espace de caractéristiques, car l'espace des distances entre les vecteurs de caractéristiques est privilégié. La distance entre deux points du même espace de caractéristiques est représentée par un point dans un espace des distances. Cet espace est unidimensionnel. Un point dans cet espace des distances représente la distance au sens d'une métrique entre deux points de l'espace de caractéristiques à n dimensions. Le représentant $I^*_{M_F,C}$ d'une classe C au sens d'une métrique M dans un espace de caractéristiques F définit un

référentiel dans lequel se situent toutes les classes $C' \in ClassSet$ dans $Space_F$.
Grâce à la métrique M_F, on peut calculer les distances entre C et toutes les autres classes dans ce référentiel.
Soit $\tilde{I}$ une image de document dont on cherche la classe. Soit $\tilde{C}$ sa classe réelle.
La position de $\tilde{I}$ dans le référentiel est alors déterminée en calculant selon M_F la distance $D_{(M_F,C)}(I)$ entre $\tilde{I}$ et C. Bien qu'une seule distance soit à ce moment calculée, on obtient une information concernant l'appartenance de $\tilde{I}$ à toutes les classes. En effet, plus $\tilde{I}$ est éloignée d'une classe dans ce référentiel, plus la probabilité qu'elle en soit un membre est faible.
De façon plus formelle, pour chaque classe C', la liste des distances $DX_{M_F,C}(C')$ entre $I^*_{M_F,C}$ le représentant de la classe C au sens de la métrique M dans l'espace de caractéristiques F et les images de la classe C' est construite. Soit $D_{(M_F,C)}(I)$ la distance entre le représentant de C et $\tilde{I}$. On peut alors dire que $\tilde{I}$ est susceptible d'appartenir à toutes les classes C'' telles que $D_{(M_F,C)}(I)$ est proche des distances $DX_{M_F,C}(C')$.
Ceci est l'idée principale de notre approche. Dans les méthodes classiques, $\tilde{I}$ est considérée comme étant membre de $\tilde{C}$, la classe la plus proche dans l'espace de caractéristiques. Dans notre approche, la distance d entre deux points R et $\tilde{I}$ est calculée. La classe attribuée à $\tilde{I}$ est une classe dont la distance à R vaut d à ϵ près. On peut considérer R comme un point de vue dans l'espace des caractéristiques.
Le module de classification dispose de $Card(MetricSet) \times Card(ClassSet)$ couples (M_F,C). On peut associer un point de vue à chacun de ces couples. La phase d'apprentissage de notre module, détaillée en section 5 consiste alors à calculer les distances entre les images de document des différentes classes de la base d'apprentissage.
La classification d'une image de document inconnue revient à choisir une séquence de couples (M_F,C) et à exploiter les différentes hypothèses de classification pour éliminer au fur et à mesure des classes de la liste des classes candidates jusqu'à la détermination de la classe réelle. Le choix du meilleur couple (M_F,C) revient à choisir à chaque étape du processus, l'espace de caractéristiques et la métrique la plus adaptée pour discriminer les classes candidates. La phase de classification est détaillée en section 6.

5 Apprentissage

L'apprentissage est réalisé pour tous les espaces de caractéristiques $Space_F$ et toutes les métriques correspondantes M_F. Il est effectué en deux temps.
La première phase de l'apprentissage a pour but de déterminer $I^*_{M_F,C}$, une image de document de $ImageSet_C$ représentant la classe C pour la métrique M_F.
Pour ce faire, on calcule, pour chaque classe C et pour chaque métrique M_F, les distances intra-classe $D_{M_F}(I,J)$ entre toutes les images I et J, $\forall (I,J) \in ImageSet_C \times ImageSet_C$.
On déduit, pour chaque couple (M_F,C), $I^*_{M_F,C}$ de ces distances. Il y a différentes possibilités pour effectuer ce choix. Ce peut être le centre de gravité de la classe, l'image qui minimise la plus grande distance intra-classe, l'image qui minimise l'écart-type des distances intra-classe, l'image qui minimise la largeur de l'intervalle des distances intra-classes. Ce choix pourrait être l'objet de tests futurs. Notre choix s'est actuellement porté sur l'image qui minimise l'écart-type des distances intra-classe.
La seconde phase de l'apprentissage consiste à donner à chaque classe la connaissance de la distance qui la sépare des autres classes au sens de M_F. Pour ce faire, on peut évaluer les distances entre chaque document de classe C et chaque document de la classe C'. Cependant, afin de réduire le temps d'apprentissage, seules les distances séparant l'image représentant de C et les images de document des autres classes sont calculées. Ces distances permettent de constituer les listes $DX_{M_F,C}(C')$ et $DI_{M_F,C}$ de distances inter-classe et intra-classe définie par la notation 3.6.
Chaque classe connait maintenant la distance pour chaque métrique M_F qui sépare son représentant des autres classes. Une fonction d'appartenance est définie à partir des données calculées pendant les étapes précédentes. En effet, selon chaque métrique M_F, chaque classe connait la distance qui la sépare des autres classes.
Donc, si une image de document à classer est représentée par un point dans l'espace des distances dont l'origine est le représentant de C, et que la projection de ce point dans l'espace des distances se situe dans l'intervalle défini par $[\min(DX_{M_F,C}(C')), \max(DX_{M_F,C}(C'))]$, alors cette image de document est susceptible d'être membre de la classe C'. D'un autre côté, plus le point est éloigné de l'intervalle, plus la valeur de la fonction d'appartenance est faible. Cette fonction d'appartenance vaut 1 pour les points situés dans l'intervalle. Pour les autres points, la fonction d'appartenance est décroissante et tend vers zéro quand le point s'éloigne de l'intervalle, en prenant également en compte la largeur de l'intervalle.

6 Classification

La classification est effectuée par un filtrage itératif d'une liste de classes candidates. Cette liste est initialisée par l'ensemble des classes du problème.
Dans un premier temps et lors de chaque itération, un couple (M_F,C) est choisi. Le couple choisi est celui pour lequel les intervalles $[\min(DX_{M_F,C}(C')), \max(DX_{M_F,C}(C'))]$ sont les plus dispersés pour toutes les classes C' de la liste des classes candidates. Le critère de dispersion consiste à minimiser les intersections et à maximiser les distances entre les intervalles pris deux à deux. Ces intervalles dans l'espace des distances correspondent à des couronnes dans l'espace des caractéristiques (cf. figures 1, 2 et 3).
En d'autres mots, le couple (M_F,C) offrant le meilleur point de vue est celui pour lequel l'intersection entre les couronnes est la plus faible. Ainsi, une image $\tilde{I}$ quelconque représentée dans l'espace des caractéristiques par un point situé à une distance $D_{(M_F,C)}(\tilde{I})$ du point de vue sélec-

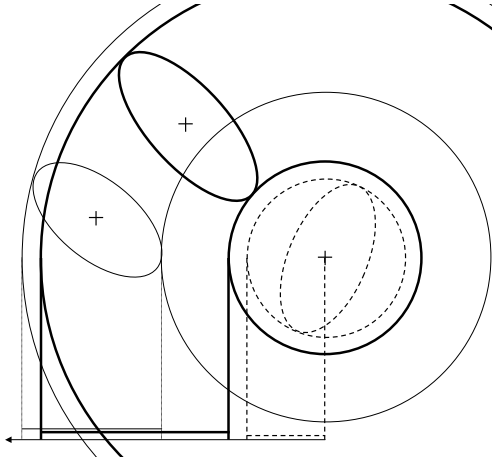


FIG. 1 – *Choix du meilleur point de vue (1)*

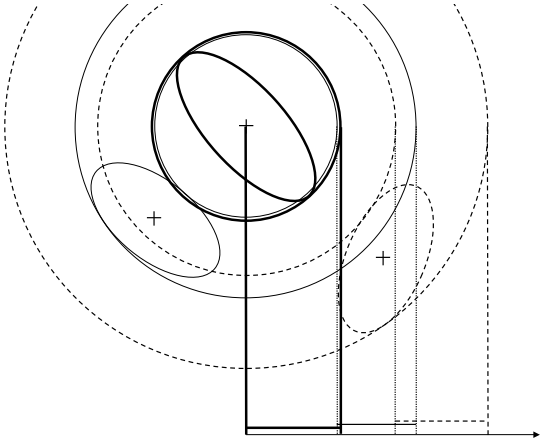


FIG. 2 – *Choix du meilleur point de vue (2)*

tionné appartiendra au plus petit nombre de couronnes.
Les figures 1, 2 et 3 représentant un problème à trois classes, illustrent le fait qu'un point de vue est meilleur qu'un autre au sein d'un même espace de caractéristiques. En effet, lorsque le représentant de la classe représentée en pointillés est choisi comme point de vue (Fig. 1), la couronne représentée en trait fin et celle en trait épais présentent une intersection. Cette intersection se traduit par un chevauchement des intervalles dans l'espace à une dimension des distances. Des tels chevauchements existent également si le représentant de la classe représentée en trait épais est choisi comme point de vue (Fig. 2). En revanche, ils disparaissent si le représentant de la classe en trait fin est choisi (Fig. 3), si bien que tout point de l'espace de caractéristiques appartient au plus à une seule couronne de l'espace des caractéristiques, sa projection appartient au plus à un seul des intervalles de l'espaces des distances.
‘ La distance $D_{(M_F,C)}(\tilde{I})$ entre l'image de classe inconnue $\tilde{I}$ et le représentant $I^*_{M_F,C}$ de la classe C pour la métrique

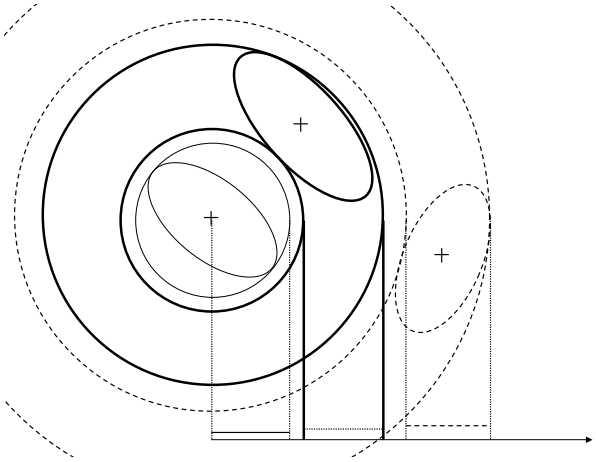


FIG. 3 – *Choix du meilleur point de vue (3)*

M_F est calculée. Si la métrique utilisée est la distance euclidienne, ce calcul de distance correspond au fait de placer un cercle dans l'espace de caractéristiques $Space_F$ centré sur $I^*_{M_F,C}$ et de rayon $D_{(M_F,C)}(\tilde{I})$. Les classes dont les nuages se situent sur ce cercle sont conservées dans la listes des classes candidates pour itération suivante, les autres classes sont rejetées. Afin de ne pas rejeter à tort la classe réelle de l'image à classer, le critère enoncé plus haut est assoupli. Seules sont rejetées les classes pour lesquelles la fonction d'appartenance est inférieure à un seuil.
Le processus itératif s'interrompt quand il reste moins de 2 classes candidates. S'il reste une classe dans la liste des classes candidates, celle-ci est retenue comme étant la classe de l'image. S'il n'en reste pas, l'image est rejetée.
La selection d'un point de vue correspond à choisir l'espace de caractéristiques, la métrique et la position à partir de laquelle seront calculées les distances permettant de discriminer au mieux les classes candidates. Ce choix est effectué à chaque étape de l'algorithme itératif, qui revient alors à déterminer un parcours dans un arbre de décision [14] construit de façon dynamique.
On peut remarquer que le critère permettant le choix du meilleur point de vue ne prend pas en compte les listes de distances $DX_{M_F,C}(C')$ engendrées par les classes C' rejetées lors des itérations précédentes de l'algorithme. En revanche, ces classes offrent toujours des points de vue potentiels puisqu'elles peuvent engendrer des listes de distances $DX_{M_F,C'}(C)$ à des images de document dont la classe C reste candidate.

Algorithme 6.1
Candidates $\leftarrow ClassSet$
TantQue Card(Candidates) > 1
choisir le meilleur couple (M_F,C)
calculer d=$D_{(M_F,C)}(I)$
Pour toutes les classes calculer la fonction d'appartenance de I
Éliminer de la liste des classes

```
  candidates celles pour lesquelles
  la fonction d'appartenance est
  inférieure à un seuil ε
finTantQue
si Card(Candidates) > 0
 then C=Candidates[0]
 else reject I
```

7 Résultats expérimentaux

La version actuelle du module de classification dispose de 6 extracteurs de caractéristiques (relatives à l'image [5] [13] et des éléments de structure du document [10] [1]). Toutes les métriques sont basées sur la distance euclidienne. Dans les tests présentés dans cette section, le point de vue sélectionné est celui minimisant l'écart-type de la distance intra-classe.

Les tests effectués portent sur la classification d'images de remises de chèques provenant de différentes banques. Ces images se caractérisent par le fait que les informations des établissements (logos, encarts publicitaires) sont localisées dans la partie supérieure de l'image. Ces informations varient beaucoup d'une banque à l'autre. Les informations manuscrites que le système doit lire se trouvent généralement dans la partie basse de l'image. La position et l'étiquette logique de ces champs définissent les classes de documents.

Aux contraintes évoquées en section 2, s'ajoutent d'autres difficultés. La binarisation des images, effectuée par la chaine de numérisation de la banque, n'est de ce fait pas maîtrisable. Elle engendre une grande variabilité. Certaines images sont presque blanches. Sur d'autres, d'épaisses lignes noires apparaissent. D'autres part, les utilisateurs finaux du système ne sont en mesure de fournir que 5 à 10 images par classe pour l'apprentissage.

Pour illustrer l'influence de la taille de la base d'apprentissage, nous avons réalisé un test sur un problème à 6 classes. Lorsque la base d'apprentissage est contituée de 5 images par classe apprise, le taux de reconnaissance globale était de 80%. Il passait à 92% lorsque la base d'apprentissage contenait 10 images de document par classe.

Sur ce même test, nous avons remarqué que les images provenant d'une des classes étaient reconnues à 100%. Les images de 4 autres classes étaient reconnues avec un taux équivalent au taux global, soit environ 80% dans le cas d'un apprentissage à 5 images par classe et environ 92% dans le cas d'un apprentissage à 10 images par classe. En revanche, la dernière classe n'était reconnue qu'à 50 ou 70% selon la taille de la base d'apprentissage. Pour cette classe, nous avons remarqué que pour toutes les caractéristiques utilisées, le nuage de point dans l'espace des caractéristiques était inclus dans celui de la classe avec laquelle elle était confondue. Cela traduit le fait que les caractéristiques utilisées n'étaient pas suffisament discriminantes pour séparer ces deux classes.

Au pire, 3 itérations de l'algorithme étaient nécéssaires pour effectuer la classification d'une image de document. Le temps de traitement était alors de 1 seconde par image de document, extraction de caractéristiques incluse.

D'autres tests ont été effectuées sur 3 problèmes différents. Il est à noter que les caractéristiques utilisées étaient les mêmes que précédemment. Aucune reconfiguration du système n'a été refaite. Seule la base d'apprentissage diffère d'un problème à l'autre.

7.1 Test A

Les images représentent des bordereaux de remise de chèques. Il s'agit d'un problème qui contient uniquement deux classes mais pour lesquelles les structures de documents varient à l'intérieur même d'une classe. Cette variabilité intra-classe vient du fait que les parties graphiques telles que les logos, les cadres et les tableaux sont fortement dégradés par les prétraitements successifs (numérisation, binarisation). La figure 4 montre quelques exemples d'images des classes $A1$ et $A2$.

La base d'apprentissage est constituée de 10 images pour chacune des deux classes. Les tests ont été réalisés sur une base de 25 125 images. Les résultats figurant dans le tableau 1 montre que le problème est parfaitement résolu avec notre approche.

7.2 Test B

Dans ce problème à 6 classes, les images sont également des images de remises de chèques. La figure 5 montre un exemple d'image pour chacune des classes. Les principales difficultés de ce problème résident dans la proximité des classes et la variabilité intra-classe.

La base d'apprentissage est constituée de 60 images de document (10 par classe). Les tests ont été réalisés sur une base de 65 689 images. Les résultats sont présentés sur le tableau 2. Le taux de reconnaissance global sur cette base est de 91,6%, le taux d'erreur étant donc de 8,4%.

7.3 Test C

Ce problème est également un problème à 6 classes. La figure 6 présente des exemples d'images de document de ce problème. Cette figure illustre le fait que les classes diffèrent entre-elles un peu plus que sur le problème précédent. La base d'apprentissage est également constituée de 10 images de document par classe. Les tests portent sur 8 770 images. Le taux de classification global sur ce problème est de 96% avec 4% d'erreur. Ces résultats sont détaillés sur le tableau 3.

7.4 Test B&C

Afin d'examiner le comportement de notre classifieur sur des problèmes où le nombre de classes est plus important, nous avons constitué des bases d'apprentissage et de test à partir des images provenant des problèmes B et C.

La base d'apprentissage est comme précédemment constituée de 10 images par classe et les tests ont été réalisés sur 10 000 documents.

Les résultats sont présentés sur le tableau 4. Le fait que les résultats aient quelque peu chuté (90% de classification,

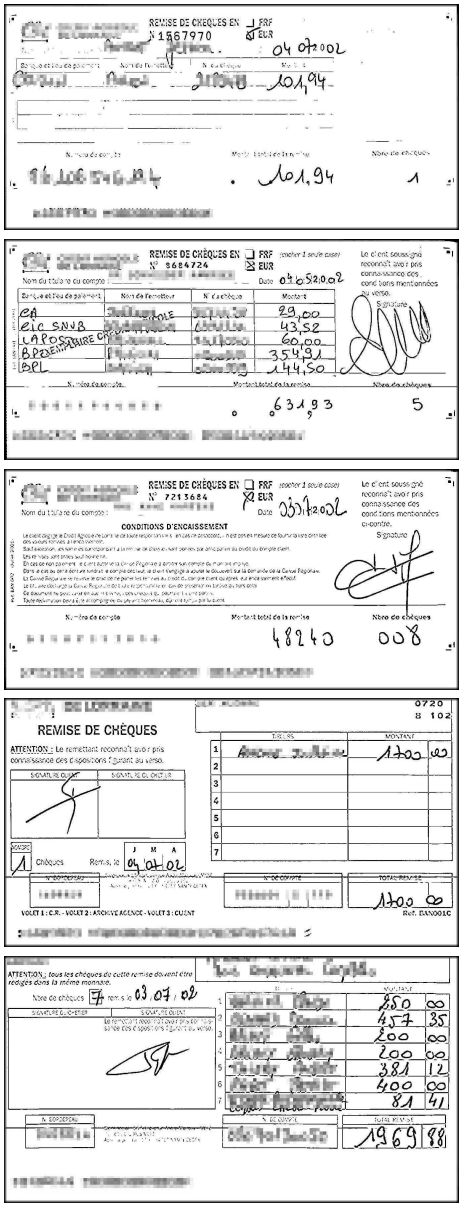


FIG. 4 – *Exemples d'images du test A*

	classe attribuée	A1	A2
classe réelle	A1	**100**%	0 %
	A2	0 %	**100**%

TAB. 1 – *Matrice de confusion pour le test A*

	classe attribuée	B1	B2	B3	B4	B5	B6
classe réelle	B1	**89**%	9%	2%	0%	0%	0%
	B2	0%	**100**%	0%	0%	0%	0%
	B3	0%	4%	**96**%	0%	0%	0%
	B4	2%	11%	1%	**77**%	9%	0%
	B5	0%	6%	8%	4%	**82**%	0%
	B6	0%	2%	0%	0%	0%	**97**%

TAB. 2 – *Matrice de confusion du test B*

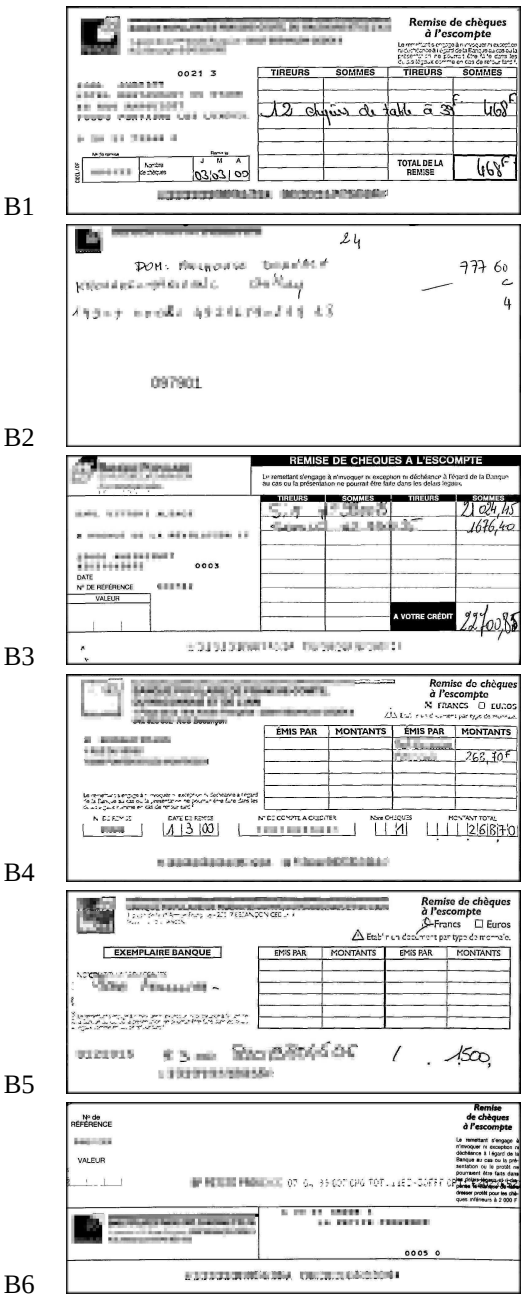
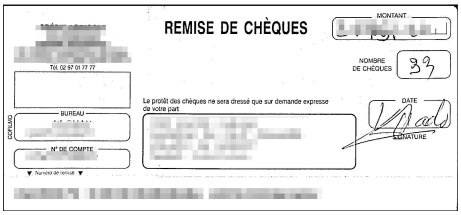


FIG. 5 – *Exemples d'images du test B*

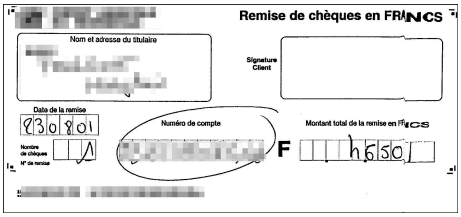
	classe attribuée	C1	C2	C3	C4	C5	C6
classe réelle	C1	**100**%	0%	0%	0%	0%	0%
	C2	1%	**98**%	1%	0%	0%	0%
	C3	2%	4%	**94**%	0%	0%	0%
	C4	3%	0%	0%	**97**%	0%	0%
	C5	6%	0%	0%	0%	**94**%	0%
	C6	2%	0%	0%	0%	1%	**97**%

TAB. 3 – *Matrice de confusion du test C*

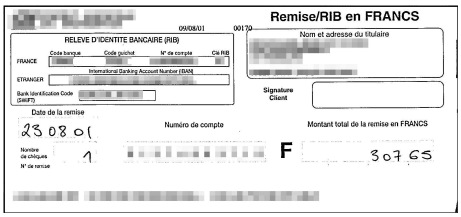
C1



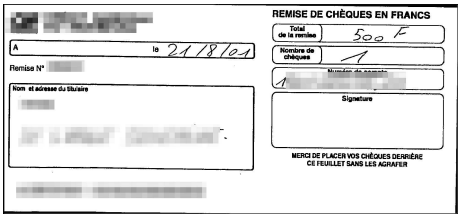
C2



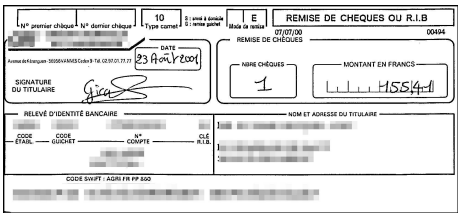
C3



C4



C5



C6

FIG. 6 – *exemples d'images du test C*

9% d'erreur et 1% de rejet) s'explique par le fait que les images des problèmes B et C sont également proches. On voit, en effet, apparaître des confusions entre classes des deux problèmes. En revanche, nous avons constaté que le temps de classification était resté dans les mêmes ordres de grandeur. Une itération suplémentaire de l'algorithme a été nécéssaire pour une partie des images.

7.5 Analyse des résultats

Les tests réalisés sur les problèmes de 6 et 12 classes ont donné des résultats de classification variant de 90% à 96% avec des taux d'erreur de 4% à 9%. Si ces résultats peuvent apparaître modestes, ceci s'explique par la difficulté du problème (proximité entre les classes et variabilité intra-classe). Ces résultats s'expliquent également par le fait que les caractéristiques utilisées ne sont pas dédiées au problème. Elles ont volontairement été choisies de telle sorte qu'elles puissent être utilisées pour d'autres problèmes de classification d'images binaires de document.

D'autre part, les résultats devraient pouvoir facilement être améliorés en effectuant une reconfiguration du module en fonction des souhaits de l'utilisateur final. On a remarqué en particulier que des erreurs étaient commises parce que la classe réelle du document était éliminée des classes candidates dès la première itération. On pourrait alors au détriment du temps de classification assouplir le critère conduisant à l'élimination des classes candidates. D'autre part, on peut également résoudre le problème du taux d'erreur rendant plus strict le critère de rejet sur les dernières itérations, même si cela doit dans le même temps réduire le taux de classification.

Enfin, il est possible d'ajouter au module de classification de nouveaux extracteurs de caractéristiques (éventuellement dédiés à la discrimination des classes du problème particulier). L'augmentation du nombre d'extracteurs de caractéristiques ne pénalise que sur le temps d'apprentissage, elle est même susceptible d'améliorer le temps de classification puisque des caractéristiques plus discriminantes filtrent davantage la liste des classes candidates.

7.6 Comparaison avec les résultats d'un classifieur k-ppv

Nous avons souhaité comparer les résultats obtenus par le module de classification se basant sur notre stratégie avec une méthode de classification couramment utilisée. Nous avons ainsi implanté un classifieur des k plus proches voisins pour chacun des extracteurs de caractéristiques disponibles. Pour chacun de ces classifieurs, la classe attribuée à l'image à classer était la classe majoritairement représentée dans les k plus proches voisins. Les classifieurs implantés n'avaient donc pas l'aptitude de rejeter des images de documents. Différentes valeurs de k ont été testées (1, 3, 5 et 10). Les tests ont été appliqués aux images des tests B, C et B&C. Les bases d'apprentissage étaient constituées pour chacun des tests de 10 images par classe, soit 60 images pour les problèmes à 6 classes et 120 images pour le pro-

blème à 12 classes.
Si le paramètre k ne semble pas influer énormément, les résultats ont été très différents d'un classifieur à l'autre lorsque les jeux de caractéristiques utilisés étaient différents. En effet, on a constaté des résultats supérieurs à 97% de reconnaissance pour un des jeux de caractéristiques. À l'opposé, d'autres caractéristiques ont donné des résultats inférieurs à 40% de reconnaissance.
Même si, dans certaines configurations, les classifieurs k-ppv se sont montrés meilleurs, leurs résultats permettent de mettre en valeur notre stratégie de classification sur plusieurs aspects. Tout d'abord, en terme de complexité, là où le classifieur k-ppv doit calculer toutes les distances entre l'image à classer et les images de la base d'apprentissage, notre stratégie ne nécessite que le calcul d'une distance par itération. Or, le nombre maximal d'itérations constaté sur les tests effectués est 4 sur le problème à 12 classes, là où le classifieur k-ppv doit en calculer 120.
Les résultats disparates des classifieurs k-ppv en fonction du jeu de caractéristiques utilisé montrent la nécessité de mener une étude pour déterminer pour chaque problème quelles sont les caractéristiques les plus discriminantes.
Cette étude n'est pas nécessaire dans notre approche dans laquelle les jeux de caractéristiques les plus discriminants sont déterminés dynamiquement en fonction de l'état de la liste des classes candidates. En examinant les traces des extracteurs de caractéristiques appelés par notre module de classification, nous avons d'ailleurs remarqué que, seuls ceux présentant de bons résultats parmi les classifieurs k-ppv étaient utilisés. Nous pouvons donc en conclure que cette comparaison avec les résultats de classifieurs de type k-ppv valide notre stratégie de sélection de caractéristiques.

8 Conclusion

Cet article présente une stratégie originale de classification appliquée au problème de la classification d'images de documents. Cette stratégie est implantée comme module au sein d'un produit dont le but est la lecture automatique de champs manuscrits sur des images de documents hétérogènes provenant de différentes classes.
Les données produites pendant la phase d'apprentissage permettent de choisir un point de vue (espace de caractéristique, métrique et origine de la comparaison) qui permettent de discriminer au mieux les classes candidates. La distance, au sens d'une métrique et d'un espace de caractéristique, entre le représentant d'une classe et une image de classe inconnue, donne non seulement une information sur l'appartenance de l'image à la classe considérée, mais également sur l'appartenance aux autres classes.
La classification s'opère en filtrant de façon itérative un ensemble de classes candidates jusqu'à ne plus avoir que 0 (le document est rejeté) ou 1 classe (à laquelle le document est attribuée).
Cette stratégie présente un certain nombre d'avantages par rapport aux méthodes classiques. D'abord, le processus de classification est une fonction sublinéaire du nombre de classes du problème, si bien qu'il n'est pas nécessaire de calculer la distance du document à toutes les classes. Ceci permet d'utiliser cette stratégie dans un contexte industriel où le temps de traitement est primordial. D'autre part, le module de classification peut facilement être configuré. On peut, par exemple, implanter différentes règles permettant de filtrer la liste des classes candidates (éliminer une certaine proportion et donc fixer le nombre maximum d'itérations, éliminer par rapport à la valeur de la fonction d'appartenance...). Enfin, de nouveaux extracteurs de caractéristiques et les différentes métriques associées peuvent facilement être intégrés. Ces nouvelles caractéristiques (éventuellement dédiées à la résolution d'un problème) sont rendues disponibles. Elles seront utilisées ou pas dans le processus de classification sans que cela n'ait d'influence sur le temps de classification. On pourrait même imaginer analyser pour chaque problème quelles ont été les caractéristiques utilisées et dans quel contexte. Par exemple, on pourrait en conclure qu'un jeu de caractéristiques n'est pas utile, ou qu'il n'est utile que pour discriminer deux classes particulières difficilement séparables par les autres caractéristiques.
Le principal avantage de raisonner dans un espace de distance est que cela permet d'utiliser des caractéristiques de différentes natures. Il n'y a pas de restriction quant à la taille des vecteurs de caractéristiques.
D'autre part, des caractéristiques numériques, syntaxiques ou structurelles peuvent être utilisées au sein d'un même vecteur de caractéristiques si on peut lui associer, au moins, une métrique permettant de calculer une distance entre deux vecteurs de caractéristiques.
La phase d'apprentissage est incrémentale. Si une nouvelle classe doit être ajoutée, les distances calculées précédemment peuvent être conservées ; seules les distances avec les nouveaux documents doivent être calculées. L'ajout de données d'apprentissage pour une classe déjà connue entraine davantage de modifications puisque le représentant de cette classe est susceptible d'être modifié. Enfin, de nouveaux vecteurs de caractéristiques ou de nouvelles métriques peuvent également être ajoutées sans remise en cause des données extraites des autres caractéristiques.
Il semble que cette stratégie puisse être appliquée à un large panel de problèmes de classification supervisée.
La stratégie présentée dans cet article est un travail naissant. Les premiers résultats semblent intéressants. Ils pourraient cependant être améliorés sur plusieurs aspects (critère de choix du représentant, détermination dynamique du meilleur point de vue...)
D'autres perspectives d'amélioration concernent les vecteurs de caractéristiques utilisés. De nouveaux vecteurs de caractéristiques pourraient être obtenus par combinaison de ceux existants. Les plus discriminants pourraient être déterminés par exemple par l'utilisation d'algorithmes génétiques appliqués sur les données de la base d'apprentissage. Les caractéristiques pourraient être associées à différentes zones de l'image. Les algorithmes génétiques effec-

tueraient une sélection des zones les plus discriminantes.

Références

[1] A. S. Azokly. *Approche uniforme pour la reconnaissance de la structure physique des documents composites basée sur l'analyse des espaces blancs*. PhD thesis, Université de Fribourg, 1995.

[2] A. Belaid and Y. Belaid. *Reconnaissance des Formes : Méthodes et Applications*. InterEditions, Paris, 1992.

[3] J.-P. Benzécri. *L'Analyse de Données : la Taxinomie*. Dunod, Paris, 1973.

[4] H. Bunke. Syntactic and structural pattern recognition theory and applications. *Series in computer science*, 7, 1990.

[5] É. Clavier. *Stratégies de tri : un système de tri des formulaires*. PhD thesis, Université de Caen, 2000.

[6] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transaction on Information Theory*, 13(1):21–27, 1967.

[7] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley-Interscience Publication, 1973.

[8] K. S. Fu. *Syntactic Pattern Recognition and Applications*. Prentice Hall, 1982.

[9] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press Inc,, 2^{nd} edition, 1990.

[10] P. Héroux, S. Diana, A. Ribert, and É. Trupin. Classification methods study for automatic form class identification. In *14th IAPR International Conference on Pattern Recognition ICPR'98*, pages 926–928, Brisbane, Australie, 1998. International Association on Pattern Recognition.

[11] M. Milgram. *Reconnaissance des Formes. Méthodes Numériques et Connexionnistes*. Armand Colin, Paris, 1993.

[12] G. Tou. *Pattern Recognition Principles*. Addison-Wesley, 1974.

[13] M. Unser, A. Aldroubi, and C. R. Gerfen. A multiresolution image registration procedure using spline pyramids. In *Wavelet Applications in Signal and Image Processing*, volume 2034, pages 160–170. SPIE, 1993.

[14] P. Vannoorenberghe and T. Denoeux. Handling uncertain labels in multiclass problems using belief decision trees. In *IPMU'2002*, 2002.

[15] R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173, 1974.

classe réelle \ classe attribuée	C1	C2	C3	C4	C5	C6	B1	B2	B3	B4	B5	B6	Rejet
C1	**88**%	0%	0%	0%	0%	0%	0%	4%	8%	0%	0%	0%	0%
C2	0%	**97**%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%
C3	0%	0%	**99**%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%
C4	0%	0%	7%	**93**%	0%	0%	0%	0%	0%	0%	0%	0%	0%
C5	0%	0%	2%	0%	**93**%	0%	0%	4%	1%	0%	0%	0%	0%
C6	0%	0%	0%	0%	1%	**98**%	0%	1%	0%	0%	0%	0%	0%
B1	0%	0%	1%	0%	0%	0%	**89**%	9%	1%	0%	0%	0%	0%
B2	0%	0%	0%	0%	0%	0%	0%	**100**%	0%	0%	0%	0%	0%
B3	0%	0%	0%	0%	0%	0%	0%	4%	**96**%	0%	0%	0%	0%
B4	0%	0%	2%	0%	0%	0%	1%	10%	1%	**77**%	9%	0%	4%
B5	0%	0%	3%	0%	0%	0%	1%	2%	15%	4%	**74**%	0%	5%
B6	0%	0%	0%	0%	1%	0%	0%	24%	0%	0%	0%	**75**%	0%

TAB. 4 – *Matrice de confusion du test B&C*