Dix-septième colloque GRETSI, Vannes, 13-17 septembre 1999

Apprentissage et Reconnaissance Automatique de types de Formulaires par une Méthode Statistique

Said RAMDANE, Bruno TACONET, Abderrazak ZAHOUR, Saddok KEBAIRI

Laboratoire d'Informatique du Havre
Place Robert Schuman, 76610 Le Havre, France

r_said@hotmail.com, bruno.taconet@iut.univ-lehavre.fr, abdel.zahour@iut.univ-lehavre.fr, saddok.kebairi@iut.univ-lehavre.fr

Résumé - Cet article présente une méthode statistique de reconnaissance automatique des types de formulaires imprimés, comportant des champs manuscrits. Les blocs principaux rectangulaires qui définissent la structure physique du formulaire, sont fournis par un algorithme de segmentation automatique. La difficulté réside dans le fait que, pour plusieurs échantillons d'un même modèle, les blocs obtenus ne sont pas forcément stables (phénomène de fusionnement et/ou de fragmentation de blocs). Lors de la phase d'apprentissage, la probabilité d'occurrence de chaque bloc est comptabilisée. Dans la phase d'identification, nous tenons compte de cette probabilité. Une nouvelle distance, que nous avons appelée distance statistique pondérée, conçue spécialement pour résoudre ce problème d'instabilité, est inspirée de la distance de Mahalanobis, mais elle est enrichie par une pondération de pénalisation affectée à chaque bloc. La méthode a été appliquée à une base d'apprentissage, et de test d'une cinquantaine de classes, avec 20 échantillons par classe.

Abstract - This article presents a statistical method of automatic recognition of printed forms types, comprising handwritten fields. The main rectangular blocks which define the physical structure of the form, are provided by an automatic algorithm of segmentation. The difficulty lies in the fact that, for several sample of a same model, the obtained blocks are not inevitably stable (phenomenon of merging and/or fragmentation of blocks). During the phase of training, the probability of occurrence of each block is computed. In the phase of identification, we take into account this probability. A new distance, that we have called weighted statistical distance, conceived especially to solve this problem of instability, is derived from the distance of Mahalanobis, but it is enriched by a weighting of penalization for each block. The method was applied to a base of training and test of about fifty classes, with 20 samples in a class.

1. Introduction

Le traitement de l'écrit et du document occupe une très grande place dans les congrès et les magazines internationaux spécialisés dans la reconnaissance de formes. On peut citer Neschen [1] qui a présenté un système pour la lecture automatique de formulaires bancaires allemands, comprenant une unité de segmentation dynamique, un classifieur basé sur l'approche de plus proche voisin (kPPV) ainsi qu'une unité de correction sémantique effectuant des recherches dans de grandes banques de données. D'autres auteurs [2], ont présenté une étude de trois classifieurs dans leur utilisation pour l'identification automatique de classes de formulaires. Ces classifieurs sont répartis en deux catégories. La première comprend le classifieur des k-Plus Proches Voisins (kPPV) et le Perceptron Multi-Couches (PMC).

2. Description générale du système d'identification de formulaires

La structure générale du système est décrite dans la figure 1. Le système est conçu de manière interactive : l'utilisateur peut, s'il connaît le type du formulaire occulter la phase d'identification du formulaire, en introduisant directement la référence correspondante. Lors de la phase d'apprentissage [3], le professeur localise les zones rectangulaires d'insertion des données manuscrites, ainsi que la nature et les attributs du support des données manuscrites (cadre rectangulaire, lignes de référence continues ou pointillées, peigne,...). Un algorithme complexe, basé sur une méthode de rectangulation [4] et [5], permet d'extraire les blocs rectangulaires englobant les zones d'inscription, et en fournit la liste. Dans le cas où le type de formulaire est inconnu, un modèle statistique vectoriel de chaque classe doit être construit [6]. Ce modèle doit prendre en compte la structure physique du document. La méthode, consistait à retenir une liste de dimension fixe des plus grands blocs rectangulaires enveloppant les zones inscrites parfaitement stables, et à classifier le formulaire selon le calcul de la distance de **Mahalanobis,** appliqué aux composantes du vecteur qui sont les caractéristiques de position et de taille des principaux blocs. Le formulaire est affecté à la classe la plus proche au sens de cette distance.

En réalité, les essais ont montré que la méthode [6] était inapplicable telle quelle dans la plupart des cas. La méthode fonctionne correctement seulement lorsque le document est constitué de blocs de textes ou de graphiques séparés par de très grands espaces vierges : la segmentation est stable. Cependant, dans le cas général, la disposition des blocs est très variable, ainsi que les espaces de séparation. Une autre difficulté s'ajoute : il s'agit des variations de la position et de la taille d'informations textuelles introduites par les différents scripteurs. Cette variabilité engendre deux phénomènes (cf. figure 2) : fusionnement de blocs et fragmentation de blocs. Induisant ainsi plusieurs configurations pour un même type de formulaire. Des erreurs de correspondance entre les blocs peuvent surgir. Ces mises en correspondance erronées

provoquent directement de grandes erreurs dans le calcul de la distance et donc une classification de mauvaise qualité.

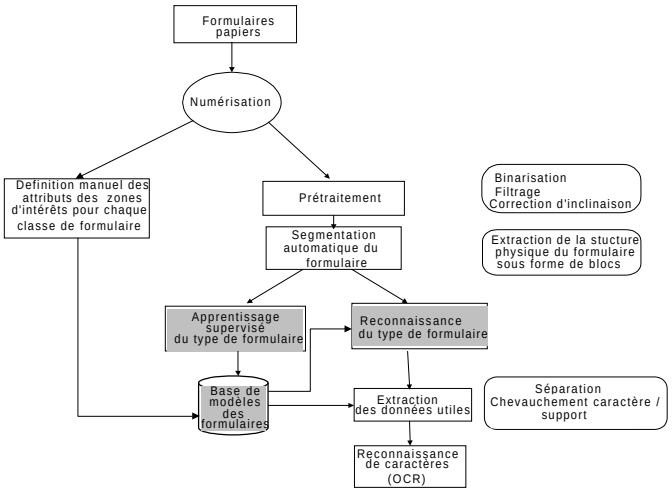

FIG. 1 : organisation du système de lecture automatique de formulaires

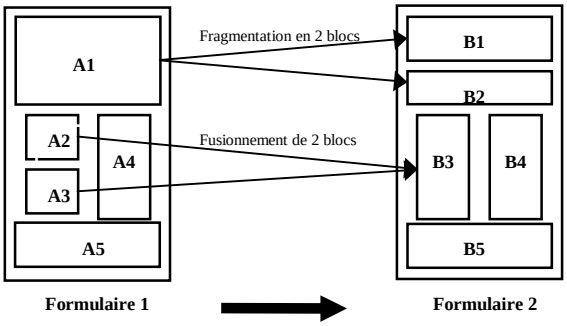

FIG. 2 : le phénomène de fusionnement et de fragmentation

3. Apprentissage du type de formulaire

3.1 Voisinage d'un bloc

Pour résoudre le problème de correspondance erronée des blocs, nous allons définir le 8-voisinage d'un bloc, puis nous associerons à chaque bloc un vecteur d'attributs qui caractérise la relation avec les 8-voisins. Nous nous sommes inspirés des travaux de **James F. Allen** [7], repris par **Hanno Walischewski** [8]. Tout bloc possède 8-voisins, les bords de l'image étant parfois considérés comme une bordure d'un voisin, le cas échéant. Un bloc voisin peut se trouver dans une des huit directions indiquées dans la figure 3. La figure 4 montre les 13 positions du côté horizontal du bloc supérieur selon [7], et la figure 5 montre la restriction à 9 positions que nous avons apportée. Les figures 6 à 12 montrent les positions relatives des 7 autres voisins. Les attributs du vecteur sont de deux types(cf. figure 13) : - distances séparatrices entre un bloc et ses 8-voisins, au nombre de douze, - hauteurs respectivement (longueurs) des blocs voisins selon la direction verticale respectivement (horizontale), au nombre de douze.

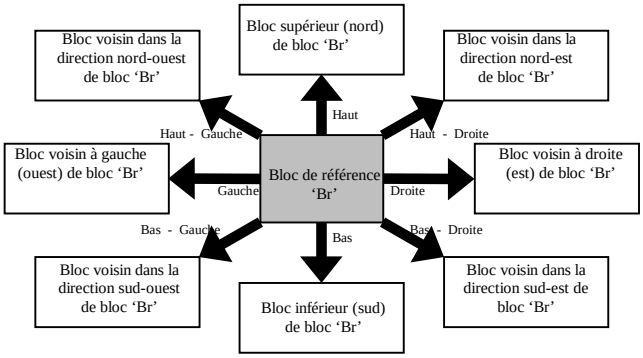

FIG. 3 : les 8-voisins d'un bloc

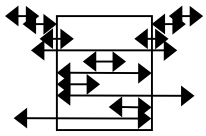

FIG. 4 : les 13 positions relatives du bloc supérieur selon [7]

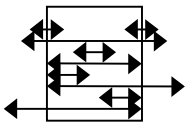

FIG. 5 : les 9 positions relatives du bloc supérieur

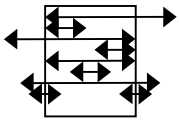

FIG. 6 : les 9 positions relatives du bloc inférieur

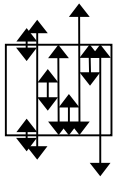

FIG. 7 : les 9 positions relatives du bloc voisin gauche

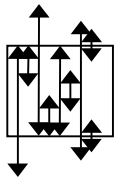

FIG. 8 : les 9 positions relatives du bloc voisin droit

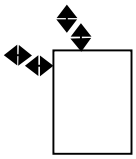

FIG. 9 : les 4 positions relatives du bloc nord-ouest

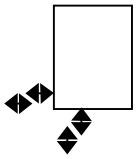

FIG. 10 : les 4 positions relatives du bloc sud-ouest

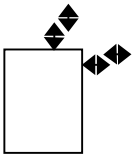

FIG. 11 : les 4 positions relatives du bloc nord-est

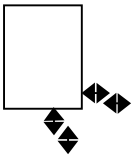

FIG. 12 : les 4 positions relatives du bloc sud-est

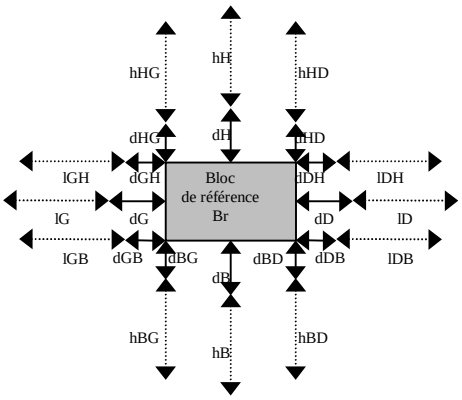

FIG. 13 : attributs du vecteur associé à un bloc caractérisant ses relations avec son 8-voisinage

3.2 Construction d'un modèle

A partir d'un certain nombre N de formulaires appartenant à la même classe, remplis par différents scripteurs sans autre contrainte que celle d'écrire dans les champs manuscrits prévus à cet effet, on construit un modèle **statistique**. Chacun de ces formulaires est décrit par un ensemble de blocs issus de l'opération de segmentation **automatique** du formulaire. Rappelons que le nombre de ces blocs n'est pas forcément identique d'un formulaire à un autre à cause des problèmes évoqués ci-dessus (fusionnement et fragmentation des blocs). Il est à remarquer que ce modèle n'est pas la réunion de toutes les configurations, mais chaque bloc ayant apparu dans un échantillon d'apprentissage, au moins, figure dans le modèle. Un bloc du modèle sera caractérisé par un coefficient de stabilité et un vecteur de grandeurs géométriques statistiques (moyenne et écart-type). La figure 14, illustre la phase d'apprentissage appliqué à un ensemble réduit de 4 échantillons pour former le modèle d'une classe.

Les étapes suivantes décrivent le processus de construction de modèles des formulaires :

3.2.1 Appariement des blocs

La correspondance entre blocs de différents formulaires d'une même classe est établie selon les critères suivants : la **distance euclidienne** entre les centres de deux blocs appariés est minimale et, éventuellement le **même comportement** avec ses voisins lors de fusionnement ou de fragmentation, le cas échéant.

3.2.2 Coefficient de stabilité d'un bloc

Un coefficient de stabilité de chaque bloc est calculé (pour une même classe) : $C_s = \frac{N_a}{N_t}$

avec : N_a : le nombre d'échantillons où le bloc est apparu, N_t : le nombre total des échantillons de la classe considéré.

Le coefficient de stabilité ainsi défini se confond avec la probabilité d'occurrence du bloc, relativement à un modèle donné (figure 14).

3.2.3 Attributs statistiques des blocs du modèle

Les grandeurs géométriques qui caractérisent un bloc rectangulaire sont : coordonnées du centre (x,y), longueur (l), hauteur (h). On associe à chaque bloc qui formera le modèle, un vecteur V_c de caractéristiques statistiques :

$$V_c = \left\{ x_m, y_m, l_m, h_m, \sigma_x, \sigma_y, \sigma_l, \sigma_h \right\}$$

avec :
$$x_m = \frac{1}{N_a}\sum_{i=1}^{N_a} x_i \quad , \quad y_m = \frac{1}{N_a}\sum_{i=1}^{N_a} y_i \quad ,$$

$$l_m = \frac{1}{N_a}\sum_{i=1}^{N_a} l_i \quad , \quad h_m = \frac{1}{N_a}\sum_{i=1}^{N_a} h_i \quad ,$$

$$\sigma_x^2 = \frac{1}{N_a}\sum_{i=1}^{N_a}\left(x_i - x_m\right)^2 \quad , \quad \sigma_y^2 = \frac{1}{N_a}\sum_{i=1}^{N_a}\left(y_i - y_m\right)^2 \quad ,$$

$$\sigma_l^2 = \frac{1}{N_a}\sum_{i=1}^{N_a}\left(l_i - l_m\right)^2 \quad , \quad \sigma_h^2 = \frac{1}{N_a}\sum_{i=1}^{N_a}\left(h_i - h_m\right)^2$$

où : x_m et y_m sont les moyennes des coordonnées du centre de gravité, des blocs appariés entre eux, l_m et h_m sont les moyennes des longueurs des côtés horizontaux et verticaux respectivement, des blocs appariés entre eux, σ_x et σ_y sont les écarts-types des coordonnées du centre de gravité, des blocs appariés entre eux, σ_l et σ_h sont les écarts-types des longueurs des cotés horizontaux et verticaux respectivement, des blocs appariés entre eux.

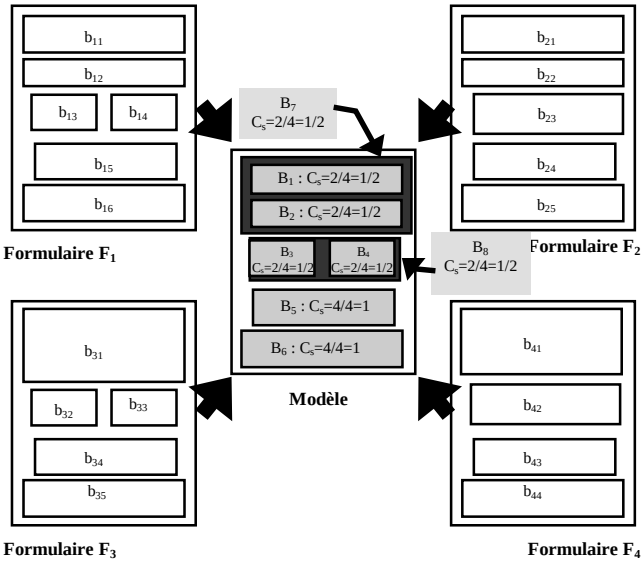

FIG. 14 : formation d'un modèle à partir de 4 échantillons

4. Reconnaissance

Pour identifier un formulaire inconnu, on exécute les opérations suivantes :

4.1 Sélection des modèles candidats

Pour pouvoir calculer la distance, on doit retenir seulement les modèles qui ont un nombre de blocs supérieur ou égal à celui du formulaire inconnu, car la phase d'apprentissage a pris en compte tous les blocs qui apparaissent au moins une fois.

4.2 Appariement des blocs

Chaque bloc du formulaire inconnu doit être apparié à un bloc du modèle sélectionné, de façon bijective. Cet appariement est effectué en minimisant la distance euclidienne entre le vecteur des grandeurs géométriques (position et taille) du formulaire inconnu et les grandeurs géométriques moyennes du modèle

$$d\left(b_i, b_j^k\right) = \min_j \left[\left(x_i - x_{mj}\right)^2 + \left(y_i - y_{mj}\right)^2 + \left(l_i - l_{mj}\right)^2 + \left(h_i - h_{mj}\right)^2 \right]^{1/2}$$

avec : b_i : le bloc d'étiquette i du formulaire inconnu,

b_j^k : le bloc d'étiquette j du k ième modèle.

Le calcul de cette distance ne se fait pas seulement selon la position du centre, mais aussi selon la taille du bloc, ce qui rend l'appariement insensible aux problèmes de décalage d'origine, de fusionnement et de fragmentation.

4.3 Calcul de la distance statistique pondérée

Donc, pour une classification plus fiable, il est logique de prendre en compte en premier lieu les blocs les plus stables. Dans le cas où plusieurs modèles présentent une similitude avec le formulaire inconnu, on procède au choix des blocs les moins stables. Autrement dit, nous faisons une sélection hiérarchique à partir des blocs les plus stables aux blocs moins stables. Ce raisonnement nous a conduit à modifier l'expression de la distance de **Mahalanobis** en multipliant le terme relatif à chaque bloc par l'inverse de son coefficient de stabilité C_{si} : (une démonstration a été faite dans [9], [10])

$$d\left(F, M_k\right) = d\left(F, C\right) = \left[\sum_{i=1}^{N} \left(\frac{\left(x_i - x_{mi}\right)^2}{\sigma_{xi}^2} + \frac{\left(y_i - y_{mi}\right)^2}{\sigma_{yi}^2} + \frac{\left(l_i - l_{mi}\right)^2}{\sigma_{li}^2} + \frac{\left(h_i - h_{mi}\right)^2}{\sigma_{hi}^2} \right) * \frac{1}{C_{si}} \right]^{1/2}$$

où F : le formulaire inconnu, M_k : modèle d'étiquette k, C : configuration appariée du modèle k, N : nombre de blocs dans le formulaire à identifier.

La décision d'affectation à une classe est prise selon un double critère : la distance au modèle représentant la classe doit être la plus petite, cette distance doit être assez petite pour éviter le rejet.

5. Expérimentation et conclusion

La base d'apprentissage est constituée de 50 classes. Chacune des classes comprend 20 formulaires remplis par des scripteurs différents. La reconnaissance a été testée sur une autre base comprenant les 50 mêmes classes (Chaque classe comprend 4 exemplaires remplis par des scripteurs différents) et 4 classes supplémentaires, non apprises. Tous les exemplaires des 4 classes supplémentaires ont été rejetés. Nous avons obtenu un taux de reconnaissance de : 97%. La figure 15 montre deux échantillons de la base de test.

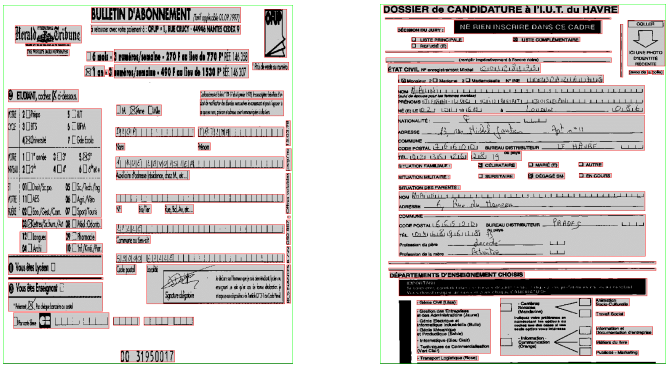

FIG. 15 : deux échantillons de formulaires utilisés pour les tests

Références

[1] M. Neschen, "Reconnaissance de Formulaires Manuscrits Basée sur la Quantification Vectorielle Hiérarchique", CNED'96, PP 245-250, Nantes, France, 1996.

[2] P. Héroux, S. Diana, A. Ribert, E. Trupin, "Etude de Méthodes de Classification pour l'Identification Automatique de Classes de Formulaires", CIFED'98, PP 463-472, Québec, Canada , mai 1998.

[3] S.Kebairi, B. Taconet, "A System of Automatic Reading of Forms", *International Conference of Pattern Recognition and Information Analysis, PRIP'97*, PP 264-270, Minsk, Biélorussie, 20-23 Mai 1997.

[4] L. Boukined, B. Taconet "Recherche de la Structure Physique d'un Document Imprimé par Rectangulation", Proc. RFIA 91, pp. 1027-1031, 1991.

[5] S. Kebairi, A. Zahour, B. Taconet, L. Boukined, "Segmentation of Composite Documents Into Homogenous Blocks", proc. IGS'98, pp. 111-112, 1997.

[6] S.Kebairi, B. Taconet, A.Zahour, P. Mercy, "Détection Automatique du Type de Formulaire Parmi un Ensemble Appris et Extraction des Données Utiles", CIFED'98, PP 255-264, Québec, Canada , mai 1998.

[7] James F. Allen. "Maintaing Knowledge About Temporel Intervals", Communication of the ACM, 26 (11), PP 832-843, Novembre 1983.

[8] H. Walischewski, "Automatic Knowledge Acquisition for Spatial Document Interpretation", Ulm, Germany, Proc. of ICPR'97, PP 243-247, Avril 1997.

[9] S. Ramdane, "Apprentissage et Reconnaissance Automatique de Type de Formulaires par une Méthode Statistique", Rapport de DEA (Instrumentation et Commande), Juin 1998.

[10] S.Kebairi, B. Taconet, A. Zahour, S. Ramdane, "A Statistical Method for an Automatic Detection of Form Types", Proc. DAS'98, PP.109-118, Nagano, Japan, November 4-6, 1998.