Identification de formulaires par modèles de Markov cachés planaires

Saïd RAMDANE[1, 2], Bruno TACONET[2], Abderrazak ZAHOUR[2], Alain FAURE[1]

[1]G.R.E.A.H. , [2]G.E.D. , Université du Havre

Place Robert Schuman, 76610 Le Havre, France

ramdan@iut.univ-lehavre.fr

Résumé - Nous présentons une méthode de modélisation de la structure physique de formulaires avec champs manuscrits, au moyen de modèles de Markov cachés pseudo-bidimensionnels (PHMMs). La description obtenue est ensuite utilisée pour la classification automatique des types de formulaires. La méthode étudiée s'appuie plus précisément sur la détection des rectangles principaux qui contiennent les zones de textes ou d'images séparées par des bandes blanches horizontales et verticales. Par la nature même du document, qui comporte des champs manuscrits, la position et les dimensions des rectangles sont variables. De plus, les phénomènes de fusionnement et de fragmentation (figure 1), résultant de la segmentation, induisent une variabilité supplémentaire dans le nombre des rectangles qui caractérisent la structure physique d'une classe de formulaires. En raison de la double variabilité des rectangles, qui présente un caractère manifestement aléatoire, et du fait du caractère 2D intrinsèque à l'image, la modélisation par PHMMs nous paraît un outil tout à fait adapté aux problèmes posés par la classification automatique des formulaires. Toutes les phases de traitement de formulaires, depuis leur saisie jusqu'à la construction des modèles, sont complètement automatiques, contrairement aux travaux publiés jusqu'à présent. En particulier, l'apprentissage des super-états est effectué par le découpage automatique en bandes de l'image du document. Le nombre d'états du modèle markovien intra-bande est déterminé par apprentissage non supervisé. La méthode des "k-means" permet ensuite d'obtenir tous les paramètres de chaque modèle.

Abstract *– We present a method of modelling physical structure of forms with hand-written fields, by means of Pseudo-bidimensional Hidden Markov Models (PHMMs). This description is then used for automatic classification of types of forms. The present method is based more precisely on the detection of the main rectangles, which contain the zones of texts or images separated by horizontal and vertical white stripes. By the nature of the document, which comprises hand-written fields, the position and dimensions of the rectangles are variable. Moreover, the phenomena of merging and fragmentation (figure 1), resulting from the segmentation, induce an additional variability in the number of the rectangles, which characterise the physical structure of a class of forms. Because of the double variability of the rectangles, which presents an obviously random character, and because of the intrinsic 2D image feature, modelling by PHMMs appears as a tool completely fitting to the problems arising from the automatic classification of the forms. All the data processing runs of forms, from their data entry to the construction of the models, are completely automatic, contrary to the work published until now. In particular, automatic cutting in stripes of the image of the document carries out the training of the super-states. The number of states of the intra-striped Markovian model is determined by unsupervised training. Then, the method of the " k-means " makes it possible to obtain all the parameters of each model.*

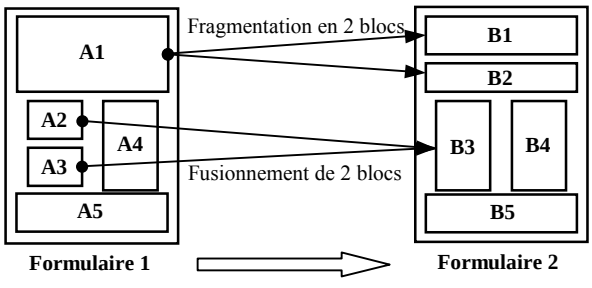

FIG. 1 : Le phénomène de fusionnement et de fragmentation

1. Architecture générale du modèle markovien planaire

Puisque le document traité est composé de pavés noirs sur fond blanc, nous observons fréquemment des ensembles de lignes successives identiques ; une super-ligne décrit un tel ensemble. De façon à comprimer la représentation, un document sera décrit par un tableau de super-lignes, composées de super-segments noirs. Nous avons opté pour une architecture à modèle principal vertical [1,2] ; l'image d'un document doit donc être découpée en bandes horizontales homogènes (dont les lignes sont semblables). Chaque bande horizontale est modélisée par un modèle secondaire (HMM-1D) de type gauche droite (figure 2).

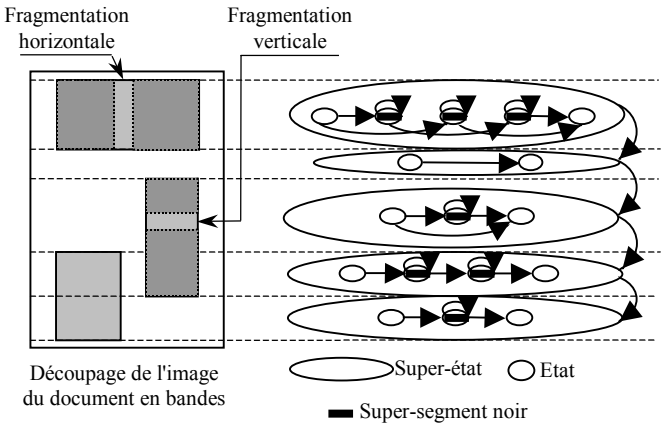

FIG. 2 : Architecture générale du modèle markovien

2. Modèle vertical

Par construction le modèle vertical est un modèle de Bakis orienté du haut vers le bas. Le nombre de super-états a été fixé, lors de l'apprentissage, par le découpage en bandes (ici, ce sont des super-lignes) du modèle majeur. A priori, il n'y a pas de saut d'état puisque un rectangle majeur peut être fragmenté mais ne disparaît pas. La notion de transition d'un état vers lui-même est remplacée par la notion de durée, qui semble mieux adaptée pour régler le passage vers l'état suivant. La fonction de durée possède deux paramètres statistiques : la hauteur moyenne (μ) de la bande du super-état considéré (nombre de lignes moyen de la bande), et l'écart type (σ) calculé sur la hauteur de la bande. La fonction de durée vaut 1 lorsque la hauteur de la ligne courante est inférieure à la hauteur moyenne, et décroît à partir de ce moment en suivant la gaussienne [3,4,5] (figure 3).

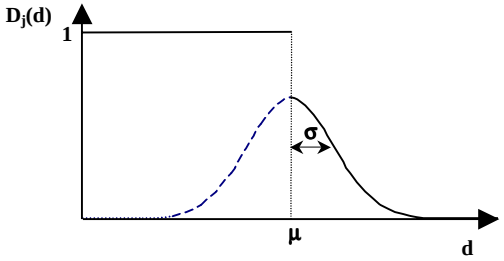

FIG. 3 : Allure de la fonction durée associée à un super-état.

Soit j : le super-état, d : la position de la dernière ligne de la super-ligne dans le super-état, μ_j la hauteur moyenne de la bande, σ_j l'écart type sur la hauteur ; la fonction de durée D_j s'écrit :

$$D_j(d) = \begin{cases} 1 & \text{pour } 1 \le d \le \mu_j \\ \dfrac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\dfrac{(d-\mu_j)^2}{2\sigma_j^2}\right) & \text{pour } d > \mu_j \end{cases}$$

3. Modèles secondaires

Puisque l'image est composée de rectangles noirs sur un fond blanc, et puisque un super-état correspond à une bande composée de super-lignes, une représentation compressée consiste à coder la position et la longueur des super-segments noirs. Donc une observation de HMM secondaire est un segment noir paramétré par la position du milieu et par la longueur. Comme les paramètres varient continûment, l'observation est continue.

Conformément à la recommandation de Dours [6], nous avons construit les états en leur attribuant une signification physique : à chaque segment noir distinct du modèle moyen, compris dans la bande du super-état, on associe un état. Nous avons ajouté un état initial et un état final, sans observation. Une bande blanche (pas d'observation) est représentée par une transition de l'état initial vers l'état final (figure 2).

Les modèles secondaires du PHMM proposé sont des modèles markoviens continus d'ordre 1, de type gauche-droite. L'apprentissage des HMMs est fait par l'algorithme des k-means [7], pour répartir le mieux possible les observations dans les états. Les phénomènes de fragmentation horizontale sont naturellement pris en compte par les transitions entre états des HMMs secondaires. Les phénomènes de fragmentation verticale, se produisant à l'intérieur d'un même super-état, sont également (et paradoxalement) absorbés par les transitions entre états du modèle secondaire (figure 4).

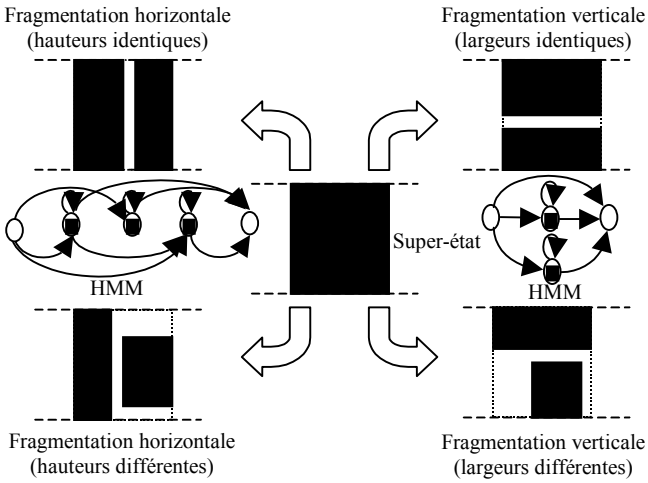

FIG. 4 : Fragmentation horizontale ou verticale dans un super-état et HMM associé

4. Phase d'apprentissage

4.1 Construction du modèle moyen complet

A partir d'un ensemble de N formulaires représentatif de la même classe, remplis par différents scripteurs sans autre contrainte que celle d'écrire dans les champs manuscrits prévus à cet effet, on construit le modèle moyen complet. 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 (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 une fois, figure dans le modèle. La correspondance entre blocs de différents formulaires d'une même classe est établie selon les critères suivants : d'une part la distance euclidienne entre les centres de deux blocs appariés est minimale, et d'autre part chaque bloc a le même comportement avec ses voisins lors d'un éventuel fusionnement ou d'une possible fragmentation [8,9]. La figure 5, illustre la phase pour former le modèle moyen complet d'une classe appliqué à un ensemble réduit de 4 échantillons.

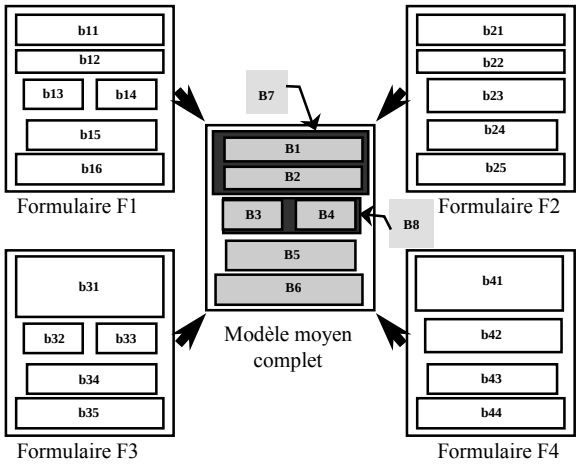

FIG. 5 **:** Formation du modèle moyen complet à partir de 4 échantillons

4.2 Construction du modèle majeur

Le modèle majeur s'obtient à partir du modèle moyen complet en conservant les rectangles les plus grands pour former une configuration possible (figure 6). Ainsi une configuration quelconque du modèle moyen sera rapportée au modèle majeur seulement par fusionnements éventuels. Le modèle majeur n'est pas forcément la configuration la plus probable, mais il est la configuration la moins fragmentée : aucun bloc du modèle majeur ne peut être fusionné avec un autre. Toute configuration peut être considérée comme résultant de fragmentations éventuelles du modèle majeur.

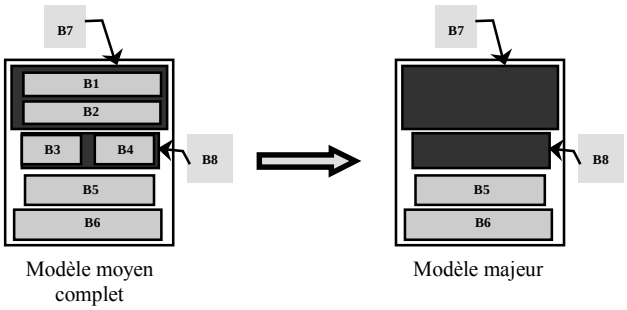

FIG. 6 **:** Du modèle moyen complet au modèle majeur

Chaque super-ligne du modèle majeur détermine un super-état ; en effet, l'ouverture ou la fermeture d'un pavé noir se traduit par le début ou la fin d'une super-ligne (figure 7). L'utilisation du modèle majeur élimine la possibilité de tout fusionnement vertical pour le découpage en super-état. Le nombre des super-états relatifs à une classe est ainsi fixé. Par conséquent, la fragmentation d'un rectangle majeur dans le sens vertical se produira à l'intérieur d'un super-état, et sera traité par le modèle secondaire associé. Le modèle majeur ainsi découpé en super-états va servir de référence pour le découpage automatique des échantillons d'apprentissage.

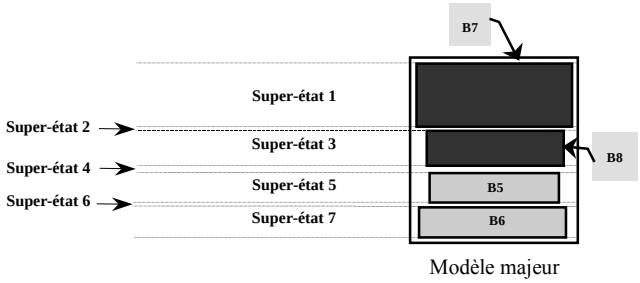

FIG. 7 **:** Modèle majeur découpé en super-états

4.3 Découpage en bandes de super-état

Le modèle majeur découpé en super-états sert de référence. Une bande de super-état d'un échantillon sera composée d'une ou de plusieurs super-lignes. L'affectation d'une super-ligne à un super-état de référence est faite par une méthode de programmation dynamique : on dresse un tableau de distances entre une super-ligne de l'échantillon portée en abscisse et un super-état de référence (représenté par la super-ligne associée) porté en ordonnée. La première super-ligne est affectée au premier super-état de référence, et la dernière super-ligne au dernier super-état de référence. La progression dans le chemin optimal se fait soit horizontalement, soit vers le super-état d'indice supérieur. Il joint la première case à la dernière case en minimisant la somme des distances (figure 8).

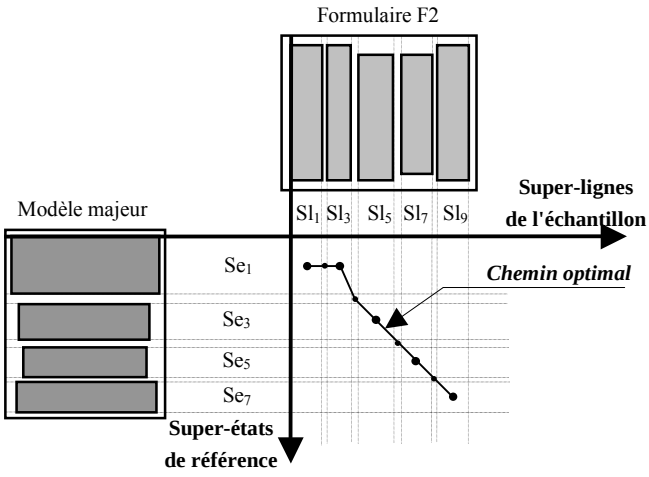

FIG. 8 : affectation des super-lignes par programmation dynamique

La distance entre la super-ligne courante Sl_c de l'échantillon et la super-ligne Sl_r représentant le super-état de référence Se_r est calculée de la façon suivante :

$$d(Sl_c, Sl_r) = \frac{\sum_i l_i}{\sum_j L_j}$$

avec $\sum_j L_j$: longueur cumulée des super-segments noirs, $\sum_i l_i$: longueur cumulée des disparités (cf. (figure 9) : $\sum_j L_j = L_1 + L_2 + L_3$; $\sum_i l_i = l_1 + l_2 + l_3$).

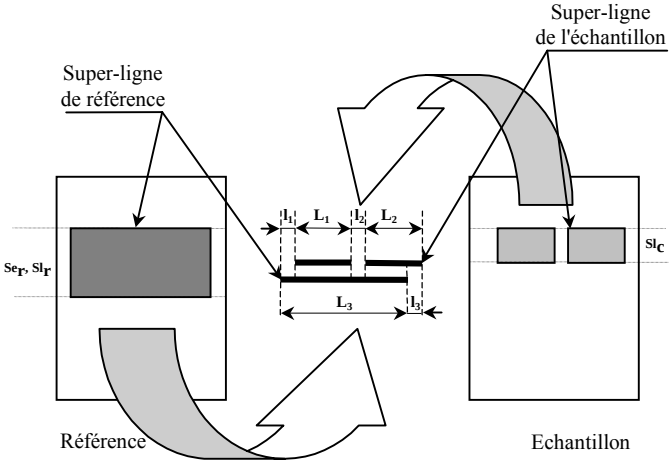

FIG. 9 : Appariement entre une super-ligne de l'échantillon et la super-ligne représentant le super-état de référence

5. Phase de reconnaissance

Une fois effectuée la phase d'apprentissage, un PHMM modélise chaque classe. L'échantillon à reconnaître X est décrit par un tableau de super-lignes d'observations. Une observation est un segment noir représenté par ses paramètres de position et de longueur. L'objectif de traitement est double : il s'agit d'apparier l'échantillon X avec le modèle de la classe C, et de calculer la probabilité conditionnelle *Prob(X/C)*. Le critère bayésien de maximum de vraisemblance (maximiser *Prob(X/C).Prob(C)*) permet de prendre la décision d'affectation. L'algorithme de Viterbi est exécuté de façon imbriquée en deux étapes selon les deux directions (verticale, horizontale). Dans la direction verticale, la probabilité du modèle est évaluée en s'appuyant sur la fonction "durée" associée à chaque super-état. La fonction "durée" joue le rôle d'un facteur de pondération, qui accentue la probabilité de transition vers le super-état suivant lorsque la durée dépasse la valeur moyenne apprise. Le passage au super-état suivant a lieu lorsque la probabilité de rester dans le super-état présent (elle est calculée par le modèle secondaire correspondant), pénalisée par la fonction de durée, devient inférieure à la probabilité de passer au super-état suivant (calculée par le modèle secondaire suivant).

A ce stade, remarquons que la modélisation par PHMM donne un résultat approché : la probabilité d'un échantillon peu fragmenté verticalement sera trouvée supérieure à celle d'un échantillon semblable, mais très fragmenté verticalement, les lignes qui séparent ces fragments étant en faible nombre au moment de l'apprentissage.

6. 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 10 exemplaires remplis par des scripteurs différents). Nous avons obtenu un taux de reconnaissance de : 97,6%.

Références

[1] S. Ramdane, B. Taconet, A. Zahour et A. Faure. *Planar Hidden Markov Models for the Classification of Forms.* ICISP'2001, Agadir, Morocco, 3-5 May 2001.

[2] S. Ramdane, B. Taconet, A. Zahour et A. Faure. *Modélisation Pseudo Bidimensionnelle pour la Reconnaissance Automatique de Types de Formulaires avec Champs Manuscrits.* CIFED'2000, pp. 61-70, Lyon, France, juillet 3-5, 2000.

[3] N. Ben Amara et A. Belaïd. *Printed PAW Recognition Based on Planar Hidden Markov models.* ICPR'96, pp. 220-224, 25-29 août 1996.

[4] N. Ben Amara, A. Belaïd et N. Ellouze. *Modélisation Pseudo Bidimensionnelle pour la Reconnaissance de Chaînes de Caractères Arabes Imprimés.* CIFED'98, pp. 131-140, 11-13 mai 1998.

[5] N. Ben Amara. *Utilisation des Modèles de Markov Cachés Planaires en Reconnaissance de l'Ecriture Arabe Imprimée.* Thèse de doctorat, Université Tunis II, 1999.

[6] C. Dours. *Contribution à l'Etude du Décodage Acoustico-phonétique pour la Reconnaissance Automatique de la Parole.* Thèse de doctorat, Université Paul Sabatier de Toulouse, 1989.

[7] B. H. Juang and L. R. Rabiner. *The Segmental k-means algorithm for estimating the parameters of Hidden Markov Models.* IEEE Trans. Accoust., Speech, Signal Processing, vol. ASSP-38, no. 9, pp. 1639-1641, sept. 1990.

[8] S. Ramdane, B. Taconet, A. Zahour et S. Kebairi. *Apprentissage et Reconnaissance Automatique de Types de Formulaires par une Méthode Statistique.* GRETSI'99, pp. 111-114, Vannes, France, septembre 13-17, 1999.

[9] S. Kebairi, B. Taconet, A. Zahour et S. Ramdane. *A Statistical Method for an Automatic Detection of Form Types*. Proc. DAS'98, pp. 109-118, Nagano, Japan, november 4-6, 1998.