Système TDNN/HMM de reconnaissance de mots cursifs en ligne à apprentissage simplifié

Emilie Poisson* — C. Viard-Gaudin* — P.M. Lallican**
** Institut de recherche en Communication et Cybernétique de Nantes,*
Unité mixte CNRS n°6597 / Ecole Centrale de Nantes, Université de Nantes
Ecole polytechnique de l'université de Nantes
Equipe Image et VidéoCommunication
La Chantrerie – rue Christian Pauc – B.P. 50609
F-44 306 Nantes Cedex 3
{Emilie.Poisson,Christian.Viard-Gaudin}@polytech.univ-nantes.fr
*** Vision Objects, 9 rue du pavillon, 44980 Sainte Luce/Loire*
pmlallican@visionobjects.com

RÉSUMÉ. *Un système hybride, combinant un réseau de neurones à convolution et des modèles de Markov à états cachés est proposé dans cet article. Il étend à la reconnaissance mot un précédent système dédié à la reconnaissance de caractères isolés [Poi 02]. L'architecture global du système est présentée ici, ainsi que l'étape d'apprentissage qui se fait directement au niveau mot bien que l'approche soit de type analytique avec construction dynamique des modèles-mots par concaténation des modèles-lettres. Pour cela, une fonction objectif maximisant l'information mutuelle entre le mot à reconnaître et les mots reconnus est utilisée. Il est possible dans ces conditions de rétropropager dans le réseau de neurones, le gradient de cette fonction de coût sans avoir à disposer d'exemples étiquetés au niveau caractères. Les expériences réalisées montrent la convergence du système global même avec une initialisation aléatoire du réseau de neurones qui sert à reconnaître les caractères.*

ABSTRACT. *A hybrid system, combining convolutional neural network and Hidden Markov Models is proposed in this paper. It extends to word recognition a previous system dedicated to isolated character recognition. We introduce here the global architecture of the system and we define the training stage which is performed directly at the world level although it is an analytical based approach with a dynamic lexicon construction. The weight adaptation of the neural network is derived from a cost function defined at the word level based on a Maximum Mutual Information criteria. With such an approach, it is not required to label the training database at the character level. The experiments that have been conducted show that the global system is able to converge even with a random initialisation of the neural network which acts as a character recogniser.*

MOTS-CLÉS : *Système hybride neuro-markovien, TDNN, HMM, reconnaissance de mots cursifs en ligne.*

KEYWORDS: *Neuro-markovian hybrid system, TDNN, HMM, on line handwritten cursive words recognition.*

1. Introduction

Nos travaux s'intègrent dans le contexte de la reconnaissance de l'écriture en ligne destiné aux systèmes mobiles communicants (assistant numérique personnel, ardoise électronique, smart-phone). Dans ce domaine, il importe encore d'améliorer les performances de reconnaissance tout en respectant des contraintes fortes sur l'occupation mémoire et la vitesse de traitement. Dans cette optique, nous avons opté pour l'étude de réseaux de neurones à convolution. Ces architectures bénéficient de propriétés intéressantes tant au niveau de leurs performances que de leur relative faible encombrement, c'est pourquoi, elles sont tout à fait dignes d'intérêt [Sim 03][Lec 01], et en particulier pour des systèmes embarqués. Ainsi, dans [Poi 02], nous avons étudié différents réseaux de neurones à convolution appliqués à la reconnaissance de caractères manuscrits et montré que ces systèmes avaient de très bonnes performances en terme de taux de reconnaissance et des architectures à complexité limitée les rendant compatibles avec des systèmes portables de faibles capacités tels que les assistants personnels ou téléphones mobiles. Ces premiers résultats nous ont amené à étendre l'utilisation de ces réseaux dans un système hybride de reconnaissance mot basé réseaux à décalage temporel (Time Delay Neural Network : TDNN) et modèles de Markov Cachés (HMM) avec apprentissage global. En effet, les systèmes existants sont encore très lourds soit dans la complexité et coût de leur architecture soit dans leur mode d'apprentissage. On peut citer : les systèmes à base de méthodes à noyaux comme les machines à vecteurs supports (Support Vector Machine : SVM) [Rag 03] qui demandent un nombre de vecteurs supports important ou encore ceux basés sur des réseaux de neurones classiques [Lec 01] [Tay 02] où le nombre de poids reste très pénalisant. Par ailleurs, les méthodes d'apprentissage nécessitent différentes étapes, quelquefois alternées sur plusieurs itérations, liées au type de segmentation choisi avec souvent d'abord un apprentissage au niveau lettre ou graphèmes puis un apprentissage au niveau mot [Jae 00][Wim 00].

Dans cet article, nous proposons une méthode de reconnaissance de mots cursifs saisis en ligne avec une architecture simplifiée de type reconnaissance analytique avec segmentation implicite et apprentissage global au niveau mot. Le système de reconnaissance que nous développons actuellement est illustré par la figure 1. C'est un système hybride qui fait coopérer un modèle de Markov caché (HMM) et un réseau à convolution (RC). Cette coopération consiste à utiliser le HMM pour effectuer la tâche de reconnaissance au niveau mot en trouvant le meilleur alignement temporel (programmation dynamique) sur les différents modèles mots du lexique à partir des probabilités d'observations estimées par le RC dans sa tâche de classification des entités élémentaires –liées aux fenêtres de convolution, qu'il balaie régulièrement le long de la trajectoire.

Les données fournies en entrée du RC proviennent directement du signal dynamique $e(t) = (x(t), y(t))$ dans le cas d'un TDNN ou bien sont issues de l'image statique $I(x,y)$ reconstruite à partir de $e(t)$. Dans ce dernier cas, les convolutions se

font dans le domaine spatial et le RC est de type Space Displacement Neural Network (SDNN). Ces deux architectures peuvent être combinées en une architecture unique par un simple produit des couches de sortie ou via la dernière couche de l'étage classifieur, on a nommé cette architecture SDTDNN [Poi 02].

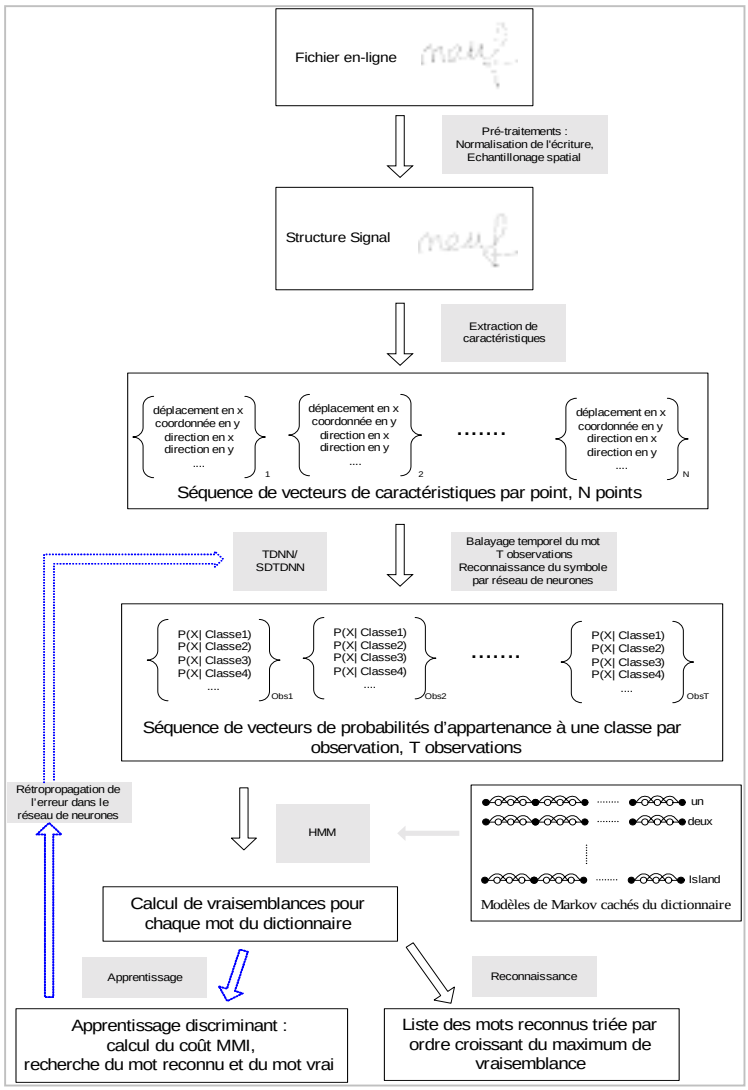

Figure 1. Modélisation du système de reconnaissance de mots en-ligne proposé.

Avec une telle approche, une segmentation implicite du mot en lettres voire graphèmes est proposée grâce à la fenêtre d'observations glissante sur laquelle est

calculée l'estimation des probabilités d'observations des entités élémentaires. La dernière étape, à savoir l'étape de reconnaissance est quand à elle, basée sur l'utilisation de modèles de Markov, elle fournit en sortie une liste des scores décroissants obtenus en fonction du dictionnaire utilisé. Pour à la fois, simplifier le processus d'apprentissage, et améliorer les taux de reconnaissance au niveau mot, nous avons opté pour un apprentissage global du système hybride avec un critère discriminant de type MMI (Maximum Mutual Information).

C'est un point important de la méthode proposée. De cette façon, il n'est pas nécessaire de disposer d'une base de mots étiquetée au niveau caractère. En effet, avec les systèmes hybrides classiques, il convient soit de disposer d'un système annexe de reconnaissance permettant d'initialiser le système de reconnaissance pour permettre la segmentation et l'étiquetage en caractères (post-segmentation par l'algorithme de Viterbi) soit d'entraîner le réseau sur une base de caractères isolés qui ne peut être très représentative de l'écriture cursive. De plus, le comportement d'un réseau de neurones vis-à-vis des formes non apprises (hypothèse de segmentation ne correspondant pas à un vrai caractère) est toujours un problème délicat. Ici, il n'y a pas explicitement d'apprentissage au niveau caractère mais optimisation du réseau pour satisfaire la fonction objectif définie au niveau global des mots.

2. Présentation générale du système

2.1. Entrées et prétraitements

Les bases de mots choisies pour évaluer notre système sont des bases de mots manuscrits saisis en ligne dans un contexte omni scripteur non contraint au format UNIPEN [Gui 94]. Nous testerons ce reconnaisseur sur deux bases de référence soit les bases IRONOFF [Via 99] et UNIPEN [Gui 94]. Les données fournies correspondent finalement à une suite de points de coordonnées (x,y) du mot avec son label.

L'écriture manuscrite a des styles très variés qui dépendent de plusieurs sources différentes comme l'identité des auteurs, l'environnement et de la situation pendant l'écriture, l'appareillage d'écriture et les médias aussi bien que le but de l'écriture. La plupart des variations sont non pertinentes voire perturbatrices, il est donc préférable de les éliminer à l'étape initiale pour soulager la difficile tâche d'identification. C'est le but principal de l'étape de prétraitement des données de système, illustrée à la figure 2. Dans un premier temps, la hauteur de la zone centrale du mot est déterminée à partir de l'extraction de quatre lignes de référence (haute, corps, base, basse), cf. illustration du mot « quand » sur la figure 2. Pour cela, un algorithme de type EM (Expectation-Maximization) utilisant les extremums du contour du mot cherche à optimiser la position et l'orientation de quatre lignes parallèles [Ben 94]. Ces lignes servent alors pour normaliser le mot en orientation, en taille et pour caractériser implicitement les extensions hautes et basses de

l'écriture selon nos règles primaires de production de l'écriture. Ensuite une procédure de rééchantillonnage spatial par interpolation linéaire du tracé est mise en œuvre pour s'affranchir des variabilités de vitesses d'écriture. Une suite de vecteurs de sept paramètres, appelés caractéristiques, est extraite du tracé point par point. Ces caractéristiques sont celles les plus standards : position, direction et courbure du tracé, information de posé ou levé de stylet.

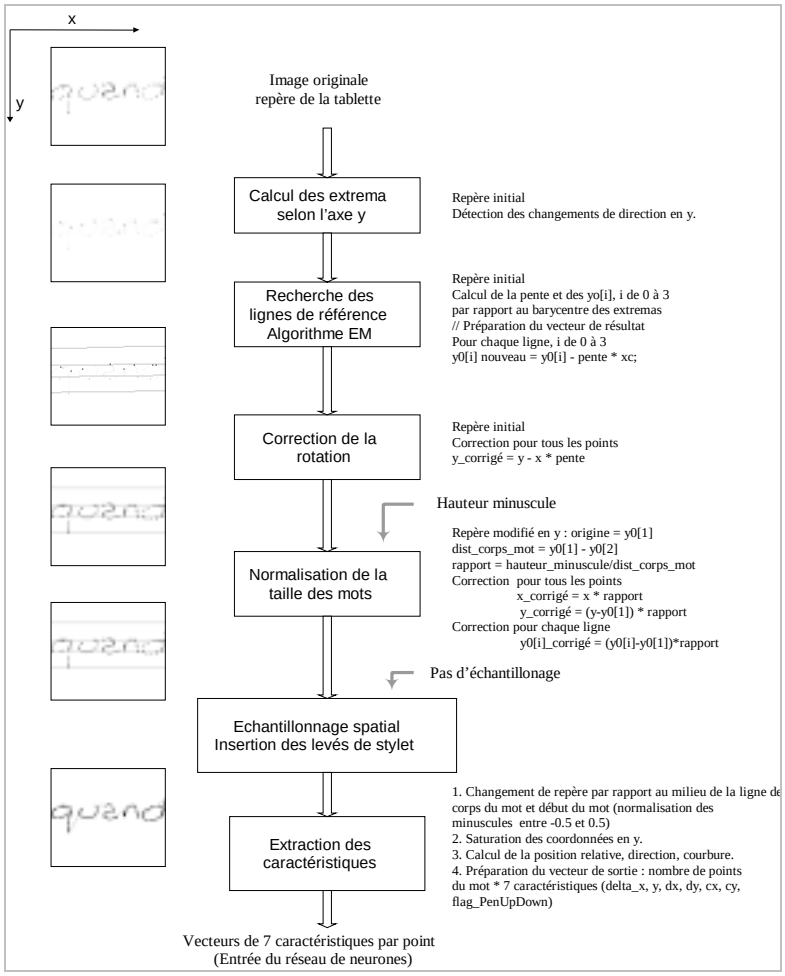

Figure 2. Schéma des prétraitements effectués pour l'entrée d'un TDNN, contexte dynamique uniquement.

2.2. Segmentation dynamique et reconnaissance par RC

Un mot manuscrit est donc représenté par une séquence temporelle, de longueur variable en fonction de la taille du mot, chaque élément de la séquence étant un vecteur de sept composantes, cf figure 3. Cette séquence va être balayée par le RC avec une fenêtre de taille fixe pour laquelle il calculera un vecteur O_t correspondant aux probabilités a posteriori des classes associées à ses différentes sorties (représentées par des classes caractères sur la figure 3). A l'intérieur de chaque fenêtre, des fonctions de convolution sur des sous-fenêtres glissantes à poids partagés sont opérées.

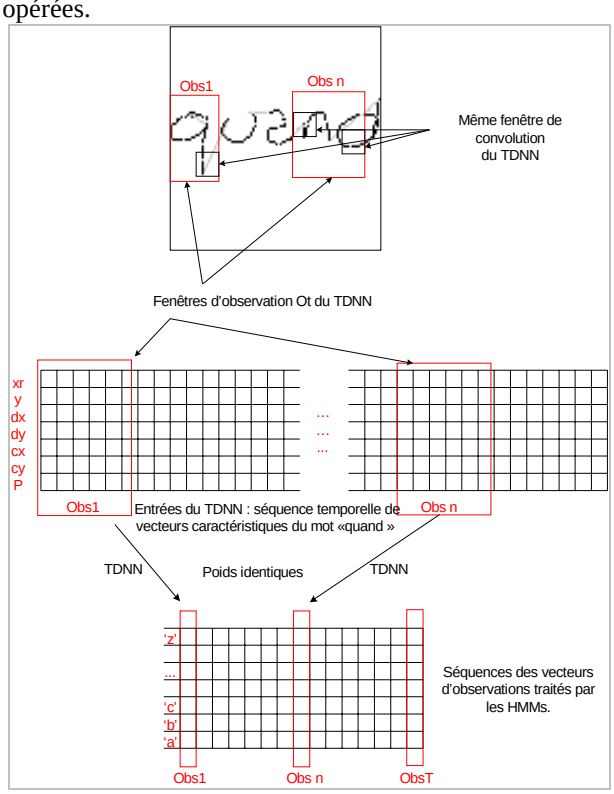

Figure 3 . Représentation de l'entrée du TDNN et de la segmentation dynamique par le TDNN

Nous avons utilisé pour cela un TDNN défini dans [Poi 01] avec une couche cachée de taille de fenêtre de convolution 7 (sous-fenêtre) et d'un délai égal à 2 et dont la couche de sortie utilise des fonctions d'activations de type Softmax. Après une analyse statistique sur la base IRONOFF de la taille moyenne d'un caractère, nous avons fixé la fenêtre d'observation à 20 unités avec un décalage de deux unités de temps.

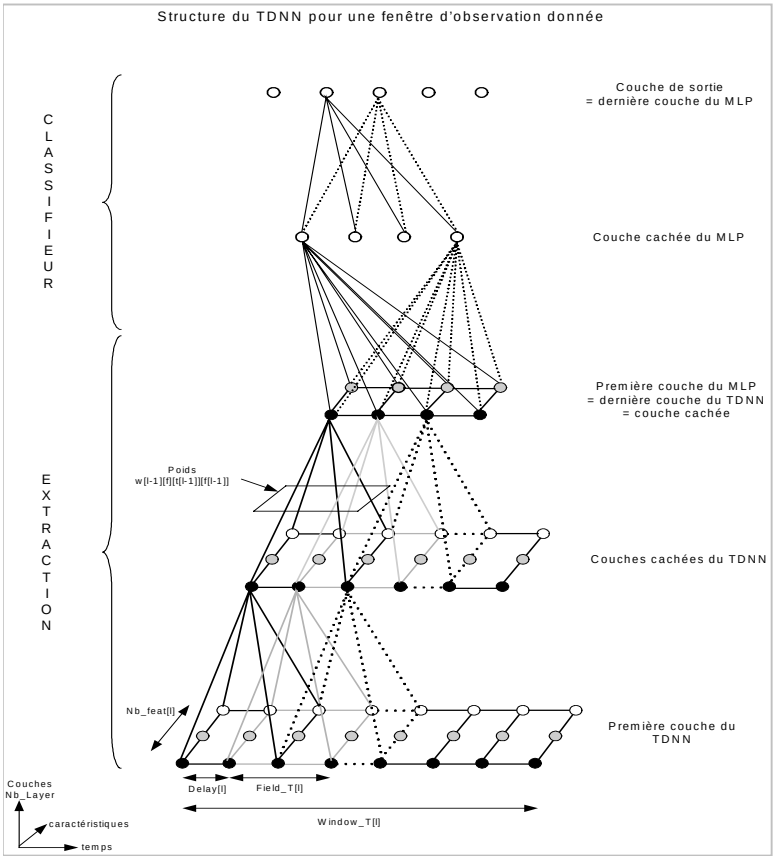

Figure 4. Structure du TDNN sur une observation donnée O_t

2.3. Analyse du treillis temporel par HMM

Pour chaque mot présent dans le lexique utilisé, un modèle HMM-mot est construit de manière dynamique au moment de la reconnaissance par concaténation de modèles HMMs-lettres, modèles gauche droite [Lal 99]. Dans l'exemple de la figure 5, un état unique est associé à chaque caractère, chacun de ces états correspond à une sortie du réseau de neurones. Les probabilités d'observation de chaque état émetteur des modèles-lettres $b_j(O_t)$ sont fournies par les sorties du réseau de neurones. Les probabilités de transitions entre états ne sont pas apprises, elles sont considérées constantes et équiprobables. Ainsi le système global calcule la vraisemblance d'un mot en multipliant toutes les probabilités d'observation le long du meilleur chemin du graphe obtenu par l'algorithme de Viterbi [Rab 89]. Lors de la reconnaissance, ce module de programmation dynamique fournit en sortie une liste des candidats mots dans l'ordre décroissant du score obtenu.

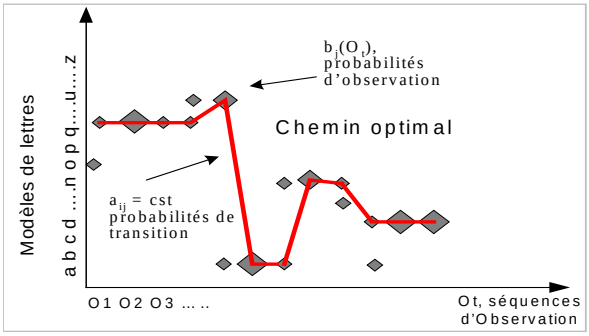

Figure 5. Segmentation dynamique du mot « quand »

3. Apprentissage global au niveau mot

L'objectif ici est de ne plus entraîner le réseau de neurones avec comme fonction objectif l'optimisation de la reconnaissance caractère, action très coûteuse – segmentation, étiquetage,… - mais de rétro-propager directement l'erreur au niveau mot dans le réseau de neurones pour mettre à jour ses paramètres. Pour cela, nous avons choisi une approche basée sur la minimisation de l'entropie croisée calculée sur le modèle vis-à-vis des données d'apprentissage. La fonction objectif définit un critère de type MMI (Maximum Mutual Information) simplifié, équation [1] prenant en compte la différence des logarithmes des vraisemblances entre le vrai HMM λ^{τ}, et le meilleur HMM λ^{*}.

$$L=\log P\left(O\middle|\lambda^{\tau}\right)-\log P\left(O\middle|\lambda^{*}\right) \qquad [1]$$

De cette façon, si λ^{*} se trouve être le vrai HMM, alors le critère est nul et les coefficients du réseau de neurones ne seront pas affectés. Sinon, la rétro-propagation du gradient de la fonction objectif est appliquée pour mettre à jour les poids *W* du réseau. On considère pour adapter notre réseau la règle de dérivation en chaîne d'où l'équation suivante :

$$\frac{\partial L}{\partial W_{ji}}=\sum_{t}\frac{\partial L}{\partial v_j(O_t)}\cdot\frac{\partial v_j(O_t)}{\partial W_{ji}} \qquad [2]$$

où *j* est l'indice du neurone concerné et *i* un neurone associée de la couche inférieure, *t* l'indice temporel d'observation et $v_j(O_t)$ le potentiel synaptique du neurone *j* considéré pour l'observation *t*. En posant $\delta_{j,t}$ le terme d'erreur à calculer pendant la phase de rétropropagation pour chaque neurone du réseau, on obtient :

$$\frac{\partial L}{\partial W_{ji}}=\sum_{t}\frac{\partial L}{\partial v_j(O_t)}\cdot x_i(O_t) \;=\; \sum_{t}\delta_{j,t}\cdot x_i(O_t) \quad avec \quad \delta_{j,t}=\frac{\partial L}{\partial v_j(O_t)} \qquad [3]$$

La propagation dans les couches cachées du réseau suit l'algorithme standard (eq. [4]) avec juste la prise en compte des fenêtres de convolution du TDNN.

$\delta_{j,t}=\sum_k \delta_{k,t}.W_{kj}.f'(v_{j,t})$ où f' est la dérivée de la fonction sigmoïde. [4]

Pour la couche de sortie du réseau, on utilise comme fonction de transfert la fonction Softmax d'où

$$\delta_{j,t}=b_j(O_t)\left(\frac{\partial L}{\partial b_j(O_t)}-\sum_k b_k(O_t).\frac{\partial L}{\partial b_k(O_t)}\right) \quad [5]$$

$$\text{or } \frac{\partial L}{\partial b_j(O_t)}=\frac{1}{b_j(O_t)}\left(\frac{P(O,q_t=j|\lambda^\tau)}{P(O,|\lambda^\tau)}-\frac{P(O,q_t=j|\lambda^*)}{P(O,|\lambda^*)}\right) \quad [6]$$

où $P(O,q_t=j|\lambda)$ est calculé à l'aide d'un algorithme de programmation dynamique.

Ainsi pour chaque observation O_t, les gradients positifs du HMM vrai et négatifs du HMM reconnu sont rétro-propagés dans les poids du réseau de neurones.

Par ailleurs, le système permet d'ajouter facilement une classe « rejet » entraînée indirectement via la contrainte de la fonction Softmax imposant que les sorties somment à 1, de cette façon celle-ci récupère les probabilités correspondant aux formes qui sont mal expliquées par les classes « lettre ».

4. Conclusion

Nous avons proposé dans cet article une amélioration de l'apprentissage d'un système hybride de reconnaissance de mots manuscrits en-ligne via une méthode d'apprentissage discriminant avec une fonction globale au niveau mot et la rétro-propagation directe sur les poids d'un réseau de neurones à convolution intégrant directement la phase de segmentation souvent très coûteuse. Un système hybride MLP/HMM [Tay 02] pour la reconnaissance de mots hors-ligne a déjà démontré l'intérêt de la méthode d'apprentissage globale : meilleurs résultats que l'apprentissage classique en deux étapes, l'un au niveau lettre, l'autre au niveau mot. Notre système ayant déjà apporté des améliorations significatives en reconnaissance de caractères manuscrits en ligne [Poi 02], nous pensons que les différentes expériences menées actuellement confirmeront le gain en taux de reconnaissance pour de la reconnaissance mot. Pour le moment, nous avons pu valider la bonne convergence de l'algorithme sans aucune initialisation préalable du réseau de neurones. Il reste à optimiser la structure des modèles HMMs, celle-ci a été choisie volontairement initialement très simple, avant de pouvoir comparer cette méthode en terme de taux de reconnaissance avec notre précédent système [Lal 99].

5. Bibliographie

[BEN 94] Bengio Y., LeCun Y., "Word Normalization for On-Line Handwritten Word Recognition", *Actes de International Conference on Pattern Recognition , ICPR'94*, Jerusalem, Israel, Octobre 1994, vol. 2, pp. 409-413.

[GUY 94] Guyon, I., Schomaker, L., Plamondon, R., Liberman, M. & Janet, S. "UNIPEN project of on-line data exchange and recognizer benchmarks", ICPR'94, pp. 29-33, Jerusalem, Israel, October 1994. IAPR-IEEE.

[JAE 00] Jaeger S., Manke S., Reichert J. , Waibel A., "On-Line Handwriting Recognition: The NPen++ Recognizer", *International Journal on Document Analysis and Recognition, IJDAR'00*, volume 3, p. 169-180, 2000.

[LAL 99] Lallican P.M., Extraction de la chronologie d'un tracé manuscrit a partir d'une image statique et exploitation pour la reconnaissance de caractères, Thèse doctorat, Universite de Nantes et Ecole Centrale de Nantes, Novembre 1999.

[LEC 01] LeCun Y., Bottou L., Bengio Y., Haffner P., "Gradient-Based Learning Applied to Document Recognition", *Intelligent Signal Processing*, p. 306-351, 2001.

[POI 01] Poisson E., Viard-Gaudin C., Réseaux de neurones à convolution : reconnaissance de l'écriture manuscrite non contrainte, Valgo 2001, ISSN 1625-9661, num. 01-02, 2001.

[POI 02] Poisson E., Viard-Gaudin C., LALLICAN P.M., « Réseaux de neurones à convolution : reconnaissance de l'écriture manuscrite non contrainte », *CIFED 2002*, Hammamet, Tunisie, Oct. 2002.

[RAB 89] Rabiner L.R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of IEEE*, volune 77, p. 257-285, 1989

[RAG 03] Ragot N., Anquetil E., "A generic hybrid classifier based on hierarchical fuzzy modeling experiments on on-line handwritten character recognition", IEEE Proceedings, *Actes des Seventh International Conference on Document Analysis and Recognition, ICDAR'2003*, Edinburgh, UK, vol. 2, p. 963-967, 2003.

[SIM 03] Simard S.Y., Steinkrauss D., Platt J.C., "Best Practices for Convolutional Neural Networks Appied to Visual Document Analysis", IEEE Proceedings, *ICDAR'2003*, Edinburgh, UK, vol. 2, p. 958-962, 2003.

[TAY 02] Tay Y.H., Offline handwriting recognition using artificial neural network and hidden Markov models, Thèse de doctorat, Ecole polytechnique de l'université de Nantes et Université de Technologie de Malaisie, Mars 2002.

[VIA 99] Viard-Gaudin C., Lallican P.M., Knerr S., Binter P., «The IRONOFF Dual Handwriting Database", *ICDAR'99*, pp. 455-458. Bangalore, , Sept. 1999.

[WIM 00] Wimmer Z., Garcia-Salicetti S., Lifchitz A., Dorizzi B., Gallinari P., Artières, T., « REMUS », http://www-connex.lip6.fr/~lifchitz/Remus/ .