# Université Paris-Sud

Ecole Doctorale Informatique de Paris-Sud
Laboratoire d'Informatique pour la Mécanique
et les Sciences de l'Ingénieur


Discipline : Informatique

## Thèse de doctorat

Soutenue le 13 mai 2015 par


**Théodore Bluche**


# Deep Neural Networks for Large Vocabulary Handwritten Text Recognition

## Résumé en Français

| | | |
|---|---|---|
| **Directeur de thèse :** | M. Hermann Ney | Professeur (RWTH, Aachen ; LIMSI, Paris) |
| **Co-directeur de thèse :** | M. Christopher Kermorvant | Docteur (Teklia, Paris ; A2iA, Paris) |

**Composition du jury :**

| | | |
|---|---|---|
| Président du jury : | | () |
| Rapporteurs : | M. Christophe Garcia | Professeur (INSA Lyon, LIRIS/IMAGINE) |
| | M. Enrique Vidal | Professeur (UPV, Espagne) |
| Examinateurs : | M. Balazs Kegl | Docteur (IN2P3, Paris) |
| | M. Thierry Paquet | Professeur (LITIS, Rouen) |

Groupe Traitement
de la Langue Parlée
LIMSI-CNRS
B.P. 133
91403 Orsay Cedex, France

A2iA
39, rue de la Bienfaisance
75008 Paris, France

ED IPS
Université Paris-Sud
UFR Sciences Orsay
Bât. 650
Orsay, France

# Introduction

Nous vivons dans un monde digital, où l'information est stockée et traitée automatiquement. L'écriture manuscrite n'échappe pas à la règle. Les enjeux de sa reconnaissance automatique sont multiples, allant du traitement automatique des chèques et de la dématérialisation de formulaires administratifs, à l'extraction automatique d'information d'archives, de registres, ou de courriers entrants au sein de grandes administrations et entreprises.

Le domaine de la reconnaissance automatique de l'écriture manuscrite a suscité l'intérêt de chercheurs depuis plus de soixante ans, et est toujours un champ de recherche très actif, comme l'attestent les nombreux projets financés et la tenue de conférences internationales sur le sujet. D'abord concentrés sur la reconnaissance de caractères isolés, les systèmes ont peu à peu évolué vers la reconnaissance de mots isolés, puis de lignes segmentées en mots, jusqu'à la reconnaissance directe de lignes aujourd'hui.

Les systèmes *"classiques"* de reconnaissance comportent aujourd'hui plusieurs composantes. Les lignes de textes doivent tout d'abord être extraites des pages, et nous supposons dans le cadre de cette thèse que cette étape est déjà faite. Ensuite, certaines sources de variabilités dans les images sont éliminées lors d'une étape de pré-traitement des images. Les étapes suivantes dépendent surtout de la granularité des modèles. Certains systèmes, peu adaptés au larges vocabulaires, modélisent les mots entiers directement, ce qui nécessite une segmentation de la ligne de texte en mots. Les systèmes modélisant les caractères passent plus facilement à l'échelle, mais la nature cursive de l'écriture manuscrite rend la segmentation en caractères difficile (Sayre, 1973). Alors que certains tentent une segmentation explicite et heuristique de l'image en caractères ou en partie de caractères (graphèmes ; Baret (1990); Knerr et al. (1998)), la plupart des systèmes actuels se reposent sur une segmentation implicite, obtenue comme produit de la reconnaissance. L'une des approches principale, choisie dans cette thèse, consiste à extraire séquences de caractéristiques de l'image en la parcourant avec une fenêtre glissante (Kaltenmeier et al., 1993). La transformation de ces séquences d'observations en séquences de caractères est généralement obtenue avec des Modèles de Markov Cachés (MMCs ; Rabiner & Juang (1986)). Les probabilités d'émission des MMCs peuvent être par exemple estimées avec des mixtures de Gaussiennes, ou avec des Réseaux de Neurones (RNs ; Bourlard & Morgan (1994)). L'autre possibilité consiste à de fournir à des réseaux de neurones avec une architecture appropriée, par exemple à convolutions, l'image de la ligne directemement, ce qui est fait dans les réseaux de neurones récurrents multi-dimensionnels (Graves & Schmidhuber, 2008). Un modèle de langue est ensuite appliqué aux séquences prédites afin de les contraindre à être plausibles.

Par certains aspects, ce type de systèmes peut être mis en parallèle avec ceux utilisés dans d'autre domaines. Par exemple, ils ressemblent aux modèles de reconnaissance de la parole où des séquences de caractéristiques sont extraites et tranformées en texte, également avec des MMCs et des modèles de langues. Les entrées étant des images, certaines composantes peuvent se rapprocher des systèmes de reconnaissance d'objets et de vision articifielle.

Dans ces deux domaines, on constate ces dernières années une réduction considérable des taux d'erreur dûe à l'utilisation de réseaux de neurones profonds. En reconnaissance d'écritures, ce type de réseaux est limité, soit à des architectures à convolutions, soit à des sous-problèmes, comme la localisation de mots-clés (Thomas et al., 2013) ou la reconnaissance de caractères isolés (Ciresan et al., 2010; Cireşan et al., 2012).

De plus, quand des réseaux de neurones profonds sont appliqués, on constate, notamment en reconnaissance de parole et d'objets, que l'extraction de caractéristiques devient moins cruciale.

D'autre part, des méthodes d'entrainements de ce type de réseau ont émergés ou regagné de l'intérêt ces dernières années, comme la technique de régularisation dite de *dropout*, ou les critères discriminants au niveau des séquences complètes.

Enfin, toujours concernant les réseaux de neurones, on constate en reconnaissance d'écriture un abandon progressifs des perceptrons multi-couches, pourtant toujours détenteurs d'excellents résultats en parole, au profit des réseaux de neurones récurrents, qui font partie de tous les meilleurs systèmes et gagnent la plupart des évaluations internationales.

Dans cette thèse, nous proposons d'étudier les réseaux de neurones profonds pour la reconnaissance d'écriture manuscrite, dans le cadre de systèmes hybrides RN/MMC. Les observations précédentes donnent naturellement lieu aux questions suivantes, auxquelles nous tentons d'apporter une réponse :

— *Avec des réseaux de neurones profonds pour la reconnaissance d'écriture, est-il toujours important de choisir et d'implémenter des caractéristiques pré-définies, ou bien les simples intensités de pixels sont-elles suffisantes ?*

— *Des réseaux de neurones profonds, comportant plus qu'une ou deux couches cachées, peuvent-ils apporter des améliorations significatives en reconnaissance d'écriture, et dans quelles situations ?*

— *Comment les perceptrons multi-couches (profonds) se comparent-ils aux désormais très populaires réseaux de neurones récurrents, très répendus en reconnaissance d'écriture manuscrite et produisant des résultats à l'état de l'art ?*

— *Quelles caractéristiques des réseaux de neurones récurrents les rendent si appropriés à la reconnaissance d'écriture manuscrite ?*

— *Quelles sont les bonnes stratégies d'entrainement de réseaux de neurones pour la reconnaissance d'écriture ? Est-ce que la Classification Temporelle Connexionniste (CTC, Graves et al. (2006)) peut être appliquée à d'autre types de réseaux de neurones que des réseaux récurrents ? Quelles améliorations peuvent être observées en utilisant un critère d'entrainement discriminant au niveau des séquences ?*

## Système de Base

Nous avons mené nos expériences sur trois bases de données publiques, Rimes (Augustin et al., 2006) , IAM (Marti & Bunke, 2002), et Bentham (Sánchez et al., 2014),

afin de comparer nos résultats à l'état de l'art. Elle représentent deux langues (l'anglais et le français), et deux époques (contemporaine et XIXème sciècle). Les conditions de collection de ces bases sont plus ou moins contraintes, en terme de contenu. Une description détaillée se trouve en Annexe A.

Dans cette thèse, nous nous focalisons sur le modèle optique de systèmes hybrides réseaux de neurones profonds / modèles de Markov cachés. Néanmoins, le sysème de reconnaissance complet inclut un certain nombre d'autres composantes, que nous fixons à l'avance. Le choix de ces composantes est validé sur l'ensemble de développement de chaque base, en mesurant le taux d'erreurs mots d'un modèle de Markov caché gaussien (cf. Chapitre 4).

**Pré-traitement des images.** Tout d'abord, les inclinaisons de la ligne de texte (Bloomberg et al., 1995), et des caractères (Buse et al., 1997) sont corrigées. Le contraste est renforcé (Roeder, 2009). Les régions horizontales correspondantes aux ascendants et descendants sont détectées (Vinciarelli & Luettin, 2001) et normalisées à une hauteur de 24 pixels.

**Extraction de caractéristiques.** Des séquences de vecteurs de caractéristiques sont extraites à l'aide de fenêtres glissantes (Kaltenmeier et al., 1993). Nous avons utilisé et comparé deux types de caractéristiques :

— **pré-définies :** 56 caractéristiques statistiques et géométriques (configurations de pixels, centre de gravité, intensités moyennes de pixels, histogrammes de gradients), extraites avec une fenêtre glissante de largeur 3px et de pas 3px (Bianne-Bernard, 2011; Bianne et al., 2011)

— **intensités de pixels :** vecteur contenant tous les niveaux de gris d'une fenêtre de taille $20 \times 32$px, résultant d'un redimensionnement d'une fenêtre glissante de largeur 45px et de pas 3px ($25 \times 32$px et 57px pour la base Bentham)

**Modèles de langue** Nous avons utilisé des modèles de langue de type "*n-gram*", estimant la probabilité d'un mot, conditionnée aux $n-1$ mots précédents, et limitée à un vocabulaire pré-défini.

Pour *IAM*, nous avons défini un modèle trigramme, estimé sur les corpus LOB, Brown et Wellington, avec un lissage Kneser-Ney modifié (Kneser & Ney, 1995). Les passages du corpus LOB qui apparaissent dans les ensembles de test d'IAM ont été retirés. Le vocabulaire est limité au 50,000 mots les plus fréquents. Le taux de mots hors vocabulaire et la perplexité de ce modèle de langue sont de 4.3% et 298 (*resp.* 3.7% et 329) sur l'ensemble de validation (*resp.* test).

Pour *Rimes*, nous avons défini un modèle quadrigramme, estimé sur les annotations de l'ensemble d'entraînement, avec un lissage Kneser-Ney modifié. Le vocabulaire contient 5,000 mots, et le taux de mots hors vocabulaire et la perplexité de ce modèle de langue sont de 2.9% et 18 (*resp.* 2.6% et 18) sur l'ensemble de validation (*resp.* test).

Pour *Bentham*, nous avons défini un modèle quadrigramme, estimé sur les annotations de l'ensemble d'entraînement, avec un lissage Kneser-Ney modifié. Les débuts et

fins de mots sont ajoutés aux unigrammes pour prendre en compte le grand nombre de coupure de mots présents dans cette base. Le taux de mots hors vocabulaire et la perplexité de ce modèle de langue sont de 5.6% et 103.1 sur l'ensemble de validation.

**MMCs gaussiens** La topologie retenue pour les MMCs comporte plusieurs états par caractères (5 pour Rimes, 6 pour IAM et Bentham), et des transitions uniquement d'un état vers lui-même et vers l'état suivant. L'espace est modélisé par un MMC à deux états. Nous avons entrainé des MMCs gaussiens, afin de valider les choix de prétraitement, de fenêtres glissantes, de modèle de langues et de topologie de MMCs. Les taux d'erreurs atteints sur les trois bases sont comparables à l'état de l'art en matière de MMC gaussiens.

## Systèmes Hybrides Perceptrons Multi-Couches Profonds / MMC

Alors qu'en reconnaissance de la parole, où les systèmes standards ressemblent beaucoup à ceux que nous utilisons (extraction de caractéristiques, modélisation par MMCs et modèle de langue), les perceptrons multi-couches (PMCs) sont devenus une composante essentielle, en particulier les PMCs profonds, en reconnaissance d'écriture ils tendent à être délaissés ces dernières années au profits des réseaux de neurones récurrents.

Dans le Chapitre 5, nous nous intéressons aux perceptrons multi-couches profonds, combinés aux MMCs. Nous entrainons des PMCs pour les trois bases et les deux types de caractéristiques, avec une à sept couche cachées, en utilisant un critère de classification local, et un critère discriminant au niveau de la séquence complète.

En premier lieu, nous avons étudié l'influence du contexte en entrée du réseau. Dans les systèmes hybrides RN/MMC, il est courant de concaténer plusieurs trames successives. Nous confirmons cette hypothèse : les réseaux utilisant un concaténation de la trame centrale avec un nombre fixe de trames précédentes et suivantes donnent de bien meilleurs résultats, comparés à ceux n'utilisant que la trame centrale. La différence est particulièrement visible sur les performances du réseau seul, sans MMC et modèle de langue. Avec ces derniers, qui ajoutent des contraintes sur les séquences d'états reconnus, les améliorations sont moindres quoique visible.

Ensuite, nous avons entrainé des PMC de différentes profondeurs. Nous montrons qu'en passant d'une à plusieurs couches cachées, il est possible d'atteindre une reduction relative du taux d'erreur au niveau trame de l'ordre de 20%, pour le réseau seul, sans contrainte linguisique. En ajoutant le MMC et le modèle de langue, les améliorations sont moins visibles, surtout pour les caractéristiques pré-définies, mais tout de même non négligeables, d'environ 5% relatifs.

Quand les entrées sont les intensités de pixels, la différence entre une et plusieurs couches cachées est bien plus nette. Le simple passage à deux couches cachées permet

d'obtenir une amélioration relative du taux d'erreur mots, incluant l'application du modèle de langue, allant jusqu'à 10%. Avec des réseaux plus profonds, la réduction observée monte à 17%. Ces résultats confirment l'hypothèse des modèles profonds que des représentations intermédiaires, de complexité croissante, sont calculées dans les couches basses du réseau, qui permettent une classification plus robuste.

En entrainant également des réseaux plus larges, c'est-à-dire avec plus d'unités dans les couches cachées, nous avons vérifié que ce n'est pas uniquement l'augmentation du nombre de paramètres libres, résultant de l'ajout de couches dans le PMC, qui sont à l'origine de l'amélioration des performances.

De plus, nous avons comparé les taux d'erreur mots obtenus avec les caractéristiques pré-définies et les intensités de pixels (cf. Table 1). L'écart de performance entre les systèmes utilisant ces deux types d'entrée, d'environ 10% pour des PMCs à une couche cachée en faveur des caractéristiques, est réduit de plus que de moitié quand on considère des réseaux profonds. Ce résultat suggère que l'effort mis en oeuvre pour sélectionner et implémenter les caractéristiques adaptées au problème et leur extraction n'est pas forcément nécessaire, et qu'un système ne faisant aucun hypothèse, considérant directement le signal d'entrée, est suffisant pour obtenir des performances raisonnables.

Table 1 – Comparaison du taux d'erreurs mots (%) obtenus avec des PMCs utilisant des caractéristiques pré-définies et des intensités de pixels

| | **Caractéristiques** | **Pixels** | |
|---|---|---|---|
| **Rimes** | | | |
| PMC à une couche | 14.0% | 15.3% | (+9.3%) |
| PMC profond | 13.5% | 14.0% | (+3.7%) |
| +sMBR | 12.5% | 12.6% | (+0.8%) |
| **IAM** | | | |
| PMC à une couche | 12.4% | 13.6% | (+9.7%) |
| PMC profond | 11.8% | 12.3% | (+4.2%) |
| +sMBR | 10.9% | 11.7% | (+7.3%) |
| **Bentham** | | | |
| PMC à une couche | 21.5% | 28.8% | (+34.0%) |
| PMC profond | 20.1% | 22.4% | (+11.4%) |
| +sMBR | 18.6% | 19.4% | (+4.3%) |

Enfin, nous avons appliqué un critère d'entrainement des réseaux discriminant au niveau des séquences de mots. Ce type de critères vise à augmenter la probabilité *a posteriori* de la séquence de mots correcte, étant donnée la séquence d'observations, ce qui reflète mieux le but final des systèmes de reconnaissance. En pratique, pour des raisons computationnelles, cela consiste à extraire des treillis de reconnaissance, incluant le réseau de neurones et le modèle de langue, et à distinguer la séquence correcte des séquences concurrentes. L'entrainement discriminant au niveau des séquences

est classique en reconnaissance de la parole, où il permet d'obtenir des améliorations relatives des taux d'erreurs entre 5 et 10%.

Nous avons choisi le critère sMBR (Kingsbury, 2009), qui donnait de meilleurs résultats que d'autres sur une tâche de reconnaissance de la parole. Nous montrons qu'affiner un PMC entrainé au niveau trame avec sMBR permet d'atteindre des améliorations significatives, entre 5 et 13% en taux d'erreur mots (cf. Table 1) et jusqu'à 20% en taux d'erreur caractères. Nous montrons également l'importance du modèle de langue dans le calcul de ce critère, bien qu'un simple unigramme semble suffisant pour obtenir ces améliorations.

## Systèmes Hybrides Réseaux de Neurones Récurrents Profonds / MMC

Le Chapitre 6 est consacré aux réseaux de neurones récurrents (RNRs) profonds, combinés aux MMCs. Les neurones de ces réseaux sont des unités Longue Mémoire à Court Terme (LMCT, en anglais *Long Short-Term Memory*, Hochreiter & Schmidhuber (1997)). Les couches récurrentes contiennent en fait deux couches, pour parcourir les séquences dans les deux directions. Nous entrainons ces réseaux avec le critère de Classification Temporelle Connexionniste (CTC), défini par Graves et al. (2006). Ce type de réseau est populaire en reconnaissance d'écriture, et ses bonnes performances en ont fait un standard du domaine (Graves & Schmidhuber, 2008; Bluche et al., 2014; Moysset et al., 2014; Kozielski et al., 2014; Doetsch et al., 2014).

Afin de faire rentrer ces RNRs dans le cadre des systèmes hybrides RN/MMC, nous définissons des MMCs de caractères à un état, ainsi qu'un MMC pour le symbole *blank* de la CTC, optionnel entre deux caractères. Nous entrainons des RNRs pour les trois bases et les deux types de caractéristiques, avec une à treize couche cachées, alternant des couches récurrentes et des couches standards.

Tout d'abord, nous avons reproduit l'expérience du chapitre précédent, consistant à concaténer les trames en entrée du réseau pour lui fournir un contexte plus grand. Nous montrons que cette stratégie, bénéfique pour les PMCs, tend à déteriorer les résultats des RNRs. Nous vérifions que la récurrence de ces réseaux leur permettent d'apprendre effectivement les dépendences et le contexte nécessaire pour prédire les caractères. Nous observons que ce contexte dépasse souvent la zone d'un caractère, s'étendant parfois au mot entier, suggérant qu'un large contexte est parfois nécessaire pour prédire de façon fiable un caractère, et supportant le choix des unités LMCT.

Ensuite, nous montrons que la récurrence dans ces réseaux, avec l'entrainement CTC, est particulièrement importante dans les couches supérieures. Les couches basses, en particulier quand les entrées sont les intensités de pixels, servent plus à extraire des caractéristiques élémentaires du signal, rendant la récurrence moins cruciale. Cette observation pourrait permettre de simplifier les RNRs, en n'implémentant la récurrence, plus chère en temps de calcul, que dans les couches hautes du réseau.

Nous reportons également des améliorations significatives grâce à l'augmentation de la profondeur des réseaux. Nous nous sommes limités à treize couche cachées. Nous avons observé des baisses de taux d'erreur au niveau caractère, quand le réseau est utilisé seul, sans modèle de langue, entre 35 et 50% pour les caractéristiques pré-définies, et entre 70 et 80% pour les pixels, quand nous passons d'un réseau à une couche cachée à un réseau profond.

Les améliorations des taux d'erreurs ne sont pas aussi impressionnantes quand les réseaux sont inclus dans la chaîne de reconnaissance complète, incluant les contraintes liguistiques. Cela est certainement dû au fait que les erreurs corrigées par des réseaux plus profonds étaient également corrigées par le modèle de langue. Ceci dit, nous observons quand même des baisses de taux d'erreurs, entre 10 et 15% pour les caractéristiques, et aux alentours de 40% pour les pixels, ce qui n'est pas négligeable.

Comme nous l'avons fait pour les PMCs, nous avons vérifié que ces résultats n'étaient pas simplement dûs à l'augmentation du nombre de paramètres libres du modèle. Nous avons comparé des RNRs plus profonds à des RNRs plus "larges", c'est-à-dire avec plus d'unités dans les couches cachées, et obtenu de meilleurs résultats en augmentant la profondeur.

De plus, nous confirmons l'efficacité du *dropout*, une technique de régularisation des réseaux consistant à ignorer les réponses de certains neurones lors de l'entrainement. Cette méthode, introduite par Hinton et al. (2012), a récemment été appliquée aux RNRs (Pham et al., 2014; Zaremba et al., 2014). En particulier, avec une architecture de réseaux proche de la nôtre, mais pour des RNR multi-dimensionnels, Pham et al. (2014) ont fait le choix d'appliquer le *dropout* uniquement en sortie des couches LMCT.

En explorant de manière exhaustive les différentes positions pour le dropout, nous avons d'une part confirmé que cette technique permettait pratiquement toujours d'améliorer les performances des RNRs seuls. La position du dropout relative aux couches LMCT semble particulièrement importante, et nous reportons des améliorations significatives par rapport à la méthode proposée par Pham et al. (2014) quand le dropout est appliqué en entrée des LMCTs plutôt qu'en sortie.

Cependant, dans le système complet, l'ensemble des sorties est passé au décodeur, et il apparaît que, dans ce cas, le dropout sur les poids de classification soit bénéfique. Finalement, les meilleurs résultats ont été obtenu en utilisant le dropout avant les couches récurrentes basses, et après les couches récurrentes hautes.

Enfin, nous montrons que l'écart de performance entre les systèmes traitant les caractéristiques pré-définies et les simples valeurs de pixels est largement réduit quand les réseaux sont profonds, passant d'une soixantaine à une dixaine de pourcents, et devenant pratiquement négligeables quand le dropout est utilisé en entrainement (cf. Table 2). Comme pour les PMCs, la nécessité de trouver et d'implémenter une extraction de caractéristiques pertinantes disparaît, au profit d'un apprentissage automatique, sans a priori spécifique, des caractéristiques par le réseau.

Table 2 – Comparaison du taux d'erreurs mots (%) obtenus avec des RNRs utilisant des caractéristiques pré-définies et des intensités de pixels

| | **Caractéristiques** | **Pixels** | |
|---|---|---|---|
| **Rimes** | | | |
| LMCTB-RNR à une couche | 14.9% | 24.1% | (+65.1%) |
| LMCTB-RNR profond | 12.9% | 14.0% | (+7.9%) |
| +dropout | 12.7% | 12.7% | (+0.0%) |
| **IAM** | | | |
| LMCTB-RNR à une couche | 13.4% | 83.2% | (+520.0%) |
| LMCTB-RNR profond | 11.4% | 12.8% | (+12.3%) |
| +dropout | 11.2% | 11.4% | (+0.9%) |
| **Bentham** | | | |
| LMCTB-RNR à une couche | 20.6% | 33.8% | (+64.1%) |
| LMCTB-RNR profond | 18.0% | 20.3% | (+12.8%) |
| +dropout | 16.0% | 17.0% | (+6.3%) |

## Une Comparaison Expérimentale de l'Entrai-nement CTC et au Niveau Trame

Au Chapitre 7, nous nous intéressons plus particulièrement aux sorties des réseaux et à la méthode d'entrainement. En effet, dans les expériences précédentes, les PMCs prédisent des états de MMC, six par caractères, et sont entrainés au niveau trame, tandis que les RNRs sont entrainés avec la CTC à prédire directement des caractères, ainsi qu'un symbole spécial, *blank*, correspondant à une absence de prédiction de caractère. Les deux types de système obtiennent de bons résultats.

L'entrainement CTC définit des transitions possibles entre les différents labels, et considère toutes les segmentations possibles du signal d'entrée en prédictions, ce qui rappelle l'entrainement Baum-Welsch des MMCs. Dans un premier temps, nous montrons que l'entrainement CTC est très similaire à l'entainement global des systèmes hybrides RN/MMC par une procédure *forward-backward*, déjà proposée par plusieurs groupes de recherche dans les années 90 (Senior & Robinson, 1996; Hennebert et al., 1997; Yan et al., 1997). Les principales différences sont l'absence de probabilités de transition et de probabilités à priori d'états dans la CTC, ainsi qu'une topologie spécifique incluant un symbole *blank*.

Nous avons étudié l'influence de la topologie, en faisant varier le nombre d'états des MMCs de caractères, et la présence ou non d'un MMC de blank à un état. Les mixtures de gaussiennes et les PMCs obtiennent de meilleurs résultats avec plusieurs états par caractère, et pas de blank, alors que les RNRs atteignent leur meilleure performance avec la topologie définie par la CTC.

D'autre part, l'entrainement CTC n'étant pas théoriquement limité aux RNRs, nous l'avons appliqué à un PMC mais les résultats obtenus sont relativement décevants. Par contre, en gardant plusieurs états par caractères, sans *blank*, l'entrainement CTC modifié permet une légère amélioration des résultats du PMC, quoique limitée, ce qui est cohérent avec les conclusions de Hennebert et al. (1997) : si l'ensemble d'entrainement est suffisamment grand, il y a peu ou pas de différence entre l'entrainement au niveau trame, avec des étiquettes de trames fixes, et l'entrainement consistant à considérer toutes les segmentations possibles en états de MMC.

Nous avons ensuite poussé plus loin la comparaison entre la CTC et l'entrainement au niveau des frames. Nous avons entrainé des PMCs et des RNRs avec les deux critères, en faisant également varier la topologie. Nous avons trouvé que le cadre strict de la CTC, avec un état par caractère et le *blank*, était particulièrement adapté aux RNRs. Dans toutes les situations sauf celle-ci, le taux d'erreur diminuait avec le nombre d'état (cf. Figure 1). Pour les PMCs, nous avons obtenu les meilleurs résultats finaux avec six états par MMC, et sans blank, ce qui correspond à la topologie choisie pour les chapitres précédents.

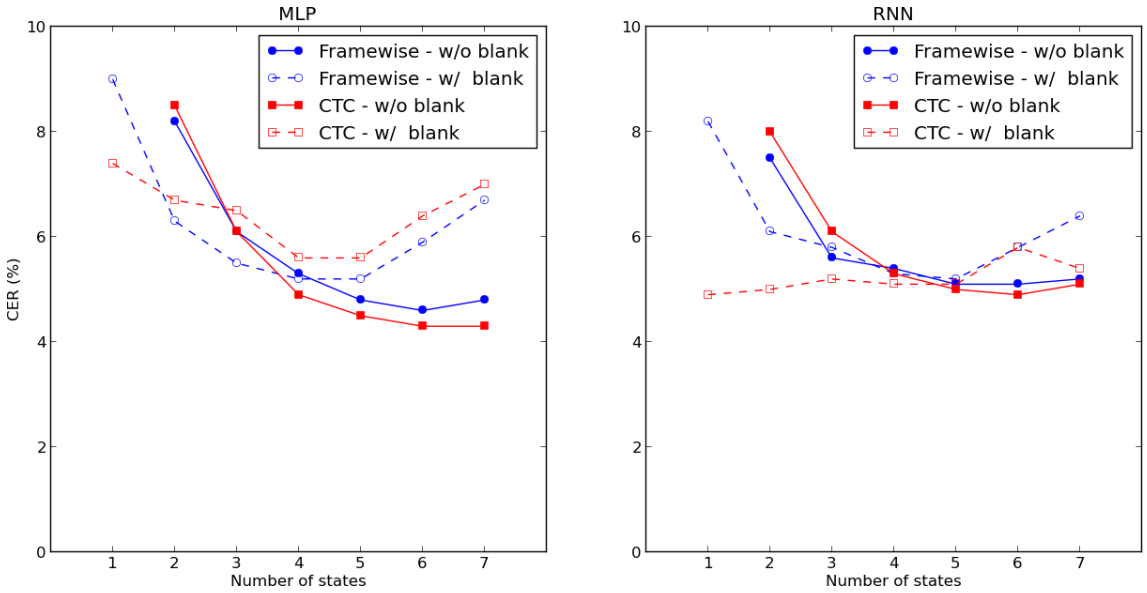


FIGURE 1 – Comparaison des taux d'erreur caractères (CER%) avec l'entrainement CTC et au niveau trame, avec et sans le symbole spécial *blank*. (à gauche : PMC ; à droite : RNR).

De plus, quand les réseaux sont considérés seuls, sans modèle de langue, l'erreur de classification augmente avec le nombre d'états, à l'inverse du taux d'erreur dans la chaine complète. Cela confirme néanmoins que le problème de classification est plus simple avec un nombre limité d'états. En revanche, pour l'entrainement CTC sans blank, l'ajout d'états dans les modèles fait baisser le taux d'erreur, ce qui suggère que le choix d'une longueur de MMC adaptée permet d'aider l'algorithme à produire de bons alignements pendant la procédure d'entrainement.

Nous avons étudié l'interaction de l'entrainement CTC avec le *blank*. Nous avons observé qu'en utilisant directement la CTC sans blank, l'entrainement ne convergeait pas, ou vers des solutions sous-optimales. Dans ces cas-là, nous avons dû faire une première passe d'entrainement au niveau trame, avant de poursuivre avec la CTC. Cela suggère que, quand les caractères sont représentés par seulement quelques états, le blank peut aider dans la procédure d'alignement. En fait, nous expliquons que la présence du blank, associé à la CTC provoque dans les premiers instants de l'entrainement une tendance du réseau à ne prédire que ce symbole. Les prédictions de caractères apparaissent ensuite sous forme de pics localisés, qui peuvent être intéressant pour avoir un décodage riche et rapide quand les contraintes linguistiques sont introduites.

Enfin, concernant cette partie, nous concluons que la CTC, incluant la procédure *forward-backward*, mais plus spécifiquement le *blank* et la représentation des caractères par une seule sortie de réseau, ou un seul état de MMC, est particulièrement adaptée aux RNRs, et leur permet d'obtenir relativement facilement de très bonnes performances. Quant aux PMCs, nous avons confirmé que la topologie des MMCs choisies pour cette thèse était bonne. La procédure *forward-backward*, que l'on trouve également dans l'entrainement discriminant au niveau des séquences, ne permet qu'une légère amélioration.

## Combinaisons et Résultats Finaux

Dans le dernier chapitre, nous étudions les systèmes de reconnaissance complets, incluant le modèle hybride RN/MMC, et le décodage avec les contraintes linguistiques (vocabulaire et modèle de langue). Nous comparons directement les différents modèles optiques (RNRs et PMCs), et les différents types d'entrées (pixels et caractéristiques). Nous présentons également les résultats de nos systèmes sur les ensembles de test de chaque base de données, afin de les comparer aux meilleurs résultats publiés. Nous donnons enfin dans ce chapitre une description détaillée des systèmes soumis à l'évaluation HTRtS 2014, ainsi que leurs résultats.

Nous confirmons nos observations des chapitres précédents, à savoir que les pixels permettent d'obtenir des résultats compétitifs avec ceux atteints avec les caractéristiques pré-définies, pour des modèles optiques comportant des réseaux de neurones profonds. D'autre part, nous montrons que des PMCs profonds atteignent des performances similaires à celles des RNRs, malgré la domination de ces derniers dans la littérature récente en reconnaissance d'écriture.

Cependant, comme vu précédemment, les PMCs requièrent un ajustement du nombre d'états dans les MMCs, et de la quantité de contexte à donner en entrée des réseaux. De plus, ce niveau de performance a été atteint grâce à l'entrainement discriminant au niveau des séquences. A l'inverse, les RNRs, avec l'entrainement CTC, modélisent les séquences de caractères directement, et sont plus simples à optimiser pour atteindre facilement les résultats présentés. Ils gèrent la séquence d'entrée, et notamment la taille de contexte nécessaire aux prédictions directement, et il n'y a pas besoin d'alignements

forcés des trames avec les états de MMCs pour l'entrainement. Ceci dit, pour une comparaison plus juste de ces modèles, nous prévoyons d'appliquer la technique du *dropout* dans les PMCs, ainsi que la procédure d'entrainement en séquence aux RNRs.

Nous soulignons également l'importance des paramètres de décodage. Le facteur optique, qui contrôle l'influence du score du modèle optique par rapport à celui du modèle de langue, a le plus grand impact sur les résultats finaux. A travers les différentes expériences et configurations, nous avons trouvé que 0.1 semble être la meilleure valeur pour les systèmes basés sur des PMCs, tandis que 1.0 est le meilleur choix pour les RNRs. L'ajustement de la pénalité d'insertion de mot et du facteur appliqué aux probabilités a priori d'états de MMC peuvent également apporter quelques pourcents d'amélioration.

Le contraintes linguisiques, c'est-à-dire le vocabulaire et le modèle de langue, sont cruciaux pour atteindre les résultats présentés dans cette thèse. Sans eux, les taux d'erreur sont largement plus haut, en particulier pour les PMCs, pour lesquels la moitié des mots seraient alors mal reconnus. En n'appliquant qu'un vocabulaire, sans donner un score aux séquences de mots, plus d'un mot sur quatre serait encore faux.

Les deux types de réseau de neurones et d'entrées nous donnent quatre systèmes de reconnaissance, qui obtiennent tous des résultats comparables, mais se comportent différemment, et font des erreurs différentes. En combinant ces systèmes avec deux méthodes, au niveau des transcriptions (ROVER ; Fiscus (1997)), et au niveau des treillis de reconnaissance (Xu et al., 2011), nous montrons la complémentarité des différents modèles. En effet, nous avons obtenu plus qu'un pourcent d'amélioration absolue du taux d'erreur mots, comparé à celui du meilleur système. La méthode basée sur les treillis de reconnaissance, qui considère un plus grand nombre d'alternatives, produit de meilleurs et plus robustes résultats.

Table 3 – Résultats sur la base Rimes (taux d'erreur mots - WER% - et caractères - CER%).

| | | **Dev.** | | **Eval.** | |
|---|---|---|---|---|---|
| | | **WER%** | **CER%** | **WER%** | **CER%** |
| GMMC | Caractéristiques | 17.2 | 5.9 | 15.8 | 6.0 |
| PMC | Caractéristiques | 12.5 | 3.4 | 12.7 | 3.7 |
| | Pixels | 12.6 | 3.8 | 12.4 | 3.9 |
| RNR | Caractéristiques | 12.8 | 3.8 | 12.6 | 3.9 |
| | Pixels | 12.7 | 4.0 | 13.8 | 4.6 |
| | Combinaison ROVER | 11.3 | 3.5 | 11.3 | 3.7 |
| | Combinaison de treillis | 11.2 | 3.3 | **11.2** | 3.5 |
| | Pham et al. (2014) | - | - | 12.3 | **3.3** |
| | Doetsch et al. (2014) | - | - | 12.9 | 4.3 |
| | Messina & Kermorvant (2014) | - | - | 13.3 | - |
| | Kozielski et al. (2013) | - | - | 13.7 | 4.6 |
| | Messina & Kermorvant (2014) | - | - | 14.6 | - |
| | Menasri et al. (2012) | - | - | 15.2 | 7.2 |

Table 4 – Résultats sur la base IAM (taux d'erreur mots - WER% - et caractères - CER%).

| | | Dev. | | Eval. | |
|---|---|---|---|---|---|
| | | **WER%** | **CER%** | **WER%** | **CER%** |
| GMMC | Caractéristiques | 15.2 | 6.3 | 19.6 | 9.0 |
| PMC | Caractéristiques | 10.9 | 3.7 | 13.3 | 5.4 |
| | Pixels | 11.4 | 3.9 | 13.8 | 5.6 |
| RNR | Caractéristiques | 11.2 | 3.8 | 13.2 | 5.0 |
| | Pixels | 11.8 | 4.0 | 14.4 | 5.7 |
| | Combinaison ROVER | 9.6 | 3.6 | 11.2 | 4.7 |
| | Combinaison de treillis | 9.6 | 3.3 | **10.9** | **4.4** |
| | Doetsch et al. (2014) | **8.4** | **2.5** | 12.2 | 4.7 |
| | Kozielski et al. (2013) | 9.5 | 2.7 | 13.3 | 5.1 |
| | Pham et al. (2014) | 11.2 | 3.7 | 13.6 | 5.1 |
| | Kozielski et al. (2013) | 11.9 | 3.2 | - | - |
| | Messina & Kermorvant (2014) | - | - | 19.1 | - |
| | Espana-Boquera et al. (2011) | 19.0 | - | 22.4 | 9.8 |

Enfin, nous avons appliqué nos modèles sur les ensembles de test de chaque base, afin de comparer leur résultats à ceux publiés par d'autres sur les mêmes données. Nous les rappelons ici, Table 3 pour Rimes, Table 4 pour IAM, et Table 5 pour Bentham. Nous voyons que sur Rimes et IAM, tous nos systèmes simples atteignent un niveau de performance comparable à l'état de l'art. De plus, la combinaison des reconnaisseurs égale ou dépasse tous les meilleurs résultats publiés sur les trois bases : Rimes, IAM et Bentham.

# Conclusions et Perspectives

En conclusion, nous présentons dans cette thèse une étude poussée des réseaux de neurones profonds pour la reconnaissance d'écriture manuscrite et de leur intégration dans des systèmes hybrides RN/MMC.

Nous nous sommes intéressés aux entrées des réseaux, et avons montré qu'avec la profondeur des réseaux, l'écart de performance entre des caractéristiques extraites prédéfinies et les simples valeurs de pixels se réduisait, diminuant ainsi l'intérêt de la création et de l'implémentation d'une bonne extraction de caractéristiques. D'autre part, nous avons étudié la taille du contexte à fournir en entrée du réseau, et montré que ce paramètre était crucial pour les PMCs, alors que les résultats n'étaient pas améliorés par l'introduction d'un plus grand contexte pour les RNRs, qui peuvent apprendre à utiliser un contexte arbitrairement grand à travers leurs connexions récurrentes. Pour les valeurs de pixels, nous n'avons pas expérimenté beaucoup de tailles différentes des fénêtres glissantes, et des améliorations peuvent être attendues si une attention plus fine est portée à ce paramètre.

Table 5 – Résultats sur la base Bentham (taux d'erreur mots - WER% - et caractères - CER%).

(a) Tâche restreinte

| Model | | WER% | CER% |
|---|---|---|---|
| PMC profond | *Caractéristiques* | 18.6 | 7.5 |
| | *Pixels* | 20.9 | 8.2 |
| RNR profond | *Caractéristiques* | 16.2 | 5.4 |
| | *Pixels* | 16.9 | 5.9 |
| Combinaison de treillis | | **14.1** | **5.0** |
| **CITlab** | | 14.6 | - |
| Notre système (Compétition) | | 15.1 | - |

(b) Tâche non restreinte

| Model | | WER% | CER% |
|---|---|---|---|
| PMC profond | *Caractéristiques* | 13.2 | 4.9 |
| | *Pixels* | 14.4 | 6.1 |
| RNR profond | *Caractéristiques* | 11.2 | 4.0 |
| | *Pixels* | 11.5 | 4.4 |
| | uRNR1 | 10.9 | 4.0 |
| | uRNR2 | 10.5 | 3.7 |
| | uRNR3 | 10.2 | 3.6 |
| Combinaison de treillis | | **8.6** | **3.1** |
| **A2iA (système de production)** | | **8.6** | - |
| Notre système (Compétition) | | 11.1 | - |

Les nombreuses expériences réalisées nous permettent d'affirmer que des gains significatifs de performance résultent d'une plus grande profondeur de réseau, c'est-à-dire d'un plus grand nombre de couches cachées. Cela constitue la majeure contribution de notre travail, sachant qu'aujourd'hui, en reconnaissance d'écriture manuscrite, très peu de travaux ont considéré des architectures à plusieurs couches cachées, mises à part les architectures à convolutions.

Il est important de signaler que les réseaux à convolutions sont une excellente alternative aux PMCs et obtiennent de très bons résultats en vision artificielle, de même que les RNRs multi-dimensionnels sont une excellente alternative aux RNRs présentés, et constituent l'état de l'art en reconnaissance d'écriture. Dans un souci de cohérence, nous ne discutons pas dans cette thèse de nos travaux sur les réseaux à convolutions (Bluche et al., 2013a,b) ou utilisant des RNRs multi-dimensionnels (Bluche et al., 2014; Moysset et al., 2014; Pham et al., 2014), mais il va de soi qu'une étude comparative les incluant doit être menée.

Enfin, nous avons étudié l'influence des sorties des réseaux (notamment le nombre d'états de MMCs par caractère et la présence d'un symbole *blank*), et la méthode d'en-

trainement des réseaux, qui a en particulier souligné la bonne interaction entre la CTC, le *blank*, et les RNRs. Des gains intéressants ont été atteints avec un entrainement en séquence, et l'utilisation du dropout, et une poursuite logique de ces travaux consisterait à appliquer aux RNRs l'entrainement en séquence, et aux PMCs la technique du *dropout*.

Parmi les directions futures de recherche, outre l'application de tels systèmes à d'autres types de données, telles que des scripts différents comme l'arabe, où des données plus difficiles comme celles de l'évaluation MAURDOR (Brunessaux et al., 2014), il semble aujourd'hui important de trouver des pistes pour améliorer les modèles de langues, et pour passer de la reconnaissance de lignes, nécessitant une segmentation explicite, à une reconnaissance de bout en bout, qui transcrirait les images des documents entiers, et serait dans la veine de l'évolutions des systèmes faisant de moins en moins d'hypothèses.

# Bibliographie


Augustin, E., Carré, M., Grosicki, E., Brodin, J.-M., Geoffrois, E., & Preteux, F. (2006). RIMES evaluation campaign for handwritten mail processing. In *Proceedings of the Workshop on Frontiers in Handwriting Recognition*, 1.

Baret, O. (1990). *Régularités singulairtés de représentations et leur complémentarité : application à la reconnaissance de l'écriture manuscrite non contrainte*. Ph.D. thesis.

Bianne, A.-L., Menasri, F., Al-Hajj, R., Mokbel, C., Kermorvant, C., & Likforman-Sulem, L. (2011). Dynamic and Contextual Information in HMM modeling for Handwriting Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *33*(10), 2066 – 2080.

Bianne-Bernard, A.-L. (2011). *Reconnaissance de mots manuscrits cursifs par modèles de Markov cachés en contexte*. Ph.D. thesis, Telecom ParisTech.

Bloomberg, D. S., Kopec, G. E., & Lakshmi Dasari (1995). Measuring document image skew and orientation. *Proc. SPIE Document Recognition II*, *2422*(302), 302–316.

Bluche, T., Louradour, J., Knibbe, M., Moysset, B., Benzeghiba, M. F., & Kermorvant, C. (2014). The A2iA Arabic Handwritten Text Recognition System at the Open HaRT2013 Evaluation. In *11th IAPR International Workshop on Document Analysis Systems (DAS)*, (pp. 161–165). IEEE.

Bluche, T., Ney, H., & Kermorvant, C. (2013a). Feature Extraction with Convolutional Neural Networks for Handwritten Word Recognition. In *12th International Conference on Document Analysis and Recognition (ICDAR)*, (pp. 285–289). IEEE.

Bluche, T., Ney, H., & Kermorvant, C. (2013b). Tandem HMM with convolutional neural network for handwritten word recognition. In *17th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 2390–2394). IEEE.

Bourlard, H., & Morgan, N. (1994). *Connectionist speech recognition : a hybrid approach Chapter 7*, vol. 247 of *The Kluwer international series in engineering and computer science : VLSI, computer architecture, and digital signal processing*. Kluwer Academic Publishers.

Brunessaux, S., Giroux, P., Grilhères, B., Manta, M., Bodin, M., Choukri, K., Galibert, O., & Kahn, J. (2014). The Maurdor Project : Improving Automatic Processing of Digital Documents. In *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*, (pp. 349–354). IEEE.

Buse, R., Liu, Z. Q., & Caelli, T. (1997). A structural and relational approach to handwritten word recognition. *IEEE Transactions on Systems, Man and Cybernetics*, *27*(5), 847–61.

Ciresan, D. C., Meier, U., Gambardella, L. M., & Schmidhuber, J. (2010). Deep big simple neural nets for handwritten digit recognition. *Neural computation*, *22*(12), 3207–3220.

Cireşan, D. C., Meier, U., Gambardella, L. M., & Schmidhuber, J. (2012). Deep big multilayer perceptrons for digit recognition. In *Neural Networks : Tricks of the Trade*, (pp. 581–598). Springer.

Doetsch, P., Kozielski, M., & Ney, H. (2014). Fast and robust training of recurrent neural networks for offline handwriting recognition. (pp. –).

Espana-Boquera, S., Castro-Bleda, M. J., Gorbe-Moya, J., & Zamora-Martinez, F. (2011). Improving offline handwritten text recognition with hybrid HMM/ANN models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *33*(4), 767–779.

Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates : Recognizer output voting error reduction (ROVER). In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU1997)*, (pp. 347–354). IEEE.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks. In *International Conference on Machine learning*, (pp. 369–376).

Graves, A., & Schmidhuber, J. (2008). Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. In *Advances in Neural Information Processing Systems*, (pp. 545–552).

Hennebert, J., Ris, C., Bourlard, H., Renals, S., & Morgan, N. (1997). Estimation of global posteriors and forward-backward training of hybrid HMM/ANN systems.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv :1207.0580*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Kaltenmeier, A., Caesar, T., Gloger, J. M., & Mandler, E. (1993). Sophisticated topology of hidden Markov models for cursive script recognition. In *Document Analysis and Recognition, 1993., Proceedings of the Second International Conference on*, (pp. 139–142). IEEE.

Kingsbury, B. (2009). Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, (pp. 3761–3764). IEEE.

Knerr, S., Augustin, E., Baret, O., & Price, D. (1998). Hidden Markov model based word recognition and its application to legal amount reading on French checks. *Computer Vision and Image Understanding*, *70*(3), 404–419.

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1, (pp. 181–184). IEEE.

Kozielski, M., Doetsch, P., Hamdani, M., & Ney, H. (2014). Multilingual Off-line Handwriting Recognition in Real-world Images. (pp. 1–1).

Kozielski, M., Doetsch, P., Ney, H., et al. (2013). Improvements in RWTH's System for Off-Line Handwriting Recognition. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, (pp. 935–939). IEEE.

Marti, U.-V., & Bunke, H. (2002). The IAM-database : an English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, *5*(1), 39–46.

Menasri, F., Louradour, J., Bianne-Bernard, A.-L., & Kermorvant, C. (2012). The A2iA French handwriting recognition system at the Rimes-ICDAR2011 competition. In *Document Recognition and Retrieval Conference*, vol. 8297.

Messina, R., & Kermorvant, C. (2014). Surgenerative Finite State Transducer n-gram for Out-Of-Vocabulary Word Recognition. In *11th IAPR Workshop on Document Analysis Systems (DAS2014)*, (pp. 212–216).

Moysset, B., Bluche, T., Knibbe, M., Benzeghiba, M. F., Messina, R., Louradour, J., & Kermorvant, C. (2014). The A2iA Multi-lingual Text Recognition System at the second Maurdor Evaluation. In *14th International Conference on Frontiers in Handwriting Recognition (ICFHR2014)*, (pp. 297–302).

Pham, V., Bluche, T., Kermorvant, C., & Louradour, J. (2014). Dropout improves recurrent neural networks for handwriting recognition. In *14th International Conference on Frontiers in Handwriting Recognition (ICFHR2014)*, (pp. 285–290).

Rabiner, L., & Juang, B.-H. (1986). An introduction to hidden Markov models. *ASSP Magazine, IEEE*, *3*(1), 4–16.

Roeder, P. (2009). *Adapting the rwth-ocr handwriting recognition system to french handwriting*. Ph.D. thesis, Master's thesis, Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Aachen. Germany.

Sánchez, J. A., Romero, V., Toselli, A., & Vidal, E. (2014). ICFHR 2014 HTRtS : Handwritten Text Recognition on tranScriptorium Datasets. In *International Conference on Frontiers in Handwriting Recognition (ICFHR)*.

Sayre, K. M. (1973). Machine recognition of handwritten words : A project report. *Pattern recognition*, *5*(3), 213–228.

Senior, A., & Robinson, T. (1996). Forward-backward retraining of recurrent neural networks. *Advances in Neural Information Processing Systems*, (pp. 743–749).

Thomas, S., Chatelain, C., Paquet, T., & Heutte, L. (2013). Un modèle neuro markovien profond pour l'extraction de séquences dans des documents manuscrits. *Document numérique*, *16*(2), 49–68.

Vinciarelli, A., & Luettin, J. (2001). A new normalisation technique for cursive handwritten words. *Pattern Recognition Letters*, *22*, 1043–1050.

Xu, H., Povey, D., Mangu, L., & Zhu, J. (2011). Minimum Bayes Risk decoding and system combination based on a recursion for edit distance. *Computer Speech & Language*, *25*(4), 802–828.

Yan, Y., Fanty, M., & Cole, R. (1997). Speech recognition using neural networks with forward-backward probability generated targets. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 4, (pp. 3241–3241). IEEE Computer Society.

Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent neural network regularization. *arXiv preprint arXiv :1409.2329*.