

# PROPOSITION DE STAGE

*Campagne 2011*



Laboratoire L3i

## Sujet Stage :

Conception d'une méthode de reconnaissance d'écriture manuscrite cursive adaptative

## Résumé du travail proposé :

Le but de stage est de développer, à partir d'une méthode d'identification de scripteur et d'une méthode de reconnaissance d'écriture manuscrite cursives développées au laboratoire, une méthode de reconnaissance d'écriture manuscrite cursive adaptative. L'idée de cette méthode est de personnaliser le moteur de reconnaissance de mots cursifs en fonction de l'identité du scripteur ou de son style d'écriture.

## Mots clés :

Reconnaissance d'écriture manuscrite cursive, identification du scripteur, identification du type d'écriture d'un scripteur, personnalisation.

## Informations complémentaires :

**Encadrant(s)** : Sophea Prum, Muriel Visani (encadrant(s) scientifique(s)).

**Axe thématique** :

« Données complexes, Images et Documents »

**Axe stratégique** : « Environnement »

**Cadre de coopération** : projet Reconomad

**Date de début du stage** : Février 2011

**Durée du contrat** : 5 mois

## Contexte de l'étude:

Actuellement, les documents manuscrits proviennent principalement des archives (documents anciens par exemple ou documents administratifs). Chaque jour, des millions de formulaires papier sont remplis. Le traitement de ces formulaires représente un travail considérable de gestion : transmission, indexation, stockage, etc... Pour traiter automatiquement ces formulaires par le biais de systèmes informatiques, on a besoin de saisir préalablement les contenus des différents champs dans le système informatique. Pour éviter cette étape de saisie à la fois fastidieuse et coûteuse, le projet RecoNomad vise à proposer un système de reconnaissance automatique ou semi-automatique du formulaire utilisé et des contenus remplis dans le formulaire. Les formulaires

sont remplis en posant le formulaire papier sur une tablette électro-magnétique captant et enregistrant l'écriture, sans dispositif supplémentaire (pas de stylo spécial), cf. Figure 1.



Figure 1 : Tablette électro-magnétique.

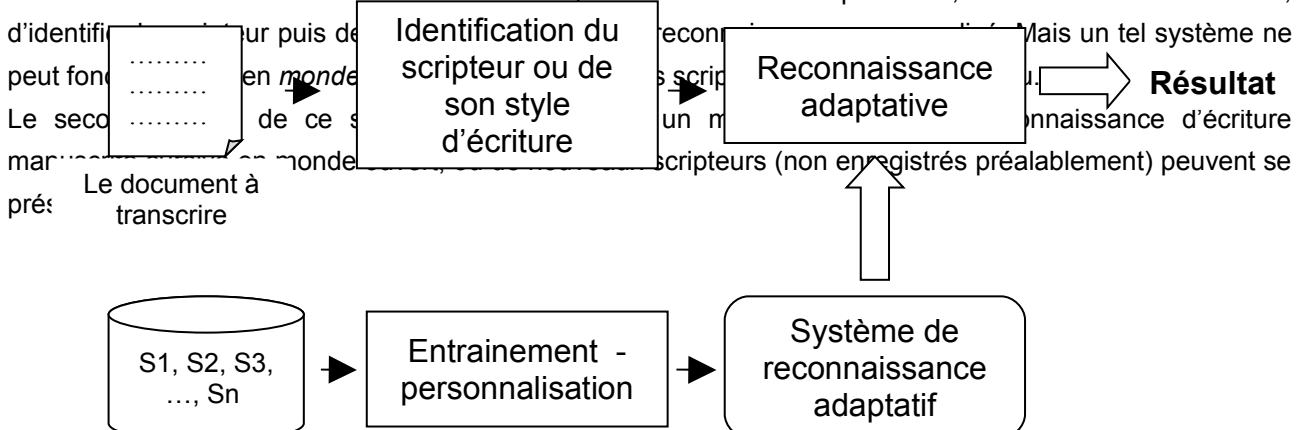
## Description du sujet :

Les documents manuscrits sont produits par les humains avec d'innombrables variations, provenant en partie des variations d'écriture entre scripteurs, mais aussi des variations entre deux occurrences d'un même mot écrit par un même scripteur. Les caractéristiques de l'écriture dépendent en particulier de l'origine du scripteur, du contexte et du contenu. Cette variabilité très importante constitue un problème majeur pour la reconnaissance d'écriture manuscrite cursive.

Une manière de limiter les variations est de mener, préalablement à la phase de reconnaissance d'écriture, une phase d'identification de scripteur (cf. Figure 2.). Ainsi, nous pouvons concevoir un modèle de reconnaissance d'écriture dédié à chacun des scripteurs dans la liste (writer-dependent). Le système de reconnaissance d'écriture manuscrite cursive personnalisé ainsi créé devra être plus performant que le moteur de reconnaissance d'écriture manuscrite cursive générique. Un exemple de méthode de reconnaissance de mots cursifs writer-dependent est présenté dans [1].

Mlle Sophea PRUM, dans le contexte de sa thèse, a déjà développé un système de reconnaissance d'écriture manuscrite cursive performant [2,3]. Dans le cadre de son stage de Master 2, M. Quang Anh BUI a développé un système d'identification de scripteur performant [4, 5]. Les caractéristiques utilisées sont à la fois des caractéristiques *en-ligne* capturées par la tablette et des caractéristiques *hors-ligne* reconstruites à partir de ce signal en-ligne

Le premier objectif de ce stage sera de concevoir, à partir de ces deux systèmes, un outil de reconnaissance personnalisé en fonction du scripteur ou style de l'écriture (cf. Figure 2). Ce système est facilement réalisable par combinaison des deux systèmes existants, puisqu'il suffit d'entraîner un modèle de reconnaissance d'écriture manuscrite cursive pour chacun des scripteurs et, en mode reconnaissance,



Deux pistes sont possibles :

- La première piste consiste à améliorer le système d'identification de scripteur existant pour permettre le rejet, c'est-à-dire la détection d'un scripteur qui n'aurait pas été préalablement enregistré. Tandis que, pour un scripteur connu, on pourra charger le moteur de reconnaissance personnalisé correspondant, dans le cas d'un scripteur inconnu il faudra utiliser le moteur de reconnaissance générique.
- La seconde piste consiste à modifier le système d'identification de scripteur de manière à ce qu'il permette non pas d'identifier le scripteur, mais son style d'écriture. On passe alors d'un problème d'apprentissage supervisé à un modèle d'apprentissage non supervisé. Il faudra dans ce cas modifier les caractéristiques utilisées pour choisir des caractéristiques plus adaptées à la caractérisation du style d'écriture. L'une des principales difficultés consistera à trouver une méthode automatique permettant de sélectionner automatiquement le nombre optimal de styles d'écritures à prendre en compte.

**Figure 2** : schéma du processus du système de reconnaissance adaptative. Les documents S1 à Sn sont des exemples de l'écriture des scripteurs 1 à n.

### Prérequis et contraintes particulières :

- Programmation en C/C++
- Bonnes bases en data-mining et/ou apprentissage machine
- Rudiments d'analyse d'images

### Références bibliographiques :

- [1] Marcus Liwicki, Andreas Schlapbach, and Horst Bunke : Writer-Dependent Recognition of Handwritten Whiteboard Notes in Smart Meeting Room Environments. The Eighth IAPR Workshop on Document Analysis Systems
- [2] S.Prum, M.Visani, JM.Ogier, On-line Handwriting word recognition using a bi-character model. 20th International Conference on Pattern Recognition, Istanbul, 2010.
- [3] S.Prum, M.Visani, JM.Ogier, Cursive on-line Handwriting word recognition using a bi-character model for large lexicon applications 12<sup>th</sup> International Conference on Frontiers in Handwriting

Recognition, Kolkata 2010.

[4] BUI Quang Anh, Identification du scripteur pour la reconnaissance de l'écriture manuscrite cursive. Mémoire de fin d'étude Master 2, 2010.

[5] Q.A. Bui, M. Visani, S. Prum and J.M. Ogier. Writer Identification using TF-IDF for Cursive Handwritten Word Recognition. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR) Beijing, China, pp. 844-848, September 2011.

## **Contacts – liens :**

**Email :** [sophea.prum@univ-lr.fr](mailto:sophea.prum@univ-lr.fr)

**Lien vers le fichier de description :** (PDF)

---

## **Présentation libre :**

---

### **Argumentaire interne (non visibles vis-à-vis de l'extérieur) :**

### **Enjeu du stage par rapport à la politique scientifique du laboratoire :**

Ce sujet de stage s'inscrit parfaitement dans le cadre de l'axe thématique « Données complexes, Images et Documents ». Il est très lié aux deux premiers verrous scientifiques identifiés dans cette thématique, à savoir :

- l'extraction d'informations dans des contextes complexes ou dégradés. En effet, le sujet écrit sur une tablette sur laquelle un formulaire a été posé, mais les conditions de capture nomades font qu'un certain nombre de bruits peuvent apparaître, tels que des mouvements parasites du stylo, des irrégularités dans l'écriture ou des mouvements du formulaire sur la tablette. Tous ces facteurs dégradent fortement l'écriture et rendent la reconnaissance d'écriture manuscrite cursive (qui constitue en soi une tâche compliquée) encore plus difficile dans notre contexte.
- La gestion de la complexité pour le passage à l'échelle, notamment au regard de la multiplication des attributs. En effet, l'écriture manuscrite cursive est composée de motifs complexes caractérisés par une très grande variabilité. Elle nécessite donc l'utilisation de nombreuses caractéristiques pour sa reconnaissance. Cela pose le problème de la malédiction de la dimensionnalité (liés au fait que les espaces en grandes dimensions sont presque vides), et nous impose de concevoir des solutions adaptées (réduction de la dimensionnalité, sélection de variables, méta-apprentissage...).

L'adéquation avec l'axe stratégique « Environnement » se fait au travers du contexte applicatif de ce sujet. En effet, dans ce domaine, il existe de nombreux cas de figure où l'utilisation d'un formulaire papier posé sur une tablette électromagnétique portable pour l'analyse automatique d'écriture manuscrite cursive peut être très utile. Surtout dans un contexte parfois rural où les utilisateurs sont réticents vis-à-vis des PDA. On pense notamment aux gardes forestiers chargés

de faire des recensements d'espèces naturelles, au suivi des progressions d'insectes parasites de la culture du riz en Asie du Sud-Est, etc...

## **Positionnement du travail dans le cadre de la politique de capitalisation du laboratoire (1) :**

Les choix technologiques faits dans le contexte de ce stage sont totalement cohérents avec la politique de capitalisation du laboratoire. En effet, le module développé par l'étudiant sera intégrée dans la plateforme Reconomad, déjà capitalisée au sein du laboratoire. L'étudiant devra programmer en langage C++.

## **Mise à disposition de livrables :**

Après 2 mois, l'étudiant devra fournir un prototype de système de reconnaissance d'écriture manuscrite cursive adaptative en monde fermé utilisant les deux modules déjà développés au laboratoire (identification de scripteur et reconnaissance d'écriture manuscrite cursive).

Après 4 mois, l'étudiant devra fournir un prototype de système de caractérisation et de reconnaissance du type d'écriture du scripteur.

Avant la fin de son stage, l'étudiant devra fournir un prototype de moteur de reconnaissance d'écriture manuscrite cursive adaptative fonctionnant en monde ouvert.

La mise à disposition des livrables se fera dans le cadre de la politique de propriété intellectuelle du projet Reconomad.

*(1) Rappel : les ressources communes du laboratoire ne seront désormais attribuées que si elles rentrent dans le cadre de la politique de capitalisation du laboratoire.*