International Workshop IHMIM'14
03-06 May 2014 Hammamet - TUNISIA

Vers un système de reconnaissance optique des caractères dans des documents multilingues : Français-Amazighe

Khadija EL GAJOUI[*], Fadoua ATAA ALLAH[**]

[*]*Laboratoire de recherche en Informatique et Télécommunications*
Faculté des sciences – Rabat
khadija.gajoui@gmail.com

[**] *Centre d'Etude Informatique des Systèmes d'Information et de Communication*
Institut Royal de la Culture Amazighe
ataaallah@ircam.ma

Résumé: La reconnaissance optique de caractères est un processus qui permet de convertir un texte présenté par une image numérique en un texte modifiable.
Le problème de l'OCR a été exploré en profondeur pour l'alphabet latin ainsi que pour d'autres langues. Cependant, il n'y a pas beaucoup de systèmes OCR fiables disponibles pour la langue amazighe. Les études concernant les systèmes existants d'OCR pour la langue amazighe se sont intéressées à l'écriture amazighe en alphabet tifinaghe.
Cependant, cet alphabet n'a été généralisé que récemment avec la création de l'Institut Royal de la Culture Amazighe en 2001. D'où l'intérêt de traiter les documents amazighes écrits en alphabet latin et arabe, qui représentaient les alphabets les plus utilisés au Maroc. Nous focalisons notre étude dans cet article sur les documents amazighes transcrits en latin.

Mots clés: Amazigh, document multilingue, image texte, reconnaissance optique des caractères.

INTRODUCTION

La reconnaissance optique de caractères ou OCR (en anglais : Optical Recognition Character) est une technologie qui permet de convertir différents types de documents tels que les documents papiers scannés, les fichiers PDF ou les photos numériques vers des formats modifiables et exploitables [Eikvil 93]. Sur le plan méthodologique, l'OCR propose des approches différentes suivant le mode d'écriture : manuscrit ou imprimé. Deux domaines distincts sont considérés, il s'agit de la reconnaissance statique, dite encore «hors-ligne », qui travaille sur un instantané d'encre numérique (sur une image) et la reconnaissance dynamique « en-ligne » où les symboles sont reconnus au fur et à mesure qu'ils sont écrits à la main. La technologie d'OCR a été appliquée ces dernières années à travers tout le spectre d'industries en train de révolutionner le processus de gestion des documents.

Les systèmes d'OCR ont permis à des documents numérisés de se transformer en documents entièrement consultables avec le contenu du texte qui est reconnu par les ordinateurs.

Cependant, après plus de deux décennies de recherche sur la numérisation des documents, ces systèmes peuvent encore laisser quelques imperfections pour parvenir à une réédition du document ce qu'il peut être dû aux différents problèmes dont la qualité du document et de l'impression, la discrimination de la forme, le type d'acquisition, les variations des dimensions, le nombre de scripteurs, la taille du vocabulaire, etc.

Dans le reste de l'article, nous définissons dans la section1, l'architecture d'un système OCR et nous présentons les approches développées pour chaque module du système. Nous introduisons l'écriture amazighe dans la section 2 et nous décrivons, dans la section 3, l'approche adoptée pour la reconnaissance de l'écriture de la langue amazighe par le biais du caractère latin et l'architecture du système proposé.

Finalement nous finissons dans la section 4 avec une conclusion et les perspectives.

1. SYSTEME OCR

L'objectif d'un système OCR est de reconnaître le texte et puis le convertir en une forme modifiable. La reconnaissance optique de caractères implique la traduction du texte dans l'image en codes de caractères modifiables tel que l'ASCII.

Les systèmes OCR proposés par les différents chercheurs sont composés d'un ensemble de modules. L'architecture du système varie d'un système à un autre en fonction des besoins.

Le système suivant peut être une généralisation de tous les systèmes proposés.

Figure 1. *Système de reconnaissance des caractères multilingues.*

Un système de reconnaissance fait appel généralement aux étapes suivantes :

Acquisition qui permet la conversion du document papier sous la forme d'une image numérique, prétraitement qui consiste à préparer les données issues du capteur à la phase suivante, identification de langues qui permet d'identifier les différentes langues dans un document, segmentation pour extraire les éléments du document (ligne, mot, caractère,...), extraction des caractéristiques, classification des éléments extraits, suivies éventuellement d'une phase de post-traitement.

Plusieurs approches et techniques ont été développées au niveau de chaque module par exemple la binarisation [Noor 05], la correction d'inclination [Noor 05] [Belaïd 92], l'encadrement [BOUSLIMI 06], l'élimination de bruits [BOUSLIMI 06] [Noor 05] [Charles & al. 12], la normalisation [BOUSLIMI 06] et la squelettisation [Muaz 10] dans la phase de prétraitement et la projection horizontale [Abu Obaida 11] [Tan & al. 99],la Densité optique [Tan & al. 99] et le Bounding box [Tan & al. 99] pour la détection de langues. Au niveau de la segmentation deux approches sont distinguées, une approche globale qui se base sur la description du mot comme étant une entité et l'approche analytique se basant sur la segmentation du mot en parties.

Pour la phase de classification il existe deux approches, l'approche statistique utilisant l'approche bayésienne, la méthode du plus proche voisin, Modèle de Markov Caché [Muaz 10], Réseau de neurones [El Ayachi & al. 11] [Noor 05] [Park & al. 08] et l'approche structurelle qui se base sur les relations entre des éléments simples ou primitives.

2. ÉCRITURE AMAZIGHE

La langue amazighe, ou tamazight, est présente à l'heure actuelle dans une dizaine de pays de l'ensemble Maghreb-Sahara-Sahel : Maroc, Algérie, Tunisie, Libye, Égypte, Niger, Mali, Burkina-Faso et Mauritanie. Mais l'Algérie et le Maroc sont, de loin, les deux pays qui comptent les populations amazighophones les plus importantes. Signe fort de l'identité amazighe, la langue amazighe est riche d'une tradition orale qui a su intégrer les médias modernes. De plus, la renaissance volontariste de l'alphabet traditionnel, le tifinaghe, a permis de suppléer à la mémoire collective, de traduire les œuvres majeures du patrimoine mondial et développer une littérature amazighe qui répond à une forte demande.

Les amazighs possèdent donc depuis l'antiquité un système d'écriture qui leur est propre [GACI 11].

Mais, depuis l'aube de l'histoire, lorsqu'il s'agit de rédiger des documents consistants, les amazighs ont eu recours aux langues et/ou aux alphabets des peuples dominants avec lesquels ils étaient en contact : punique, latin puis arabe ou français.

Pour transcrire l'amazighe, trois systèmes d'écritures sont utilisés [Skounti & al. 03] :

- ✓ Le tifinaghe, comme alphabet authentique attesté dans les inscriptions libyques depuis l'antiquité.
- ✓ L'alphabet arabe, suite à l'arrivée des arabes à la fin du 6ème siècle.
- ✓ Le latin, dès la fin du 19ème siècle, par des savants coloniaux et plutard par des chercheurs nationaux.

Dans ce travaille, nous nous intéressons à la transcription de la langue amazighe en latin.

Suite à l'exploration d'une ensemble de documents amazighes transcrits en latin, il s'agit de « CHOIX DE VERSION BERBÈRES PARLER DU SUD-OUEST MAROCAINE » d'Arsène roux [Roux 51], « MOTS ET CHOSES BERBÈRES » de Emile Laoust [Laoust 20] et « THE ARGAN TREE AND ITS TASHELHIYT BERBER LEXICON » de Harry Stroomer [Stroomer 08], nous avons constaté que le latin utilisé dans la transcription est un mélange entre le latin, le latin étendu A et le latin étendu additionnel.

Le tableau ci-dessous représente un exemple de caractères utilisés pour la transcription de la langue amazighe. Ces caractères sont composés de caractères latins et des diacritiques, qui représentent des signes accompagnant une lettre ou un graphème. Ces diacritiques peuvent être placés au-dessus (diacritique suscrit), au-dessous (diacritique souscrit) ou après

(diacritique adscrit).

Ā	ā	Ă	ă	A^c	a^c	B^w	b^w	B^c	b^c
Ḅ	ḅ	Ḍ	ḍ	D^c	d^c	Ĕ	ĕ	F^w	f^w
Ġ	ġ	G^w	g^w	Ḡ	ḡ	Ḥ	ḥ	Ḫ	ḫ
K^w	K^w	L^c	L^c	Ḷ	ḷ	M^w	m^w	Ŏ	ŏ
Ṛ	ṛ	Ṡ	ṡ	T^c	t^c	Ṭ	ṭ	Ů	ů
Ŭ	ŭ	Ẓ	ẓ						

Table 1. *Un exemple des caractères amazighes transcrits en latin.*

3. SYSTÈME ADOPTÉ

Le but de notre travaille consiste à élaborer un système de reconnaissance des caractères dans un document multilingue contenant du texte français et amazighe transcrit en latin. La vue globale du système est présentée sur la figure ci-dessous.

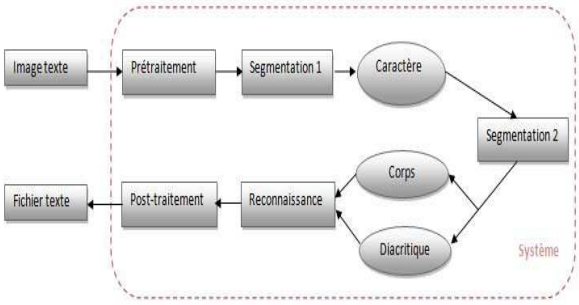

Figure 2. Vu globale du système adopté

L'image est entrée dans le système et subie un ensemble de prétraitements. Elle est segmentée d'abord en lignes et en caractères. Ensuite, une extraction du corps et du diacritique d'un caractère est effectuée. Cette étape aura comme sortie le diacritique et le corps ainsi que la position du diacritique par rapport au corps pour le cas du caractère avec diacritique et retournera l'image d'entrée dans le cas d'un caractère sans diacritique. L'étape de reconnaissance consiste à reconnaitre les éléments extraits. Le post-traitement est composée de deux étapes : d'abord une combinaison entre le diacritique et le corps reconnus pour former le caractère final en se basant sur la position du diacritique puis l'étape de vérification par l'utilisateur.

3.1. Acquisition

Dans la phase d'acquisition, nous utilisons un scanner optique pour scanner les pages du document et les mettre sous forme d'une image sous le format jpg, gif ou tif (types supportés par le système).

3.2. Prétraitement

Dans la phase de prétraitement, nous choisissons d'appliquer quatre différents traitements sur l'image afin d'améliorer sa qualité. Ces quatre traitements sont :

- La binarisation : permet de transformer l'image d'entrée en image binaire. Dans cette étape nous effectuons un seuillage par méthode d'OTSU [Noor 05], par la moyenne de l'histogramme et par la valeur de 127 qui représente la mi-valeur des niveaux de gris, puis une binarisation par rapport au seuil résultant.
- L'encadrement :

C'est une technique qui sert à encadrer le caractère dans une image afin d'éliminer la partie vide de l'image. Ainsi, nous avons proposé un algorithme pour l'encadrement qui consiste à parcourir l'image et éliminer sa partie vide pour retourner une image ne contenant que le caractère.

- La normalisation : Cette technique a comme objectif de normaliser la taille de l'image. La phase de segmentation donne comme sortie des images de taille différentes d'où l'intérêt de faire une normalisation permettant d'avoir une taille fixe pour toutes les images. La taille choisie pour notre système est de 64 * 64 pixels. D'autre part, nous avons développé une méthode qui normalise la taille du caractère à partir de l'image encadrée en ajoutant des pixels blancs pour avoir des images de même taille sans toucher à la forme du caractère.
- La correction d'inclination : C'est une technique qui vise à corriger l'inclination d'écriture. Ce problème se produit généralement dans la phase d'acquisition. La méthode la plus connue par sa performance pour corriger l'inclination est la méthode de Hough [El Ayachi & al. 11]. Ainsi, nous l'avons utilisée dans notre système.

3.3. Segmentation

La phase de segmentation, dans notre système, consiste dans un premier temps à séparer les lignes, puis à extraire les caractères. Dans ce cadre, nous avons utilisé un histogramme vertical pour segmenter l'image texte en lignes, puis un histogramme horizontal pour extraire les mots et les caractères.

Ainsi, nous avons structuré le résultat en une matrice dont le nombre de lignes est le nombre de caractères dans l'image et le nombre de colonnes est 4 qui correspond au cordonnées des quatre coins de chaque caractère dans l'image.

Dans un deuxième temps, nous avons effectué une segmentation par extraction des composants connectés dans le but de séparer entre le corps et le diacritique d'un caractère.

Nous avons fait la distinction entre le corps et le diacritique en comparant la taille de chaque élément. L'élément dont la taille est la plus grande correspond au corps.

3.4. Extraction des caractéristiques

La phase d'extraction des caractéristiques est une étape très importante dans le système d'OCR. Dans notre système, nous avons choisis des caractéristiques locales du type numérique / statistique. Ces caractéristiques sont le moment du Hu [El Ayachi & al. 11], la transformation de walsh [ElKonyaly & al. 97] et l'histogramme de LBP (Local Pinary Patterns) [Pietikäinen & al. 11].

3.5. Classification

L'approche adoptée dans notre système est l'approche statistique. Elle correspond aux caractéristiques extraites qui sont du type numérique / statistique.

La méthode choisie dans cette approche est la méthode de reconnaissance par réseau de neurone. La capacité des réseaux de neurones à généraliser et apprendre à partir des données et des exemples s'apparente à notre capacité d'apprendre à partir de l'expérience.

Les réseaux de neurones se basent sur deux phases : la phase d'apprentissage et la phase de reconnaissance.

L'apprentissage consiste à descendre le réseau de façon itérative en ajustant les poids à chaque passage selon le calcul d'erreur jusqu'à ce qu'il n'y ait plus d'amélioration. Pour cela, un algorithme de rétro-propagation de l'erreur est mis en œuvre dont le principe est de :

- initialiser la matrice des poids au hasard.
- choisir un exemple en entrée.
- propager le calcul de cette entrée à travers le réseau.
- calculer la sortie de cette entrée.
- mesurer l'erreur de prédiction par différence entre sortie réelle et sortie prévue.
- calculer la contribution d'un neurone à l'erreur à partir de la sortie.
- déterminer le signe de modification du poids.
- corriger les poids des neurones pour diminuer l'erreur.

Le processus recommence ainsi, à partir du choix de l'exemple en entrée, jusqu'à ce qu'un taux d'erreur minimal soit atteint.

3.6. Post traitement

La première étape dans la phase de post-traitement consiste à former le caractère par combinaison entre le corps et le diacritique puis composer le mot. La seconde étape dans notre système est manuelle. L'utilisateur se charge de confirmer le résultat obtenu.

4. Conclusion

Notre travail consiste à concevoir un système de reconnaissance des caractères de la langue amazighe transcrits en caractères latins.

Nous avons présenté le système OCR et décrit ses modules comportant les étapes de l'acquisition, le prétraitement, la segmentation, l'extraction des caractéristiques, la classification et le post-traitement ainsi que les approches développées pour chaque module. Ensuite nous avons menu une étude concernant la langue amazighe et les caractères utilisés dans la transcription en Latin.

Nous avons conçu un système contenant les différentes phases précédemment cités. Dans chaque phase nous utilisons les traitements et les méthodes adéquates. Nous nous sommes basés sur l'idée que ces caractères sont composés de caractères Latin et de diacritiques pour diminuer le nombre de classes de reconnaissance et par conséquent augmenter le taux de reconnaissance.

Le travail réalisé nous ouvre plusieurs perspectives. Nous allons implémenter le système conçu et munir des tests permettant de comparer différentes méthodes afin d'améliorer le taux de performance du système. Nous allons aussi créer le corpus permettant d'effectuer les tests.

REFERENCES

[Belaïd 92] A. Belaïd, *Reconnaissance automatique de l'écriture et du document*, 1992, Campus scientifique, Vandoeuvre-Lès-nancy, 2001.

[Eikvil 93] Line Eikvil, "*OCR, Optical Character Recognition*", Norsk Regnesentral, 1993.

[CHAKER & al. 11] I. CHAKER & R. BENSLIMANE, Nouvelle approche pour la reconnaissance des caractères arabes imprimés, *Revue Méditerranéenne des Télécommunications*, 2011.

[Park & al. 08] S. Park W. Jung, Y. Shin, D. Jang, Optical Character Recognition System Using BP Algorithm, IJCSNS *International Journal of Computer Science and Network Security*, VOL.8 No.12, 2008.

[Charles & al. 12] P.Charles V.Harish M.Swathi CH. Deepthi, A Review on the Various Techniques used for Optical Character Recognition, *International Journal of Engineering Research and Applications* 2012.

[El Ayachi & al. 11] R. El Ayachi, M. Fakir and B. Bouikhalene, Recognition of Tifinaghe Characters Using Dynamic Programming & Neural Network, *'Recent Advances in Document Recognition and Understanding'*, 2011.

[BOUSLIMI 06] R. BOUSLIMI, Système de reconnaissance hors-ligne des mots manuscrits arabe pour multi-scripteurs, Memoire de mastère 2006.

[Muaz 10] A. Muaz, Urdu Optical Character Recognition System, Thesis, 2010.

[Noor 05] N. Noor, BANGLA OPTICAL CHARACTER RECOGNITION, Thesis, 2005.

[Abu Obaida 11] M. Abu Obaida, J. Begum , Sh. Alam, Multilingual OCR (MOCR) : An Approach to Classify Words to Languages, *International Journal of Computer Applications*, 2011.

[Tan & al. 99] Ch. Tan, P. Leong, Sh. He, Language Identification in Multilingual Documents, *International Symposium on Intelligent Multimedia and Distance Education* , 1999.

[GACI 11] Z. GACI, Quel système d’écriture pour la langue berbère (le Qabyle), Mémoir de magister, 2011.

[Belaïd 06] A. Belaïd et H. Cecotti, *La numérisation de documents : Principe et évaluation des performances*, 2006.

[DARGENTON 94] P. DARGENTON, Contribution à la Segmentation et à la Reconnaissance de l’Ecriture Manuscrite,thèse, 1994.

[ElKonyaly & al. 97] W. Abd-almageed, E. ElKonyaly and S. Saraya, Point *Feature Matching Adopting Walsh Transform*, 1997.

[Pietikäinen & al. 11] M. Pietikäinen, A. Hadid, G. Zhao, T. Ahonen, Local Binary Patterns for Still Images, *Computer Vision Using Local Binary Patterns Computational Imaging and Vision* Volume 40, 2011, pp 13-47.

[Skounti & al. 03] A. Skounti, A. Lemjidi, M. Nami, Tirra aux origines de l’écriture au Maroc, Publications de l’Institut Royal de la Culture Amazigh, 2003, Rabat.

[Roux 51] A. Roux, *CHOIX DE VERSION BERBÈRES PARLER DU SUD-OUEST MAROCAINE*, France, 1951.

[Laoust 20] E. Laoust, *MOTS ET CHOSES BERBÈRES*, Paris, 1920.

[Stroomer 08] H. Stroomer, *THE ARGAN TREE AND ITS TASHELHIYT BERBER LEXICON*, Université de Leyde, Etudes et document berbères, 2008.