Aperçu sur les techniques d'extraction des informations des manuscrits

Manal Boualam, Youssef Elfakir, Ghizlane Khaissidi, Mostafa Mrabti

Aperçu sur les techniques d'extraction des informations des manuscrits

[1*] Manal BOUALAM, boualam.manal@gmail.com , [1]Youssef ELFAKIR, [1] Ghizlane KHAISSIDI, [1] Mostafa MRABTI.

[1] LIPI/ ENS, FES, MAROC

I. Introduction :

La bibliothèque nationale de Rabat est une source riche des anciens documents manuscrits Arabes, ils sont présents sous format papier en une seule version, leur manipulation risque de les détériorer, d'où la nécessité de réaliser un système d'extraction automatique des informations pertinentes des anciens manuscrits pour faciliter leur manipulation. Plusieurs travaux ont été réalisés pendant des années pour construire ce system, cinq étapes sont nécessaires pour concrétiser cette opération: prétraitement, segmentation, extraction des informations, classification et la reconnaissance de mots.

Mots-clés: manuscrite, repérage/recherche de mots, technologies.

II. Repérage/ reconnaissance de mots contexte (En ligne et Hors-ligne) :

Il existe en générale deux phases pour la reconnaissance d'écriture :

Hors ligne : Cette phase a pour but de traiter le document source, par la suite construire une base de donnée de référence pour les différents mots existants dans le document, cette phase nécessite un traitement gourmand en matière de temps et de mémoire ;

En ligne : la phase en ligne est un traitement en temps réel, il permet la détection du mot recherché dans le document entier en se basant sur les données de sortie de la phase en ligne

Les phases principales sont :

1. Prétraitement :

Son but est d'éliminer les défauts liés à : la chaine de numérisation (inclinaison, luminosité, bruit, …), la qualité intrinsèque du document (les tâches d'humidité, apparition du verso, des trous, ...). Selon le type de défaut, des opérations sont effectué sur l'image pour améliorer sa qualité [1] [2]

2. Segmentation :

Il existe en général deux approches « based-segmentation » et « free-segmentation » [3] [4]. Le choix de l'approche Based ou free segmentation se base sur la

complexité et la qualité du document, si le document est bruité donc free segmentation est la méthode la plus efficace.

3. **Extraction des caractéristiques :**

Dans cette étape chaque caractère est représenté sous forme de vecteur de caractères, qui est considéré par la suite comme son identité. Le but principal d'extraction des caractéristiques est d'extraire un ensemble de fonctionnalité pour maximiser le taux de reconnaissance avec le minimum d'éléments.

4. **Reconnaissance et décision :**

Elle se base sur le calcul de la similarité des vecteurs de caractéristiques extraites de la requête avec les mots de la base de données du document, ce qui permet d'identifier les images de mots des documents qui sont similaires à une image de mot de requête donnée.

5. **Discussion et conclusion :**

Ce document présent un aperçu sur les techniques d'extraction des informations des documents manuscrits. Les dernières recherches se sont focalisées sur les techniques de free-segmentation pour les textes complexes, et la plus part des recherches actuelles se concentrent sur les réseaux de neurones pour la classification (deep learning).

Le tableau suivant présente un résumé de ces recherches :

Ref	classification	langue	Base de données	Type du document	Précision
[5]	HMM	francais	RIMES	handwritten	60%
[6]	MMC+ RN	arabe	-	handwritten	91,77
[7]	Réseau de neurone	-	GW	handwritten	76.72 %.
[8]	Lowe	francais	793 images	-	90.7%-100%
[9]	HMM	arabe	-	handwritten	93.1%
[10]	Réseau de neurone	-	150 pages	handwritten	91.2% - 98.2%
[11]	SVM	Farsi	Hoda	handwritten	97.90-99.80%

Tableau1. Exemples de recherches existantes

Plusieurs chercheurs ont abordé le document latin [5,8,11…], par contre, peu de travaux ont traité les documents manuscrits Arabes [6,9]. Pour cela, nous nous proposons de les numériser et créer des bibliothèques électroniques afin de pouvoir exploiter cette richesse et diffuser la connaissance tout en préservant ces documents.

Références

1. L. Likforman-Sulem, Apport du traitement des images à la numérisation des documents manuscrits anciens, 2003.
2. M. K. Dalel Ketata, Un survol sur l'analyse et la reconnaissance de documents : imprimé, ancien et manuscrit, 2010.
3. D. H. M.Arif, A Review on Feature Extraction and Feature Selection for Handwritten Character Recognition, Johor Bharu,Malaysia., 2015.
4. H. A. Sherif Abdel Azeem, Effective Technique for the Recognition of Writer Independent Off-line Handwritten Arabic Words, Egypt, 2012.
5. C. C. L. H. T. P. Simon Thomas, An Information Extraction model for unconstrained handwritten documents, Turkey, 2010.
6. A. E. M. S. Abdallah Benouareth, Reconnaissance des Mots Manuscrits Arabes par Combinaison d'une Approche Globale et une Approche Analytique, Algerie, 2014.
7. M. C. Safa ABA, Reconnaissance Des Mots Arabes Manuscrites, Tébessa, 2016.
8. J. Camillerapp, Utilisation des points d'interêt pour rechercher des mots imprimes ou manuscrits dans des documents anciens, france, 2012.
9. S. K. Fouad Slimane, Modèles de Markov Cachés et Modèle de Longueur pour la Reconnaissance de l'Ecriture Arabe à Basse Résolution, France, 2009.
10. J. L. M. G. R. G. Brijesh Verma, A Feature Extraction Technique for Online Handwriting Recognition, Australia, 2004.
11. A. B. Abdelhak Boukharouba, Novel feature extraction technique for the recognition of handwritten digits, algeria, 2015.