Identification of Arabic/French - Handwritten/Printed Words using GMM - Based System

Anis Mezghani[*] — Fouad Slimane[**] — Slim Kanoun[*] — Volker Märgner[**]

**MIRACL Lab*
ISIMS, University of Sfax, Sfax, Tunisia
{anis.mezghani, slim.kanoun}@gmail.com

*** Institute for Communications Technology (IFN)*
Braunschweig Technical University, Braunschweig, Germany
{slimane, Maergner}@ifn.ing.tu-bs.de

ABSTRACT. *The discrimination between languages is one of the first steps in the problem of automatic documents text recognition. In many documents, such as bank checks and application forms, printed and handwritten texts are mixed. In this paper, an automatic identification system of Arabic and French words in both handwritten and printed script based on Gaussian Mixture Models (GMMs) was presented. A fixed-length sliding window was used for the feature extraction. Experiments using some parts of the freely available AHTID/MW, APTI and RIMES databases show a remarkable performance of the proposed approach.*

RÉSUMÉ. *La discrimination entre les langues est l'une des premières étapes dans le problème de reconnaissance automatique des documents de textes. Dans de nombreux documents, tels que les chèques bancaires et les formulaires, les textes imprimés et manuscrits sont mélangés. Dans cet article, nous proposons un système d'identification automatique des mots arabes et français dans les deux formes: manuscrite et imprimée. Ce système est basé sur les modèles de mélanges gaussiens (GMMs). Pour l'extraction des caractéristiques, nous utilisons une fenêtre glissante de longueur fixe. Des expérimentations utilisant quelques parties des bases gratuitement disponibles AHTID/MW, APTI et RIMES montrent une performance remarquable de l'approche proposée.*

KEYWORDS: *Handwritten and printed text, Arabic and French words, GMMs, local features, language identification.*

MOTS-CLÉS : *Texte manuscrit et imprimé, mots Arabes et Français, GMMs, caractéristiques locales, identification des langues.*

1. Introduction

Script and language identification is a main step for the automation of the Optical Character Recognition (OCR) procedure. Different scripts, printed and handwritten types are often met in application forms, especially in the international administrative environments. An automatic script identification system leads to reduce the search space of the OCR, consequently facilitating to determine proper recognition and preprocessing algorithms in an early process. Few systems are interested at the same times in Arabic/Latin and Printed/Handwritten script identification ((Kanoun *et al.*, 2002), (Ben Moussa *et al.*, 2008), (Benjelil *et al.*, 2009)). The study of the presented systems shows that the identification at word level is the most complicated since it is possible to analyze pages which mix in the same line different type of words. Also, the state of the art on the script and language identification shows that no work deals with the Arabic/French identification between the multi-font printed texts and multi-scripter handwritten texts at the word level. In this context, we propose a new approach dealing with the problem of Arabic and French word identification in both handwritten and printed script based on Gaussian Mixture Models (GMMs).

The rest of the paper is organized as follows: In Section 2, we present the proposed identification approach. Experimental results are reported in Section 3. Finally, conclusion and future work are drawn in Section 4.

2. Proposed approach

The proposed writing type/language identification system is based on GMMs for the computation of likelihood estimates of different classes. The main advantage of this approach is that no a priori segmentation into characters is needed, which is an important feature for French handwritten and Arabic text where characters are tied to each other and sometimes difficult to separate.

Different features are considered for classification of a given word image into one of four classes: Arabic-Handwritten, Arabic-Printed, French-Handwritten and French-Printed. Each word image is normalized to a size of 30 pixels height and then transformed into a sequence of feature vectors computed from a fixed-length analysis window sliding from right to left over the word image. Each window is represented with a 35 features whose distribution is captured by GMMs. These features, proposed by Flusser and Suk (1993) and Heutte (1994), and described in (Mezghani *et al.*, 2014) include:

- affine moment invariants computed from moments of objects on images which are invariant under general affine transformation;
- fourteen features corresponding to the number and the X–Y position of the top and the bottom extrema;

- cumulated horizontal projection values at 10 equal parts divided by the height of the image;
- five features corresponding to the maximal amplitude obtained from the difference between the top and the bottom profiles at particular locations divided by the height of the image;

The proposed writing type and language identification system includes two main parts (training and classification) similarly to the system presented by Slimane *et al.* (2013) for Arabic font recognition. In the proposed system, four models (Arabic-Handwritten, Arabic-Printed, French-Handwritten and French-Printed) are used. To increase the number of Gaussian mixtures through the training procedure, we apply a simple binary splitting procedure after each N iterations. The GMM issuing the highest likelihood score is selected to determine the writing type/language category hypothesis. To maximize the identification performance, we have chosen to use 64 Gaussian mixtures and 10 iterations. Performances are evaluated using an unseen set of word images.

3. Tests and identification results

Our data corpus consists of four datasets: AHTID/MW (Mezghani *et al.*, 2012), RIMES (Grosicki *et al.*, 2009), APTI (Slimane *et al.*, 2009) and a French printed dataset using the generation procedure of APTI. Our experiments are carried out using ten fonts for Arabic (Andalus, Arabic Transparent, AdvertisingBold, Diwani Letter, DecoType Thuluth, DecoType Naskh, Tahoma, Traditional Arabic, Simplified Arabic and M Unicode Sara) and ten fonts for French (Arial, Monotype Corsiva, ComicSansMS, Edwardian Script ITC, Times New Roman, French Script MT, Impact, Georgia, Arial Black and Tahoma).

To evaluate the classifier performance, we used 20,000 word images: 5,000 for each class (Arabic-Handwritten/Arabic-Printed/French-Handwritten and French-printed). The learning dataset contains 80% of all word images, and the test dataset contains the rest. The average identification rate is about of 99.10%. To show the efficiency of the proposed approach, we performed the experiments using different fonts in training and test datasets. This is done by dividing the dataset into two parts: Dataset1 and Dataset2. These datasets have mutually exclusive font characters. Dataset1 has first seven French and first seven Arabic fonts and Dataset2 has remaining three French and three Arabic fonts. The average identification accuracy is about 98.92%.

The analysis of the results shows that the proposed system identifies in a reliable way French and Arabic handwritten scripts with an accuracy of 100%. On the other hand, there are some confusion between the printed Arabic and the printed French scripts. This error rate is due to the use of multi-font printed words that cover various complexities of shapes.

4. Conclusion and future work

In this work, a new approach dealing with the problem of Arabic and French word identification in both handwritten and printed script is presented. It is based on Gaussian Mixture Models for the estimation of script category likelihoods. For feature extraction, a fixed-length sliding window from right to left is used with no need of a priori segmentation into characters. For experimental evaluation, the developed system was tested using some parts of three public image databases of Arabic and French content. The results are fairly encouraging. Based on our reported results, we plan to implement a cascading system working in two steps: script identification and word recognition.

5. References

Kanoun, S., Ennaji, A., Alimi, A., Lecourtier, Y., « Script and Nature Differentiation For and Latin Text Images », *International Workshop on Frontiers in Handwriting Recognition*, 2002, p. 309-313.

Ben Moussa S., Zahour A., Benabdelhafid A., Alimi A.M., « Fractal-Based System for Arabic/Latin, Printed/Handwritten Script Identification », *International Conference on Pattern Recognition*, 2008, p. 1-4.

Benjelil M., Kanoun S., Mullot R., Alimi A.M., « Arabic and Latin script identification in printed and handwritten types Based on Steerable Pyramid Features », *International Conference on Document Analysis and Recognition*, 2009, p. 591-595.

Flusser J., Suk T., « Pattern recognition by affine moment invariants », *Pattern Recognition* vol.26, 1993, p. 167-174.

Heutte L., 1994, « Reconnaissance de caractères manuscrits: Application à la lecture automatique des chèques et des enveloppes postales », PhD thesis, University of Rouen.

Mezghani A., Slimane F., Kanoun S., Märgner V., « Printed/Handwritten Arabic Script Identification using Local Features and GMMs », *International Conference on Information and Communication Technologies Innovations and Applications*, 2014.

Slimane F., Kanoun S., Hennebert J., Alimi M.A., Ingold R., « A study on font-family and font-size recognition applied to Arabic word images at ultra-low resolution », *Pattern Recognition Letters*, vol.34(2), 2013, p. 209-218.

Mezghani A., Kanoun S., Khemakhem M., El Abed H., « A Database for Arabic Handwritten Text Image Recognition and Writer Identification », *International Conference on Frontiers in Handwriting Recognition*, 2012, p. 399-402.

Grosicki E., Carré M., Brodin J.M., Geoffrois E., « Results of the RIMES Evaluation Campaign for Handwritten Mail Processing », *International Conference on Document Analysis and Recognition*, 2009, p. 941-945.

Slimane F., Ingold R., Kanoun S., Alimi A. M., Hennebert J., « A New Arabic Printed Text Image Database and Evaluation Protocols », *International Conference on Document Analysis and Recognition*, 2009, p. 946-950.