2009 10th International Conference on Document Analysis and Recognition

A Multi-Lingual Recognition System for Arabic and Latin Handwriting

Yousri KESSENTINI[1, 2], Thierry PAQUET[1], AbdelMajid BEN HAMADOU[2]
[1] *Laboratoire LITIS EA 4108, university of Rouen France*
[2] *Laboratoire MIRACL, university of Sfax Tunisia*
{yousri.kessentini, thierry.paquet}@univ-rouen.fr
abdelmajid.benhamadou@isimsf.rnu.tn

Abstract

Generally, handwritten word recognition systems use script specific methodologies. In this paper, we present a unified approach for multi-lingual recognition of alphabetic scripts. The proposed system operates independently of the nature of the script using the multi-stream paradigm. The experiments have been carried out on a multi-script database composed of Arabic and Latin handwritten words from the IFN/ENIT and the IRONOFF public databases and show interesting recognition performances with only 1.5% of script confusion and an overall word recognition rate of 84.5% using a multi-script lexicon of 1142 words.

1. Introduction

In today's business world, Latin script is mostly used, but with the increasing communication between the different world communities more scripts are getting integrated into information systems. In recent years various handwriting recognition systems have been developed for the recognition of different scripts (Latin, Arabic, Indian, Chinese …). In the case of multi-lingual documents, where handwritten scripts are present in several different languages on the same document, the combination of multiple systems has been envisaged. Generally, document contents are first classified according to the type of script before a specific script dependant system is applied to recognize textual information [1, 2]. Following this framework, multi-script recognition systems cumulate the complexity of each individual recognition system.

Some other approaches propose a unified recognition framework able to deal with multi-script handwriting. Such systems do not use any script specific methodology and therefore appear to be more general and applicable to a wider range of multi-lingual applications. However, few works have investigated this strategy until now. In [3], the authors propose a handwriting recognition system of various scripts such as Latin, Devanagari, and Kanji. In [4] the authors propose an OCR system able to read two Indian language scripts: Bangla and Devanagari. In the same spirit in [6] the authors investigate the recognition of keywords in oriental printed scripts. In [5] the authors propose a unified approach to deal with various printed scripts.

To the best of our knowledge, no multi-script recognition system has been proposed for Arabic and Latin handwritten scripts. In this context, we propose a unified approach for multi-script handwritten word recognition. As we want the approach to be script independent, the system proceeds without explicit segmentation of handwriting into graphemes. This is because explicit segmentation methods generally rely on script specific rules to find segmentation points. According to the proposed strategy it is therefore mandatory that the system can operate on low level frame features such as directional, contour or pixel densities. In order to achieve good discriminative power of such low level features, the proposed approach is based on the multi-stream paradigm [7, 8] that provides a promising way of combining individual feature streams. In [7], we have tested our handwriting recognition system on two different scripts and significant results have been reported comparable to the best results reported in the literature. In this paper, we extend this previous work by proposing a mutli-lingual recognition system. The input images are fed to the recognizer that provides an ordered list of words belonging to the multi-lingual lexicon. The system shows interesting performances for both the recognition of handwritten words us well us script identification.

The paper is organized as follows: section 2 describes some specificities of Arabic script. Section 3 describes the multi-stream modeling of the multi-script handwriting used for this study. In section 4, the multi-

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.55

script recognition scheme is presented and the different stages are described. Section 5 shows the experimental results of the proposed approach and discusses the results. Conclusion and the future works follow in section 6.

2. Arabic script

The Arabic script is written from right to left, and includes 28 basic letters. Words are segmented into PAWs, "parts of Arabic words", each consisting of a group of letters. A word is composed of one or more PAWs (Fig. 1). There is no difference between upper and lower case characters. The character shape is context sensitive, i.e. it depends on its position within a word. For instance, letter غ has 4 different shapes: isolated 'غ' as in 'بلوغ', beginning as in 'غروب', middle as in 'نغم' and end as in 'نبغ'. Some combinations of two or three letters have special shapes called "ligatures", they work exactly as Latin ligatures such as **œ** and **æ**.

Figure 1: One Arabic script example with 2 PAWs

3. Multi-script modeling

The proposed system is based on a multi-stream framework providing a convenient formalism to combine several information sources, namely feature streams, using cooperative hidden Markov models (HMM) (see Figure 2).

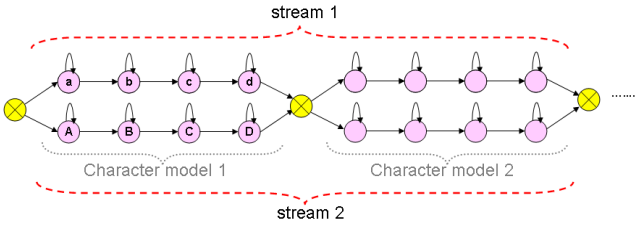


Figure 2 : Example of a multi-stream HMM with 2 streams and 4 states in each character model.

An HMM based framework offers several advantages allowing us to propose a unified multi-script recognition system. These advantages are mainly due to automatic training of character models on non-segmented words (embedded training), and the segmentation-free recognition paradigm that fits particularly well to a multi-script approach.

In order to model the Latin characters, we built 26 uppercase character models and 26 lowercase character models. In the case of Arabic characters, we built up to 159 character models. An Arabic character may actually have different shapes according to its position within the word. Other models are specified with additional marks such as "shadda".

In both Latin and Arabic script, each character is modeled by a 2-stream HMM. These are later dynamically composed to word-HMMs.

Each character-HMM is composed of 4 emitting states. The observations probabilities are modeled with Gaussian Mixtures (3 per state).

The training step on the multi-script recognition system is achieved independently for each script. Embedded training is used where all character models are trained in parallel using Baum-Welch algorithm applied on word examples.

The multi-script recognition scheme is presented as following.

4. The recognition scheme for multiple scripts

Apart from character shapes difference, Latin and Arabic script also differ from their writing direction. Our aim is to propose a unified multi-script recognition system working similarly on Arabic and Latin words. For this reason, features vectors are extracted according to a single direction for both scripts (left to right) and a lexical string inversion is applied to consider the inverted writing direction of the Arabic script. Doing this way, our system will process Arabic word in the opposite of their natural reading direction.

We describe in Figure 3 the different stages of our multi-script recognition system. In the first step, pre-processing is applied to the word image. The next step is the feature extraction where two types of features are considered: (i) contour based features and (ii) density based features. The last step is the recognition during which the HMM models are simultaneously decoded according to the multi-stream formalism. The system output is a ranked list of the words producing the best likelihood on the input image. These steps are detailed in the subsequent sections.

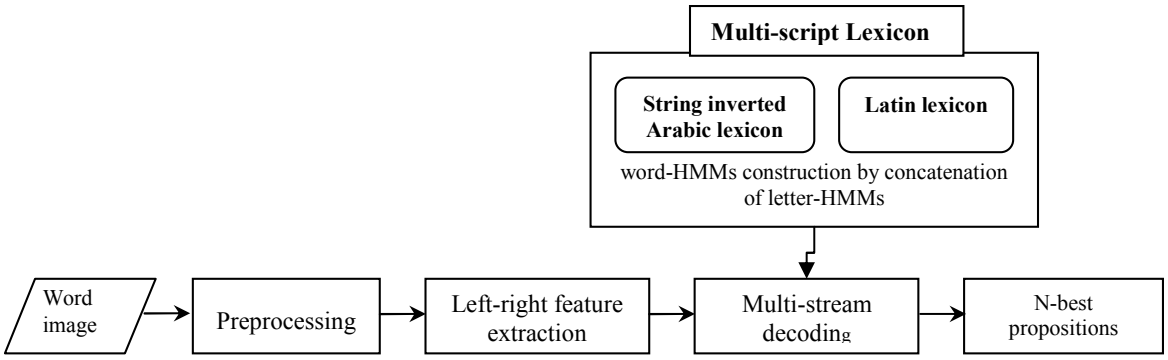


Figure 3: The multi-script recognition system description

4.1. Preprocessing

Preprocessings are applied to word images in order to eliminate noise and to simplify the procedure of feature extraction. It is worth noticing that these preprocessings are script independent.

- Normalization: In an ideal model of handwriting, a word is supposed to be written horizontally and with ascenders and descenders aligned along the vertical direction. In real data, such conditions are rarely respected. We use slant and slope correction so as to normalize the word image.
- Contour smoothing: Smoothing eliminates small blobs on the contour.
- Base line detection: Our approach uses the algorithm described in [9] based on the horizontal projection curve that is computed with respect to the horizontal pixel density (see Figure 4). Baseline position is used to extract baseline dependent features that emphasize the presence of descenders and ascenders.

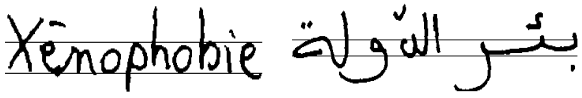


Figure 4 : Baseline detection

4.2. Feature extraction

In order to build the feature vector sequence, the image is divided into vertical overlapping windows or frames. The sliding window is shifted along the word image left to right and a feature vector is computed for each frame position. Two feature sets are used in this work. The first one is based on directional density features and is described in [7]. This kind of features has proved to be discriminative for off-Line handwriting recognition, especially for Latin script. The second one is based on foreground (black) pixel densities and is detailed in [10]. So that, each word image is represented by two script independent feature streams.

4.3. Multi-stream decoding

The multi-stream paradigm is to model independently each stream between two pre-determined synchronization points, using multiple single-stream HMMs (two in this study). The synchronization states are the character boundaries (see Figure 2). Decoding based on this integration method would require the computation of the best state sequence for both streams verifying the synchronous recombination rule at the same time. To avoid the computation of two best state paths, the model can be formulated as a composite or product HMM (see Figure 5) where each state is built by merging a K-tuple of states from the K stream HMMs (here, $K=2$).

The topology of this composite model is defined so as to represent all possible state paths given the initial HMM topologies. Decoding under such a model requires computing a single best path using the well known Viterbi decoding algorithm.

In this model, the observation log-likelihood conditioned on a composite state a-A, that is composed of the stream states a and A respectively, is computed using a weighted sum of log-likelihood by:

$$\log P(X_1(t), X_2(t) \mid a\text{-}A) = \omega \log P_1 + (1-\omega) \log P_2$$

$$\text{where } P_1 = P(X_1(t) \mid a) \text{ et } P_2 = P(X_2(t) \mid A)$$

$X_1(t)$ *(similarly* $X_2(t)$*)* is the observation vector corresponding to stream 1 (similarly stream 2) and ω the reliability of stream 1 ($0 \le \omega \le 1$).

The transition probabilities of the product HMM are derived from the transition probabilities of the 2 single stream HMMs assuming independence of the models between 2 recombination states. For example,

$$P(a\text{-}B \mid a\text{-}A) = P_1(a \mid a) \times P_2(B \mid A)$$

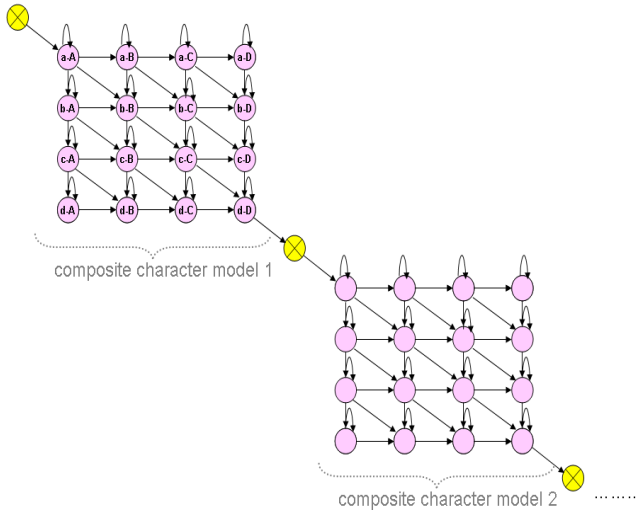


Figure 5 : The product (composite) HMM.

5. Experiments and results

To evaluate the performance of our recognition system, experiments have been conducted on a multi-script database constructed from the IFN/ENIT database [11] of Arabic words and the IRONOFF database [12] of Latin words. The training steps are carried out separately for each script. The Arabic models are trained on the 4 sets (a,b,c,d) of the IFN/ENIT database. A training dataset of 24,177 word images is used in the case of Latin script. The mutli-script recognition system is tested on a multi-script database composed of 1000 Arabic words (from set e of the IFN/ENIT database) and 1000 Latin words (from the test set of IRONOFF database). The multi-script lexicon size is 1142 words (946 Arabic words and 196 Latin words).

Table 1 shows experimental results of our multi-script recognition system on 3 tests: 1) using the multi-script test database; 2) using only the Arabic words of the multi-script test database; 3) using only the Latin words of the multi-script test database.

The obtained results are better on the Latin script than the Arabic script; this can be explained by the fact that Latin words belong to a smaller lexicon.

Table 1: The multi-script recognition system performances %

Tests	TOP			
	1	2	5	10
Test 1	84.5	90.5	94.6	96.3
Test 2	80.2	88	93.4	95.2
Test 3	88.8	93	95.8	97.4

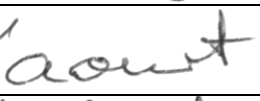
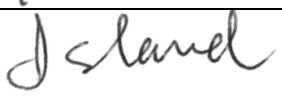
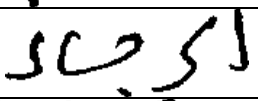
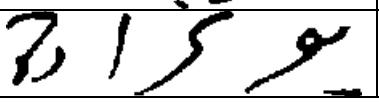
To analyze these results, we present in table 2 the script confusion rate of the multi-script recognizer. We define the Latin confusion rate (similarly Arabic confusion rate) by the percentage of Latin words that have been identified as Arabic words (similarly the percentage of Arabic words that have been identified as Latin words). The multi-script confusion rate is the sum of the Arabic and the Latin confusion rates.

Table 2: The script confusion rate % of the multi-script recognition system

Confusion rates	TOP			
	1	2	5	10
Multi-script	1.5	3.7	7.4	11.8
Arabic	0.3	0.3	0.6	0.8
Latin	1.2	3.4	6.8	11

The results of table 2 show that the multi-script recognition system identifies correctly the script of most of the test images. The confusion rate is only 1.5% on Top 1. Some confusion cases are reported in table 3. Images 1 and 3 are confused because the words “treize” and “Island” are written with tow PAWs. The confusion can be explained by the presence of loops and descenders as in images 1, 2 and 3. The confusion in images 4, 5 and 6 is not very understandable; possibly, this can be explained by the bad quality of the writing.

Table 3: Example of samples that are confused by the multi-script recognizer

N°	Word test image	Recognized word
1		بزمة
2		جلمة
3		شط السلام
4		tout
5		sous
6		six

6. Conclusion

We have proposed in this paper a multi-lingual recognition system. It proceeds without explicit segmentation of handwriting into graphemes which generally rely on script specific rules to find segmentation points. Preprocessing and feature sets are script independent. In order to process similarly on Latin and Arabic images, the Arabic lexicon has been inverted so as to have a single reading direction during the recognition phase. The experiments have been carried out on a multi-script database composed of Arabic and Latin handwritten words and show an interesting recognition performances with only 1.5% of script confusion. This multi-script system can easily integrate any other alphabetic script and this can be considered for our future works.

7. References

[1] G. Nagy, S. Seth, and X. Zhang, Multi-Character Field Recognition for Arabic and Chinese Handwriting, *Proceedings of the Summit on Arabic and Chinese Handwriting Recognition*, 2006, College Park, MD, pp. 93-100.

[2] J.J. Lee, M. Nakajima and J. Kim, A Hierarchical HMM Network-based Approach for On-line Recognition of Multi-Lingual Cursive Handwritings, *IEICE Transactions on Information and Systems*, 1998, vol. E81-D, No. 8, pp. 881-888.

[3] A. Malaviya, C. Leja, L. Peters, Multi-script handwriting recognition with FOHDEL. *New Frontiers in Fuzzy Logic and Soft Computing Biennial Conference of the North American Fuzzy Information Processing Society- NAFIPS*. IEEE, Piscataway, NJ, USA, 1996, pp. 147-151

[4] B.B. Chaudhuri, U. Pal, An OCR system to read two Indian language scripts: Bangla and Devnagari (Hindi), *ICDAR '97,* vol 2, 1997 , pp. 1011 -1015.

[5] J. Makhoul, R. Schwartz, C. Lapre, I. Bazzi, A Script-independant Methodology for Optical Character Recognition, *Pattern Recognition*, vol. 31, No. 9, pp. 1285-1294, 1998.

[6] Jason Zhu, Tao Hong, Jonathan J. Hull, Image-based Keyword Recognition in Oriental Language Document Images, *Pattern Recognition*, vol. 30, No. 8, pp. 1293-1300, 1997.

[7] Y. Kessentini, T. Paquet, A. Benhamadou, A Multi-Stream HMM-based Approach for Off-line Multi-Script Handwritten Word Recognition. *ICFHR'08*, 2008.

[8] Y. Kessentini, T. Paquet, A. Benhamadou. A Multi-stream Approach to Off-Line Handwritten Word Recognition, *ICDAR'07*, vol. 1, pp. 317-321, 2007.

[9] A. Vinciarelli, S.Bengio, H.Bunke, Offline Recognition of Unconstrained Handwritten Texts Using HMMs and Statistical Language Models, *IEEE Transactions on PAMI*, vol. 26, No 6, pp. 709-720, 2004.

[10] R. El-Hajj, L. Likforman-Sulem, C. Mokbel, Arabic handwriting recognition using baseline dependent features and Hidden Markov Modeling, *ICDAR 2005*, Seoul, South korea, vol 2, pp. 893- 897.

[11] M. Pechwitz, S. Maddouri, V. Maegner, N. Ellouze, IFN/ENIT–DataBase for Handwritten Arabic words, *CIFED'02*, pp. 129-136, 2002.