Journal of Data Mining and Digital Humanities

CursiveArabic Handwriting Recognition System Without Explicit Segmentation Based on Hidden Markov Models

Mouhcine RABI[1*], Mustapha AMROUCH[1], Zouhair MAHANI[2]

1laboratory IRF-SIC, faculty of sciences IbnZohr University, Agadir, Morocco
2 High School of Technology, IbnZohr University, Agadir, Morocco

*mouhcineh@gmail.com
m.amrouch@uiz.ac.ma
zouhir.mahani@gmail.com

Abstract
In this paper we present a system for offline recognition cursive Arabic handwritten text which is analytical without explicit segmentation based onHidden Markov Models (HMMs). Extraction features preceded by baseline estimation are statistical and geometric to integrate both the peculiarities of the text and the pixel distribution characteristics in the word image. These features are modelled using hidden Markov models. The HMM-based classifiercontains a training module and a recognition module. The training module estimates theparameters of each of the character HMMs uses the Baum-Welchalgorithm. In the recognition phase, feature vectors extracted from an image are passed to a network of word lexicon entries formed of character models. The character sequence providing the maximumlikelihood identifies the recognized entry. If required, the recognition can generate N best output hypotheses rather than just the single best one. To determine the best outputhypotheses, the Viterbi algorithm is used.The experiments on images of the benchmark IFN/ENIT database show that the proposed system improves recognition.

Keywords
Recognition; handwriting; Arabic text; HMMs.

I INTRODUCTION

The recognition of cursive Arabic handwriting is an active area of pattern recognition research. The variability of words, letter shapes are context sensitive, inter and intra word spaces, thecursive nature of Arabic handwriting, the skew and slant ofcharacters and words makes the construction of offlinerecognition system a challenging task.
Researches have tried various approaches for text recognition employing various techniques for pre-processing, featuresextraction and classification [A.Lawgali 2015].

The subject of this article concerns the recognition of cursiveArabic handwriting [M.T Parvez 2013] [AL-Shatnawi 2011].Several systems are available based on two approaches; a globalapproach that considers the word as non-divisible base entityavoiding the segmentation process and its problems.This approach is reliable and applicable for vocabularies oflimited size. Against, the analytical approach is based on thedecomposition of the word sequence into characters orgraphemes proceeding by a segmentation phase. The latter canbe explicitly based on a priori division of the image into subunits(letters or grapheme) or implicitly based on a recognitionengine to validate and rank the segmentation hypothesis.The approach used in our system is analytical based on implicitsegmentation; segmentation and recognition are carried outjointly.

The first step of a handwriting recognition system afterpreprocessing is the extraction features. The objective of thisphase is the selection of primitives relevant for the next steps ofclassification and recognition. The performance of a recognitionhandwritten system largely depends on the quality and therelevance of the extracted features. In our system after thebaselines estimation, the extracted features are statistics actingon the densities of pixels and structural extracted from therepresentation of the character shapes.
Hidden Markov models (HMMs) are used for classification [S.Azeem 2013] [AlKhateeb 2011][A. Maqqor 2014]. There aremany reasons for success of HMMs in text recognition includingavoidance of the need to explicitly segmentation. In addition,HMMs have sound mathematical and theorical foundations.Each word is described by a model built by concatenating themodels of the component character. The system performstraining and recognition of words and characters.

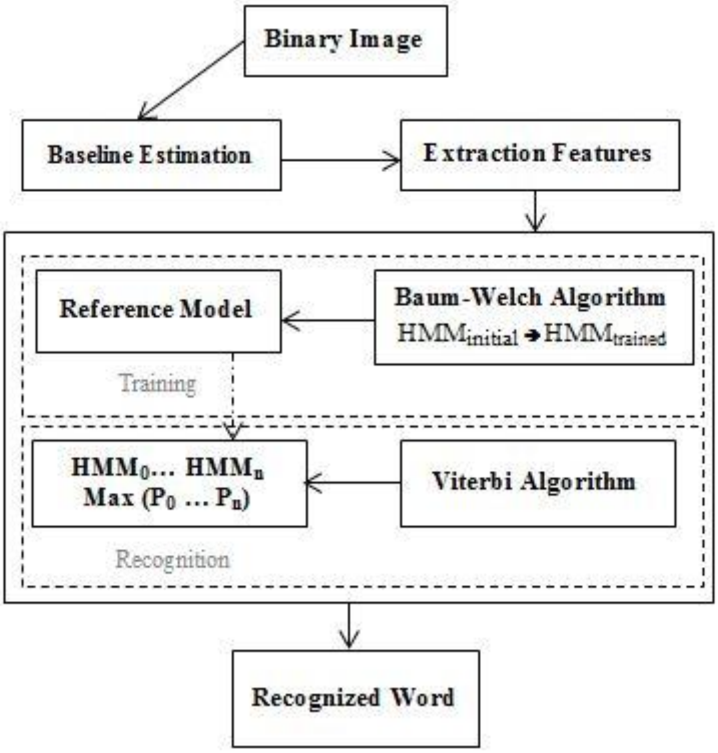

Figure 1 presents the synopsis of the proposed system.

The remainder of this paper is organized as follow. Section 2presents a detailed description of the features extraction precededby baselines estimation. Section 3 is focused on classificationstep. The performance of therecognition system has been experimented on the benchmarkdatabase IFN/ENIT and the obtained experimental results areshown and analysed in section 4. The paper finally concludeswith some conclusions and perspectives.

IIEXTRACTION FEATURES

1.1 Baseline Estimation

The goal is to find, for a given word, the positions of the two following parallel lines (Figure 2):

- Lower baseline (LB),
- Upper baseline (UB).

These baselines divide the image into lower, upper and middle zone.

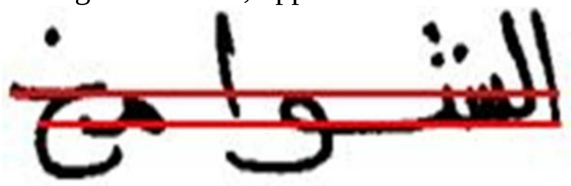

Figure 2. Lower and upper baselines estimation

Several techniques exist for estimating these two lines,horizontal projection histogram, the Hough transform, theminima of the bottom contours [Farooq et al, 2005], theneighborhood approach and components, PCA (PrincipalComponent Analysis) [B. Su, X. Ding 2013] [O. Morillot 2013].

The approach used is based on the horizontal projection curvethat is computed with respect to the horizontal pixel density,knowing that the skew and slant correction of words are made inpreprocessing step to harmonize the course of the slidingwindows in the extraction features.

LB corresponds to the maximum of the projection profile curve,then, the algorithm scans the image from top to bottom to findthe upper baseline, which corresponds to the first line with aprojection value higher or equal to the average row density.Thus, the handwritten variability of word in both zones, upperand lower are considered.

1.2Extraction Features

The features extraction method used in our system is inspired by work of El-Hajj [El-Hajj 2009] with some modifications; theused technique has shown excellent results in several researches[A-L. B-Bernard 2012][T.Bluche 2015]. The features extraction stage consists of extracting a sequence of characteristics vectorby dividing the word image into vertical frames. The sliding windows are shifted in the direction of writing (right to left). The width of each window is a parameter to set, the height of awindow varies according to the dimension of the word image.

In each window we extract a set of 28 features represent thedistribution features based on foreground pixels densities andconcavity features. Each window is divided into a fixed numbern of cells. Some of these features are extracted from specificareas of the image delimited by the word baselines.

In our experimentation the parameters are set to n = 20 cells andthe width = 8 pixels. This leads to a total of Nf= 28 to calculatein each frame.

Let:

- n(i) : the number of foreground pixels in cell i
- r(j) : the number of foreground pixels in the jth row of a frame.
- b(i) : the density level of cell i : b(i)=0 if n(i)=0 else b(i)=1.

The extracted features are the following:

f1: density of foreground (black) pixels.

$$f1 = \frac{1}{H \times w} \sum_{i=1}^{n_c} n(i)$$

f2: number of transitions black/white between two consecutive cells.

$$f2 = \sum_{i=2}^{n_c} \left| b(i) - b(i-1) \right|$$

ISSN 2416-5999, an open-access journal

f3: difference's position of gravity centers of foreground pixels in two consecutive frame(current and previous)

$$f3 = g(t) - g(t-1)$$

f4: normalized vertical position of the center of gravity of the foreground pixels in the whole frame with respect to the lower baseline.

$$f4 = \frac{g - L}{H}$$

f5, f6: represent the density of foreground pixels over and under the lower baselines.

$$f5 = \frac{\sum_{j=L+1}^{H} r(j)}{H.w} \quad ; \qquad f6 = \frac{\sum_{j=1}^{L-1} r(j)}{H.w}$$

f7: number of transitions black/white between two consecutive cells of different density levels above the lower baseline.

$$f7 = \sum_{i=k}^{n_c} \left| b(i) - b(i-1) \right|$$

f8: zone to which the gravity center of blck pixels belongs (lower zone f8=3, middle zone f8=2, upper zone f8=1)

f9,..., f14: the concavity features are defined as:

$$f9 = \frac{C_{left-up}}{H} \quad ;$$

$C_{left-up}$: the number of background pixels that have neighbor black pixels in the two directions (left and up)
The same applies to f9, …, f14 in six directions left-up, up-right, right-down, down-left, vertical and horizontal.

f15, …, f20 : the baseline dependent features related to the core zone are defined as :

$$f15 = \frac{CM_{left-up}}{d} \, ;$$

$CM_{left-up}$: the number of background pixels in the configuration left-up
The same applies to f16, …, f20 in six directions left-up, up-right, right-down, down-left, vertical and horizontal.

f21,…, f28 : represent the density of foreground pixels in each vertical column in a frame.

In each frame 28 features vector are extracted, these features are statistical and geometric to integrate both the peculiarities of the text and the pixel distribution characteristics in the word image, which capture the type of strokes (curved, oriented, vertical, and horizontal).

IIIMODELING

2.1Hidden Markov Models

A Hidden Markov Model (HMM) is a doubly stochastic processwith an under-lined stochastic process (Markov chain) that is notobservable (it is hidden), but can only be observed throughanother set of stochastic processes that produce the sequence ofobserved symbols [Rabiner, 1989].In order to define an HMM completely, following elements areneeded, $\lambda = (N, M, A, B, \Pi)$:

- N: The number of states of the model,
- M: The number of observation symbols in the alphabet. If the observations are continuous then M is infinite,
- A : A set of state transition probabilities,
- B: A probability distribution in each of the states,
- Π: The initial state distribution.

The use of HMM aims to resolve three problems:
❖ The Evaluation
Given an HMM λ and a sequence of observations O=o1,o2,…,oT,what is the probability that the observations are generated by themodel, P{O|λ}
❖ The Decoding
Given a model λ and a sequence of observations O=o1,o2,…,oT,what is the most likely state sequence in the model that producedthe observations?
❖ The Training
Given a model λ and a sequence of observations O=o1,o2,…,oT,how should we adjust the model parameters {Π, A, B} in orderto maximize P{O|λ}

2.2Character and word models

The used approach is analytical and based on character modelingby HMM. Each character model has a right-left topology andparameters: number of hidden states, state transition probabilitiesand observation probabilities.
There is no specific theory to set these parameters, so thesolution is empirical. Many considerations can be taken intoaccount in setting these parameters and in particular thetechnique used in the generation of sequences of observations. Inour system we used a model with four states for each characterwith three transitions for each state (Figure 3).

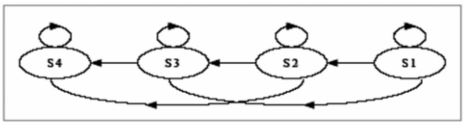

Figure 3. Character HMM topology

Word model is built by concatenating the appropriate character models (figure 4).

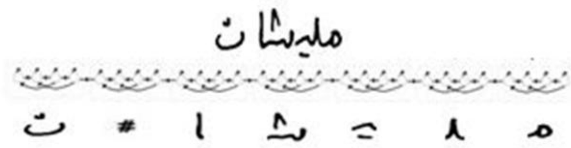

Figure 4. HMM model for Arabic word :مليشات

IV. EXPERIMENTATIONS AND RESULTS

In order to investigate the potential of our systemfor offline cursive handwriting recognition, the benchmarkdatabase IFN/ENIT is used [Peschwitz 2003], that contains atotal of 26459 handwritten words of 946 Tunisian town/villagesnames written by different writers.We used the toolbox HTK (Hidden Markov Model Toolkit[S.Young 2006]) to model the characters and words.

The table below shows the experimental results of our systemcompared to other recognition systems using the samebenchmarking database IFN/ENIT, divided into four sets a, b, cfor training and d for testing :

Table 1: Recognition results of various systems

System	***Models***	***Recognition rate %***
[Kessentini et al 2010]	Multi Stream HMM	79.80
[Alkhateeb et al. 2011]	HMM and dynamic Bayesian network	86.73
[Parvez et al. 2013]	FATF with set medians	79.58
[A.Maqqor 2014]	Multiple Classifiers	76.54
Proposed System	HMM	**87.93**

Table1 shows the results of recognition rates for various offlinesystems recognition of cursive Arabic handwritten text usingvarious models and the same database to compare rates and inferthe effectiveness of the proposed method.

[Alkhateeb et al. 2011] are presented a comparative study ofapproaches for recognizing handwritten Arabic words usingHidden Markov Models (HMM) and Dynamic Bayesian Network (DBN) classifiers, and The recognition rate achievedwas 86,73%. [Kessentini et al 2010] [Parvez et al. 2013][A.Maqqor 2014] are presented systems using respectively amulti-stream Hidden Markov Models, Fuzzy Attributed TurningFunction (FATF) with set-medians; the recognition rate for theresults of the systems mentioned does not exceed 80%. Whereasthe proposed system outperforms the results and achieve87.93%.
Finally, the performance of handwritten Arabic recognition systemis significantly improved using our system based on HMMs.It remains to boost the rate usingannexes improvements (Post-Processing: language models)

V. CONCLUSION & PERSPECTIVES

In this paper, we present a recognition system of Arabic cursivehandwriting based on hidden Markovmodels. The extracted features are based on the densities offoreground pixels, concavity and derivative features usingsliding window, some of these features depends on baselinesestimation. The modelling proposed has improved recognition,and shown encouraging results to be perfect later.Many points are yet to be achieved, firstly modeling a characterallows deformations related to its context (next and previouscharacter). To account possible deformations, contextualmodeling of characters is opted. The word is no longer seen as asuccession of independent characters, but as a sequence ofcharacters in context. Word models are the concatenation ofcontext-dependent characters models: the trigraphe, thismodelling will allow building more accurate and more efficientmodels. Taking into account the characters environment allowsmore precise and more effective models to be built. However,this implies a multiplication of HMM parameters to be learned, itwould be the focus of our next work. Then language models willbe incorporated to refine and improve the results and lead to amore efficient system.

References

A.Lawgali : “A Survey on Arabic Character Recognition” International Journal of Signal Processing, Image Processing and Pattern

Recognition. Vol. 8, No. 2 (2015), pp. 401-426.

ThéodoreBluche, Hermann Ney, Christopher Kermorvant (2015) The LIMSI Handwriting Recognition System for the HTRtS 2014 Contest. In International Conference of Document Analysis and Recognition (ICDAR).

Ahmad, I., Fink, G.A., Mahmoud, S.A.: Improvements in sub-character HMM model based Arabic text recognition. In: ICFHR (2014)

M.T Parvez and S.A Mahmoud “ Offline Arabic handwritten text recognition : A survey “ ACM ComputSurvvol 45 no 2 pp 23-35 May 2013

S. Azeem and H. Ahmed, “Effective technique for the recognition of offline Arabic handwritten words using hidden

Markov models,” Int. J. Doc. Anal.Recognit., vol. 16, no. 4, pp.399– 412, 2013.

B. Su, X. Ding, L. Peng, and C. Liu, “A Novel Baseline independent Feature Set for Arabic Handwriting Recognition,” in Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, 2013, pp. 1250–1254.

C. Choisy et A. Belaïd « Apprentissage croisé en reconnaissance analytique de l’écriture manuscrite » colloque international francophone sur l’écrit et le Document 2000.

S. Young, al., the HTK Book V3.4 Cambridge University Press, Cambridge UK, 2006

L.R. Rabiner : “A tutorial on Hidden Markov Models and selected aplications in speech recognition”. Proceedings of the IEEE, 77(2) :257- 286, 1989.

O. Morillot “New baseline correction algorithm for text-line recognition with bidirectional recurrent neural networks” Journal of

Electronic Imaging 22(2), 023028 (Apr–Jun 2013)

A. Maqqor, A. Halli, K. Satori, and H. Tairi, Off-line recognition Handwriting combination of mutiple classifiers, In 3rd International IEEE Colloquium on Information Science and Technology, IEEE CIST’14, October 2014.

M. T. Parvez, A. M. Sabri, Arabic handwriting recognition using structural and syntactic pattern attributes, in Pattern Recognition 46 (2013), pp. 141-154.

Anne-laureBianne-bernard, Fares Menasri, Laurence Likforman-sulem, ChaficMokbel, Christopher Kermorvant (2012) Variable length and context-dependent HMM letter form models for Arabic handwritten word recognition. In Document Recognition and Retrieval Conference (DRR)

AL-Shatnawi, M. Atallah, AL-Salaimeh, Safwan, AL-Zawaideh, Farah Hanna, Omar, Khairuddin, Offline arabic text recognition an overview, World Comput. Sci. Inform. Technol. J. 1 (5), 184–192, 2011.

J. AlKhateeb, H., Ren, Jinchang, Jiang, Jianmin, Al-Muhtaseb, Husni, Offline handwritten arabic cursive text recognition using hidden markov models and re-ranking, Pattern Recognition Lett. 32 (8), 1081–1088, 2011.

J. H. Alkhateeb, O. Pauplin, J. Ren, J. Jiang, Performance of hidden Markov model and dynamic Bayesian network classifiers on handwritten Arabic word recognition, In Knowledge-Based Systems 24 (2011), pp.680-688.

Y. Kessentini, T. Paquet, A. Ben Hamadou, Off-line handwritten word recognition using multistream hidden Markov models, in: Pattern Recognition Letters 31 (2010), pp. 60-70

L.R. Rabiner : “A tutorial on Hidden Markov Models and selected applications in speech recognition”. Proceedings of the IEEE, 77(2) :257- 286, 1989.

Pechwitz M. and Maergner V., "Baseline estimation for Arabic handwritten words", in Proc. of the 8th IWFHR 2002, pp 479-484, Canada, August (2002).

Farooq F., Govindaraju V., Perrone M., "Pre-processing methods for handwritten Arabic Documents", ICDAR’05.pp: 267-271, Seoul (2005).

EL-Hajj R R. Likforman-Sulem L. Mokbel C., “ Combining Slanted-Frames Classifiers for Improved HMM-Based Arabic Handwriting Recognition”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009.

ISSN 2416-5999, an open-access journal