Distribution, Directional, structural and concavity features for historical Arabic handwritten recognition: a comparative study

Meriem GAGAOUA, Hamza Ghilas, Abdelkamel TARI
Département d'informatique
Université de Bejaia
Bejaia, Algérie
{gagaouameriem; hamzaghilas}@yahoo.fr
tarikamel59@gmail.com

Mohamed CHERIET
Synchromedia Laboratory for Multimedia Communication in Telepresence Ecole de Technologie Supérieure, Montréal (QC), H3C 1K3
Mohamed.Cheriet@etsmtl.ca

Abstract[1]

In the process of automatic handwritten recognition especially in Arabic historical documents, the feature extraction is an important step, which find a set of measured values that accurately discriminate the input handwritten words or characters. In this paper, we try to determine how features designed for Arabic handwritten recognition can be efficient in Arabic historical documents by conducting a comparative study of four types of features (distribution, directional, structural and concavity features). The recognition process is based on Hidden Markov models with HTK toolkit and sliding window features. Words HMMs are learned using embedded training based on character HMM models. Experiments are performed on the benchmark Iben Sina database of Arabic historical documents.

CCS CONCEPTS

• **Applied computing → Document management and text processing• Computing methodologies → Artificial intelligence;** machine learning; Computer vision.

KEYWORDS

Arabic handwritten recognition; historical documents; feature extraction; HMM; HTK toolkit.

ACM Reference format:

M.GAGAOUA, H.GHILAS, M.CHERIET, A.TARI 2017. SIG Proceedings Paper in Word Format. In Proceedings of ACM ICCES conference, Istanbul, Turkey, July 2017 (ICCES '17), 6 pages. https://doi.org/10.1145/3129186.3129200

1

ICCES '17, July 22–24, 2017, Istanbul, Turkey

ACM ISBN 978-1-4503-5309-0/17/07…$15.00
https://doi.org/10.1145/3129186.3129200

1. INTRODUCTION

Ancient Arabic manuscript documents are very important in particular for historians, sociologists, researchers and students. These documents contain a very precious knowledge, which is available for consultation through digital libraries. The information retrieval in these libraries is very difficult because of the difficulty of the manual search in an enormous set of digital pages (images). Hence the need to make use of optical character recognition systems (OCR in historical documents) to facilitate this task.

The automatic recognition of Arabic handwritten in historical documents is a challenging problem and difficult task, this difficulty is related to the deteriorated quality of the documents, the large handwriting variability (the cursive nature of Arabic Writing (ascenders, descenders, loops…) (Figure.1) and a word can be written with different manners) (Figure.2)).

Figure 1. The cursive nature of Arabic Writing

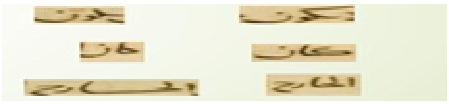

Figure 2. The Arabic handwriting variability

In the process of automatic handwritten recognition, the feature extraction is an important step, which find a set of measured values that accurately identify the input handwritten words or characters. The features can be categorized into two main types: structural and statistical features.

One of the first attempts for Arabic handwriting recognition using structural features is proposed in[1]. First, the words are segmented into strokes. Those strokes are then classified using their geometrical and topological characteristics (using start-

point, end-point and cross-point). The position of the classified strokes are examined, and the strokes are combined in several steps into a string of characters that represents the recognized word. Structural modeling of Arabic handwriting based on fuzzy polygonal approximation of Arabic text contours is used in [2]. The resulting models called Fuzzy Attributed Turning Functions (FATF), tolerate variations of the handwritten text of different writers. In [3] the input image is decomposed into several images representing the left diagonal, right diagonal, vertical and horizontal edges in the image. A set of features representing the densities of the foreground pixels in the various edge images is extracted using sliding windows.

Chain code direction frequencies were used in [2-4] in [4] a sliding window from left to right was used to extract the histogram of chain-code directions of the image strips. Gradient features for numeral recognition were used in [2, 5-7]. In [8-10] authors used as features the Hu moments invariants. Zernike moments and normalized central moments for character recognition are used in[11], segmented letters were partitioned into main body and secondary components. Then moment features were extracted from the main body and the secondary components as well as from the whole letter. Efficient feature subsets were selected by using Multi-objective genetic algorithm. M-band packet wavelet transform, proposed in [12] was used to recognize Persian/Arabic handwritten words. Vertical and horizontal projections, centroid distances, image zoning and other statistical features like Kirsch features, transitions count, , number of end- points, branch- points and cross-points, etc. are used in [2] [13].

Other features include wavelet transforms [9, 13],.Gabor filters [14, 15], DCT coefficients [9] and distribution features [16] which captures global and local features.

The GSC (gradient; structure; concavity) feature is a combination of structural and statistical features. It is used in [17, 18], in [18] it is computed as follow; first, gradient direction at each pixel is estimated to construct a gradient map from the normalized image. The Gradient features are obtained by counting the pixels which have almost the same gradient. The structural features detects complex patterns of the contour. The concavity features are caculated as follow, pixels in certain special regions such as holes and strokes are detected; then the image is divided into bins and the number of such pixels in each bin is counted. Cheng-Lin Liu et al in [19] combines structural and statistical features. The features include chain code feature, gradient feature, profile feature, and peripheral direction contributivity. The presence of diagonal, vertical, horizontal, and curved strokes and their orientation is collected into the feature vector. In [20] both structural and statistical features are used , end-points and intersection points were detected on a skeleton and then used to partition it into primitives. By using the average and variance of X and Y changes in each primitive, the direction and curvature of the skeleton can be statistically described. In [21] a combination of four feature sets is used: pixel description, structural description (ascender, descender, loup and dots), Gabor filter and Fourier descriptors.

In [22] a comparative Study of Persian/Arabic Handwritten Character Recognition is conducted. Authors used different feature sets and classifiers. Feature sets used in this study are computed based on gradient, directional chain code, shadow, under-sampled bitmap, intersection/junction/endpoint, and line-fitting information. Support Vector Machines (SVMs), Nearest Neighbour (NN), k-Nearest Neighbour (k-NN) are used as different classifiers. They evaluated their systems on a standard dataset of Persian handwritten characters.

In this paper, we try to determine how features designed for Arabic handwritten recognition can be efficient in Arabic historical documents by conducting a comparative study. To do four types of features (distribution, directional, structural and concavity features) which belong to one of the two categories of features (structural, statistical) and which are well used in the recognition of the Arabic handwritten are choosed for the comparative study.The work in this paper deal with CCs of Iben Sina database with embedded training of character HMMs.

2. METHODOLOGY

In Arabic handwriting recognition, the design of relevant feature is a very important and challenging task. Four types of features (distribution, directional, structural and concavity features) which belong to the two categories of features (structural, statistical) and which are well used in the recognition of the Arabic handwritten are choosed to be used for the recognition of Arabic handwritten in historical documents. In this paper, the preprocessing step is not the aim of the study; then the work deals with the cleaned and binarized CCs images of the Iben Sina database and the features are extracted from the mirror of the CCs images.

2.1. Distribution features:

The first type of features has been proposed by Marti and Bunke [16]. Many researchers have used these features for handwritten text recognition with HMM. Nine geometrical quantities are computed from the foreground pixels in each image column to determine the features. The first three features are the number of black pixels (2), the center of gravity (3) and the second order moment (4).

$$p(x,y) = \{1 \quad if\ pixel\ noir \\ 0 \quad if\ pixel\ blanc\} \tag{1}$$

$$f_1(x) = \sum_{y=1}^{m} p(x,y) \tag{2}$$

$$f_2(x) = \frac{1}{m}\sum_{y=1}^{m} y \cdot p(x,y) \tag{3}$$

$$f_3(x) = \frac{1}{m^2}\sum_{y=1}^{m} y^2 \cdot p(x,y) \tag{4}$$

The remaining features give mor detail of the writing. The forth and fifth features give the position of the upper and lower profile, the sixth, seventh, eighth and ninth features give the fraction of foreground pixels between the upper and lower profiles, the number of foreground to background transitions, and the gradient of the upper and lower profile with respect to the previous column, which provides dynamic information. To

extract a sequence of features from the CCs, first the CCs images are divided into four horizontal blocs, which allow the capture of the spatial distribution of the information, then a left to right sliding window is used.

2.2. Directional features:

The second type of features has been first proposed by favata et al [23] and used for Arabic handwritten by[17]. After the image is divided horizontally into four blocs, the gradient is extracted for each segment of the CCs. The gradient features are computed by convolving two 3x3 Sobel operators with the binary image. The gradient of a center pixel is computed as a function of its eight nearest neighbors. In this work, the gradient direction is used in the computation of a feature vector of the gradient feature map. The direction is split into 12 non-overlapping regions (1°–30°, 31°–60°…331°–360°). In each sampling region, a histogram of gradient directions is estimated, resulting in 12 gradient features for each image segment. This corresponds to the count of each gradient direction in the region. These counters are concatenated to give the gradient features of a CC.

2.3. Structural features

The third type of features (structural) proposed by favata et al [23] and used for Arabic handwritten by [17]. It captures the "mini-strokes" of the image. A set of 12 rules are applied to each pixel. Each rule examines a particular pattern of the neighboring pixels for allowed gradient ranges. The full list of rules and the neighborhood definitions are defined in [23]. These features represent the structural feature vector. The feature vector contains 12 types of structural features according to the rules.

2.4. Concavity features

The fourth feature (concavity) is defined in [24]. The concavity feature are used to calculate the average distance scanning pixel by pixel from the cell border to the first met stroke (Figure.3). There are four concavity features. Let ri, li, ui and di respectively be the distance from the right, left, up and bottom cell border to the first met stroke along left, right, up and down directions. Then (5), (6), (7) and (8) can calculate the four concavity features respectively.

$$f4 = \frac{\sum_{i=1}^{w} di}{w} \quad (5)$$

$$f5 = \frac{\sum_{i=1}^{w} ui}{w} \quad (6)$$

$$f6 = \frac{\sum_{i=1}^{h} ri}{h} \quad (7)$$

$$f7 = \frac{\sum_{i=1}^{h} li}{h} \quad (8)$$

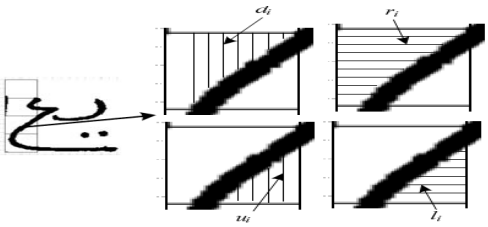

Figure3. The four types of concavity features [24]

3. RECOGNITION WITH HMM

The recognition process is based on Hidden Markov models with HTK toolkit. For details about the HMMs refer to LR.Rabiner [25]. The hidden Markov model toolkit (HTK) [26] is a portable toolkit for building and manipulating hidden Markov models. The HTK tools are divided into four types: data preparation, training, testing and result analysis tool. With HTK HMMs can be built with any topology using simple text files.

Our recognition system's model is built by using embedded HMM which is constructed by concatenating characters HMM models. Each character is defined by an HMM. The HMM character model's topology is a three states left-right. In each state the observation's probability of density are modeled as a mixture of 50 to 60 Gaussian. A 31 character HMMs are used. The 28 Arabic character basic shapes and 3 types of diacritics.

The training tools of HTK adjust parameters of HMMs by using an embedded training version of the Baum-Welch algorithm [25] for likelihood probability's maximization of the training data.

In the recognition phase, the Viterbi Algorithm is implemented in the HTK's recognition tool [25]. The recognition tool takes as input the vector of the extracted features a network describing the allowable word sequences, a dictionary defining each word and a set of HMMs.

The performance of the recognition system is evaluated by the HTK result analysis tool by matching the recognizer output data with the original reference transcription. The number of substitution (S), deletion (D) and insertion (I) are counted [26].

4. EXPERIMENTS AND RESULTS

4.1. Iben Sina database

Our approach is evaluated on the Ibn Sina database [27] for ancient Arabic documents. The Ibn Sina database is based on an important philosophical work by the famous Persian scholar Ibn Sina. This database consists of 60 pages (Figure.4) and approximately 25,000 Arabic subword shapes written in the Naskh style. The document images were binarized with a dedicated algorithm to preserve the shape's topology. Each page contains approximately 500 subword shapes. There are 1200 different classes (different subword (CC) shape) (Figure.5) , but the distribution of the database is highly unbalanced; some classes have up to 5000 entries, while others have fewer than five [28].

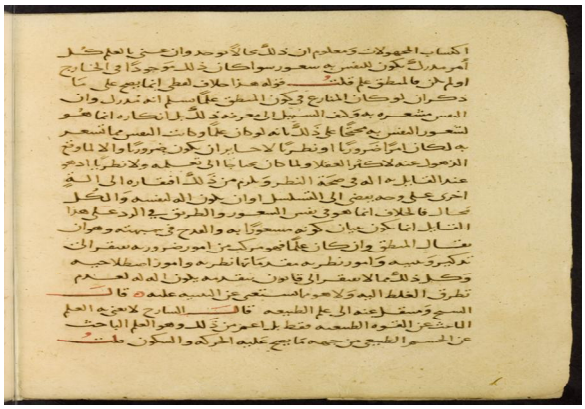

Figure.4. Page 0010 of the Iben Sina database.

Figure.5. Example of binarised subword shapes (CCs) of Iben Sina database

4.2. Results discussion

In our experiments, the CCs images were divided into four horizontals blocs, which allows the capture of the spatial distribution of the information, and the feature extraction is done by using sliding window. The frame width from 1pixel to 2 pixels is used and an overlap of 1 pixel with respect to the minimal size of a character (a CC) which can be equal two pixels.

Ten pages (5137 CCs) of the Iben Sina database were used in the training phase and five pages (2628 CCs) were used in the test phase.

In Table.1 are shown the recognition rates resulted in the application of different types of features (distribution, gradient, structural and concavity) and the combination of a part of these features (structural and statistical). The results show that distribution features give the best rate when using frame width 1 pixel without overlap, unlike the other types of features wich work well with frame width two pixels with one pixel overlap. The results also show that the GSC features gives best rate with

50 mixtures of Gaussian in contrary of the other features which works well with 60 mixtures of Gaussian.

In Figure.6 are given the recognition rates of the whole CC as well as the characters returned by the system with the use of the different combinations of the features

The results showed that the combination of the directional (gradient) and concavity features outperforms the other types of features in recognition of the whole CC with a recognition rate of 45.36%. They also show that the combination of directional (gradient), structural and concavity features give the best recognition rate 63.88% in the recognition of the characters in the CCs. The recognition rate of character is greater than that of the CCs; this is due to the recognition of a part of CC instead of the whole CC.

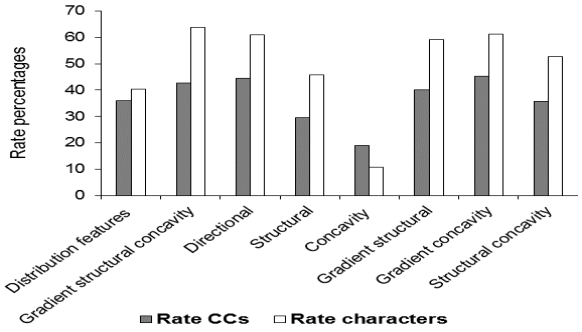

Figure 6. Recognition rate using different types of features.

Through these results, we observe that the combination of directional, concavity and structural features are more suitable for the recognition of the Arabic handwritten in the historical documents; however, the recognition rate 45.36% is low and not satisfactory.

Table.1. Experimental results

Features	Frame width	overlap	Hmm topology	Mixture of Gaussian	Rate CCs %	Rate Characters %
Distribution features	1 pixel	0	31 HMMS 1 HMM with 3 states 1 hmm per character	60	35,93%	40,27%
	2 pixels	1		60	29.85 %	30.58%
Directional (gradient)	1 pixel	0		60	39,54%	56,03%
	2 pixels	1 pixel		60	44,54%	60,89%
GSC (gradient, concavity ,structural)	1 pixel	0		50	41,58%	61,25%
	2 pixels	1 pixel		50	42,79%	**63,88%**
Structural	1 pixel	0		60	28,27%	42,15%
	2 pixels	1 pixel		60	29,56%	45,74%
GS (gradient, structural)	2 pixels	1 pixel		60	40,17%	59,17%
GC (gradient, concavity)	2 pixels	1 pixel		60	**45,36%**	61,29%
Concavity	2 pixels	1 pixel		60	18,84%	10,69%
SC (structural, concavity)	2 pixels	1 pixel		60	35,59%	52,74%

5. CONCLUSION

In this paper, a comparative study on features extraction methods (distribution, directional, structural and concavity features) for the recognition of Arabic handwritten in historical documents is done.

The results show that the combination of directional and concavity features have reached a high recognition rate in the recognition of a whole CC but in the recognition of characters, the combination of directional, structural and concavity features permitted to achieve the best recognition rate.

This comparative study shows that the combination of concavity and directional features are well suited for the Arabic handwritten recognition in historical documents but the results are low and still not satisfactory, hence the need for a novel feature extraction method that takes into a count the complex nature of Arabic handwritten and the degradation in the historical documents.

6. REFERENCES

[1] Almuallim, H. and S. Yamaguchi, A Method of Recognition of Arabic Cursive Handwriting. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1987. PAMI-9(5): p. 715-722.

[2] Abdleazeem, S. and E. El-Sherif, Arabic handwritten digit recognition. International Journal of Document Analysis and Recognition (IJDAR), 2008. 11(3): p. 127-141.

[3] Alaei, A., P. Nagabhushan, and U. Pal. Fine Classification of Unconstrained Handwritten Persian/Arabic Numerals by Removing Confusion amongst Similar Classes. in 2009 10th International Conference on Document Analysis and Recognition. 2009.

[4] Dehghan, M. Faez, K. Ahmadi, M. Shridhar, M., Handwritten Farsi (Arabic) word recognition: a holistic approach using discrete HMM. Pattern Recognition, 2001. 34(5): p. 1057-1065.

[5] ABDELAZEEM, S., Comparing arabic and latin handwritten digits recognition problems, W.A. Sci. and Engin., Editors. 2009, Technol.

[6] Awaidah, S.M. and S.A. Mahmoud, A multiple feature/resolution scheme to Arabic (Indian) numerals recognition using hidden Markov models. Signal Processing, 2009. 89(6): p. 1176-1184.

[7] Surinta, Olarik Karaaba, Mahir F. Schomaker, Lambert R. B. Wiering, Marco A., Recognition of handwritten characters using local gradient feature descriptors. Engineering Applications of Artificial Intelligence, 2015. 45: p. 405-414.

[8] Abd, M.A. and G. Paschos, Effective Arabic Character Recognition Using Support Vector Machines, in Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, T. Sobh, Editor. 2007, Springer Netherlands: Dordrecht. p. 7-11.

[9] AlKhateeb. Jawad H, Jiang. Jianmin Ren, Jinchang Khelifi, Fouad Ipson and Stan S, Multiclass classification of unconstrained handwritten arabic words using machine learning approaches. The Open Signal Process., 2009. 2: p. 21-28.

[10] M. Hamdani, H. E. Abed; M. Kherallah and A. M. Alimi ,Combining Multiple HMMs Using On-line and Off-line Features for Off-line Arabic Handwriting Recognition. in 2009 10th International Conference on Document Analysis and Recognition. 2009.

[11] Gheith, A. and A. Nasser, Novel Moment Features Extraction for Recognizing Handwritten Arabic Letters. Journal of Computer Science, 2009. 5(3): p. 226-232.

[12] Broumandnia, A., J. Shanbehzadeh, and M. Nourani. Handwritten Farsi/Arabic Word Recognition. in 2007 IEEE/ACS International Conference on Computer Systems and Applications. 2007.

[13] Abandah, G.A., K.S. Younis, and M.Z. Khedher, Handwritten Arabic character recognition using multiple classifiers based on letter form, in Proceedings of the Fifth IASTED International Conference on Signal Processing, Pattern Recognition and Applications. 2008, ACTA Press: Innsbruck, Austria. p. 128-133.

[14] Jin Chen, Huaigu Cao, Rohit Prasad, Anurag Bhardwaj and Prem Natarajan., Gabor features for offline Arabic handwriting recognition, in Proceedings of the 9th IAPR International Workshop on Document Analysis Systems. 2010, ACM: Boston, Massachusetts, USA. p. 53-58.

[15] Mahmoud, S.A. and W.G. Al-Khatib, Recognition of Arabic (Indian) bank check digits using log-gabor filters. Applied Intelligence, 2011. 35(3): p. 445-456.
[16] U., V.M. and H. Bunke, Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system,. International Journal of Pattern Recognition and Artificial Intelligence, 2001. 15(01): p. 65-90.
[17] Mahmoud, S.A. and M.A. Sameh Recognition of off-line handwritten arabic (indian) numerals using multi-scale features and support vector machines vs. hidden markov models. The Arabian Journal for Science and Engineering, 2009. 34(2B): p. 429-444.
[18] S. Saleem, H. Cao, K. Subramanian, M. Kamali, R. Prasad, P. Natarajan. Improvements in BBN's HMM-Based Offline Arabic Handwriting Recognition System. in 2009 10th International Conference on Document Analysis and Recognition. 2009.
[19] Liu Cheng-Lin, Nakashima Kazuki, Sako Hiroshi and Fujisawa Hiromichi., Handwritten digit recognition: benchmarking of state-of-the-art techniques. Pattern Recognition, 2003. 36(10): p. 2271-2285.
[20] Mozaffari, S., K. Faez, and M. Ziaratban. Structural decomposition and statistical description of Farsi/Arabic handwritten numeric characters. in Eighth International Conference on Document Analysis and Recognition (ICDAR'05). 2005.
[21] S. Haboubi S Maddouri, N. Ellouze and H. El-Abed. Invariant Primitives for Handwritten Arabic Script: A Contrastive Study of Four Feature Sets. in 2009 10th International Conference on Document Analysis and Recognition. 2009.
[22] Alaei, A., U. Pal, and P. Nagabhushan. A comparative study of persian/arabic handwritten character recognition. in Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on. 2012. IEEE.
[23] John , T.F. and S. Geetha, A Multiple Feature/Resolution Approach to Handprinted Digit and Character Recognition. International Joumal of Imaging Systems and Technology, 1996. 7: p. 304-311.
[24] Xiang Dong, Yan Huahua, Chen Xianqiao and Cheng Yanfen. Offline Arabic handwriting recognition system based on HMM. in Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on. 2010. IEEE.
[25] Rabiner, L.R., A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 1989. 77(2): p. 257-286.
[26] Young Steve, Evermann Gunnar, Gales Mark, Hain Thomas, Kershaw Dan, Liu Xunying, Moore Gareth, Odell Julian, Ollason Dave and Povey Dan The Htk Book. 2001, Cambridge University Engineering Dept.
[27] Farrahi Moghaddam Reza, Cheriet Mohamed, Adankon Mathias M, Filonenko Kostyantyn, and Wisnovsky Robert, IBN SINA: A database for research on processing and understanding of Arabic manuscripts images", P.o. DAS'10, Editor. 2010. p. 11-18.
[28] Chherawala, Y. and M. Cheriet, W-TSV: Weighted topological signature vector for lexicon reduction in handwritten Arabic documents. Pattern Recognition, 2012. 45(9): p. 3277-3287.