Journal of Pattern Recognition Research 1 (2016) 55-60

Received April 15, 2016. Revised October 10, 2016. Accepted October 17, 2016.

Arabic OCR Using a Novel Hybrid Classification Scheme

Abdul Mueed Hafiz *mueedhafiz@yahoo.com*
Department of Electronics & Communication, University of Kashmir
Srinagar, J&K 190006, India

Ghulam Mohiuddin Bhat *drghmbhat@gmail.com*
University Science Instrumentation Center, University of Kashmir
Srinagar, J&K 190006, India

Abstract

Hidden Markov Models or HMMs, are a relatively recent phenomenon for Arabic handwriting recognition. They are robust and efficient in classification. In this paper, an effort has been made to further boost the recognition capability of HMM Based Arabic Optical Character Recognition Systems, by using a two-tier hybrid classification scheme. The first tier consists of Part of Arabic Word or PAW, Based HMMs, and the second tier is a k-Nearest Neighbor Classifier or KNN Classifier. The second tier receives its inputs from the first tier. A second novel Hybrid Scheme is also examined. The recognition accuracies of the proposed schemes have been compared to contemporary techniques and they show an improvement in classification accuracy. HMMs have been implemented using the HTK Toolkit. The database used has been obtained from the IFN-ENIT Database of Arabic Words.

Keywords: Hidden Markov Models, Hybrid Classifiers, Arabic Text Recognition, IFN-ENIT Database

1. Introduction

In the field of Arabic OCR, many different classifiers have been used for both printed and handwritten text, e.g. Hidden Markov Models (HMMs) [1, 2, 3, 4], Recurrent Neural Networks[5], Dynamic Bayesian Networks (DBNs) [6], K-Nearest Neighbor (KNN) classifier [7], Support Vector Machines (SVM) [8], etc. Recently, hybrid systems [9] have been implemented [10, 11, 12, 13, 14] and have shown better performance than single classifiers, e.g. for HMM/ANN hybrid classifier [13]. However, in hybrid classifiers, the fusion technique and decision making are usually complex.

Where Artificial Neural Networks (ANNs) are used as one of the classifiers, many parameters have to be taken into account for the same and behavior for such varies for the dataset used and the instance of training. Sometimes the network converges quickly and on other instances it may take a long time to converge. When SVM is used, the training of the same, places limits of error rate of the hybrid system.

HMM-KNN hybrid classifiers have been used previously for handwritten cheque recognition [15]. However, the recognition is holistic i.e. segmentation free and uses a modified KNN classifier. HMM-KNN classifier combination has also been used for facial expression recognition [16]. Here, the number of variables passed on to the KNN classifier is very small.

In this paper, an HMM-KNN hybrid classifier has been introduced for Arabic language OCR. As per the literature survey, this is the first attempt to use a hybrid HMM-KNN classifier for Arabic OCR. The hybrid system combines pre-HMM-classification with post-KNN-classification. The approach is segmentation based. Also, the KNN stage borrows large number of variables from the HMM stage. The merits of the proposed hybrid classifier are high accuracy, robustness and simplicity, due to the use of post-KNN-classification scheme.

The proposed scheme consists of a two-tier system. The first-tier of the hybrid system is a Part-of-Arabic-Word (PAW) [17] based HMM Classifier, which generates the corresponding log-probabilities for each PAW image. Once all the PAW images have been classified by the HMM classifier, the sequence and number of emitted PAWs for the sample word is stored in an array. The emitted PAW vector is converted into an integer vector, by assigning to each scalar, a serial number after a simple serial-wise look-up in the overall PAW list for the dataset. This process is performed for both training set as well as testing set. This is followed by the second-tier classifier, viz. the KNN Classifier [18, 19, 20], which classifies the (first-tier output) testing-set integer vectors as neighbors of the (first-tier output) training-set integer vectors.

An improvement in recognition was noted by following the proposed hybrid scheme over that while using word-based HMMs for classification of the same dataset. The noted improvement stood firm also when a novel two-tier HMM-HMM system was used for classification of same dataset. The word images used were obtained from the IFN/ENIT Arabic Database of Arabic Words [21]. The HMM Classifiers were implemented by using HTK Toolkit [22, 23]. The KNN Classifier was also implemented using MATLAB.

2. Experimental Investigation

2.1 Preprocessing

The sample words taken from the IFN/ENIT database of Arabic words comprise of three sets of randomly selected images. Set A had 600 word images (15 words with a total of 39 PAWs). Set B had 400 word images (10 words with a total of 32 PAWs). Set C had 200 word images (5 words with a total of 19 PAWs). Thirty images were used per word for training and ten images were used per word for testing. The word images were binarized by using grayscale thresholding having a grayscale range of 0.0 to 1.0 (threshold used was 0.5). The images were de-slanted and resized to 100x500 pixels dimensions. They were thinned using the Zhang-Suen Thinning Algorithm [24], after which they were dilated.

The word images were hand-segmented into their respective PAW images. An assumption of a fairly consistent Arabic Word-to-PAW image segmentation preprocessing system was made, which is not unusual. Boosting the HMM Classifier was investigated thus. Next, horizontal flipping of the word images was done. The pre-processed images were divided in 6-pixel wide non-overlapping vertical frames for feature extraction. The features [25] extracted for each frame were as follows:

1. f_1 : Density of foreground pixels.

2. f_2 : Number of black/white transitions between two consecutive frame cells going from top to bottom:

$$f_2 = \sum_{i=2}^{n_c} |b(i) - b(i-1)|. \tag{1}$$

   b(i) is the density level belonging to the i-th cell of the frame. It is equal to 1 if the cell contains at least one foreground pixel and is equal to 0 otherwise.

3. f_3 to f_8 : Each feature is the sum of foreground pixels in one of the six vertical columns in each frame.

4. f_9 to f_{14} : Six concavity features - N_{lu}, N_{ur}, N_{rd}, N_{dl}, N_v and N_h. Each is the number of background pixels which have neighboring black pixels in left-up, up-right, right-down, down-left, vertical, and horizontal directions respectively.

5. f_{15} to f_{35} : These are 21 percentile features. At each y-position, the number of black pixels from the top of the frame was computed. This number was normalized from 0 to 100, thus equating it from 0% to 100% of the total blackness of the frame. y-axis was also normalized from 0 to 1, also equating it to 0% to 100% of the total height of the frame. Finally, the total blackness was divided into 21 percentiles. The corresponding values of percentile height of the frame were the percentile features.

6. f_{36} to f_{695} 660 raw pixel binary-value based features per frame.

Figure 1 illustrates the pre-processing process.

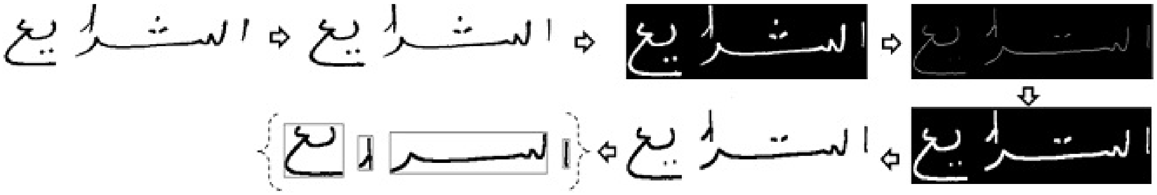

Fig. 1: Pre-processing stages for a sample word image. Clockwise: de-slanting, binary inversion, skeletization, dilation, binary inversion, segmentation into PAWs after removing dots.

2.2 Classification

The HMM Classifier was implemented using HTK. The number of states given to every PAW HMM varied from 3 to 8, where the number of states was the sum of 3 and the number of letters in the PAW image. If the sum was more than 8, still 8 states were assigned to the PAW HMM. This was done in view of the fact that increasing states per HMM above 8 did not have any appreciable effect on the overall error rate. After sequential recognition, once the respective highest scoring PAW-based HMMs gave their PAW predictions, the latter were converted into an integer vector by a simple table-lookup procedure assigning a serial number to each PAW from the overall list of PAWs in the dataset.

The training set as well as the testing set images were passed through the trained PAW-based HMMs for obtaining the lookup vectors. After this, the testing set vectors were classified using k-NN classifier against the training set vectors. Figure 2 illustrates the proposed Handwriting Recognition System.

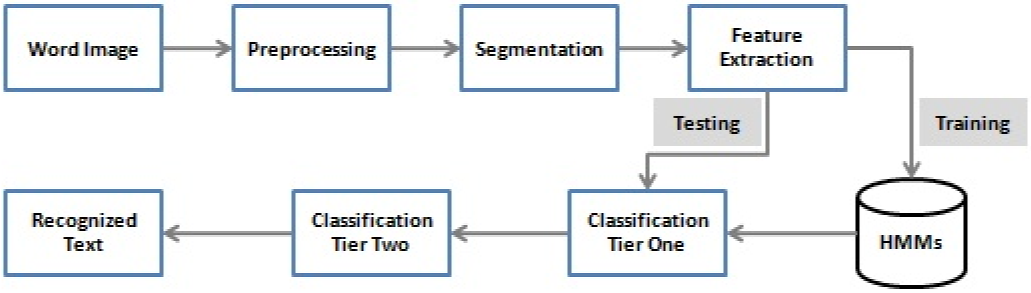

Fig. 2: Hybrid Handwriting Recognition System

For comparison against contemporary schemes, the following experiments were carried out. The hand-segmented testing set PAW images for each word were concatenated and afterwards recognized by word-based HMMs which were constructed using the individually trained PAW HMMs. Further as another experiment in this regard, the training set vectors were used to train a second set of HMMs to study a novel two-tier HMM-HMM Classification Scheme. In this scheme, the first tier recognized the PAWs and the second tier recognized the table-lookup-integer-vectors. To be noted that the second-tier HMM stage for this classification scheme was trained on the integral numbers obtained from the look-up procedure.

2.3 Experimental Results

The results of the experimental investigation are summarized in Table 1. It shows classification results for the conventional single tier classifier, HMM/ANN Hybrid Classifier and proposed classifiers, on concatenated images of hand-segmented testing set images.

The word-based HMMs constructed from their corresponding PAW HMMs were used to classify the testing set word images. The PAW-based HMMs were trained on hand-segmented PAW images belonging to the training set. The HMM/ANN hybrid classifier was implemented using the APRIL-ANN Toolkit [26] and on the lines of Boquera, et al., [13]. Hybrid HMM/ANN models, with a different number of HMM states and parameters of Multilayer Perceptron (MLP) were used. The softmax outputs [13] were used as emission probabilities of the states of the N optical models, where N was number of PAW types in sets A, B, and C, respectively. Fully connected MLPs of 540 input units (for 60-dimensional 9 feature vectors) were trained. The number of output units is determined by the total number of states of the N optical models (from Nx3 output units for 3-state HMMs to Nx8 output units for 8-state HMMs). Also, the number of hidden units was determined empirically by measuring the mean-square error (MSE) on the validation set. Parameters like the learning rate and the momentum (for training of MLPs), were empirically tuned with the validation data.

3. Conclusion

Two novel techniques were discussed to improve the classification accuracy of HMM Classifiers for Arabic OCR. The word images were taken from the IFN/ENIT Arabic Word Database. A Hybrid Classification Scheme has been proposed to boost the performance of conventional HMM Classifiers. The first tier consists of PAW-based HMMs and the second tier consists of a k-NN Classifier. The k-NN classifier assigns classes to the emitted PAWs from tier one, after they have been converted to integer numerals by a basic serial-wise look-up from the overall PAW list of the used database. The performance of the proposed approach has been found to be better in comparison to the contemporary schemes.

Table 1: Comparison of classification accuracy of proposed scheme with relevant contemporary schemes

Dataset	Classification Accuracy			
	Word-based HMM	HMM/ANN Hybrid	PAW HMMs - Integer HMM	PAW HMM-KNN
A	57.33	78.67	82.67	82.67
B	65.00	82.5	81.00	86.00
C	88.00	94.00	86.00	94.00

References

[1] M. Pechwitz, and V. Maergner, *HMM based approach for handwritten arabic word recognition using the IFN/ENIT - database.* in Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on, 2003, pp. 890-894.

[2] A. Kundu and T. Hines, *Arabic Handwriting Recognition Using VDHMM and Over-segmentation.* in Guide to OCR for Arabic Scripts, V. Mrgner and H. El Abed, Eds., ed: Springer London, 2012, pp. 507-540.

[3] I. Alkhoury, A. Gimnez, and A. Juan, *Arabic Handwriting Recognition Using Bernoulli HMMs.* in Guide to OCR for Arabic Scripts, V. Mrgner and H. El Abed, Eds., ed: Springer London,2012, pp. 255-272.

[4] I. Ahmad, and G. A. Fink, *Multi-stage HMM based Arabic text recognition with rescoring.* in Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, 2015, pp. 751-755.

[5] A. Graves, *Offline Arabic Handwriting Recognition with Multidimensional Recurrent Neural Networks.* in Guide to OCR for Arabic Scripts, V. Mrgner and H. El Abed, Eds., ed: Springer London, 2012, pp. 297-313.

[6] A. Khemiri, A. Kacem Echi, A. Belaid, and M. Elloumi, *Arabic handwritten words off-line recognition based on HMMs and DBNs.* iin Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, 2015, pp. 51-55.

[7] J. H. Alkhateeb, F. Khelifi, J. Jiani, and S. S. Ipson, *A New Approach for Off-line Handwritten Arabic Word Recognition Using KNN Classifier.* in IEEE International Conference on Signal and Image Processing Applications, 2009, pp. 191-194.

[8] M. Khalifa and Y. BingRu, *A Novel Word Based Arabic Handwritten Recognition System Using SVM Classifier.* iin Advanced Research on Electronic Commerce, Web Application, and Communication. vol. 143, G. Shen and X. Huang, Eds., ed: Springer Berlin Heidelberg, 2011, pp. 163-171.

[9] M. Woniak, M. Grana, and E. Corchado, *A survey of multiple classifier systems as hybrid systems.* Inf. Fusion, vol. 16, pp. 3-17, 2014.

[10] A. L. Koerich, Y. Leydier, R. Sabourin, and C. Y. Suen, *A hybrid large vocabulary handwritten word recognition system using neural networks with hidden Markov models.* in Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on, 2002, pp. 99-104.

[11] N. Zermi, M. Ramdani, and M. Bedda, *Arabic Handwriting Word Recognition Based on a Hybrid HMM/ANN Approach.* International Journal of Soft Computing, vol. 2, pp. 5-10, 2007.

[12] W. W. Azevedo and C. Zanchettin, *A MLP-SVM hybrid model for cursive handwriting recognition.* in Neural Networks (IJCNN), The 2011 International Joint Conference on, 2011, pp. 843-850.

[13] S. Espana-Boquera, M. J. Castro-Bleda, J. Gorbe-Moya, and F. Zamora-Martinez, *Improving Offline Handwritten Text Recognition with Hybrid HMM/ANN Models.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 33, pp. 767-779, 2011.

[14] C. Zanchettin, B. L. D. Bezerra, and W. W. Azevedo, *A KNN-SVM hybrid model for cursive handwriting recognition.* in Neural Networks (IJCNN), The 2012 International Joint Conference on, 2012, pp. 1-8.

[15] D. Guillevic, and C. Y. Suen, *HMM-KNN word recognition engine for bank cheque processing.* in Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on, 1998, pp. 1526-1529 vol.2.

[16] Q. Wang, and S. Ju, *A Mixed Classifier Based on Combination of HMM and KNN.* in 2008 Fourth International Conference on Natural Computation, 2008, pp. 38-42.

[17] A. AbdulKader, *A Two-Tier Arabic Offline Handwriting Recognition Based on Conditional Joining Rules.* in Arabic and Chinese Handwriting Recognition. vol. 4768, D. Doermann and S. Jaeger, Eds., ed: Springer Berlin Heidelberg, 2008, pp. 70-81.

[18] T. M. Cover and P. E. Hart, *Nearest Neighbor Pattern Classification.* IEEE Transactions on Information Theory, vol. 13, pp. 21-27, 1967.

[19] M. Tsypin and H. Rder, *On the Reliability of kNN Classification.* in World Congress on Engineering and Computer Science, 2007.

[20] M. Maleki, K. Eroglu, O. Aydemir, N. Manshoori, and T. Kayikcioglu, *A new method for selection optimum k value in k-NN classification algorithm.* in Signal Processing and Communications Applications Conference (SIU), 2013 21st, 2013, pp. 1-4.

[21] H. El Abed, and V. Margner, *The IFN/ENIT-database - a tool to develop Arabic handwriting recognition systems.* in Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on, 2007, pp. 1-4.

[22] S. Young, *The HTK Book V3.4. Cambridge: Cambridge University Press.* ,2006.

[23] A. Maqqor, A. Halli, K. Satori, and H. Tairi, *Using HMM Toolkit (HTK) for recognition of arabic manuscripts characters.* in Multimedia Computing and Systems (ICMCS), 2014 International Conference on, 2014, pp. 475-479.

[24] T. Y. Zhang, and C. Y. Suen, *A fast parallel algorithm for thinning digital patterns.* Commun. ACM, vol. 27, pp. 236-239, 1984.

[25] L. Likforman-Sulem, R. Al Hajj Mohammad, C. Mokbel, F. Menasri, A.-L. Bianne-Bernard, and C. Kermorvant, *Features for HMM-Based Arabic Handwritten Word Recognition Systems.* in Guide to OCR for Arabic Scripts, V. Mrgner and H. El Abed, Eds., ed: Springer London, 2012, pp. 123-143.

[26] F. Zamora-Martinez, S. Espana-Boquera, J. Gorbe-Moya, J. Pastor-Pellicer, and A. Palacios-Corella, *APRIL-ANN toolkit, A Pattern Recognizer In Lua with Artificial Neural Networks.* , 2013.