International Journal of Engineering and Future Technology,
Year 2016; Issue No. 1; Volume 2
ISSN 2455-6432

KU – Database of Handwritten Arabic Words

A.M. Hafiz[1], Z. A. Bhat[2], M. Jan[3], M. U. Bhat[4], I. B. Sofi[5], I. A. Tantray[6], G. M. Bhat[7,*]

[1]Department of Electronics and Instrumentation Technology,
University of Kashmir, Srinagar, J&K (India)
Email: mueedhafiz@yahoo.com

[2-6]Department of Electronics and Instrumentation Technology,
University of Kashmir, Srinagar, J&K (India)

[7]Department of Electronics and Instrumentation Technology,
University of Kashmir, Srinagar, J&K (India)
Email: drghmbhat@gmail.com

ABSTRACT

In this paper a new database of handwritten Arabic words has been presented. The database consists of words taken from well-known Arabic proverbs. The ground truth information is also provided. 12 writers each filled 5 forms thrice, with every form containing a total of 84 words, with Part of Arabic Word (PAW) range distributed equally (for the words) ranging from 1 to 9. The database contains a total of 3024 word images, 14616 PAWs, and 30744 characters. It has been designed for training and testing recognition systems for handwritten Arabic words. The database offers novel features like even distribution of quantity of word images with respect to word length. The Kashmir University (KU) - database is available for the purpose of research. A portion of the database was tested using a Hidden Markov Model (HMM) based Optical Character Recognition Engine and benchmarking of some important features found in literature was done.

Keywords: Arabic; Handwriting; Database; HMM; Arabic OCR;

Mathematics Subject Classification: 68T45

ACM Computing Classification System: I.7.5

1. INTRODUCTION

Arabic Optical Character Recognition (OCR) research started in 1970s (Al-Badr and Mahmoud 1995). The first work on Arabic OCR was published in 1975 (Nazif 1975). The first Arabic OCR system was released in the 1990s (Märgner and El Abed 2008). The recognition of Arabic handwritten text has some unique challenges as well as benefits for the research community (Cheriet 2008). After three decades, there has been a lack of effort in Arabic handwritten text recognition compared to that of other languages (El Abed and Margner 2008). Although, there are a few commercial Arabic OCR systems for machine-printed text (like Sakhr, IRIS, ABBYY, etc.), there is none for handwritten text. Different classifiers like Hidden Markov Models - HMMs (Likforman-Sulem et. al. 2012, Pechwitz et al. 2012, Maqqor et al. 2014), Artificial Neural Networks - ANNs (Graves 2012), Support Vector Machines - SVMs

(Pirsiavash et al. 2005, Khalifa and BingRu 2011, Amara et al. 2014), k Nearest Neighbor - kNN (Al-Jamimi and Mahmoud 2010, Rashad and Semary 2014), Bayesian Networks (Mahjoub et al. 2013), etc. have been used for Arabic text recognition. Hidden Markov Model (HMM) Based Classifiers (Rabiner 1989, Likforman-Sulem et al. 2012, Pechwitz et al. 2012) have emerged as an important classification approach for both Latin and Non- Latin texts. Arabic differs English in that it is written from right to left (i.e. in the opposite direction), it does not have uppercase letters, has diacritics, etc. HMMs are robust classifiers for Arabic Text with many important research contributions from various researchers. The various applications of OCR include both, offline recognition (Lorigo and Govindaraju 2006, Parvez and Mahmoud 2013), in the form of check processing (Knerr, Augustin et al. 1998, Mahmoud and Al-Khatib 2011), automated post office mail sorting (Fujisawa 2008), offline recognition of text corpus, recent, and old documents (Cheriet and Moghaddam 2012), as well as online recognition (Tagougui et al. 2013), for text written on hand-held devices.

Although state of the art OCR Systems promise good results for Latin text like English and many commercial OCR Systems are available in the market for the same, much work needs to be done for Arabic language which is spoken and written in many countries around the globe. Although state of the art systems have been presented in many International Competitions like International Conference on Document Analysis and Recognition (ICDAR) (Djeddi, Al-Maadeed et al. 2015), but a commercial Arabic OCR System for handwritten text is still on the cards. One of the factors limiting extent of research for Arabic OCR is the small number of databases available for Arabic Text. Although some word-based databases like AHDB (Al-Ma'adeed, Elliman et al. 2002), IFN/ENIT Database (El Abed and Margner 2007), ERIM Database , Al-Isra Database (Kharma et al. 1999), etc. are available for research, however such are limited not only by number but also public availability. Some paragraph-based databases have also been developed e.g. QUWI Database (Maadeed et al. 2012), but to use these, complete care of pre-processing aspect of OCR has to be taken into account. Hence, for research focusing on feature-extraction, classification and/or post-processing aspects of OCR, word-based databases are more suitable as the database entities are available in word format and need not be isolated from a paragraph after preprocessing techniques like document de-skewing, line detection, line segmentation, etc. Further, in the word databases available, if given, the distribution of quantity of word-images on basis of word-length (if indicated) is non-uniform. For example, the data distribution in IFN/ENIT Database (El Abed and Margner 2007) is given in Table 1.

Table 1: Data distribution in IFN/ENIT Database

Quantity of words in town names	Quantity of town name images	Quantity of PAW's	Quantity of characters
1	12992	40555	76827
2	10826	54722	98828
3	2599	20120	36004
4	42	188	552
Total	26459	115585	212211

It can be observed from Table 1, that the distribution of quantity of town image names is highly uneven and generally decreases as word-length increases. For robust OCR classifiers like HMM-based classifiers used on this database, the distribution of HMM word-models on basis of word-length is also uneven. Further, the effect of OCR techniques used after HMM emission has taken place, e.g. use of n-grams, or use of lexicons (along with search methods like Beam Search, etc.) will not be investigated properly again because of uneven distribution of quantity of word images based on word length. Keeping these indicators in mind, and to further extend research on Arabic OCR, an elaborately structured Arabic text database has been developed.

2. OVERVIEW OF THE KU-DATABASE

Kashmir University (KU) being a medium sized database, consists of 84 word-fragments taken from Arabic Proverbs. The word images consist of almost equal number of fixed PAW subsets, where the number of PAWs per subset varies from 1 to 9. The writers were students of the Post-graduate Department of Arabic, University of Kashmir, Srinagar.12 writers were employed for writing the text, and each writer wrote each sample thrice over a total of 5 forms (per writer). The total number of word images in the database is 3024, total number of PAWs is14616, and total number of characters is 30744. Figure 1 shows an example of a filled form. Figure 2 shows some words taken randomly from the KU-database.

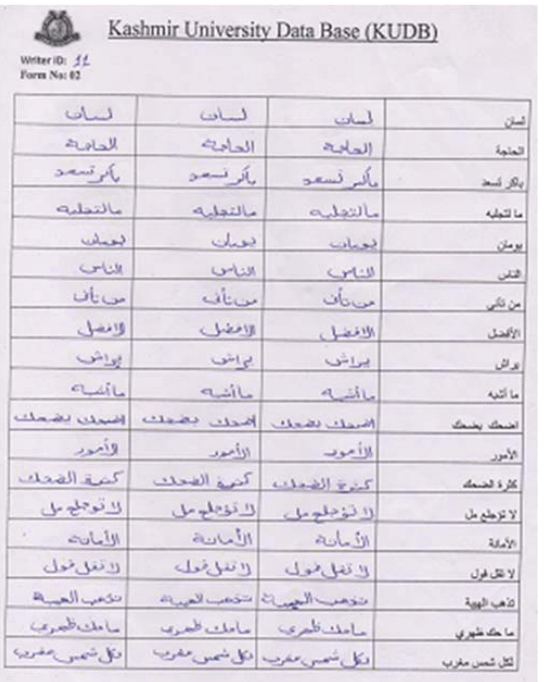

Figure 1. Example of a filled form.

Figure 2. Examples from the KU-database: 5 word images written by 3 different writers

All word images have been named such as to distinguish the Writer_ID (1 to 12), Word_ID (1 to 84) and Instance_Number (1 to 3). For example, a word image named as **w07_23_02.bmp** indicates that it has been written by the 7th writer, for the 23rd word and is the 2nd instance out of the 3 available. Ground truth is also available for all the 84 words in separate files. For example, the ground truth for 39th word, viz:

alifA-lamB_haM_alifE-jimB_taE-taB_faM_taM_qafE

indicates that the word consists of 4 PAWs (as it contains 3 PAW separators i.e. hyphens). Each letter (separated by an underscore in the ground truth) ends with an **A, B, M,** or **E,** indicating that the letter occurs either alone, or, in the beginning, middle, or end of the PAW, respectively. The letter names are found in Table 2.

Table 2: Letter denominations used in the database

Character	Isolated Form	Character	Isolated Form
Alif	ا	Dhad	ض
Ba	ب	Taa	ط
Ta	ت	Dha	ظ
Tha	ث	Ayn	ع
Jim	ج	Ghayn	غ
Ha	ح	Fa	ف

Kha	خ	Qaf	ق
Dal	د	Kaf	ك
The	ذ	Lam	ل
Ra	ر	Mim	م
Zai	ز	Nun	ن
Sin	س	He	ه
Chin	ش	Waw	و
Sad	ص	Ya	ى

The letters are combined with diacritics (signs for added pronunciation) like 'Hamza' and 'Madda', giving more variations. Since incorporating diacritics for an HMM Based OCR System tends to create problems, they are preferably removed from the word images.

The database is publically available for the purpose of research. It can be ordered by emailing one of the authors.

The data distribution of KU-database is shown in Table 3.

Table 3: Data distribution in KU-database

Quantity of paws in word names	Quantity of words	Quantity of word images	Quantity of PAW's	Quantity of characters
1	11	396	396	1548
2	10	360	720	1800
3	11	396	1188	2448
4	7	252	1008	2268
5	9	324	1620	3456
6	10	360	2160	4608
7	9	324	2268	4536
8	7	252	2016	3852
9	10	360	3240	6228
Total	84	3024	14616	30744

As evident from Table 3, the distribution of quantity of word images is generally uniform for this database which in not found in other word-based databases like IFN/ENIT (as pointed out earlier). Further, owing to the even distribution of word-images, the entire database can be conveniently used for classifiers like Support Vector Machines (SVMs), k- Nearest Neighbor (k-NN), etc. which require equal distribution of the training/sample sub-sets.

3. EXPERIMENTATION

A random sub-set of 20 words (taking all 3 instances) from 12 different writers, belonging to the KU-database was tested using a Letter Based HMM Classifier. The Training Set consisted of 20x12x2=480 word images, and the Testing Set consisted of 20x12x1=240 word images. The number of PAWs chosen were 1, 2, 3 and 5, respectively. The HMM Classifier was implemented using HTK (Young 2006, Maqqor, Halli et al. 2014).

3.1 Preprocessing

All, the word images in the training set were hand-segmented into their respective letter images. First, all letters were horizontally flipped. Then they were binarized by using grayscale thresholding. Next, the position of the letter image was located in the original word image by a binary pattern matching algorithm. The letter image was then repositioned against a blank background of same width as the segmented image and height equal to that of the cropped original binary word image, by vertical position adjustment. After this, the letter images were resized to $\{100\} \times \left\{\frac{500 \cdot r_s}{r_o}\right\}$ pixels, where r_o was the width of the original image (before resizing) and r_s was the width of the hand-segmented letter image (before resizing). This was done to keep equal size of each letter, both in resized original binary word image, as well as in the resized binary hand-segmented letter image. Then, the images were thinned using the Zhang-Suen Thinning Algorithm (Zhang and Suen 1984), after which they were dilated. For testing purposes, the words were binarized, cropped, resized to 100x500 pixels, thinned using above algorithm and then dilated. Finally, all the preprocessed images were divided into 6-pixel wide (non-overlapping) vertical frames for feature extraction. The features (Likforman-Sulem et al. 2012) extracted per frame were:

Statistical Feature Set, S:

i) f_1 : Density of foreground pixels in frame.

ii) f_2 :Number of black/white transitions between two consecutive frame cells (from top to bottom):

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

where b(i) is the density level belonging to i^{th} cell. It is equal to 1 if the cell contains at least one foreground pixel, and is 0 otherwise.

iii) $f_3 to\ f_8$:Sums of foreground pixels in each of the six vertical columns in each frame, each foreground pixel being a 1 and each background pixel being a 0.

iv) $f_9\ to\ f_{14}$:Six concavity features calculated as: Let $N_{lu}, N_{ur}, N_{rd}, N_{dl}, N_v$ and N_h be the number of background pixels which have neighboring black pixels in the following directions: left-up, up-right, right-down, down-left, vertical and horizontal, respectively.

v) $f_{15} to\ f_{35}$:Twenty-one, percentile features. At every y-position, the number of black pixels from the top of the frame were computed. This function of y was later normalized from 0 to 100, equaling 0% to 100% of the total blackness of the frame. The y-axis was normalized from 0 to 1, equaling 0% to 100% of the total height of the frame. Next, the total blackness was divided into 20 percentiles, and the corresponding values of percentile height of the frame were the percentile features.

Zernike Feature Set, Z:

vi) First 15 Zernike Features for each frame.

Raw Pixel Feature Set, R:

vii) 660 Raw pixel values for each frame.

All features extracted were independent of baseline information. Baseline information was not considered.

3.2 Hidden Markov Models

Hidden Markov Models (HMMs) (Rabiner 1989, Likforman-Sulem, Al Hajj Mohammad et al. 2012, Ahmad and Fink 2015) are well suited for classification of Arabic Text because of the connected nature of the text. HMM systems stochastically model sequences of strokes of variable length and cope with nonlinear distortions. A Markov system is a chain having different states with stochastic identities, wherein the states at a particular time are influenced by states at an earlier time. In HMMs, the states remain hidden to the observer, and the outputs are clear, hence the name, Hidden Markov Model. An HMM can be described as a set of 3 parameters, as:

$$\lambda = (\pi, A, B) \tag{2}$$

where λ is the HMM defined by π, the initial distribution of states, A is the state transition matrix, and B is the confusion matrix. If N is the number of states in a model and M is the number of distinct observation symbols per state, the state transition matrix is calculated as follows:

$$A = \{a_{ij}\} \tag{3}$$

$$a_{ij} = P[q(t+1) = S_j \mid q(t) = S_i]; \quad 1 \le i, j \le N \tag{4}$$

Thus a_{ij} is the probability of i[th] state at time t reaching j[th] state at time (t+1).Here, $S = \{S_1, S_2, S_3, \dots, S_N\}$ represent all the states of the HMM. The observation for a symbol probability distribution viz. the confusion matrix, B, is calculated as:

$$B = \{b_j(k)\} \tag{5}$$

$$b_j(k) = P[v_k \; at \; t \mid q(t) = S_j]; \quad 1 \le j \le N \;\; \text{and} 1 \le k \le M \tag{6}$$

where individual symbols are denoted by the vector $V = \{v_1, v_2, \dots, v_M\}$. The confusion matrix is the probability of emission for symbol k at state j. Finally, the last parameter of the model is the vector of the initial state probabilities, **π**, and is calculated as:

$$\boldsymbol{\pi} = \{\pi_i\} \tag{7}$$

$$\pi_i = P\{q_1 = S_j\}, \quad 1 \le i \le N \tag{8}$$

where $\boldsymbol{\pi}$ is the vector which gives the probability of being in each state at initial time t=1.
Training and testing the HMMs are handled by HTK. Initialization however needs to be done manually. For different Letter HMMs (having states from N= 3 to 5), the state matrices need have the same order A_{NxN}. As an example, the initial state transition matrix for a 5 state Left to Right HMM, was defined as follows:

$$A_{5\times5} = \begin{bmatrix} 0.3 & 0.3 & 0.4 & 0.0 & 0.0 \\ 0.0 & 0.3 & 0.3 & 0.4 & 0.0 \\ 0.0 & 0.0 & 0.3 & 0.3 & 0.4 \\ 0.4 & 0.0 & 0.0 & 0.3 & 0.3 \\ 0.3 & 0.4 & 0.0 & 0.0 & 0.3 \end{bmatrix} \tag{9}$$

Each state has a self-transition and transitions to the next two states. Two exceptions are the starting and ending states. Left to right HMMs were used because images were already flipped horizontally. Figure 3 shows a 5 state HMM.

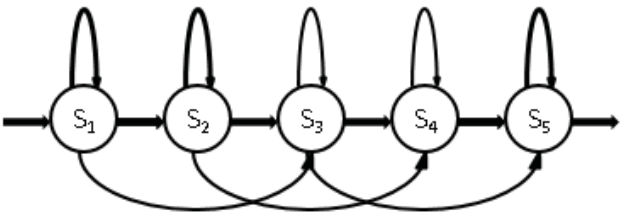

Figure 3. A 5 State HMM

3.3 Classification

The HMM Classifier was implemented using HTK. First, Letter HMMs were trained on the hand-segmented letter image features. For best results, a single HMM was used for letters of similar shape, e.g. 'ba' ب and 'ta' ت (after removing dots). This was extended to all 4 forms i.e. isolated, middle, beginning and end, wherever possible after removing dots and diacritics. This process reduced the number of letter HMMs used. Next, Word HMMs were formed from the Letter HMMs, and these were used to classify the whole word images of the testing set using their respective extracted features. Figure 4 shows the Proposed OCR System.

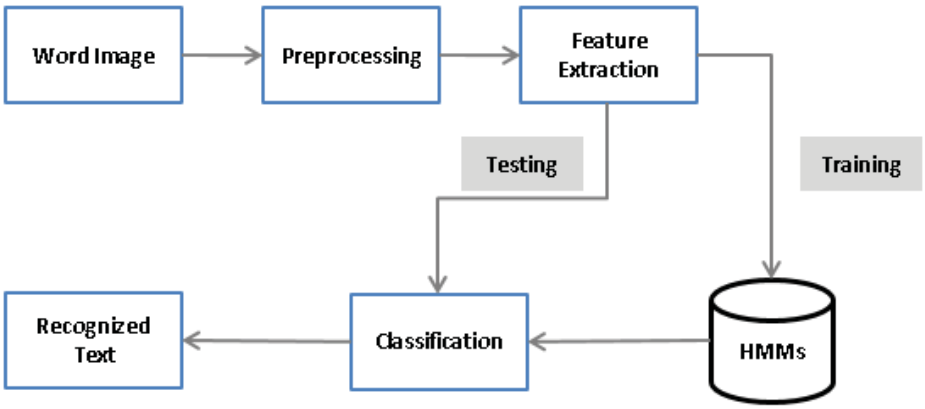

Figure 4. OCR System Used

3.4 Results

The dataset extracted from the KU-database was tested using the OCR System developed using the feature sets **S**, **Z** and **R**, and their combinations for benchmarking. The results are shown in Table 4.

Table 4: Results of Experimentation

Feature Set Used	Length of Feature Vector	Classification Accuracy (%)
R	660	74.17
S	35	75.83
Z	15	76.67
RS	660+35=695	78.75
RZ	660+15=675	72.92
SZ	35+15=50	**82.50**

From Table 4, it can be observed that when individual feature sets are used, Set Z has highest accuracy. When the feature sets are combined, SZ has highest accuracy, and this is overall highest accuracy. Also, it was observed that combining Set Z with Set R (with or without Set S), decreases the accuracy. Thus, the robustness of Zernike Features over Raw Pixel Features is observed. Also, it is observed that combining Zernike Features with Statistical Features, gives highest classification accuracy of the OCR System for the dataset used.

4. DISCUSSION AND CONCLUSION

In this paper, the KU-database has been introduced as a new database for handwritten Arabic words. The KU-database is publicly available for the purpose of research. 12 different writers filled 60 forms thrice, each having 84 words, with PAW variation from 1 to 9. 3024 word images were extracted from the forms. All words come with Ground Truths. With the KU-database it is possible to develop and test Arabic handwritten word recognition systems or parts of them. Using a portion of the database, benchmarking of some important features (used in HMM Based OCR Systems) was done. Various features and their combinations were evaluated for their performance using the implemented OCR System. Zernike Features demonstrated robustness against other features. Other researchers are kindly invited to test their systems with the KU-database.

5. REFERENCES

ERIM Arabic document database. S. G. Schlosser. Environmental Research Institue of Michigan.

Ahmad, I., Fink, G. A., 2015, Multi-stage HMM based Arabic text recognition with rescoring. *Proceedings of 13th International Conference on Document Analysis and Recognition (ICDAR), 2015.*

Al-Badr, B., Mahmoud, S. A., 1995, Survey and bibliography of Arabic optical text recognition. *Signal Processing.* **41 1,** 49-77.

Al-Jamimi, H., Mahmoud, S., 2010. Arabic Character Recognition Using Gabor Filters. *Innovations and Advances in Computer Sciences and Engineering,* Springer, Netherlands**:** 113-118.

Al-Ma'adeed, S., Elliman, D., Higgins, C. A., 2002. A data base for Arabic handwritten text recognition research. *Proceedings of 8th International Workshop on Frontiers in Handwriting Recognition, 2002.*

Amara, M., Zidi, K., Zidi, S., Ghedira, K., 2014. Arabic Character Recognition Based M-SVM: Review. *Advanced Machine Learning Technologies and Applications*. Hassanien, A., Tolba, M., Taher Azar, A., Springer International Publishing. **488,** 18-25.

Cheriet, M., 2008, Visual Recognition of Arabic Handwriting: Challenges and New Directions. *Arabic and Chinese Handwriting Recognition.* Doermann, D., Jaeger, S., Springer Berlin Heidelberg, **4768,** 1-21.

Cheriet, M., Moghaddam, R., 2012, A Robust Word Spotting System for Historical Arabic Manuscripts. *Guide to OCR for Arabic Scripts*, Märgner, V., El Abed, H.,Springer London, 453-484.

Djeddi, C., Al-Maadeed, S., Gattal, A., Siddiqi, I., Souici-Meslati, L., El Abed, H.,. ICDAR2015 competition on Multi-script Writer Identification and Gender Classification using QUWI Database. *Proceedings of 13th International Conference on Document Analysis and Recognition (ICDAR) 2015.*

El Abed, H., Margner, V., 2007, The IFN/ENIT-database - a tool to develop Arabic handwriting recognition systems. *Proceedings of 9th International Symposium on Signal Processing and Its Applications (ISSPA) 2007.*

El Abed, H., Margner, V., 2008. Arabic text recognition systems - state of the art and future trends. *Proceedings of International Conference on Innovations in Information Technology(IIT) 2008.*

Fujisawa, H., 2008, How to Deal with Uncertainty and Variability: Experience and Solutions. *Arabic and Chinese Handwriting Recognition,* Doermann, D., Jaeger, S., Springer Berlin Heidelberg. **4768**, 129-151.

Graves, A., 2012, Offline Arabic Handwriting Recognition with Multidimensional Recurrent Neural Networks. *Guide to OCR for Arabic Scripts,* Märgner, V., El Abed, H., Springer London, 297-313.

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

Kharma, N., Ahmed, M., Ward, R., 1999, A new comprehensive database of handwritten Arabic words, numbers, and signatures used for OCR testing. *Proceedings of IEEE Canadian Conference on.Electrical and Computer Engineering 1999.*

Knerr, S., Augustin, E., Baret, O., Price, D., 1998, Hidden Markov Model Based Word Recognition and Its Application to Legal Amount Reading on French Checks. *Comput. Vis. Image Underst,* **70 3**, 404-419.

Likforman-Sulem, L., Al Hajj Mohammad, R., Mokbel, C., Menasri, F., Bianne-Bernard, A.L., Kermorvant, C., 2012, Features for HMM-Based Arabic Handwritten Word Recognition Systems. *Guide to OCR for Arabic Scripts,* Märgner, V., El Abed, H., Springer London, 123-143.

Lorigo, L. M., Govindaraju, V., 2006, Offline Arabic handwriting recognition: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28 5**, 712-724.

Maadeed, S. A., Ayouby, W., Hassaïne A., Aljaam, J.M., 2012, QUWI: An Arabic and English Handwriting Dataset for Offline Writer Identification. *Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition*, IEEE Computer Society, 746-751.

Mahjoub, M. A., Ghanmy, N., Jayech, K., Miled, I., 2013, Multiple models of Bayesian networks applied to offline recognition of Arabic handwritten city names. *International Journal of Imaging and Robotics*, **9 1**, 84-105.

Mahmoud, S., Al-Khatib, W., 2011. Recognition of Arabic Indian bank check digits using log-gabor filters. Applied Intelligence, **35 3**, 445-456.

Maqqor, A., Halli, A., Satori, K., Tairi, H., 2014, Using HMM Toolkit HTK for recognition of arabic manuscripts characters. *Proceedings of International Conference on Multimedia Computing and Systems (ICMCS) 2014.*

Märgner, V., El Abed, H., 2008, Databases and Competitions: Strategies to Improve Arabic Recognition Systems. *Arabic and Chinese Handwriting Recognition*, Doermann, D., Jaeger, S., Springer Berlin Heidelberg. **4768**, 82-103.

Nazif, A., 1975, A system for the recognition of the printed arabic characters, Cairo University.

Parvez, M. T., Mahmoud, S.A., 2013, Offline arabic handwritten text recognition: A Survey. *ACM Comput. Surv*. **45 2**, 1-35.

Pechwitz, M., El Abed, H., Märgner, V., 2012, Handwritten Arabic Word Recognition Using the IFN/ENIT-database. *Guide to OCR for Arabic Scripts,* Märgner, V., El Abed, H., Springer London, 169-213.

Pirsiavash, H., Mehran, R., Razzazi, F., 2005, A Robust Free Size OCR for Omni-Font Persian/Arabic Printed Document Using Combined MLP/SVM. *Progress in Pattern Recognition, Image Analysis and Applications,* Sanfeliu, A., Cortés, M., Springer Berlin Heidelberg, **3773**, 601-610.

Rabiner, L., 1989, A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, **77 2**, 257-286.

Rashad, M., Semary, N., 2014, Isolated Printed Arabic Character Recognition Using KNN and Random Forest Tree Classifiers. *Advanced Machine Learning Technologies and Applications,* Hassanien, A., Tolba, M., Taher Azar, A., Springer International Publishing, **488**, 11-17.

Tagougui, N., Kherallah, M., Alimi, A., 2013, Online Arabic handwriting recognition: a survey. *International Journal on Document Analysis and Recognition (IJDAR),* **16 3**, 209-226.

Young, S., 2006, The HTK Book V3.4. Cambridge, Cambridge University Press.

Zhang, T. Y., Suen, C.Y., 1984, A fast parallel algorithm for thinning digital patterns. *Commun. ACM*, **27 3**, 236-239.