Automatic Recognition of Handwritten Arabic Text: A Survey

Usman Saeed

Faculty of Computing and Information Technology North Jeddah, King Abdulaziz University, Jeddah, Saudi Arabia. email: usaeed@kau.edu.sa

Abstract: Arabic script is the third most widely used writing system after Latin and Chinese, but research in Arabic Optical Character Recognition (OCR) is still nascent in comparison to Latin script. Arabic script is inherently cursive in nature, therefore techniques developed for other scripts are generally inappropriate for Arabic. In this paper we present recent progress in the field of Handwritten Arabic Text Recognition. We present the recent techniques developed, classified as steps of a general OCR system. Beginning with image acquisition and preprocessing techniques for Handwritten Arabic Text Recognition., we next present the methods for segmentation, feature extraction and classification. In the end we conclude with a discussion of future directions in Handwritten Arabic Text Recognition.

[Usman Saeed. **Automatic Recognition of Handwritten Arabic Text: A Survey.** *Life Sci J* 2014;11(3s):232-235]. (ISSN:1097-8135). http://www.lifesciencesite.com. 35

Keywords: Image Processing, Optical Character Recognition (OCR), Arabic OCR.

1. Introduction

Optical Character Recognition (OCR) combines image processing and artificial intelligence to read and understand human language. OCR systems initially focused on printed text due to its ease of recognition, but over the last few decades Handwritten Character Recognition (HCR) has become feasible. The main cause of lower recognition rates in HCR as compared to printed text is usually attributed to large variation of individual writing styles. The cursive nature and overlapping of characters in most scripts further complicates the segmentation and recognition.

The first step (fig. 1) of any OCR system is the image acquisition, where a digital image of the text is obtained, this could be done offline using a scanner or online by a digital pen/stylus. The next phase is the preprocessing of acquired images. This phase consists of several techniques used to improve the quality of images for future processing. It involves noise removal, binarization to convert gray scale or colored images to black and white image, morphological operations like opening, closing, thinning, hole filling etc. may be applied to enhance visibility and structural information of text. Slant/Skew correction may also be applied if the documents are not properly aligned. In the next phase the documents are segmented to various levels depending on the requirement of the application. Segmentation may be at blocks, lines, words or characters level. In feature extraction step we extract discriminant characteristics of the segmented components. These characteristics are then used to train and test a classifier. The last step is post processing, which is not compulsory but can improve the accuracy of recognition. Syntax analysis, semantic analysis may be applied to verify the recognized text.

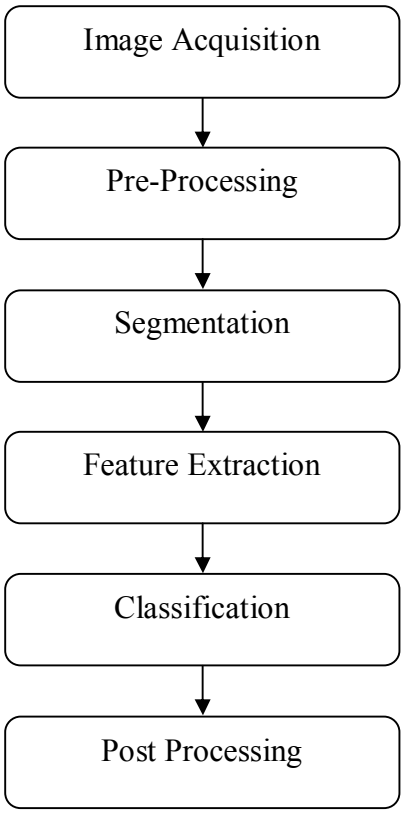

Figure. 1: Typical OCR Steps

About 300 million people in the world use Arabic as their first language and Arabic script is the second most widespread script in the world; it is used not only for Arabic but also for Persian, Urdu, Pashto and other languages. Arabic is a cursive

script with 28 characters (fig. 2). It is written from right to left, but numbers are written left to right. Its letters change their shape depending on their position in the word. A character can be represented with a vowel or diacritic mark written over or under it. Diacritics are signs that represent short vowels or other sounds, such as syllable endings. Several Arabic alphabet letters share the same shape, and are differentiated only in terms of the number and placement of dots on the letters. All these characteristics influence the processing and recognition of Arabic script, thus a simple adaptation of techniques developed for Latin character will not suffice. The two most significant steps in Arabic OCR are segmentation and feature extraction. Segmentation is especially difficult due to the cursive nature and overlapping characters in Arabic script.

2. Survey

In this section we will present the previous works in Arabic Handwritten Character Recognition (HCR). Arabic HCR began in 1975 (Nazif, 1975) and has exploded in the recent years. A comprehensive survey for the period 1975-2005 is available by (Parvez, 2013) and for the period 2005-2011 by (Lorigo, 2006). Therefore the focus of our state of the art will be the recent most period 2011-2013.We shall present the works classified as the steps involved in a standard OCR system.

2.1 Database

Majority of the studies (Leila, 2011; Azizi, 2011; Nemmour, 2011; Al Nuzaili, 2012; Al Khateeb, 2011; Parvez, 2013; Rothacker, 2012) that we came across during our literature review were based on IFN/ENIT database (Pechwitz, 2002). It consists of 26459 binary images of 937 handwritten Tunisian town/village names by 411 writers. Others (Cao, 2012) have sufficed by using self-generated database.

2.2 Preprocessing

Handwritten text in comparison to printed text is more susceptible to noise and normalization issues. Therefore pre-processing the images normally yield better results. These tasks commonly include noise removal, baseline detection, skew/slant detection and correction.

The baseline is a virtual line on which the characters of Arabic text are aligned. One of the commonly used method is the horizontal projection of pixel intensities (Azizi, 2011). The pixel intensities are combined along the horizontal axis and the maxima is selected as the baseline. Some have used more complex methods, such as (Parvez, 2013) have proposed to adapt the expectation-maximization (EM) algorithm for baseline detection. (Parvez, 2013) have also proposed a method to calculate Near-Vertical Strokes (NVSs) by multi-direction Prewitt filters to estimate the slant angle.

Name	Isolated	Initial	Medial	Final
alif	ا			ـا
baa	ب	بـ	ـبـ	ـب
taa	ت	تـ	ـتـ	ـت
thaa	ث	ثـ	ـثـ	ـث
jiim	ج	جـ	ـجـ	ـج
haa	ح	حـ	ـحـ	ـح
khaa	خ	خـ	ـخـ	ـخ
daal	د			ـد
dhaal	ذ			ـذ
raa	ر			ـر
zaay	ز			ـز
siin	س	سـ	ـسـ	ـس
shiin	ش	شـ	ـشـ	ـش
saad	ص	صـ	ـصـ	ـص
daad	ض	ضـ	ـضـ	ـض
taa	ط	طـ	ـطـ	ـط
dhaa	ظ	ظـ	ـظـ	ـظ
Ayn	ع	عـ	ـعـ	ـع
ghayn	غ	غـ	ـغـ	ـغ
faa	ف	فـ	ـفـ	ـف
qaaf	ق	قـ	ـقـ	ـق
kaaf	ك	كـ	ـكـ	ـك
laam	ل	لـ	ـلـ	ـل
miim	م	مـ	ـمـ	ـم
nuun	ن	نـ	ـنـ	ـن
haa	ه	هـ	ـهـ	ـه
waaw	و			ـو
yaa	ي	يـ	ـيـ	ـي

Figure. 2: Arabic characters and their forms

2.3 Segmentation

Segmentation of handwritten Arabic is a non-trivial problem, due to several complexities. One is the inherent nature of the Arabic script which is cursive and character overlapping is common. Second is the high degree of variation in Arabic writing styles of individuals. Classically the

segmentation techniques developed for Latin script were extended for use with Arabic script, these include; projection based and contour based methods (Parvez, 2013). Recently most researchers have started to skip the segmentation step due to its complexity and the grave number of errors that are introduced. They instead take a more holistic approach in which lines are segmented into words and complete word are used for feature extraction (Leila, 2011; Azizi, 2011; Nemmour, 2011; Al Nuzaili, 2012; Al Khateeb, 2011; Rothacker, 2012; Cao, 2012). Another recent approach has been to use Parts of Arabic Words (PAWs) (Porwal, 2012), where a sub section of the word, which is more recognizable than a character is treated as a unit of segmentation.

2.4 Feature Extraction

Features are distinct characteristics of a character/word that enables us to recognize it. There are numerous types of features but they are broadly classified into two types: structural and statistical features. Structural features represent the structure or shape, such as loops, branch points, endpoints, dots, etc. Statistical features are numerical measures of pixel intensities computed over images. They include pixel densities, directions, moments, Fourier descriptors, etc.

2.4.1 Structural Features

Simple shape based features can be easily extracted from images and that yield good results. (Azizi, 2011) have created a feature vector consisting of a several shape based features such as number of connected components, number of descenders/ascenders, dots, loop, etc. (Charfi, 2012) have proposed to extract GSC features by combining Gradient, Structural and Concavity features. This combination of features has shown to capture the local, intermediate and global information. (Parvez, 2013) have devised a fuzzy polygonal approximation with constrained collinear-points suppression of Arabic text contours, which shows tolerance to variations of the handwritten text of different writers. In (Charfi, 2012), first characteristic points such as starting, ending, branch points of a word are detected. Then a Beta-elliptic model is applied to extract features that represent the static and kinematic field for handwriting.

2.4.2 Statistical Features

One of the most common features based on pixel values are image moments. (Leila, 2011) have utilized Hu and Zernike moments extracted from images to be used as features. Transformation such as Fourier and wavelet have been used to extract features from images for some time now, but a new addition has been Ridgelet coefficients proposed by (Nemmour, 2011). The Radon transform is first computed according to several angular directions, then one-dimensional wavelet transform is applied to yield the Ridgelet coefficients. Angular and distance span method proposed by (Al Nuzaili, 2012) divide word images using predefined grids (angular, concentric circles) and then count the number of foreground pixels as feature.

Features extracted by sliding a window on image are normally used in conjunction with HMM classifiers. A sliding window of fixed size is moved on the image and features which normally consist of pixel values or their average (Al Khateeb, 2011) are extracted. (Cao, 2012) have used the same sliding window method but have extracted more complex feature which include; image intensity percentile, local angle and correlation, frame energy, gradient, concavity, and Gabor coefficients.

A much recent approach has been taken by (Rothacker, 2012), they have proposed a bag of words method for Arabic script. First points are selected by Harris corner detector, then a 128-dimensional SIFT descriptor is used to represent the corner points. These are then clustered and quantized to create a vocabulary of features.

2.5 Classification

Nearest neighbour classifier is one of the simplest and widely used classifier for image based data and has been applied to Arabic HCR by (Azizi, 2011). Support Vector Machines (SVM) initially developed as a two class classifiers have been extended to handle multi-class problems and applied to Arabic by (Azizi, 2011; Nemmour, 2011) using a Radial Basis Function kernel. Artificial Neural Networks (ANN) are another well studied classification method, and (Leila, 2011) have combined Adaptive Resonance Theory Networks with Radial Basis Function Network using majority vote, max-rule, and sum-rules. (Porwal, 2012) have used an SVM to create an ensemble of biased learners, where multiple classifiers are trained for recognizing specific classes, and later their collective opinion is used for the primary classification task.

The Hidden Markov Models (HMM) are statistical models which were effectively used in speech recognition. HMMs were extended for off-line HCR by (Al Khateeb, 2011; Rothacker, 2012; Cao, 2012). An approach quite close to HMMs is the Dynamic Bayesienne Networks, which have been applied to Arabic by (Al Khateeb, 2011). (Parvez, 2013) have proposed an integrated segmentation, feature extraction and classification approach based on fuzzy polygon matching algorithm. (Charfi, 2012) have extracted Beta-elliptic model based features and then apply a graph matching algorithm to compute distance between two trajectories. A Nearest Neighbor algorithm is used to associate the

nearest points between graph trajectories and Euclidian distance is then calculated in order to evaluate the graphic similarity.

3. Conclusions

In this paper we have presented a survey of the recent techniques and methods developed for Handwritten Arabic OCR. We have analysed the developed techniques as applied to the various steps of an OCR system. One of the major limiting factor of research in Handwritten Arabic text recognition is the absence of a standardized database that can enable meaningful comparison of results amongst various approaches. Secondly segmentation of Arabic is quite difficult due to its cursive nature and require considerable effort to improve the current state of the art. Lastly majority of the feature extraction and classification technique have been borrowed from Latin script recognition and are not adapted for Arabic script. We believe it will worthwhile to further investigate these issue and develop algorithms specifically taking into account the nature of Arabic script.

References

1. Nazif, A.: 'A system for the recognition of the printed Arabic characters'. Master's Thesis, Faculty of Engineering, Cairo University, 1975
2. Parvez, M.T., Sabri, A.M.: 'Offline arabic handwritten text recognition: A Survey', ACM Comput. Surv., 2013, 45, (2), pp. 1-35
3. Lorigo, L.M., Govindaraju, V.: 'Offline Arabic handwriting recognition: a survey', IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28, (5), pp. 712-724
4. Leila, C., Maâmar, K., Salim, C.: 'Combining neural networks for Arabic handwriting recognition'. Proceedings of 10th International Symposium on Programming and Systems (ISPS), Algiers, Algeria, 2011, pp. 74-79
5. Azizi, N., Farah, N., Sellami, M.: 'Ensemble classifier construction for Arabic handwritten recognition'. Proceedings of 7th International Workshop on Systems, Signal Processing and their Applications, Algiers, Algeria, 2011, pp. 271-274
6. Nemmour, H., Chibani, Y.: 'Handwritten Arabic word recognition based on Ridgelet transform and support vector machines'. Proceedings of International Conference on High Performance Computing and Simulation, Istanbul, Turkey, 2011, pp. 357-361
7. Al Nuzaili, Q., Mohamad, D., Ismail, N.A., Khalil, M.S.: 'Feature extraction in holistic approach for Arabic handwriting recognition system: A preliminary study'. Proceedings of IEEE 8th International Colloquium on Signal Processing and its Applications, Malacca, Malaysia, 2012, pp. 335-340
8. AlKhateeb, J.H., Pauplin, O., Ren, J., Jiang, J.: 'Performance of hidden Markov model and dynamic Bayesian network classifiers on handwritten Arabic word recognition', Know.-Based Syst, 2011, 24, (5), pp. 680-688
9. Parvez, M.T., Sabri, A.M., 'Arabic handwriting recognition using structural and syntactic pattern attributes', Pattern Recogn., 2013, 46, (1), pp. 141-154
10. Rothacker, L., Vajda, S., Fink, G.A.: 'Bag-of-Features Representations for Offline Handwriting Recognition Applied to Arabic Script'. Proceedings of International Conference on Frontiers in Handwriting Recognition, Bari, Italy, 2012, pp.149-154
11. Pechwitz, M., Maddouri, S.S., Märgner, V., *et al.*: 'IFN/ENIT-DATABASE OF HANDWRITTEN ARABIC WORDS'. Proceedings of the 7th Colloque International Francophone sur l'Ecrit et le Document, Hammamet, Tunis, 2002, pp. 129-136
12. Cao, H., Chen, J., Devlin, J., Prasad, R., Natarajan, P.: 'Document recognition and translation system for unconstrained Arabic documents'. Proceedings of 21st International Conference on Pattern Recognition, Tsukuba Science City, Japan, 2012, pp.318-321
13. Porwal, U., Shivram, A., Ramaiah, C., Govindaraju, V.: 'Ensemble of Biased Learners for Offline Arabic Handwriting Recognition'. Proceedings of 10th IAPR International Workshop on Document Analysis Systems, Queensland, Australia, 2012, pp. 322-326
14. Charfi, M., Kherallah, M., Baati, A.E., Alimi, A.M.: 'A New Approach for Arabic Handwritten Postal Addresses Recognition', International Journal of Advanced Computer Science and Applications, 2012, 3,(3), pp. 1-7

3/5/2014