http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (9): 1754-1760, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.1754.1760

Off-Line Arabic Words Classification using Multi-Set Features

[1]A.M. Al Tameemi, [1]L. Zheng and [2]M. Khalifa
[1]School of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
[2]University of Science and Technology Beijing, Information Engineering School, 100083 Beijing, China

Abstract: Optical Arabic text recognition is receiving renewed extensive research after the success in optical text recognition in many languages. In this research work a set of feature extraction methods was used to get structural and geometrical representations of Arabic words. The system was focused on employing Support Vector Machines (SVMs) as a pattern recognition tools. In this study, we assumed each shape of an Arabic word as a separate class bypassing Arabic word segmentation for characters. The proposed system was composed of four phases. The first phase performed image binarization where a word image was converted into white with black background while the next phase involved noise removal. The third phase was features extraction which identifies a set of features. The extracted features were consist of twenty sliding windows of vertical slides summation, four local maxima points in the Vertical Projection with the center of gravity, a number of connected components, positions of corners detected with end points in the word image and the mean value of the word image. The last phase was classification phase where the multi class SVMs was used, by applying a one against-all technique. The proposed method was tested using a different datasets of Arabic words which achieved high average recognition rate.

Key words: Sliding window, curvature, local maxima, connected component, support vector machine

INTRODUCTION

After the success of optical text recognition in many languages, optical Arabic text recognition is receiving renewed extensive researches. Arabic text recognition has not been researched as thoroughly as Latin, Japanese, or Chinese. Recently, it has been receiving a renewed interest not only from Arabic native researchers but also from non-Arabic natives (Zheng *et al*., 2004; Femiani *et al*., 2005). This has resulted in the improvement of the state of the art in the Arabic text recognition.

Most of the existing Arabic script recognition systems are based on characters and others based on words (Khorsheed, 2002; Alma'adeed *et al*., 2004; Alshalabi, 2005; Lotfi *et al*., 2006; Tlili-Guiassa and Tayeb, 2006; AlKhateeb *et al*., 2011). The problem of character-based-recognition system is the stage of segmenting a word to characters that gives many errors. On the other hand, the systems based on words considered as the task of recognition is similar to the general object then applied object recognition algorithms for feature extraction. Rodriguez and Perronnin (2008) have obtained a set of features by applying the Scale Invariant Feature Transform (SIFT), key point descriptor. On the other side Zhang *et al*. (2009) has used SIFT descriptor to recognize Chinese characters. While AlKhateeb *et al*. (2008) has used Discrete Cosine Transform (DCT) features for recognizing handwritten Arabic scripts.

In this study, we consider a whole word as a separate class. By extracting structural and statistical features, each subset of these features can approximate the target concept for building a high-accuracy Arabic word recognition system.

Multi-methods for extracting different types of features and the optimized Support Vector Machine (SVM) have been developed. Right here, six types of features are extracted:

- Summation of sliding windows
- The maximum and minimum four positions in Vertical Projection points
- Compute center of gravity of the word image as a feature
- Number of connected component
- Features from edge points and the end points
- Feature from statistical operation Mean

FEATURE EXTRACTION

Feature extraction is critical in any recognition system. The performance of the classifier depends directly on the features which have been extracted. In this study,

Corresponding Author: Ammar M. Al Tameemi, School of Computer Science and Technology, Harbin Engineering University, 145 Nantong Street, Nangang District, Harbin, Heilongjiang Province 150001, China Tel: +8613945131071

both structural and statistical features of an isolated Arabic word have been used to recognize each word as a separate object.

Features from sliding windows: Sliding windows are utilized to extract statistical features. By applying the sliding windows to an image word with size LW×WW, the image is divided into N vertical strips, in other words, the length of each slide strip window is the same as the word length (LW) and the width of the slide strip window is (WW/N). To generate this type of features, two main steps are needed:

Step 1: Pre-processing. A word image is binarazed into white with black background to get a binarazation image and then canny edge detector is used to extract the edge from the binary image. Finally, we clip the white word images with black background

Step 2: Let the width of the clipped image be multiple of N. In this study, we consider N = 20 (as shown in Fig. 1)

Step 3: Divide the image into twenty vertical slide strips, then extract feature from each strip by using the summation of all pixels in each strip as the first type of features

Features from the vertical projection: Vertical Projection of a word image is applied as a graphical representation, showing a visual impression of distribution pixels in the Arabic word body. Mathematically, the Vertical Projection of an image can be computed by the following equation:

$$vproj[j] = \sum_{i=0}^{m-1} I[i,j],\ m = \text{number of rows} \quad (1)$$

After obtaining the vertical projection of a word image, we divide it into four equal vertical strips. Then, the maximum and minimum values of each strip are extracted as the second type of features, thus, this type of features consists of eight elements in the feature vector (Fig. 2).

Features from the center of gravity: Another part in our features vector is the gravity of the word image. The vertical and horizontal centers of gravity are determined by the following Eq:

$$C_x = \frac{M(1,0)}{M(0,0)} \quad (2)$$

$$C_y = \frac{M(0,1)}{M(0,0)} \quad (3)$$

where, C_x and C_y are the horizontal and vertical centers of gravity, respectively, M_{pq} is the geometrical moments of rank p+q. The above two centers are considered as the third type of features.

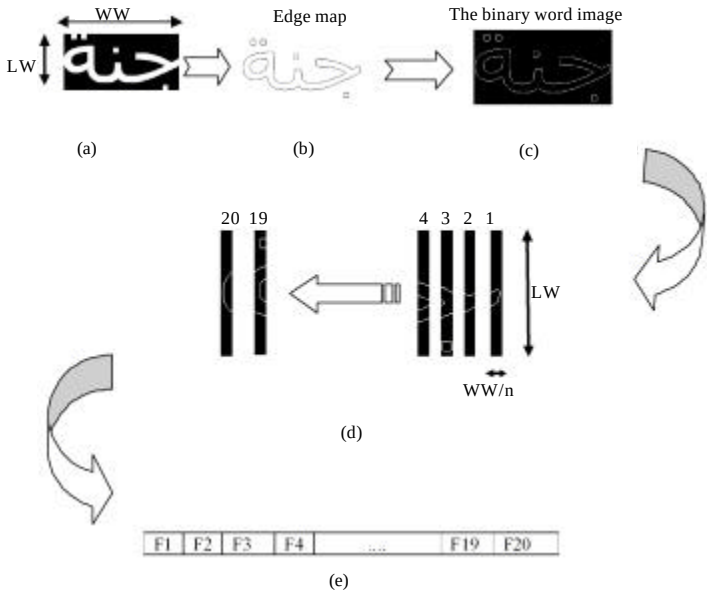

Fig. 1(a-e): The main steps of getting the first twenty features, by showing the area used for feature extraction and sliding window. (a) Binarization image, (b) Edge detection, (c) Binarization edge image, (d) Divided edge image to 20 slide and (e) Get sammation of each slide as a feature

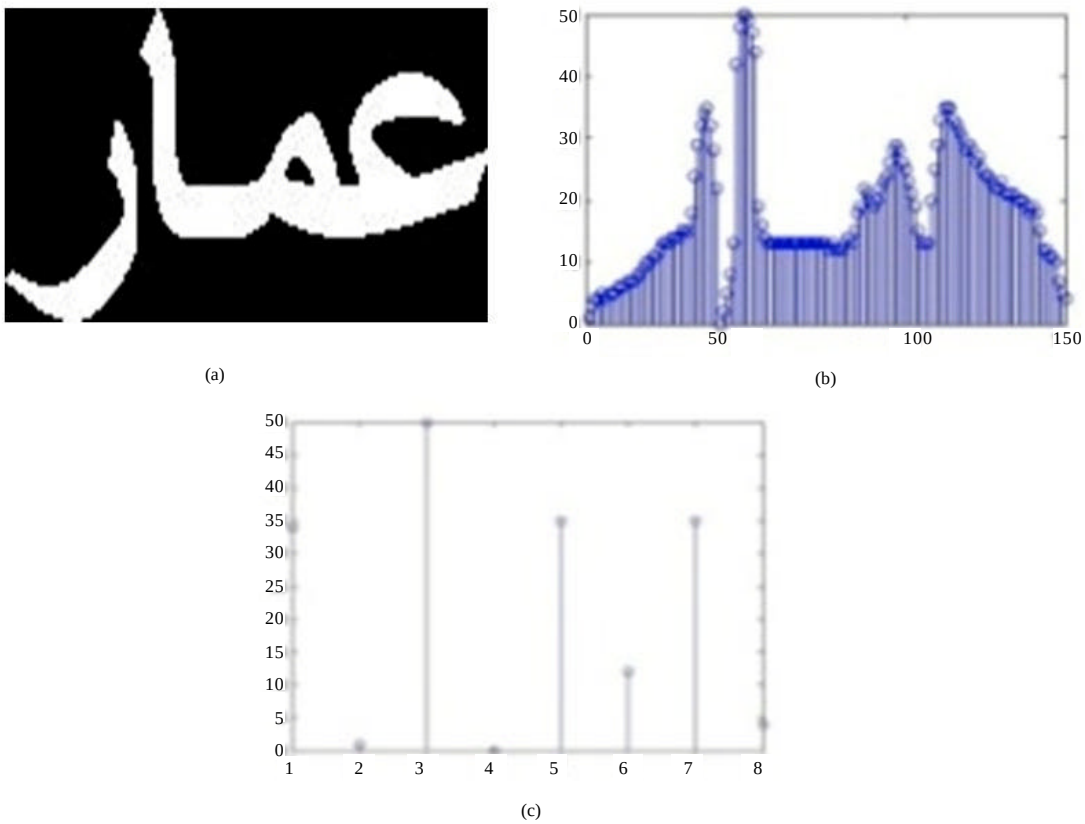

Fig. 2 (a-c): The features from the Vertical Projection, (a) Word image, (b) Vertical Projection and (c) max and min for each quarter, as example four max and min for the word (a) = [34 1 50 0 35 12 35 4]

Number of connected components: The number of connected components is an important topological invariant of a graph. We use the number of connected components as the fourth type of features to recognize Arabic words, since it plays a key role in graph toughness. In Arabic word image, the number of connected components is invariant with respect to the size and font. Figure 3 shows the number of connected components of four Arabic word images with each connected component has separated color.

The number of dots above or below the characters is an important feature to identify different Arabic words. The number of connected components can reveal this important characteristic. For example, (أحمد means: "Ahmed", NC = 2, number of sub word = 2, number of dot=0), (أحمد means: "Put down". NC = 3, number of sub word = 2, number of dot = 1).

Features from corner detectors: The complexity of Arabic text in its construction leads us to use another type of features "these features are gotten from the corner detectors". In this study we used the same detector as He and Yung (2008), because it detects both fine and course features accurately at low computational cost and performs very well in detectors on planar curves and produce relatively low localization errors.

The Canny edge detector is used to detect the edges of an Arabic word image then, all true corners of the word curvature are computed and extracted at a low scale for each contour considering all of the curvature local maxima as corner candidates. Then, eliminate rounded and false corners. The number of corners that we get from this detector is different from one word to another. Thus, we decided to select the first ten detected corners that may cover all or part of the corners in the word image.

Furthermore, we select the x-axis position of the detected corners to compute the fifth type of features, as shown in Eq. 4:

$$f_i = \frac{\text{x position of detected corner}}{\text{Length of the word image (LW)}} \times 10 \quad (4)$$

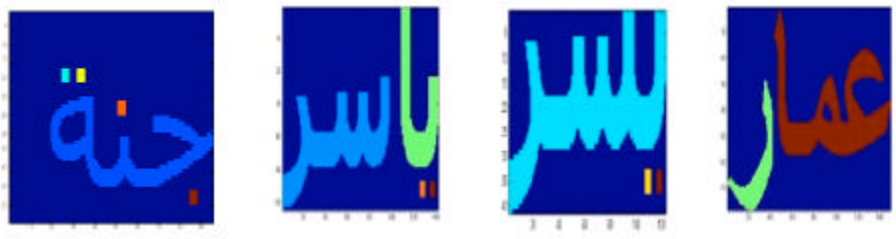

Fig. 3: Illustration number of connected component (NC) for some Arabic words give to each connected component different color

Table 1: Illustration the structure of features vector

f1 … f20	f21 … f28	f29 f30	f31	f32 …. f42	f43
Divided edge image into 20 slides windows	Four maximum and minimum values of Vertical Projection of the word image	Center of Gravity	Number of connected components	A position list of detected curvature parameter in the input image and end points in the marked image	Mean value

where, fi = 32…42 represents 10 elements in the feature vector.

Each of the first 10 "x" positions of detected corner is divided by the length of the word and then multiplied by ten to make it in the same range to other types of features.

Mean value of the word image: A statistical feature mean value is applied on Arabic word image as the last type of features which is given by:

$$\text{Mean} = \frac{1}{n}\sum_{i=1}^{n} I[x,y] \quad (5)$$

where, n is the number of pixels in the image.

The structure of features vector that we gotten from features extraction phase (Table 1).

Arabic word recognition based on the new features and SVMs: A typical recognition system consists of preprocessing, feature extraction, classification and recognition. In the preprocessing stage, 3 by 3 median filter are used on the noisy Arabic word image, then the gray scale image are converted into a binary image using thresholding technique.

An important stage in our system is the feature extraction which has been discussed earlier. Forty three features listed in Table 1 are extracted. The twenty features (f1-f20) are from sliding windows, eight features (f21-f28) are from the vertical projection and f29 to f42 are geometrical features while the last feature, f43 is statistical feature.

The last stage of the proposed system is the design of the classifier. In this paper we use an optimized Support Vector Machine (SVM) to classify each input Arabic word. The details about using SVMs in pattern recognition are found by Hsu and Lin (2002) and Izabatene *et al.* (2010).

Here, the one-against-all (1-v-all) SVM is used for the classifier as Platt *et al.* (2000), Hsu and Lin (2002) and Rifkin and Klautau (2004).

EXPERIMENTAL RESULTS

The proposed system has been tested on the printed Arabic text using public database (PATS-A01) that consists of 2766 text line images (Al-Muhtaseb *et al.*, 2008). This database was selected from two standard classic Arabic books and line images available in eight fonts: Arial, Tahoma, Akhbar, Thuluth, Naskh, Simplified Arabic Andalus and Traditional Arabic. We have selected several lines randomly from this database, each line is written in five different fonts (Andalus, Traditional Arabic, Simplified Arabic, Arial and Tahoma). We have used Vertical Projection profile to decompose a line into its words.

Moreover, some words were selected from IFN/ENIT-database. This database contains material for training and testing of Arabic handwriting recognition software, it consists of 26459 of handwritten samples of city names from 411 writers (Pechwitz *et al.*, 2002).

Group 1 contains three datasets of printed Arabic samples that helps us to test our features on the selected fonts together.

Dataset 1 contains 330 Arabic word images written by the selected five fonts, each word class is written in the same size and the same resolution. It was recognized 306 samples and produced recognition rate 92.727%.

Table 2: Recognition rates (%) of implemented SVMs model in all Datasets

Group name	Dataset named	Total sample	Approx. recognized samples	Average recognition rate (%)
Group 1 (printed)	Dataset 1	330	306	92.727
	Dataset 2	1650	1464	88.727
	Dataset 3	1650	1632	98.909
	Total	3630	3402	93.719
Group 2 (printed)	Dataset 1	330	322	97.576
	Dataset 2	330	323	97.879
	Dataset 3	330	318	96.364
	Dataset 4	330	324	98.182
	Dataset 5	330	318	96.364
	Total	1650	1605	97.273
	Dataset 6	330	327	99.091
	Dataset 7	330	325	98.485
	Dataset 8	330	323	97.879
	Dataset 9	330	329	99.697
	Dataset 10	330	324	98.182
	Total	1650	1628	98.667
Group 3 (handwrittrn)	Dataset	216	180	83.333

Table 3: Recognition rates (%) of implemented SVMs model on each selected fonts

Font name	Recognition rate of Arabic words have different size and different resolution (%)	Recognition rate of Arabic words have same size and different resolution (%)	Average of recognition rate (%)
Andalus	97.576	99.091	98.334
Traditional Arabic	97.879	98.485	98.182
Simplified Arabic	96.364	97.879	97.122
Arial	98.182	99.697	98.940
Tahoma	96.364	98.182	97.273

Dataset 2 contains 1650 Arabic word images written by the selected five fonts, each word class is written in different size and resolution. It was recognized 1464 samples and produced recognition rate 88.727%.

Dataset 3 contains 1650 Arabic word images written by the selected five fonts, each word class is written in the same size and different resolution and was recognized 1632 samples and produced recognition rate 98.909%.

Group 2 contains ten datasets that use five different fonts, the words within each dataset contains the same font. The two datasets are allocated for each font which is used for testing our Arabic word classification system among each of the fonts separately. The datasets are divided into two categories.

Each of the first five datasets contains 330 Arabic words, written in one of the five different fonts. In each dataset words are written in different size and different resolution. Under this condition when these datasets are fetched into the system, the words recognized in each of the five datasets were (322, 323, 318, 324, 318), respectively and their corresponding recognition rates (97.576, 97.879, 96.364, 98.182 and 96.364%).

In the remaining five datasets, each dataset contains 330 Arabic words which are written in one of the five different fonts and each word within the dataset contains the same size but different resolution. Under this condition when these datasets are fetched into the system, the word recognized in each of the five datasets were (327, 325, 323, 329, 324), respectively and their recognition rates (99.091, 98.485, 97.879, 99.697 and 98.182%).

In Group 3, we have used hand written Arabic words; it contains one dataset of 216 words that are selected from IFN/ENIT-database. In this database, each word is written by six different writers. When this group of data is fetched into the system, it is able to recognized 180 words which give the recognition rate of 83.333%. Table 2 shows the recognition rates obtained in each of the three groups.

To record the recognition rates of samples of Arabic words which are written in five different fonts (Andalus, Traditional Arabic, Simplified Arabic, Arial and Tahoma) separately.

First, each word class is written five times in each time the word differ in size and resolution to other four words and we got the recognition rate to each font (97.576, 97.879, 96.364, 98.182 and 96.364%), respectively. Second, each word class is written five times in each time the word has same size and differs only in the resolution compare with other four words. In this case the recognition rate of each font (99.091, 98.485, 97.879, 99.697 and 98.182%) sequentially (Table 3).

In our test, the classifier SVMs divided each dataset into two parts for considering training and testing sets during classification stage. And for achieving the accurate recognition rates, we have tested our system on each of these dataset many times to get the average of all results.

DISCUSSION

This study introduced an SVM-based system with many components to provide solutions for most of the inherent difficulties in recognizing Arabic script. An evaluation of the system shows that the features and models discriminate printed and handwritten Arabic words at higher rates.

Furthermore, these features can be used in many languages such as English, Chinese and others because it depends on several statistical and structural features. This method has been tested using both of printed and handwritten public Arabic databases and used features vector that consists of 43 features. It gives a high recognition with the robust SVM classifier. From present results, we conclude that the percentage of accurate classification rate depends on the number of samples in

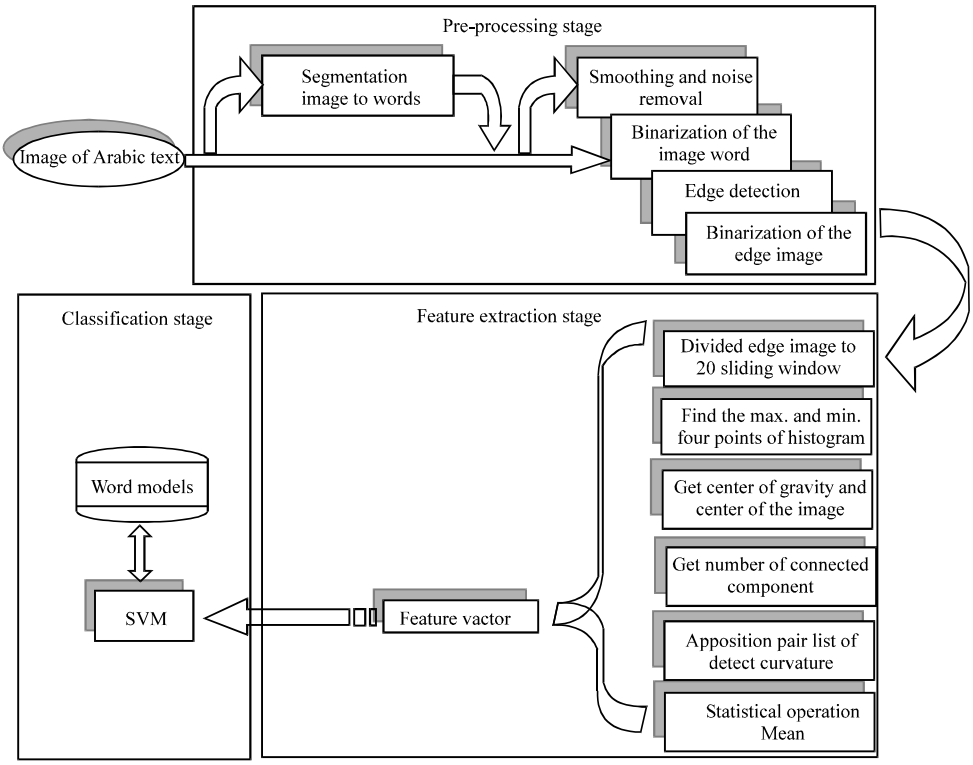

Fig. 4: Main structure of Arabic word recognition system

the dataset and the type of the images (size, resolution and font style) that used to represent each word in the testing stage. The main structure of Arabic word recognition system (Fig. 4).

CONCLUSION

This study implements linear kernel function of SVM. We expect the application of other types of kernel functions like Polynomial Kernel function or Radial Basis Function (RBF) will increase the accuracy of Arabic recognition system. Future work can include extracting more geometrical features and choose the most important descriptor point of it and can increase datasets size.

ACKNOWLEDGMENT

This work is supported in part by the National Natural Science Foundation of China, Grant No. 61003128.

REFERENCES

Al-Muhtaseb, H.A., S.A. Mahmoud and R.S. Qahwaji, 2008. Recognition of off-line printed arabic text using hidden markov models. Signal Process., 88: 2902-2912.

AlKhateeb, J.H., J. Ren, J. Jiang, S.S. Ipson and H. El Abed, 2008. Word-based handwritten arabic scripts recognition using DCT features an neural network classifier. Proceedings of the 5th International Multi-Conferenceon Systems, Signal and Devices, July 20-22, Amman, pp: 1-5.

AlKhateeb, J.H., O. Pauplin, J. Ren and J. Jiang, 2011. Performance of hidden Markov model and dynamic Bayesian network classifiers on handwritten Arabic word recognition. Knowledge-Based Syst., 24: 680-688.

Alma'adeed, S., C. Higgins and D. Elliman, 2004. Off-line recognition of handwritten Arabic words using multiplehidden Markov models. Knowledge-Based Syst., 17: 75-79.

Alshalabi, R., 2005. Pattern-based stemmer for finding arabic roots. Inform. Technol. J., 4: 38-43.

Femiani, J., M. Phielipp and A. Razdan, 2005. A system for discriminating handwriting from machine print on noisy arabic datasets. Proceedings of the Symposium on Document Image Understanding Technology, Nov. 2-4, College Park, Maryland, pp: 123-132.

He, X.C. and N.H.C.Yung, 2008. Corner detector based on global and local curvature properties. Opt. Eng., 47: 057008-1-057008-12.

Hsu, C.W. and C.J. Lin, 2002. A comparison of methods for multi-class support vector machines. IEEE Trans. Neural Networks, 13: 415-425.

Izabatene, H.F., W. Benhabib and S. Ghardaoui, 2010. Contribution of kernels on the SVM performance. J. Applied Sci., 10: 831-836.

Khorsheed, M.S., 2002. Off-line Arabic character recognition-a review. Patt. Anal. Appl., 5: 31-45.

Lotfi, F., F. Nadir and B. Mouldi, 2006. Arabic words recognition by fuzzy classifier. J. Applied Sci., 6: 647-650.

Pechwitz, M., S.S. Maddouri, V. Margner, N. Ellouze and H. Amiri, 2002. IFN/ENIT-database of handwritten Arabic words. Proceeding of 7th International Francophone Conference on Document Processing, Oct. 21-23, Hammamet, Tunis, pp: 145-152.

Platt, J., N. Cristianini and J. Shawe-Taylor, 2000. Large margin DAGs for multiclass classification. Proc. Neural Inform. Process. Syst., 12: 547-553.

Rifkin, R. and A. Klautau, 2004. In defense of one-vs-all classification. J. Mach. Learn. Res., 5: 101-141.

Rodriguez, J.A. and F. Perronnin, 2008. Local gradient histogram features for word spotting in unconstrained handwritten documents. Proceedings of the 1st International Conference on Handwriting Recognition, Aug. 19-21, Montreal, Canada, pp: 1-6.

Tlili-Guiassa, Y. and L.M. Tayeb, 2006. Tagging by combining rules-based and memory-based learning. Inform. Technol. J., 5: 679-684.

Zhang, Z., L. Jin, K. Ding and X. Gao, 2009. Character-SIFT: A novel feature for offline handwritten Chinese character recognition. Proceedings of the 10th International Conference on Document Analysis and Recognition, July 26-29, Barcelona, Spain, pp: 763-767.

Zheng, L., A. Hassin and X. Tang, 2004. A new algorithm for machine printed Arabic character segmentation. Pattern Recognit. Lett., 25: 1723-1729.