Word-Based Arabic Handwritten Recognition Using SVM Classifier with a Reject Option

Bouchra El qacimy[#1], Mounir Ait kerroum[*2], Ahmed Hammouch[#3]
[#]*Laboratory LRGE, ENSET of Rabat, Mohamed V University*
Rabat, Morocco
[1]elq.bouchra@gmail.com
[3]hammouch_a@yahoo.com
[*]*Laboratory LARIT, Ibn Tufail University,Faculty of Science, ENCG of Kenitra*
Kenitra, Morocco
[2]maitkerroum@gmail.com

***Abstract*—Arabic handwritten recognition is a challenging task due to high variability of Arabic script and its intrinsic characteristics such as cursiveness, ligatures and diacritics. This paper presents a word-based off-line Arabic handwritten recognition system based on discrete cosine transform features and SVM classifier enhanced using a reject option. The latter is based on the number of sub-words in the input word image calculated using a novel segmentation algorithm. To evaluate our proposed system, we used the IFN/ENIT database of Arabic handwritten words and the results has shown the effectiveness of our approach in enhancing the recognition performance.**

Keywords-Arabic handwriting; Word-base recognition; SVM, DCT, Reject option.

I. Introduction

Handwritten text recognition is an active topic in pattern recognition research in view of its numerous applications such as automatic postal mail sorting, bank check processing, form data entry, etc. For these applications, the accuracy and speed of recognition is crucial to the overall performance. Moreover, Arabic is the 5th spoken language world wide after Chinese, Spanish, English and Hindi. With a population exceeding 237 million, Arabic is the first language in 60 countries [1]. However, automatic recognition of Arabic handwritten script remains a challenging task due, in part, to its inherent characteristics such as cursiveness, overlapping characters, presence of diacritical marks, etc [2]. In the last few years, more efforts have been diverted, to construct different databases for Arabic handwriting text word recognition. However, only few are freely available for researchers.

Srihari et al. [3] used a database consisting of approximately 20,000 words (10 writers have written 10 pages of text, each includes between 150 and 200 words). The database is limited and is not freely available. Abdullah et al. [4] presented a database of Arabic handwritten words. It contains 12300 Arabic words written by 82 different writers. Ziaratban et al. [5] presented a Farsi Handwritten Text database (FHT). The database contains 106,600 handwritten words. Although FHT is a Farsi text database, it may be used by researchers in Arabic handwritten text recognition. The IFN/ENIT database [6] was created by the Institute of Communications Technology (IfN) at Technical University Braunschweig in Germany and the Ecole Nationale dIngenieurs de Tunis (ENIT) in Tunisia. Version 1.0 of this database consists of 26,459 images of the 937 names of cities and towns in Tunisia, written by 411 different writers. The database contains 115585 Pieces of Arabic Words (PAWs) and 212211 characters. The images are partitioned into four sets. Al-Maadeed et al. presented AHDB (Arabic Handwritten DataBase) [7], an Arabic handwritten text database of 100 writers. This database contains words used for numbers in bank checks. It also contains some of the most common words in Arabic. This database is limited in vocabulary and can be more useful in check processing applications [8].

According to the data acquisition process, we can distinguish two types of systems: on-line and off-line systems. In on-line systems, the text is recognized in real time,using handwritten trajectory collected by recording the coordinates of the pen position. On the other hand, in off-line systems, text images are scanned then used for recognition. Our work focuses on off-line Arabic handwritten text recognition. Besides, the literature distinguishes between two distinct approaches for Arabic text recognition, namely: the segmentation-based systems and segmentation free systems, or holistic systems [9]. In the latter, recognition is performed on the whole representation of a word without segmentation. Whilst in the segmentation-based systems, the text is segmented into characters or other primitives.

In this work, we address the issue of Arabic handwritten recognition using a holistic approach. We introduce an off-line Arabic handwritten recognition system based on SVM classifier and discrete cosine transform (DCT) features. This system is composed of four main stages which are: preprocessing, segmentation into sub-words, feature extraction using DCT features and classification based on SVM RBF classifier. We integrate a rejection criterion in the classification phase using the number of sub-words in the input word image. For the evaluation of our proposed system, we used the IFN/ENIT database of Arabic handwritten words to compare with state of art techniques and the results has shown the effectiveness of our approach in enhancing the recognition performance.

The rest of this paper is organized as follows: in section II,

978-1-4673-8709-5/15/$31.00

an overview of literature on Arabic handwritten recognition systems is given. Next, in section III we give a detailed description of the proposed recognition system. Finally, we discuss the results obtained on the IFN/ENIT dataset before concluding with a brief summary and addressing some perspective of future work.

II. Related Work

In the literature, numerous recognition systems were proposed by researchers for Arabic handwritten text recognition. As reported in [2], there is an intense inclination for using structural features. Indeed, Almuallim [10] and Goraine [11] used a set of structural features as loops, dots and stroke directions along with a classifier based on rules. They achieved respectively 90 and 91% as recognition rate on their particular datasets. Most systems are based on Hidden Markov Models (HMM) such as [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22] and [23]. Those systems are more likely to use structural features. Fewer systems used Support Vector machine. Khalifa et al. [24] implemented a holistic recognition system using DCT features and SVM Classifier and achieved a recognition rate of 91,70% on the IFN/ENIT Arabic Standard Database which is very competitive. Neural networks are also broadly used for handwritten Arabic text recognition [25], [26], [27]. Haboubi et al. [28] proposed a segmentation-based recognition system using pixel and structural description, Gabor Filter and Fourier descriptors. He attained 89% on the IFN/ENIT database. Finally, Parvez et al. 2013 [8] proposed a holistic recognition system based on structural and syntactic pattern attributes that were fed to a fuzzy polygon matching recognizer. They achieved a recognition rate of 92,20% on the well known IFN/ENIT database.

In this work, we propose a novel system for Arabic text recognition in a holistic approach. The novelty of this work resides in two points:

- the introduction of an efficient segmentation algorithm of words into sub-words;
- an enhanced version of the recognition phase used by Khalifa et al. [24] using a reject option.

It is also worth mentioning that there is no generally accepted database for Arabic text recognition that is freely available for researchers. Hence, different researchers of Arabic text recognition have used different data and hence the recognition rates of the different techniques may not be comparable [8].

III. Outline of the Proposed Recognition System

A typical off-line recognition system is illustrated in the block diagram in figure 1. It is composed of five main components which are:

- **Preprocessing** : The main objective of preprocessing is to remove irrelevant information that has a negative effect on the overall recognition [29]. It includes techniques such as noise removal, thresholding, thinning, skew/slant correction, baseline estimation and normalization.
- **Representation** : The acquired text image is sometimes converted into a more concise representation prior to feature extraction and recognition. In some cases, features are extracted directly from the text image. In other cases, skeleton and/or contour of the text image is extracted prior to feature extraction [8].
- **Segmentation** : Some classification techniques require the segmentation of words into sub-words, characters, strokes (graphemes), or other units. Moreover, the literature distinguishes between two distinct approaches for Arabic text recognition, namely: the segmentation-based systems and segmentation free systems also called holistic systems [9].
- **Feature extraction** : The purpose of this phase is the measurement of the attributes that are most pertinent to a given classification task. Feature extraction methods can be categorized into structural features, statistical features and feature space transformations methods. Structural features involve the geometrical and topological characteristics of an input image [30] such as strokes, end points, the number of dots, loops, etc [31]. On the other hand, statistical features are extracted from the statistical distribution of pixels and provide low complexity and high speed. [9]. As for feature space transformations, they include: Fourier Transform, Hough Transform, Gabor Transform, Wavelets, Karhunen Loeve Expression and moments.
- **Classification** : The aim of this phase is to assign each observation with a class label or membership scores to the defined classes. A number of classifiers have been used for recognition of Arabic handwritten words. These include Hidden Markov Model (HMM), Support Vector Machines (SVM), Artificial Neural Networks (ANN),k-Nearest Neighbors (kNN), and others [8].

In some cases, the representation and the segmentation phase can be merged in the processing phase before feature extraction.

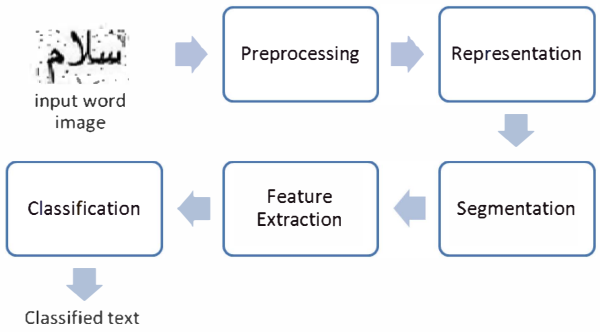

Figure 1: General model of an off-line Arabic handwritten recognition system.

Our proposed recognition system is illustrated in figure 2. First we do some preprocessing on the input word image. The images are size normalized then fed to the segmentation stage. We segment words into sub-words using a novel algorithm presented in a previous work. It is based on vertical projection along the estimated baseline and the knowledge of the charac-

teristics of Arabic script. DCT coefficients are then extracted in a zigzag fashion to form the feature vector that will be fed to an SVM classifier with a radial basis function kernel (SVM RBF). We used the number of sub-words resulting from the segmentation stage as rejection criterion to enhance the recognition performance of the SVM classifier.

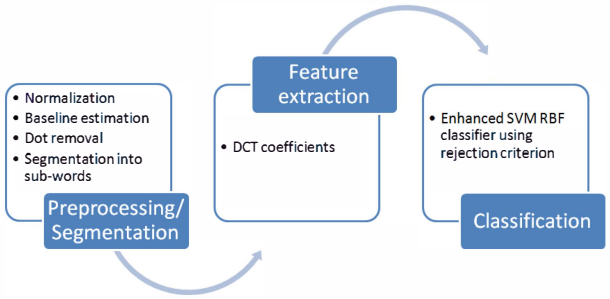

Figure 2: Overview of the proposed recognition system.

Next we describe in detail each component of our proposed recognition system.

A. Preprocessing and Segmentation

In this stage, we prepare the input images for recognition. First, the word images are size normalized, then the words are segmented into sub-words to enhance the recognition process. We used an algorithm that we developed in a previous work for segmentation.

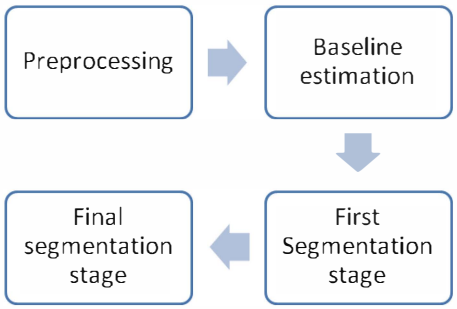

Figure 3: Block digram of the preprocessing and segmentation stage.

The general concept of our segmentation algorithm is illustrated in figure 3. At first we apply a preprocessing to the word image to remove dots which can influence the baseline estimation step. In the latter, we use the horizontal projection histogram to find the highest peak which correspond to the estimated baseline. Then, we perform a first segmentation stage based on vertical projection histogram using local minima's and the knowledge of the estimated baseline. After that, a verification segmentation stage is necessary to solve the problem of over-segmentation caused by the nature of the Arabic script. This step detects any errors that occurred during the first segmentation stage.

Figure 4 illustrates an example of an input image passing through the different steps of preprocessing and segmentation.

B. Feature extraction

After segmentation, we apply the DCT on the preprocessed image. It is a technique to convert data of the image into its elementary frequency components [32]. It clusters high value coefficients in the upper left corner and low value coefficients in the bottom right of the output matrix(m,n), where [m n] is the image size. Then we extract the N most relevant coefficients in a zigzag fashion. These retained coefficients are stored in a vector sequence in order to form the feature vectors that will be fed to the classifier for recognition. The Number N of the coefficients to retain is chosen experimentally. The higher the number of retained coefficient the better the quality of the reconstructed image characters.

C. SVM recognition

The Support Vector Machine (SVM) is a modern classifier which uses kernels to give optimal decision boundary to separate between classes in higher dimensional feature spaces. The SVM algorithm was originally introduced by Vapnik [33] in his work on structural risk minimization. It was initially designed for binary separation problems but it can be easily generalized to solve multi-class classification problems. The basic form of linear SVM classifier tries to find an optimal hyperplane that separates the set of samples belonging to different classes.

SVM was successfully evaluated on many recognition systems. Alamri et al. [34] used SVM with RBF kernel for Arabic numeral recognition. Alaei et al. [35] [36] experimented with SVM with linear, Gaussian, and polynomial kernels for numeral recognition. The authors reported that SVM with Gaussian kernel gave the best combination. Also, SVM was used for Arabic numeral recognition in Mahmoud and Olatunji [37], and Mahmoud and Owaidah [38] [8].

In this work, we use the SVM classifier with RBF kernel. The number of sub-words resulting from the segmentation phase is used as a rejection criterion to enhance the recognition process. For implementation, we used the LIBSVM package [39] that supports multi-class problem to classify the different Arabic words in a holistic approach.

IV. Experimental results

To evaluate our proposed recognition system, we used IFN/ENIT database of handwritten Arabic words. It consists of 26459 handwritten Tunisian town names written by 411 different writers [6]. Each word in this database comes with ground truth informations including: town zip-code, global word, baseline and its quality, quantity of words and characters, etc.

Experiments have been conducted in order to investigate the effectiveness of our proposed algorithm in comparison with state of art systems that use DCT features. For this purpose, we used the zip-code information to label the observations for recognition. We used four different sets of the IFN/ENIT database(a, b, c and d) for training and testing the classifier. 2000 images were randomly selected from the four datasets, 1500 for training and 500 for testing.

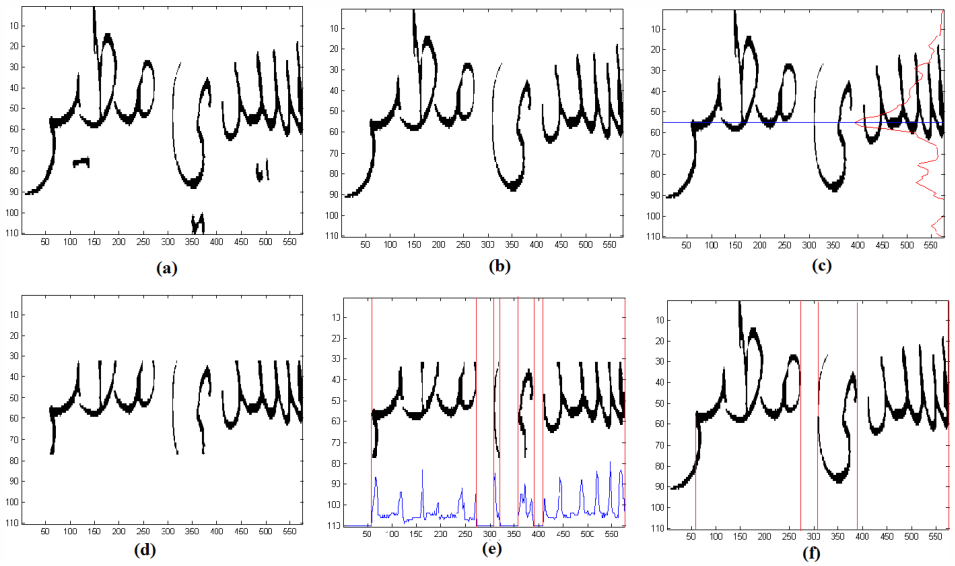

Figure 4: Example of the different steps of preprocessing and segmentation: (a)Normalized image; (b)Dot removal; (c)baseline estimation; (d) upper and lower zone removal; (e) First segmentation; (f) final segmentation.

The experiments were carried out in five steps: - First, we size normalized the word images of the IFN/ENIT database.
- Second, we segmented the words into sub-words using the algorithm described in section III-A.
- Third, we extracted the DCT features in order to constitute feature vectors that will be fed to the SVM classifier. We retained 400 DCT coefficients extracted in a zigzag manner.
- Forth, we performed a grid search with a view to find the optimal parameters c and gamma for training the SVM classifier. - Finally, we trained and tested the SVM RBF classifier using the number of sub-words as rejection criterion.

Performance of recognition from various DCT-based systems using IFN/ENIT database are summarized in table I. Figure 5 illustrates a comparative performance in terms of global classification accuracy using state of art systems using DCT features and the proposed system using SVM RBF with the rejection criterion.

System	Recognizer	Accuracy(%)
Khalifa et al. [24]	SVM RBF	91,70
AlKhateeb [40]	KNN	78,67
AlKhateeb [40]	NN	80,75
AlKhateeb et al. [41]	HMM with re-ranking	95,15
Proposed system	SVM RBF with reject option	98,06

TABLE I: Performance of recognition from various DCT-based systems using IFN/ENIT database.

As we can see from table I and figure 5, Khalifa et al. proposed a system based on DCT features and SVM RBF classifier enhanced using Recursive Feature Elimination (RFE), with Principal Component Analysis (PCA). They reported a classification accuracy of 91,70%. On the other hand, Alkhateeb investigated DCT features with various recognizers namely: K nearest neignbors (KNN), neural networks (NN) and HMM with re-ranking and achieved respectively 78,67%,

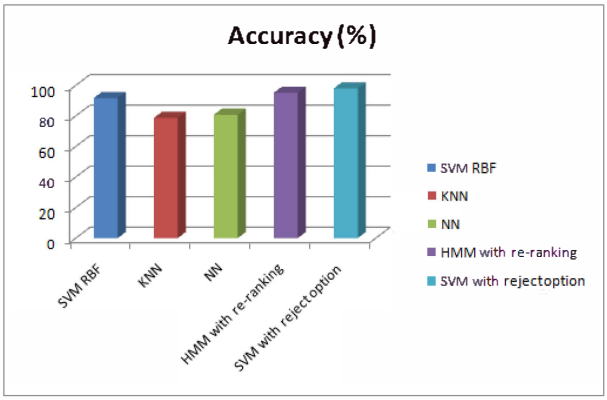

Figure 5: Comparative performance of recognition from various DCT-based systems using IFN/ENIT database.

80,75% and 95,15% in recognition rate. Whilst using SVM with the rejection criterion we attained 98% which is very competitive. And thus, validating the effectiveness of our proposed recognition system for handwritten Arabic words.

V. Conclusion

In this work, we addressed the issue of Arabic handwritten recognition using a holistic approach based on the whole word image. We proposed a word-based off-line Arabic handwritten recognition system based on DCT features and SVM classifier enhanced using a reject option. This system is composed of four main stages which are: preprocessing, segmentation into sub-words, feature extraction using DCT features and classification based on SVM RBF classifier. We integrated a rejection criterion in the classification phase using the number of sub-words in the input word image. We tested the proposed system on the IFN/ENIT database of Arabic handwritten words and compared the results to state of art DCT-based

systems. Results has shown the effectiveness of our approach in enhancing the recognition performance.

REFERENCES

[1] M. P. Lewis *et al.*, *Ethnologue: Languages of the world*. SIL international Dallas, TX, 2009, vol. 9.

[2] B. El qacimy, M. Ait kerroum, and A. Hammouch, "A review of feature extraction techniques for handwritten arabic text recognition," in *Electrical and Information Technologies (ICEIT'15), 1st International Conference on*. IEEE, 2015, pp. 241–245.

[3] S. N. Srihari, G. R. Ball, and H. Srinivasan, "Versatile search of scanned arabic handwriting," in *Arabic and Chinese Handwriting Recognition*. Springer, 2008, pp. 57–69.

[4] S. Abdulla, A. Al-Nassiri, and R. A. Salam, "Off-line arabic handwritten word segmentation using rotational invariant segments features." *Int. Arab J. Inf. Technol.*, vol. 5, no. 2, pp. 200–208, 2008.

[5] M. Ziaratban, K. Faez, and F. Bagheri, "Fht: An unconstraint farsi handwritten text database," in *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*. IEEE, 2009, pp. 281–285.

[6] M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze, H. Amiri *et al.*, "Ifn/enit-database of handwritten arabic words," in *Proc. of CIFED*, vol. 2. Citeseer, 2002, pp. 127–136.

[7] S. Al-Ma'adeed, D. Elliman, and C. A. Higgins, "A data base for arabic handwritten text recognition research," in *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*. IEEE, 2002, pp. 485–489.

[8] M. T. Parvez and S. A. Mahmoud, "Offline arabic handwritten text recognition: a survey," *ACM Computing Surveys (CSUR)*, vol. 45, no. 2, p. 23, 2013.

[9] M. S. Khorsheed, "Off-line arabic character recognition–a review," *Pattern analysis & applications*, vol. 5, no. 1, pp. 31–45, 2002.

[10] H. Almuallim and S. Yamaguchi, "A method of recognition of arabic cursive handwriting," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 5, pp. 715–722, 1987.

[11] H. Goraine, M. Usher, and S. Al-Emami, "Off-line arabic character recognition," *Computer*, vol. 25, no. 7, pp. 71–74, 1992.

[12] R. El-Hajj, L. Likforman-Sulem, and C. Mokbel, "Arabic handwriting recognition using baseline dependant features and hidden markov modeling," in *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*. IEEE, 2005, pp. 893–897.

[13] M. Dehghan, K. Faez, M. Ahmadi, and M. Shridhar, "Handwritten farsi (arabic) word recognition: a holistic approach using discrete hmm," *Pattern Recognition*, vol. 34, no. 5, pp. 1057–1065, 2001.

[14] M. Pechwitz and V. Maergner, "Hmm based approach for handwritten arabic word recognition using the ifn/enit-database," in *2013 12th International Conference on Document Analysis and Recognition*, vol. 2. IEEE Computer Society, 2003, pp. 890–890.

[15] M. S. Khorsheed, "Recognising handwritten arabic manuscripts using a single hidden markov model," *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2235–2242, 2003.

[16] S. Alma'adeed, C. Higgins, and D. Elliman, "Off-line recognition of handwritten arabic words using multiple hidden markov models," *Knowledge-Based Systems*, vol. 17, no. 2, pp. 75–79, 2004.

[17] M. Pechwitz, V. Maergner, H. El Abed *et al.*, "Comparison of two different feature sets for offline recognition of handwritten arabic words," in *Tenth International Workshop on Frontiers in Handwriting Recognition*, 2006.

[18] V. Märgner, H. El Abed, M. Pechwitz *et al.*, "Offline handwritten arabic word recognition using hmm-a character based approach without explicit segmentation," in *Actes du 9ème Colloque International Francophone sur l'Ecrit et le Document*, 2006, pp. 259–264.

[19] A. Kundu, T. Hines, J. Phillips, B. D. Huyck, and L. C. Van Guilder, "Arabic handwriting recognition using variable duration hmm," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2. IEEE, 2007, pp. 644–648.

[20] S. M. Touj, N. Ben Amara, and H. Amiri, "A hybrid approach for off-line arabic handwriting recognition based on a planar hidden markov modeling," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2. IEEE, 2007, pp. 964–968.

[21] M. S. Khorsheed, "Hmm-based system for recognizing words in historical arabic manuscript," *International Journal of Robotics and Automation*, vol. 22, no. 4, pp. 294–303, 2007.

[22] R. Al-Hajj Mohamad, L. Likforman-Sulem, and C. Mokbel, "Combining slanted-frame classifiers for improved hmm-based arabic handwriting recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 7, pp. 1165–1177, 2009.

[23] Y. Kessentini, T. Paquet, and A. Ben Hamadou, "Off-line handwritten word recognition using multi-stream hidden markov models," *Pattern Recognition Letters*, vol. 31, no. 1, pp. 60–70, 2010.

[24] M. Khalifa and Y. BingRu, "A novel word based arabic handwritten recognition system using svm classifier," in *Advanced Research on Electronic Commerce, Web Application, and Communication*. Springer, 2011, pp. 163–171.

[25] L. Souici-Meslati and M. Sellami, "A hybrid approach for arabic literal amounts recognition," *Arabian Journal for Science and Engineering*, vol. 29, no. 2, pp. 177–194, 2004.

[26] N. Farah, L. Souici, L. Farah, and M. Sellami, "Arabic words recognition with classifiers combination: An application to literal amounts," in *Artificial Intelligence: Methodology, Systems, and Applications*. Springer, 2004, pp. 420–429.

[27] M. M. M. Fahmy and S. A. Ali, "Automatic recognition of handwritten arabic characters using their geometrical features," *Studies in Informatics and Control*, vol. 10, no. 2, pp. 81–98, 2001.

[28] S. Haboubi, S. Maddouri, N. Ellouze, and H. El-Abed, "Invariant primitives for handwritten arabic script: A contrastive study of four feature sets," in *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*. IEEE, 2009, pp. 691–697.

[29] B. Q. Huang, Y. Zhang, and M. T. Kechadi, "Preprocessing techniques for online handwriting recognition," in *Proceedings of the Seventh International Conference on Intelligent Systems Design and Applications*. IEEE Computer Society, 2007, pp. 793–800.

[30] M. Cheriet, N. Kharma, C.-L. Liu, and C. Suen, *Character recognition systems: a guide for students and practitioners*. John Wiley & Sons, 2007.

[31] N. Arica and F. T. Yarman-Vural, "An overview of character recognition focused on off-line handwriting," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 31, no. 2, pp. 216–233, 2001.

[32] B. El qacimy, M. Ait kerroum, and A. Hammouch, "Handwritten digit recognition based on dct features and svm classifier," in *Complex Systems (WCCS), 2014 Second World Conference on*. IEEE, 2014, pp. 13–16.

[33] V. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 2.

[34] H. Alamri, C. L. He, and C. Y. Suen, "A new approach for segmentation and recognition of arabic handwritten touching numeral pairs," in *Computer Analysis of Images and Patterns*. Springer, 2009, pp. 165–172.

[35] A. Alaei, P. Nagabhushan, and U. Pal, "Fine classification of unconstrained handwritten persian/arabic numerals by removing confusion amongst similar classes," in *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*. IEEE, 2009, pp. 601–605.

[36] A. Alaei, U. Pal, and P. Nagabhushan, "Using modified contour features and svm based classifier for the recognition of persian/arabic handwritten numerals," in *Advances in Pattern Recognition, 2009. ICAPR'09. Seventh International Conference on*. IEEE, 2009, pp. 391–394.

[37] S. A. Mahmoud and S. O. Olatunji, "Automatic recognition of off-line handwritten arabic (indian) numerals using support vector and extreme learning machines," *International Journal of Imaging*, vol. 2, no. A09, pp. 34–53, 2009.

[38] S. A. Mahmoud and S. M. Awaida, "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*, vol. 34, no. 2B, pp. 429–444, 2009.

[39] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[40] J. H. Y. AlKhateeb, "Word based off-line handwritten arabic classification and recognition: design of automatic recognition system for large vocabulary offline handwritten arabic words using machine learning approaches," Ph.D. dissertation, University of Bradford, 2010.

[41] J. H. AlKhateeb, J. Ren, J. Jiang, and H. Al-Muhtaseb, "Offline handwritten arabic cursive text recognition using hidden markov models and re-ranking," *Pattern Recognition Letters*, vol. 32, no. 8, pp. 1081–1088, 2011.