H.A. Jeiad
Computer Eng. Dept
University of Technology
Baghdad, Iraq
hsn.uot@gmail.com

Indian Number Handwriting Features Extraction and Classification using Multi-Class SVM

Abstract_*In this paper, an Indian Number Handwriting Recognition Model (INHRM) is proposed. Mainly, the proposed model consists of four phases which are the image acquisition, image preprocessing, features extraction, and classification model. Initially, the captured images of the handwritten Indian numbers were enhanced and preprocessed to obtain the skeleton for the interested object. The extracted features of the handwritten Indian numbers were obtained by calculating four parameters for each captured number sample, these parameters are the number of starting points, the number of intersection points, the average zoning which consists of four values, and finally, the normalized chain vector of length of 10 elements. So, the resulted 16 values of the four parameters were arranged in a vectors of length of 16 elements. These features vectors were used in the training and testing processes of the proposed INHRM model. Multi-class SVM (MSVM) approach is suggested for the classification phase. An accumulation of 600 samples of various handwritten Indian numbers styles has been gathered from a group of 60 students. These samples were preprocessed, features extracted, then delivered to the classification phase by utilizing 500 samples of them for training while the remaining 100 samples were used for testing of the MSVM-classifier model. The results showed that the proposed INHRM achieved relatively high percentage of exactness of around 97%.*

Received on: 07/03/2017
Accepted on: 12/10/2017

Keywords*-Handwriting Recognition, Features Extraction, SVM, Multi-Class SVM.*

How to cite this article: H.A. Jeiad, "Indian Number Handwriting Features Extraction and Classification using Multi-Class SVM", *Engineering and Technology Journal*, Vol. 36, No. 1, pp. 33-40, 2018.

1. Introduction

The research area in pattern recognition and image processing field has been the most interesting and attracting research area in the recent years specially the character handwriting recognition. In its nature, the characters of the handwriting have wide range of variety in the style among the different peoples. For this reason, it is so difficult for the machine to be able to recognize the handwritten characters [1].

In computer field the recognition system consist of two major parts, the extraction of image content description that is known as features extraction and recognizing the subsequent features known as the classifier model. The first part is essential because of the redundant, unrelated, and unwanted data generated through capturing the digital images by the camera or scanner devices. In fact, the most sophisticated algorithms that can be used for building an efficient recognition system will be complex and slow down because of the unrelated data. So, only essential features should be considered. On other hand, having confusable and unrelated data for the training stage of the classifier makes the recognition process and decision making highly inaccurate not because of the false design of the recognition model but because of the type of data its used [2].

Alaei et al. [1]. Proposed a feature set based on modified contour chain code to achieve higher recognition accuracy of Persian/Indian numerals. Yadav et al. [2] introduced an artificial neural network-based offline English character recognition system to recognize English alphanumeric characters. Kumar et al. [3] presented a scheme for offline handwritten Gurmukhi character recognition based on SVM. Tawde, and Kundargi [4] reviewed some of the important feature extraction techniques employed for different Indian handwritten scripts. Trier et al. [5] presented an overview of feature extraction methods for off-line recognition of segmented (isolated) characters. Izakian et al. [6] proposed a chain code based algorithm along with other significant peculiarities such as number and location of dots and auxiliary parts, and the number of holes existing in the isolated character has been used in this study to identify Farsi/Arabic characters. Kumar et al [7] attempted to recognize English handwritten character using the multi-

layer feed forward back propagation neural network without feature extraction and SVM classifier. Abbas and Hashim [10] proposed automatic license plate detection and recognition system to identify cars by their license plates. The system was consisting of two parts. The first is a practical implementation to take an automatically snapshot for cars passes, while the second part is image processing to process the snapshot taken in the first to isolate the characters, numbers and words of the license plate using Otsu's and Hough transforms technique. The obtained recognition rate was about 98.245%. Rassoul et al. [11] presented signature recognition model based on transforming the signature into contours and then converting these contours into vectors. The invariant moments method and the calculation of mean was applied to the obtain vectors to extract the features of the acquired signature. The minimum distance methodis used as a classifier to identify the personal signer. Abdul Hassan and Kadhm [12] proposed Indian handwriting recognition system using the thresholding that based on fuzzy C-means clustering approach. The thresholding is used to reduce the dimensionality of image to remove the undesirable information. Mainly, the algorithm was performed by feeding the intensity of the pixel value of the image pixels into the FCM clustering algorithm. The authors pointed out that good results was obtained than the other methods.

In this paper, a classification model is proposed to recognize the Indian handwritten numbers. Mainly, the proposed model consists of four phases which are image acquisition, image preprocessing, features extraction, and classification model. Initially, the captured images of the handwritten Indian numbers were enhanced and preprocessed. The features of certain number were extracted by calculating four main parameters, these parameters are the number of starting points, the number of intersection points, the average zoning, and the normalized chain vector. The total features were equal to 16 features for each handwritten number. These features were used in the training and testing processes of the proposed model. Multi-class SVM (MSVM) approach is suggested for the classification phase. This paper is organized as follows: general introduction of handwriting recognition and review for the related works were presented in section (1). The proposed algorithm for INHRM was presented with a description for each of its stages was presented in section (3). The implementation of INHRM model was given in section (4) while the conclusions and future works were presented in section (5).

2. The Proposed Model

The model of the proposed system INHRM is shown in Figure 1. This model consists of four phases start from image acquisition and ends by the classification using MSVM. The description of each block of INHRM model was given in the next sections.

I. Image Acquisition

The image for the interested handwritten number will be scanned and acquired in this phase as an input image. This input image will be scanned at 300 dpi to obtain convenient image quality and then apply window of 67*83 pixels to have uniform image set of the same size to be processed in the next phase of INHRM model.

II. Image Preprocessing

The preprocessing phase includes two processes, which are image enhancement and noise removing to prepare the image for the next phase which is the feature extraction. Intuitively, it is assumed that a black color is used for handwriting the numbers on a white paper as a background.

Initially, the scanned image is converted to grayscale monochrome, then a contrast enhancement and equalization for the brightness level is performed by using the histogram equalization technique. The local contrast and the definition of the edges in the inputted image were improved through applying the approach of contrast limited adaptive histogram equalization (CLAHE).

The global image threshold is determined by using Otsu's method to convert the image from grayscale to binary image, and a static image zoning is used to resize the image to only the area of interest, which is the area that contains the handwritten number under consideration as shown in Figure 2. Finally, closing and opening morphology is applied to fill gaps and remove noises, and then a skeleton or thinning operation is performed to get one-pixel width object.

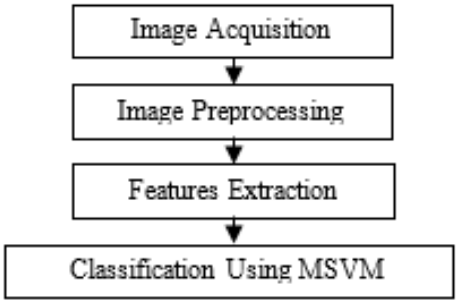

Figure 1: Block diagram of INHRM model

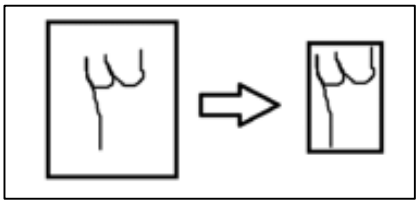

Figure 2: Static image zoning

III. Features Extraction

Determining and computing features is the most critical phase in recognition system. However, neither standard method for computing features nor a widely accepted feature set exist. In general features divided into a) structure features (time domain features) b) global transformation features (frequency domain features).

Essentially, the phase of features extraction comprises extracting four kernel features from the preprocessed image that contains the handwritten number, these features are divided into two types of features which are structural features (feature extracted from the time domain) and global transformation features (features extracted from frequency domain). The four features are as follows:

1. Finding the number of starting point(s).
2. Finding the intersections point(s).
3. Calculating the average zone of the image.
4. Determining the chain vector.

All of these features were extracted and arranged in a certain vector called a features vector with length of 16 elements to represent the features to be used in the classification and decision making phase of the proposed INHRM model.

The next subsections details each of these features and the methods of the extraction were demonstrated.

1. Starting Points

The number of starting points for a certain Indian handwritten number ranges from zero starting points as in the case of number 5 (٥) to two starting points as in numbers 7 (٧) or 8 (٨). The number of starting points was determined through finding the number and location of starting points of the object by searching the image to find pixels with only one neighbor using the 8-connectivity. Figure 3 shows a sample of Indian handwritten number 7 (٧) with two starting points.

The 8-connectivity pixels are neighbors to every pixel that touches one of their edges or corners. These pixels are connected horizontally, vertically, and diagonally. For pixel A with coordinate (*x, y*), B is considered neighbor to pixel A if its coordinate is (*x±1, y±1*). Figure 4 shows the diagram of 8-connectivity.

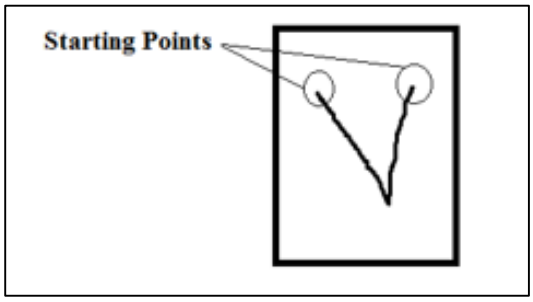

Figure 3: Determination of Starting points.

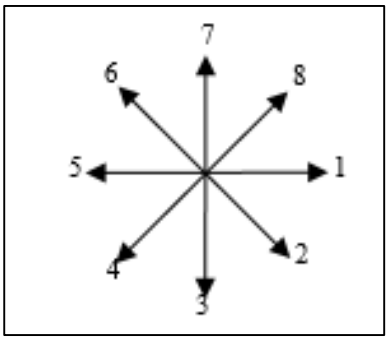

Figure 4: 8-Connectivity

2. Intersection Points:

This type of features finds the number and location of intersection points of the object by searching the image to find pixels with more than one neighbor using the 8-conectivity as shown in Figure 5.

3. Average Zone

Extracting the average zone feature needs to partition the image into four zones and calculate the average pixel values in each region to find where the object. These projections are computed using equations 1 through 4:

$$\boldsymbol{R[1]} = \frac{\sum_{i=0}^{(m/2)-1}\sum_{j=0}^{(n/2)-1} \boldsymbol{I[i,j]}}{\boldsymbol{(m/2)+(n/2)}} \quad (\mathbf{1})$$

$$\boldsymbol{R[2]} = \frac{\sum_{i=(m/2)}^{m-1}\sum_{j=0}^{(n/2)-1} \boldsymbol{I[i,j]}}{\boldsymbol{(m/2)+(n/2)}} \quad (\mathbf{2})$$

$$\boldsymbol{R[3]} = \frac{\sum_{i=0}^{(m/2)-1}\sum_{j=(n/2)}^{n-1} \boldsymbol{I[i,j]}}{\boldsymbol{(m/2)+(n/2)}} \quad (\mathbf{3})$$

$$\boldsymbol{R[4]} = \frac{\sum_{i=(m/2)}^{m-1}\sum_{j=(n/2)}^{n-1} \boldsymbol{I[i,j]}}{\boldsymbol{(m/2)+(n/2)}} \quad (\mathbf{4})$$

Where *I*[*i,j*] is an image with *m* rows and *n* columns, *R* is the average of the zone region. Figure 6 shows how the object image is partitioned to find the feature of average zone.

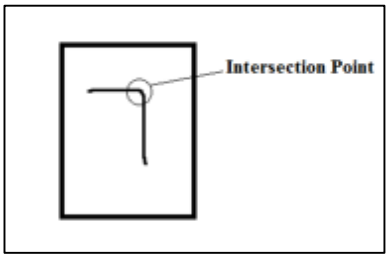

Figure 5: Determination of Intersection point

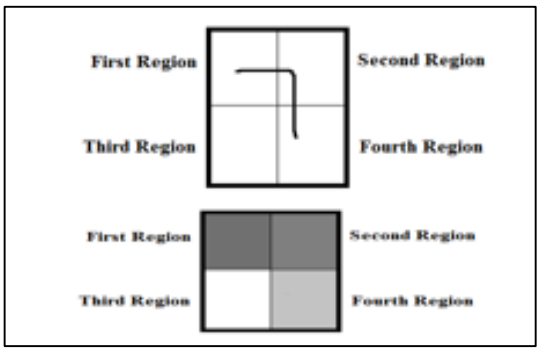

Figure 6: Partitioning the image to find the average zone

4. Chain Vector

The chain vector of the object is constructed as one of the four features through tracking the shape of object.

This is done by starting from one of the previously obtained starting points, search using 8-connectivity to find and store neighbor location, then move to next neighbor, if there is more than two neighbors then there is intersection, so store the intersection coordinate and visit one of the lines then when reach the end return to the intersection point to visit the other line. Using this method solve the problem of falling into the local minima when there is an intersection. Figure 7 demonstrates how the chain vector could be obtained for the Indian handwritten number 2(٢).

In the case of facing an intersection through applying the previously mentioned method in subsection (2.3.2), the tracking process to produce the chain vector will visit the line segment before appearing the intersection point, and after that, it will return to visit the other line segments as shown in Figure 8.

After obtaining the chain vector it is possible to reconstruct the object depending on the starting point and the chain vector. Now, to demonstrate the process of reconstruction a case study is taken here, assume that the chain vector for the handwritten Indian number 4(٤) is equal to [4 4 4 4 4 2 2 2 4 4 4 4 2 2 2]. In fact, this assumed chain may not be the same for the other samples of number 4(٤) and may be slightly differs. Now, the reconstruction for the assumed chain starts by moving in the direction of branch number 4 of the 8-connectivity structure (shown in Figure 4 above) for five dots, then toward the direction of branch number 2 for three dots, then, again, toward the direction of branch number 4 for four dots, and finally, draw three dotes in the direction of branch number 2. Figure 9 shows an illustration for the reconstruction for number ٤ that is considered. It is worth to mention here that the chain vector for number ٤ is not real and it is assumed just for simplicity of demonstration.

5. Normalization

The normalization is used to reduce the redundancy in the elements of chain vector and therefore reduce the size of feature vector, this form of dimension reduction can improve the features vector and speeds up the computation. It is obvious that the number of elements of chain vector obtained will depends on the number of neighbor pixels of the object. So, in order to reduce this number and at the same time preserve the features of the chain vector a trial and error policy is used and to determine that a chain vector of size ten elements is suitable for distinguishing different objects. So it is required to unify the chain vector element number to just ten elements regardless the number of elements produced through the tracking process for the object shape.

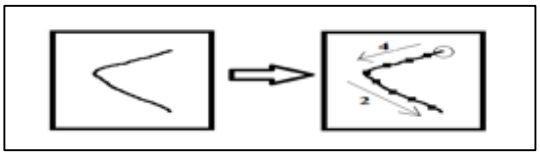

Figure 7: Tracking the shape using 8-connectivity to find the chain

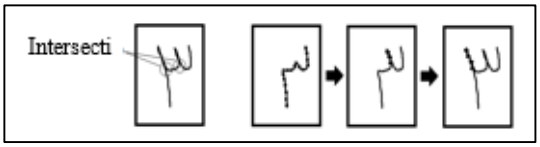

Figure 8: The case of facing an intersection point during extracting the chain vector

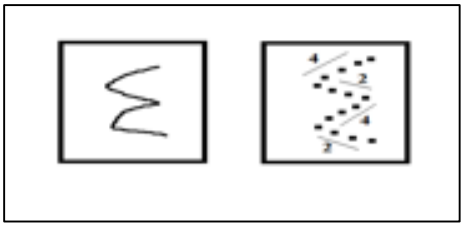

Figure 9: Reconstruction of number 4(٤).

In order to reduce the size of the chain vector the frequency of the neighbors is calculated then ignoring the neighbors with small frequency regarding to others in order to avoid the small and uncorrelated movement. The following procedure is used to calculate and normalizes the vector to the size of ten elements:

1- Calculate the frequency of each neighbor.
2- Calculate the normalized frequency of each neighbor.
3- Multiply the normalized frequency by ten and round its values.
4- Reconstruct the chain vector according to the normalized frequency vector.

The following equation was used to find the normalized frequency

$$NF_k = \frac{f_k}{F} \quad (5)$$

Where NF_k is the normalized frequency, f_k is the individual frequency of the *kth* neighbor and *F* is the accumulative frequency of all neighbors.

Equation (6) is used to round and obtain an integer value of NF_k.

$$NV_k = Round(NF_k * 10) \quad (6)$$

Where NV_k is the *kth* element of the normalized chain vector for k=1,..,10.

As an example, Table 1 demonstrates the steps of the procedure of finding the ten elements of *NV* for the handwritten Indian number 4 (٤) which gave the below chain vector prior to the normalization processes. Chain_vector= [5554444443333111111 1111445555 5444433333333222211111].

The symbol N in Table 1 denotes to the neighbor pixel, *F* is the individual frequency of each of the N neighbors, *NF* is the normalized frequency, and *NV* is the normalized vector with ten elements such as follows:

NV= [1 1 1 2 0 1 1 1 1 1]

The resulted *NV* is used for the classification after merging it with the other six features, which were pre-extracted as shown in Subsections *1*-3 to have the final features vector that consists of 16 elements. In some cases, applying the normalization equation may give size of 9 or 8 elements and in this case padding of zeroes is used.

Table 1: Normalizing the chain vector points

N	5	4	3	1	4	5	4	3	2	1
F	3	6	4	10	2	5	4	8	4	6
NF	0.05-	0.107	0.071	0.179	0.036	0.089	0.071	0.142	0.071	0.107
*NF**10	0.54	1.107	0.71	1.79	0.36	0.89	0.71	1.142	0.71	1.07
NV	1	1	1	2	0	1	1	1	1	1

IV. Classification using MSVM

SVM (Support Vector Machine) with its concepts was introduced by Vapnik and co-workers [9]. SVM became more popular due to offering a powerful features and efficient machinery to deal with the problem of classification. The statistical learning is the theory that SVM based on. The concepts of SVM targeted to make the margin of class separation as much as possible. The SVM was specified for two class problems and it seeking optimal hyper-plane, which maximized the margin between the nearest samples of both classes [9]. For each of the two classes the major part of samples is used as a training data of SVM classifier while the remaining part is used for testing to produce the classification result. In fact, Multi-class Support Vector Machine (MSVM) was introduced as an SVM-variant to deal with many classes of dataset. For more detail about SVM and MSVM one can referenced for [7- 9].

Figure 10 shows the diagram for the MSVM based classification phase of INHRM model. This phase is divided into two parts namely, training and testing part. The different samples of the pre-extracted features vectors for the ten types (classes) of handwritten Indian numbers were collected and divided into a training and testing datasets with the majority for the training. The MSVM-classifier performs the recognition on the data that is considered as a testing data due to the decisions that were determined during the training stage.

3. Implementation

The proposed classification model, INHRM, has been implemented using Matlab 2014b on a laptop with ci5 of 2 GHz. During the implementation of INHRM model a sample of handwriting Indian numbers from 0 to 9 were collected from 60 persons with total number of samples is 600, each number have 100 samples, 500 samples used for the training phase and the other 100 were used for the testing phase. In the rest of this section, the handwritten Indian number 3(٣) has been taken as an example to demonstrate how the different phases of INHRM were applied.

The image input resolution was taken to be 83x67 undergoes enhancement to remove the noise and unwanted objects then zoning is applied to get image spatial resolution of 47x31 and thinning is applied to get the Skelton of the object (one-pixel width).

Figure 11 shows the processes mentioned above. After that, INHRM begin the search in the image to find the starting point(s) that is/are one pixel with one neighbor.

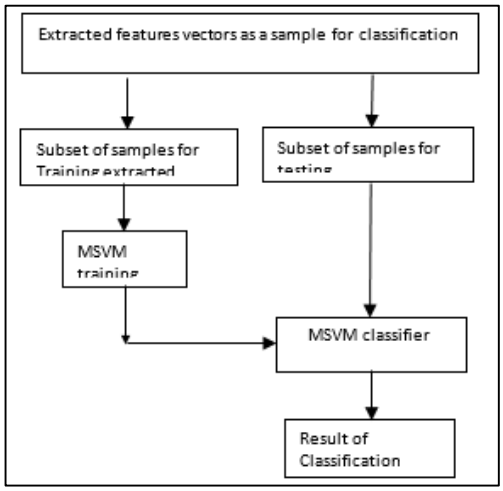

Figure 10: Block diagram of MSVM based classification phase

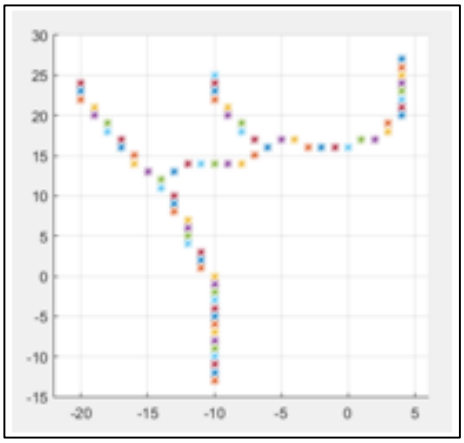

Figure 11: Preprocessing for the image of number 3(٣).

The number of starting points for the under consideration example are found to be 4 starting points at indexes (4,27), (5,14), (6,4), (44,13). So the number of starting points will be taken as a first one of the 16 features vector. Then INHRM searches for intersections which are the pixels that have more than two neighbors, and there were 2 intersections at the indexes (14,18), (18,10). After that the image is divided into four regions and the average pixel intensity values for these four regions were calculated and the values 21.434, 15.937, 14.875, 0 were obtained as shown in Figure 12.

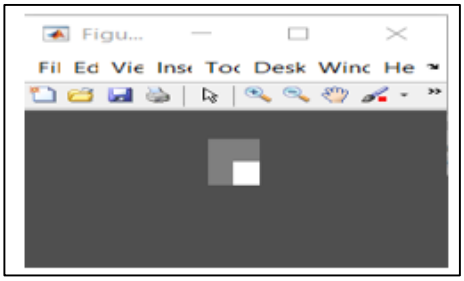

Figure 12: Average zoning for number 3(٣).9

The Calculation of the chain vector is then reached and that was done through start at one of the previously determined starting point indexes and visit all the neighbors with storing the movement directions depending on the base of 8-connectivity as shown in Figure 13. For number ٣, the size of the obtained chain vector was 99 and the frequency is computed as 51 and by using the normalization the size of the final normalized chain vector (*NV*) will be 10 elements.

Table 2 represents the accumulated data of features vector for number 3(٣) which is vector of 16 elements that will be used in the next classification phase of INHRM model. In fact, a chain vectors of the same of both of length and procedure has been calculated for each image of the other classes of handwritten Indian numbers.

For the last phase of INHRM which is the classification phase there were 10 classes according to the ten handwritten Indian numbers (٠ through ٩). The training set was consists of 500 samples with 50 samples for each class, while the testing set consists of 100 samples with 10 of them for each of the ten classes. The classification method used in this phase is one verses all. Table 3 shows the results obtained for the classification phase of INHRM model. The results shows that a percent of about 97% (96.9% to 97.5%) is achieved for the different 10 classes of handwritten Indian numbers.

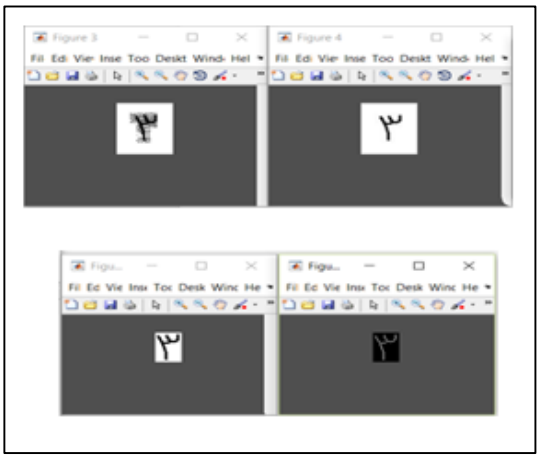

Figure 13: Tracking for number 3(٣) to find its chain vector

Table 2: The features vector for number 3

Start points	Inter-section points	Average zoning				*NV*				
4	2	21.434	15.937	14.875	0	3 4	3 3	5 3	4 3	5 3

Table 3: Result of classification phase

Indian Number (class)	No. of samples for training	No. of samples for testing	Percent of correct recognition
٠	50	10	97%
١	50	10	97.5%
٢	50	10	97.4%
٣	50	10	96.9%
٤	50	10	97.1%
٥	50	10	97%
٦	50	10	97.1%
٧	50	10	97.2%
٨	50	10	97.2%
٩	50	10	97%

4. Conclusions

This paper proposes an Indian handwriting features extraction and classification model depending on Multi-Class SVM. The proposed model consists of four main phases. The first is the image acquisition, which scans and acquires the interested handwritten number at 300 dpi to obtain a proper quality. The next phase is the image preprocessing which involve the image enhancement and noise removing through converting the image to grayscale and making a thinning to have skeleton of the interested object.

The features extraction is the third phase where a features vector of 16 elements are obtained through calculating four different factors which are the starting points of one element, the intersection points of one element, the average zone of four elements, and the normalized chain vector of ten elements.

There were 600 samples of the features vectors due to the ten classes of the handwritten Indian numbers (٠ through ٩) with 60 samples for each class. These 600 samples were delivered for the fourth phase of the proposed model, which is the MSVM based classification phase. The MSVM-classifier is trained using 500 samples while the rest 100 samples are used for the testing. The results showed the proposed model for Indian handwritten numbers achieved around 97% of the correct recognition, which is considered relatively high performance percentage. As a future work, the proposed model can be expanded to be used to classify the Indian handwritten alphabetic letters.

5. References

[1] A.R. Alaei, U. Pal and P. Nagabhushan, "Using Modified Contour Features and SVM Based Classifier for the Recognition of Persian/Arabic Handwritten Numerals," Seventh International Conference on Advances in Pattern Recognition, IEEE 2009.

[2] S.A. Yadav, S. Sharma and S.R. Kumar "A Robust Approach for Offline English Character Recognition," International conference on futuristic trend in computational analysis and knowledge management, 2015.

[3] M. Kumar and M.K. Jindal, "SVM Based Offline Handwritten Gurmukhi Character Recognition," 2011.

[4] G.Y. Tawde and J.M. Kundargi, "An Overview of Feature Extraction Techniques in OCR for Indian Scripts Focused on Offline Handwriting," IJERA, 2013.

[5] O.D. Trier, A.K. Jain and T. Taxt, "feature extraction methods for character recognition," Appeared in Pattern Recognition, Vol.29, No.4, 1996.

[6] H. Izakian, S. A. Monadjemi, B. T. Ladani, and K. Zamanifa "Multi-Font Farsi/Arabic Isolated Character Recognition Using Chain Codes," World Academy of Science, Engineering and Technology 43, 2008.

[7] P. Kumar, N. Sharma and A. Ranas "Handwritten Character Recognition using Different Kernel based SVM Classifier and MLP Neural Network," International Journal of Computer Applications (0975 – 8887) Volume 53– No.11, September 2012.

[8] M. Abdel Fattah. "The Use of MSVM and HMM for Sentence Alignmen," Journal of Information Processing Systems, Vol.8, No.2, June 2012.

[9] D. Nasien, H. Haron and S. Yuhaniz, "Support Vector Machine (SVM) for English Handwritten Character Recognition," Second International Conference on Computer Engineering and Applications (ICCEA), p.249-252, 2010.

[10] E.I. Abbas and T.A. Hashim, "Iraqi Cars License Plate Detection and Recognition System using Edge Detection and Template Matching Correlation," Eng. & Tech. Journal, Vol. 34, Part (A), No. 2, 2016.

[11] A.H. Rassoul, C.Y. Bucheet, and M.I. Abd-Almajied, "Off-Line Arabic Signature Recognition Based on Invariant Moments Properties," Eng. & Tech. Journal, Vol. 29, No. 10, 2011.

[12] A.K. Abdul Hassan, and M.S. Kadhm, "An Efficient Image Thresholding Method for Arabic Handwriting Recognition System," Eng. & Tech. Journal, Vol. 34, Part (B), No. 1, 2016.

Author's biography

Hassan Awheed Jeiad: Received the B.Sc. degree in 1989 in Electronics and Communications Engineering from University of Technology, Baghdad, Iraq. He is received the M. Sc. degree in Communication Engineering from University of Technology, Baghdad, Iraq in 1999. He is received the Ph.D. in 2006 in Computer Engineering from University of Technology, Baghdad, Iraq. He is currently a lecturer in the Department of Computer Engineering in the University of Technology, Baghdad, Iraq. His research interests include computer architecture, microprocessors, computer networks, multimedia, adaptive systems, and information systems.