A Two Stage Method for offline Farsi/Arabic handwritten word recognition systems

Elham bayesteh
Department of Electrical Engineering, Shahrood University
shahrood , Iran
Bayesteh.el@shahroodut.ac.ir

Alireza ahmadyfard
Department of Electrical Engineering, Shahrood University
Shahrood , Iran
ahmadyfard@shahroodut.ac.ir

Abstract. **In this paper a two-stage approach for offline Farsi/Arabic handwritten word recognition systems is proposed. The first stage contained cluster similar word images using upper and lower profiles, vertical projection profile and black/white transition that be used to eliminate lexicon search space .The ISOCLUS algorithm is used to cluster handwritten word images of training dataset. The initial center of clusters is determined from agglomerative hierarchical clustering algorithm.**

The second stage involved recognition of an unknown word from a subset of candidate words obtained in the first stage. The adaptive gradient feature and KNN classifier used in this stage.

Experimental results on a set of 17,000 images of 503 words show a promising result. It yields a lexicon reduction of 77% with accuracy of 94%.The proposed lexicon reduction and the adaptive gradient feature achieve 5% and 13% improvement in recognition rate respectively. *(Abstract)*

Keywords: Farsi Handwritten recognition; Isoclus algorithm; agglomerative hierarchical clustering; dynamic time warping; wavelet transform and adaptive gradient feature.

I- INTRODUCTION

Offline handwriting word recognition systems are capable to interpret handwriting word extracted from images. Some of practical applications of these systems are automatic reading of postal address, bank checks and form. Most of the researches have been mainly focused on Latin and Chinese scripts [1]. Farsi/Arabic scripts have special characteristics such as (1) these have the cursive nature even in machine printed form, (2) letter shape is sensitive to its position in the word and (3) writing style varies for the different people (4) 18 out of 32 Farsi characters have dots appearing in groups of one, two or three. These specifications make Farsi/Arabic handwritten word recognition process difficult.

The Current off line handwriting word recognition systems in order to reduced obscurity and reach a high performance are only able to recognize words in a limited list, typically comprised of 30–1000 words. This limited list is usually called lexicon. In systems with large lexicon, the recognition task becomes more difficult, as the probability of similar words' existence in the lexicon, computational complexity, and the required time for recognition are increased [2]. To reduce the problems arising from a large lexicon and improve the recognition accuracy and speed, some systems remove the number of words dissimilar the test word. This process is called lexicon reduction or lexicon pruning, and can be as an important pre-processing in the HWR systems [4].generally, the lexicon reduction used the input pattern characteristics in limiting the size of the lexicon. Mozaffari et al [3] Proposed lexicon reduction using dots for offline handwriting Farsi word recognition.

In this paper we proposed two-stage method for offline Farsi handwritten word recognition. In first stage is lexicon reduction, we cluster similar words of database using statistical features. These features are simply extracted from word images and properly categorized words into a number of clusters. The nearest clusters to the test word image are selected using DTW algorithm. The words of selected clusters are used in the next stage. This stage eliminated lexicon entries in recognition a test word and this improved the recognition accuracy and speed by removing classifier confusion. The second stage contained recognition of the test word image from the selected words. We extracted gradient feature from adaptive blocks of the words images. These features are utilized to train the K-NN classifier for classification.

II- **T**HE **P**ROPOSED **R**ECOGNITION **S**YSTEM

In this section we introduce our method proposed for recognition of Farsi handwritten words. More specifically we work on database of city names in Iran. The database included 503 city names where for each prototype a number of handwritten samples are provided. Total number of handwritten words in the database is 17,000 words. The proposed method contained two stages: lexicon reduction and recognition. To categorize similar words improve accuracy of word recognition system. Training phase of the lexicon reduction, involves three stages: pre-processing, feature extraction, and clustering. Pre-processing plays an important role in word recognition systems. After pre-processing of each word image we extract a number of features from it. We use four features vectors: vertical projection, upper and lower profiles and

vertical black/white transitions. These features described the general form of the word image. In order to measure the similarity between two words we used Dynamic Time Warping DTW [7]. We finally applied Isoclus algorithm to cluster handwritten words.

Given an input word, corresponding feature vector of word image is extracted. Distance between the extracted feature vector and the center of word clusters (provided from training stage) are measured. A number of closest clusters to the test word are chosen. The words in the selected clusters used for the recognition of an unknown word. In the second stage of algorithm, the adaptive gradient features that described detail of the word image, extracted from these word images and finally the KNN classifier is applied to decide to which class an unknown word belongs.

III- **P**RE-**P**ROCESSING

Preprocessing is an important stage of the system because all features will be extracted from the output of this step. The preprocessing consists of the following steps:

- *Binarization:* in this process the colored image is converted into gray scale and compared against a predefined threshold. The result would be a binary image. The pixels that belong to foreground are set to zero (black) and background pixels are set to one (white).

- *Noise removal:* The quality of binarized image is degraded by noise which is usually resulting of imperfect writing or scanning process. Noise appears as isolated small regions or as irregular edges on characters, which can be removed by median filter and morphological closing and opening operations with a 3x3disk as the structure element. Some of contour discontinuities are removed in this step.
- *Free space removals:* in this stage the free space between two sub-words, both vertical and horizontal lines, are removed.
- *Skew/slant correction and baseline normalize:* for reduce variation between samples of the word we detected and corrected skew/slant angle [14]. Then the baseline of word that is maximum value in the horizontal projection histogram is estimated and the image is enlarged virtually with some blank rows so that the baseline is located in the middle of the image [3].
- *Stretching:* The aim of this step is to remove the overlaps between the connected parts of a word in handwritten word images. The algorithm involves the following steps: First determining the baseline of the word. Then, by tracing the baseline and its nearest lines, the algorithm determines the connected parts that are sufficiently large and cross the baseline or its nearest lines. Afterwards, the algorithm adds spaces between the selected connected parts to increase the space between them [5]. A sample word image after stretching is illustrated in Fig.2. In script there are places where two connected parts are continuous and it is difficult to separate these connected parts.

(a) (b)

Figure1- Shows the image of one word before (a) and after (b) stretching

- *Size Normalization:* normalization is resizing images to a predefined size for proper representation and feature extraction. In this paper, we tested different normalization methods and different sizes of the images for normalization [6] and found that linear normalization with size 125*125 gives the best answer.

IV –**F**EATURE **E**XTRACTION

For clustering of handwritten words we have to extract representative features from each word. The main aim of clustering is reducing the categories of candidates for input word. The reduction speed is a critical issue for a lexicon reducer. In compare to a direct word classifier, a lexicon reducer must decide much faster. Therefore, the word features for clustering must simply being extractable and well describe the shape of word. We extract four feature vectors from each word to describe its image [7]. In our experiments we show the performance of word clustering for individual feature vectors and their combinations.

- *Vertical Projection Profile:* This profile is obtained by projecting word image vertically. By this feature we capture the distribution of ink along vertical projection. For each column of word image the value of profile is computed by counting the number of the back pixels in binary image. Thus the size of this profile is the number of image columns.
- *Upper and Lower Word Profiles:* Word profiles capture part of the outlining shape of a word. In our study, we used the upper and lower profiles. The upper (lower) profile is a vector which is calculated by moving along the upper (lower) boundary of the word's bounding box and recording for each image column the distance to the nearest "ink" pixel in that column.
- *White/Black Transitions:* To show the internal structure of a handwritten word, we used white/black transitions feature. We extracted the number of background to foreground transition from each image column. The size of this feature vector is number of image columns.

So, from each word, four features are extracted and combined into a single set of multi-variant samples defined

as follows. For each image ***I*** with height ***h*** and width ***w***, we extract a set X (I) =x_1, x_2,…,x_w, where each

$$X_i = (f_1(I,i), f_2(I,i), f_3(I,i), f_4(I,i))^T.$$
$$0 \le f_k(.,.) \le 1, k = 1,2,3,4.$$

This makes X (I) a set of 4-variate samples of length ***w***, where the f_i are the four extracted feature vectors of each image [7]. After feature extraction, we applied one dimensional Discrete Wavelet Transform (DWT) on each vector to smooth the uneven edges of image word that may be created during the writing and scanning .This process also reduce the dimensionality of feature vector. [1].therefore, in our experiments the dimension of each feature vector reduced to 32 and the dimension of each set is 4x32. The Distance between the sets obtained from the handwritten word images are computed by DTW algorithm.

V -**DTW A**LGORITHM

Although the size of the word images is equalized in preprocessing stage, but variations of the inter-character and intra-character spacing, in the handwriting of different people, is high. DTW offers to find a more flexible way to compensate for these variations [7]. So we use this algorithm to find distance between two word images in clustering process. This approach improves word clustering significantly.

VI - **C**LUSTERING

To produce a pictorial dictionary based on holistic shape of word images, we grouped handwritten word images based on holistic features extracted from word image .In this paper, we used ISOCLUS clustering algorithm to cluster word images.

- *ISOCLUS clustering algorithm***:** One of the popular clustering algorithms is ISOCLUS algorithm [10] which is based on ISODATA algorithm with minor modifications. The ISODATA method is similar in principle to the K-means procedure as the cluster centers are iteratively determined [9]. One significant advantage of ISOCLUS is that it allows the number of clusters to be automatically adjusted during the iteration by merging similar clusters, splitting clusters with large standard deviations and removing small clusters, therefore the final number of clusters may be different from the initial number of clusters. The further detail is provided in [10]. The algorithm is given a number of user-defined parameters- including:

1- The initial center of clusters
2- Maximum number of clusters
3- Minimum number of samples in a cluster for removing
4- Maximum number of samples in a cluster
5- Maximum standard deviation per cluster for splitting
6- Maximum number of pairs that can be merged per iteration
7- Minimum distance between centers of clusters for merging.

The values of these parameters in our experiments are obtained experimentally.

ISOCLUS algorithm is sensitive to initial center of clusters. Erroneous initial centers may lead the algorithm to produce invalid clusters, so, we used the agglomerative hierarchical clustering algorithm, which is independent of parameter for to determine the initial number of clusters and initial cluster centers [11].

- *Agglomerative hierarchical clustering***:** As mentioned, for initial clustering we applied hierarchical algorithm. The algorithm has different philosophy from the ISOCLUST and K-means algorithm. Specifically, instead of producing a single clustering, they produce a hierarchy of nested clustering. In agglomerative hierarchical clustering algorithm, at the first step, distances between all patterns are calculated. In the next step, the nearest samples are found to form a cluster of two samples. In the next period, the new cluster mean vector considered as input, the first and second steps are repeated. Dendrogram diagram is an effective means of representing the sequence of clusterings produced by the algorithm. The branches represent clusters that have been combined and the height of the branches shows the distance between combined clusters [12]. A level that the merged distance of next clusters is more than its previous levels can be appropriate level to cut the dendrogram. Clustering produced at this level can be the best fit with the dataset.

Representative of each cluster is a sample from the same cluster that has minimal distance with other samples.

$$\sum_{y \in C} d(m_c, y) \le \sum_{y \in C} d(z, y), \forall z \in C \qquad (1)$$

Where d is a distance measure between two samples and $m_c \in C$ is the mean center of C [12].

VII-**T**HE **R**ECOGNITION OF **U**NKNOWN **W**ORD

The second stage of algorithm is the recognition of an unknown word from the subset of candidate words obtained in the first stage. In this stage, we used gradient feature because showed more detail of the word image.

A- **G**RADIENT **F**EATURE

Gradient features are directional features and extracted from the gradient of a grayscale image. To get gradient feature, after the pre-processing phase, at first, a 2 x 2 mean filtering is applied 4 times on the mage. A Roberts filter is then applied on the normalized grayscale image to obtain gradient image. The horizontal gradient g_x and the vertical

gradient g_y an input image $I(x,y)$ were calculated as follows:

$$g_x = I(x+1,y+1) - I(x,y)$$
$$g_y = I(x+1,y) - I(x,y+1) \quad (2)$$

The gradient strength and direction of each pixel were calculated as follow:

Strength $$f(x,y) = \sqrt{g_x^{\,2} + g_y^{\,2}} \quad (3)$$

direction $$\theta(x,y) = \tan^{-1}(g_y / g_x) \quad (4)$$

The gradient directions were quantized to 32 intervals of π/16 each. The next step, the gradient image is divided into $n \times m$ grids with equal number of foreground pixels for each of *n* rows, and equal number of foreground pixels for each of *m* columns. The size blocks of grid in this approach are dependable on the distribution of black pixels of the image. Therefore, each gradient sample is segmented into *n* horizontal segments with approximately equal number of black (foreground) pixels in each segment. It then the image is segmented into *m* vertical segments with approximately equal number of black (foreground) pixels. The intersection of horizontal and vertical segmentation lines define $n \times m$ non-overlapping segments that are used to extract the features in each segment. The segment sizes and x- and y-coordinates are different for each different sample based on the sample black (foreground) pixels' distribution. Figure 2 shows the gradient images of a Farsi word and division of gradient image into 5×5, 7×7 and 5×10 grids. For each segment, the gradient strength was accumulated in 32 directions. By applying this step, the total size of the feature set in the feature vector will be $(n \times m \times 32)$. For reduction of dimension feature vector, number of direction was reduced from 32 to 16 by downsampling with a weight vector [1 4 6 4 1].

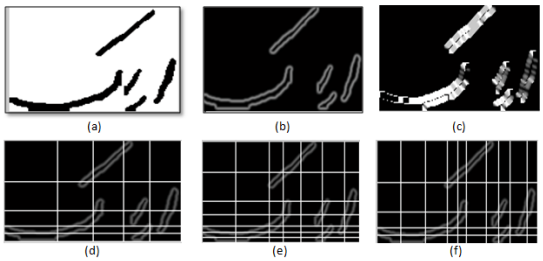

Figure 2- The word image (a), Gradient strength (b), Gradient direction(c), division of gradient image into (d) 5×5, (e) 7×7 and (f) 5×10 grids

B- THE **KNN C**LASSIFICATION

The k-NN is a fast supervised machine learning algorithm which is used to classify the unlabeled testing set with a labeled training set [13]. In order to classify a test word image, the features of the test word image is compared to the training image features of subset obtained from lexicon reduction stage based on cityblock distance. Then, prediction class of the testing image is found based on the minimum distance between the testing word image and the subset training samples.

VII- **M**EASURING **L**EXICON **R**EDUCTION **P**ERFORMANCE

Let L□ be a list of candidate words in the recognition of input image that obtained by lexicon reducer. L□ is subset corresponding lexicon(L).In a lexicon reduction system, the resulting subset does not need to be ranked .The decision if a lexicon entry x_i is in the set L□ can be very efficient in handwriting recognition system, [3]. In order to measure the performance of a lexicon reducer, Let denote the event that the truth lexicon entry X_i is contained in the reduced lexicon L□ by a random variable A, where A=1 if $X_i \in$ L□ and A=0, otherwise. The Extent of reduction is captured by random variable R, defined as R= (|L| - |L□|) / |L| [4]. Three measures of lexicon reduction performance are defined [3]:

-Accuracy of reduction: α = E (A).
- Degree of reduction: q = E(R).
- Reduction efficacy: g = αk. q.

Note that α, q, g ∈ [0, 1]. The accuracy and degree of reduction are usually related inversely to each other. The accuracy α can often be made arbitrarily close to unity at expense of q. The two measures are combined into one overall measure g. The emphasis placed on the accuracy relative to the degree of reduction is expressed as a constant k, which in turn may be determined by the particular application [4].

VIII-**E**XPERMENTAL **R**ESULT

The proposed system has been applied to the database consisting of 17000 images from 502 city name of Iran. For each word, at least 25 samples were provided. The images were divided into training (80%), and testing (20%) sets.

The proposed system for Farsi word recognition consists of two main parts: lexicon reduction and word recognition. In the former, information of global shape word is used while the details are utilized in the later. In the first experiment, performance of the proposed lexicon reduction algorithm is explored and compared to the others. Before this comparison, we selected number of clusters close to the test image. The number of clusters close to the test image *(n)* is an important parameter which the accuracy of word recognition and degree of class reduction are related to it. If we increase *n*, the degree of reduction decreases but the accuracy increases and vice versa. In small lexicons, the accuracy of reduction is more important than the degree of reduction. On the other hand, the degree of reduction can be a critical issue for medium and large databases.

Nevertheless, reduction efficiency combines these two measures and determines the overall performance of the lexicon reduction system [3]. Table I shows accuracy of reduction and degree of reduction as a function of *n*.

Table 1- Degree and accuracy of reduction for different number of nearest clusters in first experiment.

Num of nearest clusters	Degree of reduction (q)%	Accuracy of reduction(a)%	Efficiency of reduction(g)%
1	91.70	75.28	59.86
2	81.21	84.27	68.43
5	**76.87**	**94.12**	**73.35**
7	69.58	94.35	65.64
10	60.54	95.01	57.51
15	47.62	97.07	46.22
20	39.73	98.37	39.08

We compare the performance of proposed method using DTW with the proposed feature vector against the lexicon reduction system proposed in [3] using dots. The results have been shown in Table II.

Table 2- a comparison result of the proposed system and another system

method	Number of word	Degree of reduction (q)	Accuracy of reduction(a)	Efficiency of reduction(g)
sing dots[3]	200	93	85	79
Proposed system	502	77	94	73

For the second experiment, the overall performance of the proposed system was considered. The aim of this experiment was to study the effect of adaptive gradient feature and lexicon reduction on the recognition rate of the basic recognition system.
Table 3 shows effect of gradient feature with adaptive blocks in recognition rate compared with simple block.
Table 4 shows effect of lexicon reduction in recognition rate of proposed system.

Table 3-The recognition result using different feature

Feature method	Recognition rate %
Improved gradient feature	78.21
Simple gradient feature	65.44

Table 4-The recognition rate the proposed method with and without lexicon reduction

Method	Recognition rate%
Without lexicon reduction	78.21
With lexicon reduction	**82.88**

VIIII- **C**ONCLUSION

Lexicon reduction is used to improve recognition accuracy and increase the speed of the process by removing unnecessary entries in the lexicon. In this paper we presented a technique to reduce the lexicon based on clustering handwritten word images based on their holistic shape. We used horizontal projection and upper and lower profile and black/white transition features. For reduced dimension we applied 1 dimensional discrete wavelet transform. The distance between samples is computed by DTW algorithm. We employed the ISOCLUS algorithm for clustering. And the initial points obtained from hierarchical algorithm. Experiments carried out on test samples show promising performance results.

In the word recognition stage, we used adaptive gradient feature and KNN classifier for recognition of test word from subset obtained in the lexicon reduction stage.

REFERENCES

1. Zahra bahmani, Fatemh Alamdar, Reza Azmi, Saman Haratizadeh. '' Off-Line Arabic/Farsi Handwritten Word Recognition Using RBF Neural Network and Genetic algorithm '', 978-1-4244-6585 9/10/$26.00 ©2010 IEEE.
2. A. L. Koerich, R. Sabourin, C. Y. Suen,''Large vocabulary off-line handwriting recognition: A survey'', Pattern Anal Applic .2003
3. Saeed Mozaffari, Karim Faez, Volker Ma¨rgner, Haikal El-Abed, '' Lexicon reduction using dots for off-line Farsi/Arabic handwritten word recognition'', Pattern Recognition Letters 29 (2008) .
4. Safwan Wshah, Venu Govindaraju, Yanfen Cheng and Huiping Li, '' A Novel Lexicon Reduction Method for Arabic Handwriting Recognition'' 1051-4651/10 $26.00 © 2010 IEEE
5. Zaher Al Aghbari, Salama Brook,'' HAH manuscripts: A holistic paradigm for classifying and retrieving historical Arabic handwritten documents'', Expert Systems with Applications 36 (2009).

6. Cheng-Lin Liu, Kazuki Nakashima, Hiroshi Sako, Hiromichi Fujisawa, ''Handwritten digit recognition: investigation of normalization and feature extraction techniques'', Pattern Recognition 37 (2004) 265 – 279
7. Toni M. Rath and R. Manmatha,'' Word Image Matching Using Dynamic Time Warping '' 1063-6919/03 $17.00 © 2003 IEEE
8. H. Sakoe and S. Chiba: Dynamic Programming Optimization or Spoken Word Recognition. IEEE Trans. on Acoustics, Speech and Signal Processing 26 (1980) 623-625.
9. PCI Geomatics Corp.,"ISOCLUS–Isodata clustering program , " http://www.pcigeojmatics.com/cgibin/pcihlp/ISOCLUS
10. Nargess Memarsadeghi, David M. Mount, Nathan S. Netanyahu, Jacqueline Le Moigne,'' A Fast Implementation of the ISOCLUS Algorithm'',0 – 7803 – 7929 -2/03/$17.00 © 2003IEEE
11 . حسين خسروي، احسان اله كبير،" ارزيابی روشهاي بازشناسی متون فارسی بر مبناي شكل كلی زير-كلمات " نشريه مهندسی برق ومهندسی كامپيوتر
12. Sergios Theodoridis , Konstantinos Koutroumbas," Pattern Recognition'' fourth-edition . Copyright © 2009 Elsevier Inc.9781597492720.52057.
13. Jawad H AIKhateebl, Fouad Khelifil, Jianmin Jiani, Stan S Ipsonl,'' A New Approach for Off-Line Handwritten Arabic Word Recognition Using KNN Classifier'', 978- I -4244-5561-4/09/$26.00 ©2009 IEEE
14. Faisal Farooq, Venu Govindaraju, Michael Perrone," Pre-processing Methods for Handwritten Arabic Documents" 1520-5263/05 $20.00 © 2005 IEEE