Off line handwritten Farsi word recognition using improved gradient feature and KNN classifier

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

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

Abstract: **The recognition of Farsi handwriting is drawing increasing attention in recent years. This paper presents a simple and effective approach using improved gradient feature and KNN classifier for the recognition of handwriting Farsi words. The approach involved three stages: pre-processing, feature extraction and classification. Firstly the word images normalized and then gradient feature at the sub-regions level from word images extracted finally the resulting feature vectors are used to classify the words using K Nearest Neighbor classifier (KNN).**
The proposed algorithm tested on a dataset of Iran city names. We compared improved gradient feature with gradient feature. The experimental results show that the proposed algorithm improves the recognition rate about 13 percents. ***(Abstract).***

Keyword: offline handwriting word recognition, gradient feature, KNN classifier

1. 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 etc [1]. Off line Handwritten Farsi Word Recognition which is a challenging process at handwritten word recognition (HWR), has been studied in recent years. The reasons behind this difficulty are the cursive nature both in handwritten and printed forms, the sensitive letter shapes to its position in words (table .1), large variability in handwriting style from person to another person.

Basically, there are two different categories of systems for the recognition of Farsi scripts: segmentation based and segmentation free based. The segmentation based approach needs to segment words into characters or letters for recognition, and is an analytical approach. The segmentation free approach is a global approach, which uses the whole word image in recognition and does not require segmentation steps [2]. Due to the cursive style in Farsi, the classification using a global approach is more efficient and optimal. Dehghan et al [3] proposed a holistic approach for recognition of 198 handwritten Farsi words using histogram of chain-code direction of the image strips and discrete hidden markov model. In this paper, we present a simple and effective technique for handwritten Farsi word recognition based on a holistic feature.

2. The Proposed Recognition System

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 word images in the database is 17,000 images. This method contained three main stages: pre-processing, feature extraction and classification. In an HWR system, the recognition rate depends on a number of factors. Two very important factors are the quality of input images and effectiveness of pre-processing [7]. Once the sample image is acquired, pre-processing is required to enhance the signal for better performance. After pre-processing, gradient features are extracted from adaptive blocks for each word image. Finally, the KNN classifier is applied to decide to which class an unknown word belongs.

3. Pre-processing

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 result 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 [4].
- *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 Figure.1. In script there are places where two connected parts are continuous and it is difficult to separate these connected parts.

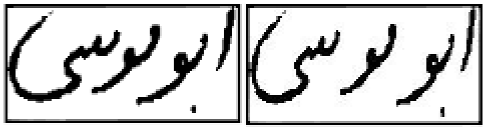

Figure 1. The image of one word before and after stretching

- 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.

4. Feature Extraction

For handwriting recognition, different features like chain codes, structural features, statistical features, curvature features, projection profile and gradient features (directional features) have been used in different studies. In our study, we extracted gradient feature from each image.

Gradient features are directional features and can be extracted from the gradient of a grayscale image for a handwritten text. A gradient vector of discrete direction is decomposed into two components by specifying a number of standard directions (chain code directions). In our gradient feature extraction phase, after the pre-processing phase, each image of size 125×125 pixels was converted back into a grayscale image using a 2×2 mean filter 4 times. 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:

$$\begin{aligned} g_x &= I(x+1, y+1) - I(x, y) \\ g_y &= I(x+1, y) - I(x, y+1) \end{aligned} \qquad (1)$$

The gradient strength and direction of each pixel $I(x, y)$ were calculated as follow:

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

direction $$\theta(x, y) = \tan^{-1}(g_y / g_x) \qquad (3)$$

In equation (3), $\theta(x, y)$ returns the direction of a vector (g_x, g_y) in the range of $[-\frac{\pi}{2}, \frac{\pi}{2}]$. These 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*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].

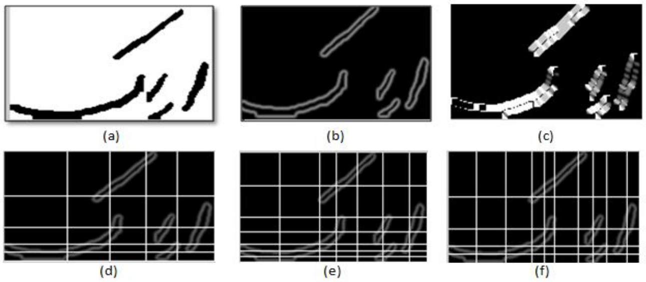

Figure 2 .The word image (a), Gradient strength (b), Gradient direction(c), division of gradient image into 5×5 (d), 7×7 (e) and 5×10 (f) grids.

5. The K_NN Classification

The k-NN is a fast supervised machine learning algorithm which is used to classify the unlabeled testing set with a labeled training set. In order to classify a word image, the features for the word image is compared to the training features based on their similarity. Then, prediction class of the testing image is found based on the minimum difference, measured by a distance criterion, between the testing word image and the training samples [7].

6. Experimental Result

In order to evaluate the performance of the proposed recognition system, several experiments were conducted on a database containing 17000 Farsi word images from 503 names of Iran cities. For each word, at least 25 samples were provided. The images were divided into training (80%), and testing (20%) sets. We experimented the algorithm with different number of divisions. Fig 3 shows symmetry of this test. It is clear from the figure that choosing a 5×10 division gives the best results with a recognition rate of 78.75.The 5×5 division achieved close results (78.21%) with considerable reduction in the number of features and processing time.

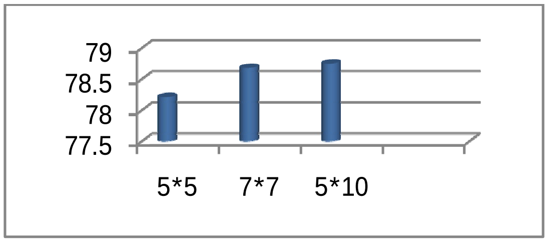

Figure 3. Recognition rate at different divisions

To test the effectiveness of improved gradient feature with adaptive segmentation, we compared that with gradient feature with simple segmentation. In the simple segmentation, size segments are same. The result classification shows in table 2.

Table 1. k-NN Recognition Results Using Different Features

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

We also showed compared results the proposed algorithm with another algorithm in table 3.

Table 2 . Recognition rate of proposed method compared to other methods

algorithm	*Size of dataset*	*%Correct top 1*
FVQ+HMM +chain code feature [10]	198	67
DHMM+translate feature[9]	200	73
Svm+gradient feature[8]	57	95
Proposed algorithm	503	78

7. Conclusion

We have proposed a system to use KNN for the classification of handwritten words. The system has been applied to the dataset of names of Iran cities. We used gradient feature and improved that by adaptive segmentation based distribution black pixel of the word image. The experimental results show that the proposed feature extraction method improved recognition rate about 13percents.

8. 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]. Jawad H. AlKhateeb, Jianmin Jiang, Jinchang Ren, Fouad Khelifi and Stan S. Ipson," *Multiclass Classification of Unconstrained Handwritten Arabic Words Using Machine Learning Approaches* "The Open Signal Processing Journal, 2009, 2, 21-28.

[3]. M. Dehghan, K. Faez, M. Ahmadi, M. Shridhar," *Handwritten Farsi (Arabic) word recognition: a holistic approach using discrete HMM*" Pattern Recognition 34 (2001) 1057.1065

[4]. Faisal Farooq, Venu Govindaraju, Michael Perrone,'' *Pre-processing Methods for Handwritten Arabic Documents*'', 1520-5263/05 $20.00 © 2005 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]. 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

[8]. Malik Waqas Sagheer, Chun Lei He, Nicola Nobile, Ching Y. Suen, " *Holistic Urdu Handwritten Word Recognition Using Support Vector Machine* ", 2010 International Conference on Pattern Recognition.

[9]. 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) .

[10]. Dehghan, M., Faez, K., Ahmadi, M., Shiridhar, M., 2001b. " *Unconstrained Farsi handwritten word recognition using fuzzy vector quantization and hidden Markov models. Pattern Recognition Lett* ". 2, 209–214.