2012 4th International Conference on Bioinformatics and Biomedical Technology
IPCBEE vol.29 (2012) © (2012) IACSIT Press, Singapore

Offline Persian Writer Identification Based on Wavelet Analysis

Mohammad Akbari[1+], Reyhaneh Eslami[2], M. H. Kashani[3]

[1]Department of engineering of Islamic Azad University, Shahr-e-Qods Branch, Tehran, Iran.().

[2]Reyhaneh Eslami was with Ferdowsi University of Mashad. She is now with Iran Argham company,Tehran, Iran.

[3]Mostafa H. Kashani is with Islamic Azad University, Islamshahr Branch.

Abstract. In this paper we introduce a new efficient approach for writer identification and verification based on written style. At first, the handwritten text image is normalized then the handwritten features are extracted by wavelet transform and KNN1 co-occurrence matrix. At last, handwritings are classified by a classifiers and the writer is being identified. Experimental results on variety of handwriting databases confirm the efficiency of this method. Writer recognition rate of this method is 93.3%.Numerical results are presented in continue.

Keywords: Writer Identification, Wavelet analysis.

1. Introduction

Writer identification has been used recently in different applications like security, courts and access control systems. The goal of writer identification is to determine the writer of a specified note among others. The goal of writer verification is to confirm if two handwritten texts have the same writer. Writer identification and verification methods are either text-dependant or text-independent.

In text-dependant methods a fixed text is used for writer verification but in text-independent methods every single text can be used for writer verification. These methods can be used either as out-line with dynamic information or as in-line with scanned images.

2. Previous Related Works

Vertical projection profile coding with Morphology method [5] widely uses horizontal and vertical projection for cantor features extraction, signature analysis and handwritten characters detection. This method uses a handwritten text image and normalizes its empty spaces and then calculates its vertical projections. Vertical projections are categorized and processed with Morphology operators so that desired feature vector is extracted.

Handwriting's features based method [7] utilizes handwriting's features that are mainly concerned as visible features like width, gradient, and height of 3 main handwritten areas. Legibility has also been used as a manner-based feature. This method uses two different categorization procedures: K-Nearest Neighbor (KNN) and feed-forward neural networks. A sub-collection of IAM database is also being utilized. The database includes 100 pages of 20 writers and each page includes 5-11 lines.

The edge direction based method [9] uses Firemaker database which includes 250 Dutch handwritings. In this study the performance of edge direction probability distribution has been compared with some non-directional features used for writer identification. The study shows that the combination of angle probability distribution between two edges with some features has better performance than other features.

[+] Corresponding author
Email: m.akbari@Qodsiau.ac.ir

In hidden Markov model based method [4] a HMM^2 is assigned to every writer and trained with their handwriting. All models have the same structure but their parameters, transfers and output probabilities are different; therefore, every HMM model can recognize a special writer like an expert.

The idea of Connected-Component Contours (CO3) based method [8] is that each writer is recognized by a coincident pattern generator that generates a set of connected components for capital letters. This method is assessed only for capital letter handwritten texts. In order to identify (the) writer, the coincident pattern generator is modeled according to a CO3 code book that has been run on a training collection of 100 writers and the probability distribution function is calculated for each writer.

3. Proposed Writer Identification Algorithm

The general structure of offered algorithm is shown in figure1. As shown in figure 1, extracted features are not context-dependent. In other words the computer is totally unaware of the context and the training text is different from the test text.

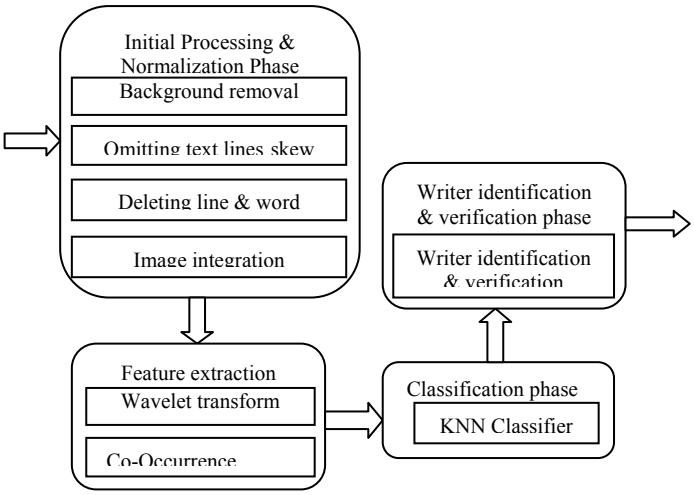

Fig. 1: General structure of suggested algorithm

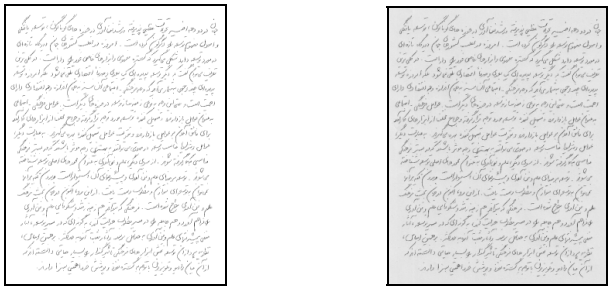

Fig. 2: Removing background of handwritten text by Otsu algorithm

Fig. 3: Elimination of handwritten text lines' skew and deviation

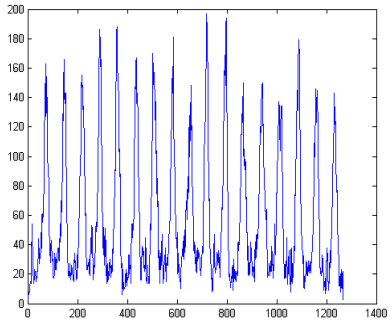

Fig. 4: Vertical projection profile of handwritten text image

3.1. Initial Processing Phase

In this step we try to reduce the fault probability of results by initial processing of handwritten text images. Factors like variation in colors, size and paper type or skew and variation of text lines are the cause of fault results. Normalization phase includes 4 separate steps.

At first the image gray scale threshold is calculated by Otsu algorithm. The Otsu algorithm is used for automatic thresholding of projection profile of images. The algorithm assumes that the image includes 2 classes of pixels: black and white (background and foreground) and calculates their disjunctive optimum threshold level. In order to eliminate every correlation to paper texture, image's background is removed by the acquired threshold.

Otsu Algorithm: In this method, we try to find a threshold that reduces inter class variance and is defined as below:

1) $\sigma_w^2(t) = w_1(t)\sigma_1^2(t) + w_2(t)\sigma_2^2(t)$

$\boldsymbol{\omega_i}$ weighs are the distribution probability of classes with t threshold and $\boldsymbol{\sigma^j_2}$ is the variance of 2 classes. Otsu proves that inter class variance decrement results in intra class variance increment.

2) $\sigma_b^2(t) = \sigma^2 - \sigma_w^2(t) = w_1(t)w_2(t)[\mu_1(t) - \mu_2(t)]^2$

In the above equation $\boldsymbol{\omega_i}$ is the probability of classes and $\boldsymbol{\mu_i}$ is the mean of classes and is calculable recursively. The algorithm steps are as below:

1. Calculating image projection profile and the probability of each color intension.
2. Initializing $\boldsymbol{\omega_i\,(0)}$ and $\boldsymbol{\mu_i\,(0)}$.
3. Updating $\boldsymbol{\omega_i}$ and $\boldsymbol{\mu_i}$ values.
4. $\boldsymbol{\sigma_b^2\,(t)}$ Calculation. Greatest $\boldsymbol{\sigma_b^2\,(t)}$ is the favorite threshold.

After calculating the gray scale level threshold of each handwritten text image, pixels that have greater value than threshold get value of 255 (equivalent to white color). As a result background pixels that are brighter than text pixels turn to white and handwritten pixels stay unchanged. Figure 2 shows the handwritten text after removing its background.

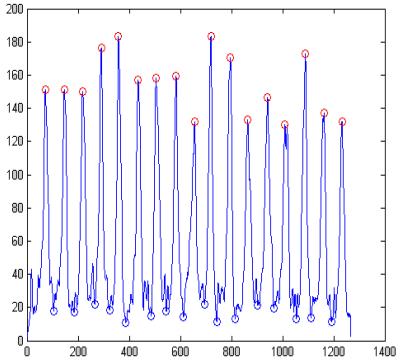

Fig. 5: Vertical projection profile after smoothing and finding local minimums and maximums

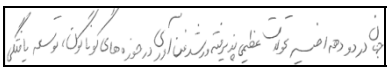

Fig.6: A text lines after segregation and omitting empty spaces

3.2. Detection and Correction of Skew Lines

The proposed Said method [1] that is based on connected components has been used to eliminate the skew and deviation of handwritten text lines. Figure 3 shows this step.

3.3. Removing Vertical and Horizontal Spaces

Texture analysis cannot be directly implemented to handwritten text images (even afterimage's background removal and elimination of text lines' skew), due to the fact that image texture has been affected by empty spaces between words and text lines. Normalization of empty spaces of text will reduce the effectiveness of these factors. The steps are described in below:

1. At first, the gray scale level of text image is binerized by Otsu algorithm and the calculated threshold. In acquired binary image, the value of black pixels is 0 and the value of white pixels is 1. Then the image complement is acquired and its vertical projection profile is calculated. The peaks are centers of text lines and the valleys are the empty spaces between lines in the vertical projection profile. Therefore, lines that represent empty spaces (zero values of projection profile) in image matrix are getting omitted. After that, the projection profile is smoothed by an average method. Finally, local minimums and maximums of projection profile are detected and text lines get segregated by them.
2. In this stage, vertical projection profile of each text lines of previous step gets calculated and empty spaces between the letters and words of a line are found. Empty spaces with greater value than 5 pixels get normalized to 5 pixels.

After text line normalization, the default value of text line spaces will be set to 5 pixels; consequently text line spaces will be 5 pixels in final normalized image. Figures 4-7 shows the steps of extra normalization.

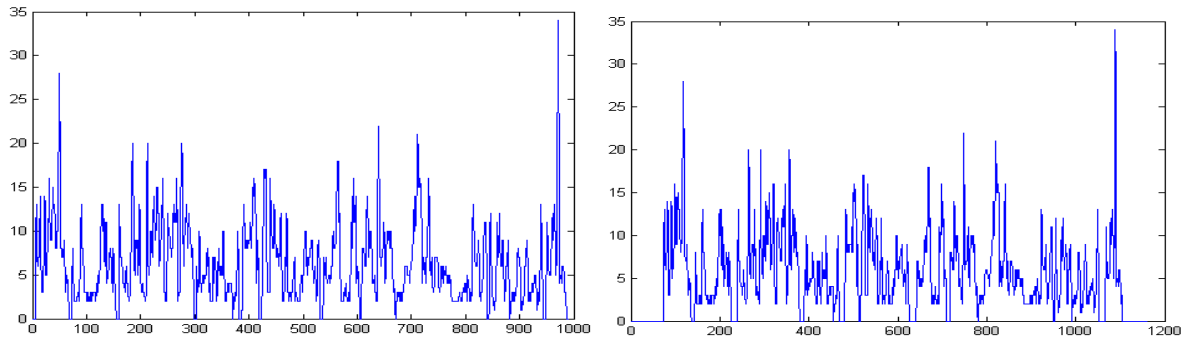

Fig. 7: Vertical projection profile of a text line after removing empty spaces

3.4. Image Integration With Parochial (Text Block) Reiteration

The purpose of this step is to create an accrete text from handwritten text. If the length of normalized line is less than 1200 pixels then the line is getting reiterated so that its length reaches 1200 pixels. If image's last parts are empty, it will be filled by reiteration of text lines so that image's length becomes 1600 pixels. After thorough normalization, the size of monotonous image will be 1600×1200 pixels. In order to extract features in next phase an 800×800 block will be segregated from normalized image.

3.5. Feature Extraction Phase

Table 1: Identification rate of the algorithm based of different features

Wavelet	Selected Feature Co-Occurrence Matrix	Length of Feature vector	Write Identification Rate
En,Md	x	6	56.67
En	x	3	60
Md	x	3	63.33
X	Con,En,Cor,Hom	16	56.67

X	Con	4	56.67
X	Hom	4	23.33
X	Cor	4	83.33
X	En	4	16.67
X	Con,Hom	8	56.67
X	Con,Cor	8	56.67
X	Con,En	8	56.67
X	Hom,Cor	8	90
X	Hom,En	8	23.33
X	Cor,En	8	83.33
X	Cor,Hom,En	12	83.33
X	Con,Hom,Cor	12	56.67
X	Con,Hom,En	12	56.67
X	Con,Cor,En	12	56.67
En	Cor,Hom,En	15	86.67
Md	Cor,Hom,En	15	73.33
En,Md	Cor,Hom,En	18	73.33
En	Cor,En	11	90
Md	Cor,En	11	70
En,Md	Cor,En	14	73.33
En	Hom,Cor	11	90
Md	Hom,Cor	11	70
En,Md	Hom,Cor	14	70
En	Cor	7	93.33
Md	Cor	7	70
En,Md	Cor	10	70

The idea of this paper is to use wavelet transform and co-occurrence matrix simultaneously in order to extract handwriting's features. Wavelet transform converts an image to 4 substitute images with a new measure. These images are: approximate image, horizontal partial image, horizontal partial image and diametrical partial image. The co-occurrence matrix is utilized to calculate and analyze different features of gray scale level images by counting the number of setting two specified pixels besides each other with a definite distance and specific angles i.e. 0º , 45 º, 90º and 135 º .In order to extract features, normalized image of handwritten text should **be** converted to approximate and partial images (horizontal, vertical and diametrical) with Haar wavelet transform. Then the 0º , 45 º, 90º and 135 º co-occurrence matrix of horizontal, vertical and diametrical partial images of wavelet should be calculated. Figure 8 illustrates this step.

3.6. Feature Extraction Phase

At first the normalized image of handwritten text is discomposed with Haar wavelet transform. Then, the wavelet energy is calculated for partial horizontal, vertical and diametrical images. The Bn image energy of wavelet transform includes Mn factor of Wn,m and is defined as below:

$$E_{B_n} = \frac{1}{M}\sum_{m=1}^{M_n} w_{n,m}^2$$

The average of Bn image factors can be calculated as below:

$$MD_{B_n} = \frac{1}{M}\sum_{m=1}^{M_n} \left| w_{n,m} \right|$$

After wavelet feature selection, symmetrical co-occurrence matrix is calculated with d=1(0° for partial horizontal, 90° for partial vertical,45º and 135º for diametrical images). Therefore, only 62 features can be defined for each image (4 directions× 14 co-occurrence matrixes + 3 partial images × 2 wavelet transform

features). Finding best collection of features that represents writer's handwriting is our aim and the best feature vector for an 800×800 normalized image contains 7 features. These features consist of the energy of each partial images plus calculated correlation feature in 4 directions on partial image factors

3.7. Classification and Writer Identification

In a 7-dimensions feature area each coordinate represents a specified feature extraction vector and every vector models a specified writer. K-Nearest Neighbor (KNN) is the most common method of data classification in pattern recognition area and is a sample based learning method. In KNN a sample is classified according to its neighbor's majority vote. It means a sample belongs to a class that has the majority of votes among K nearest neighbors. K is an integer number and if K=1, the sample belongs to nearest neighbor class. In this method, neighbors are training sample images that classification is done according to them. Although there is no explicit learning stage, first paper of handwritten text is assumed training collection.

4. Results

This paper uses a database that includes 30 handwritten texts. The specified text is a 2- A4 page and each page consists of 7-15 lines. The first page is used for system training and the second one is used for evaluation of system writer identification and verification. Writer population includes 18 men and 12 women that are aged 24-54 and have been chosen accidentally from a set of clerks.The recognition rate results and evaluation results are shown in table 1. Different combination of features are shown in table 1and as shown in the table, combination of wavelet energy feature and correlation feature of co-occurrence matrix has the maximum precession. In this study the best writer identification rate is 93.3% for 7 extracted features.

5. Conclusion

In the suggested algorithm, the handwritten text image is analyzed after normalization with multi-resolution analyzer and texture analysis methods and appropriate features for writer identification are calculated. The results of running this method on 30 handwritten text images prove the efficiency and applicability of this method.In conclusion, wavelet and texture analysis based method is a proficient method and is applicable in giant databases also. As features are specified according to data types, this method can be applied for other languages too. Besides all above, this method can be implemented text-independently if each writer has sufficient text.

6. References

[1] H. E. Said, T. N. Tan and K. D. Baker, Personal identification based on handwriting, Pattern Recognition, vol. 33, No. 1, pp. 149-160, 2000.

[2] Schomaker, L. "Advances in Writer Identification and Verification". Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference, Volume 2, page(s) 1268-1273.

[3] Ameur Bensefia, Thierry Paquet, Laurent Heutte. "A writer identification and verification system", Pattern Recognition Letters archive, Volume 26, Issue 13 (October 2005), Pages 2080 – 2092.

[4] Schlapbach and H. Bunke, UsingHMM Based Recognizers for Writer Identification and Verification, IEEE Proc. Of 9th Int. Workshop on Frontiers in Handwriting Recognition, pp. 167-172, 2004.

[5] E. N. Zois and V. Anastassopoulos, "Morphological waveform coding for writer identification" , Pattern Recognition, vol. 33, pp. 385-398, 2000.

[6] Haralick, Robert M. Shanmugam, K. Dinstein, Its'Hak. "Textural Features for Image Classification", Systems, Man and Cybernetics, IEEE Transactions on Nov. 1973, Volume 3, page(s)610-621.

[7] Marti, U.-V. Messerli, R. Bunke, H. "Writer identification using text line based features", Document Analysis and Recognition, 2001. Sixth International Conference on 2001 , page(s) 101-105.

[8] L. Schomaker, M. Bulacu and K. Franke, "Automatic Writer Identification Using Fragmented Connected-Component Contours," Proc. of the Ninth Int'l Workshop on Frontiers of Handwriting Recognition (IWFHR'04), pp 185-190, 2004.

[9] M. Bulacu, L. Schomaker, and L. Vuurpijl, Writer identification using edge-based directional features, In Seventh Int. Conf. on Document Analysis and ecognition, pp. 937- 941, 2003.