ELectronics EXpress

A writer identification method based on XGabor and LCS

Behzad Helli[a)] **and Mohsen Ebrahimi Moghaddam**[b)]

Electrical and Computer Engineering Department, Shahid Beheshti University, G.C, Tehran, Iran

a) *be_helli@std.sbu.ac.ir*

b) *m_moghadam@sbu.ac.ir*

Abstract: Writer identification is a popular research field in many languages such as English, Persian, Chinese, etc. The approaches of writer identification methods are dependent on the language because different languages letters have different pattern. In this paper, we have presented XGabor filter and proposed a language independent writer identification system. In the feature extraction phase of proposed method, Gabor and XGabor filters are used while in the classification phase, a new classification method is defined that is not based on any kind of distances among feature vectors. The proposed classifier uses the sequence similarity of the Sorted Order of Features (SOF). To measure this similarity, the Longest Common Subsequence (LCS) algorithm is employed. In simulation phase, two databases in different languages have been used. First one consisted of 100 people's Persian handwritings and second one had 30 people's English handwritings. The accuracy of the system was satisfactory in both of them.

Keywords: writer identification, LCS, Gabor, XGabor

Classification: Science and engineering for electronics

References

[1] B. Helli and M. E. Moghadam, "Persian Writer Identification using Extended Gabor Filter," *Int. Conf. Image Analysis and Recognition (ICIAR)*, 2008.

[2] Z. He, X. You, and Y. Y. Tang, "Writer identification of Chinese handwriting documents using hidden Markov tree model," *Pattern Recognition Journal*, vol. 41, pp. 1295–1307, 2008-06-15.

[3] M. Bulacu and L. Schomaker, "Text-independent writer identification and verification using textural and allographic features," *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol. 29, no. 4, pp. 701–717, April 2007. Special Issue - Biometrics: Progress and Directions.

[4] V. Eglin, S. Bres, and C. Rivero, "Hermit and Gabor transforms for noise reduction and handwriting classification in ancient manuscripts," *Int. J. Document Analysis and Recognition (IJDAR)*, vol. 9, pp. 101–122, DOI 10.1007/s10032-007-0039-z, Springer, 2007.

DOI: 10.1587/elex.6.623
Received April 05, 2009
Accepted April 13, 2009
Published May 25, 2009

[5] M. Bulacu, L. Schomaker, and A. Brink, "Text-Independent Writer Identification and Verification on Offline Arabic Handwriting," *9th Conf. Document Analysis and Recognition (ICDAR)*, 2007.

[6] M. S. Baghshah, S. B. Shouraki, and S. Kasaei, "A Novel Fuzzy Classifier Using Fuzzy LVQ to Recognize Online Persian Handwriting," *2nd IEEE Conf. Inf. Commun. Technol. (ICTTA)*, 2006.

[7] F. Shahabi and M. Rahmati, "Comparison of Gabor-Based Features for Writer Identification of Farsi/Arabic Handwriting," *10th Int. Workshop on Frontiers in Handwritten Recognition (IWFHR)*, 2006.

[8] V. S. N. Prasad and J. Domke, "Gabor Filter Visualization," Tech. Rep., University of Maryland, 2005.

[9] T. H. Cormen, C. E. Leiserson, R. L. Rives, and C. Stein, "Introduction to algorithms," second edition, The MIT Press, 2001.

[10] Z. He, X. You, and Y. Y. Tang, "Writer identification of Chinese handwriting documents using hidden Markov tree model," *Pattern Recognition Journal*, vol. 41, pp. 1295–1307, 2008-06-15.

[11] Z. He, X. You, and Y. Y. Tang, "Writer Identification using global wavelet-based features," *Neurocomputing*, vol. 71, pp. 1832–1841, Elsevier, 2008.

1 Introduction

The most important measures to determine if a biometric is usable in large scales are uniqueness and collectability of that biometric. Handwritten text is one of the few biometrics which fulfills both. However, the regular handwritten processing methods are not applicable in all languages (e.g. English); for example, the proposed methods for English handwritten are not usable in Persian, because of special characteristics of Persian handwriting such as different styles in writing that has been explained in Ref. [1],

There are different methods in writer identification in different languages. For example, a Gabor filter based method which uses the weighted Euclidian distance (WED) classifier after extracting the features have been proposed for Persian writer identification [7]. Using fuzzy feature extraction and Fuzzy Learning Vector Quantized (FLVQ) has been addressed in Ref. [6] to identify Persian writers with accuracy 90%. This method could only work on disjoint characters that are not conventional in Persian writing.

In a model based Arabic writer identification research, textual and allographic features were used and system accuracy was about 88% [5]. Zhenyu He et al. have presented an offline Chinese writer identification method which used Gabor filter and Hidden Markov Tree (HMT) in wavelet domain. They tested their system in a database with 500 writers and the accuracy in top-1 and top-15 has been reported about 40% and 100% respectively [10]. Also, these authors have used a combination of general Gaussian model (GGD) and wavelet transform on Chinese handwriting in Ref. [11] that its results are not much better than the first method. There are several approaches in English handwritten identification such as methods proposed in Ref. [3] and Ref. [4] that use allographic features and Gabor filter respectively. There

DOI: 10.1587/elex.6.623
Received April 05, 2009
Accepted April 13, 2009
Published May 25, 2009

are more other methods in English, Arabic, Chinese, and other languages to identify writers.

Most classification methods that are used in writer identification systems such as WED (Weighted Euclidian Distance) classifier, nearest neighborhood functions, artificial neural networks, Markov Tree, and fuzzy gain functions are based on closeness of any kind of distance measure between the feature vectors. The main goal of these classifiers is to find out a feature vector from training data that is the closest one to the test data feature vector. In this paper, we propose a new classification method that instead of measuring the closeness of two feature vectors, measures the sequence similarity of Sorted Order of Features (SOFs). The Longest Common Subsequence (LCS) method is used to compare sequence similarity. The proposed method uses this classifier while the feature vectors are created by Gabor and XGabor filters. The presented method has been tested on two databases with different languages. The first one was a Persian handwriting database, which included 100 writer's handwritings (5 pages per writer) and the second one, included 30 people's handwritten in English language that were selected from the IAM[1] database (7 pages per writer). The experimental results were satisfactory for both of them.

The rest of the paper is organized as follows: In section 2, the Gabor and XGabor filters are introduced. Section 3 describes the proposed LCS based classifier. In section 4 the experimental results are shown, and section 5 concludes the paper.

2 Feature extraction using Gabor & XGabor Filter

2.1 Gabor Filter

A 2D Gabor filter is obtained by modulating a 2D sinusoid with 2D Gaussian. The function definition of 2D Gabor filter centered at the origin with spatial frequency Φ and orientation θ is as follows [8]:

$$g(x, y, \theta, \phi) = \exp\left(-\frac{x^2 + y^2}{\sigma^2}\right) . \exp\left(2.\pi.\phi.i.\left(x.\cos\theta + y.\sin\theta\right)\right) \qquad (1)$$

The standard deviation of Gaussian kernel (σ) depends on the Φ value [8].

Let $I(x, y)$ denote the image and $G(x, y, \theta, \Phi)$ denote the response of a Gabor filter on the image plane, then:

$$G(x, y, \theta, \phi) = \iint I(p, q) g(x - p, y - q, \theta, \phi).dp.dq \qquad (2)$$

To extract features, the Gabor filter was generated in 36 different orientations $(0, 5, 10, 15, 20, \ldots, 175)$ in 64*64 blocks. The features were produced by:

$$f_i = \sum_{x,y} |G(x, y, 5i, \phi_0)| \qquad (3)$$

f_i is i^{th} feature where $1 \leq i \leq 36$. With $\phi_0 = 4$ the best results are achieved.

[1] http://www.iam.unibe.ch/fki/databases/iam-handwriting-database

© IEICE 2009
DOI: 10.1587/elex.6.623
Received April 05, 2009
Accepted April 13, 2009
Published May 25, 2009

2.2 XGabor Filter

A 2D XGabor filter is obtained by modulating a 2D circular sinusoid with 2D Gaussian [1]. Let $xg(x, y, \phi, r_x, r_y)$ be the function defining a 2D XGabor filter centered at the origin with Φ as the spatial frequency and r_x, r_y as the horizontal and vertical ratios. It is presented as:

$$xg(x, y, \phi, r_x, r_y) = \exp\left(-\frac{x^2 + y^2}{\sigma^2}\right).\sin\left(\phi.\frac{r_x.x^2 + r_y.y^2}{r_x + r_y}\right) \tag{4}$$

Figure 1 shows an XGabor filter sample. Generating the response for an image is done in the same manner with equation (2).

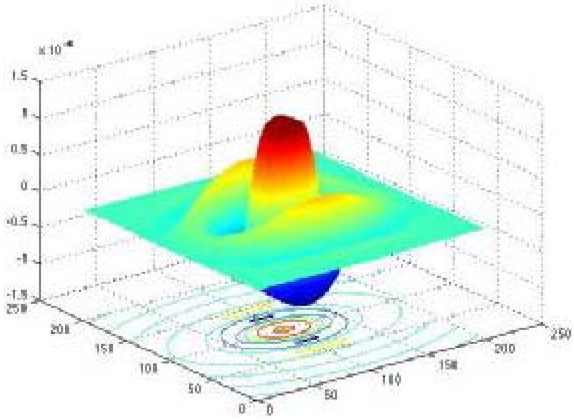


Fig. 1. XGabor filter when $\Phi = 1$, $\mathrm{r_x} = 1$, $\mathrm{r_y} = 3$.

XGabor filter models the curves in the image; therefore, by using it, most frequent curves in input image are detected [1]. In feature extraction phase, XGabor filter was used in 5 different ratios and each one of them were applied in 32*32, 64*64, 128*128 and 256*256 block sizes. In the same manner with equation (3) features have been created, therefore, 20 XGabor based features were generated.

3 The LCS-based classifier

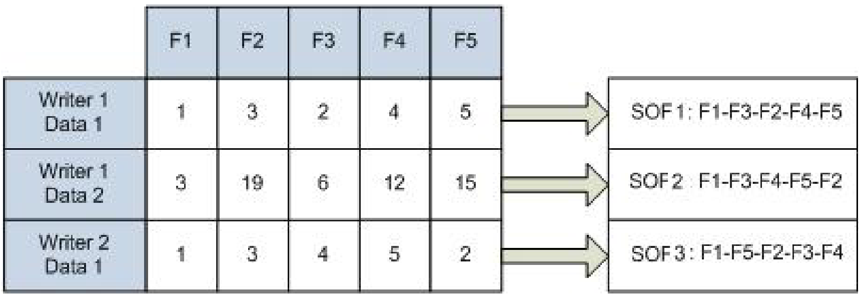
3.1 Classification Criteria

In text-independent writer identification systems, one of the problems that decrease the methods precision is the variety of the words that one writes. Depending on the words that one uses, the whole text direction may be different; for example in English language, “M” or “N” letters has more vertical direction than “S” letter, therefore, the text that has more “M” or “N” than “S” maybe more vertical, hence, the direction of whole text depends on the letters of words. If a system uses texture based features, dependent on the text that someone writes in test phase, some of the feature values may be less/more than the values in train phase, therefore, the result of any classifier that uses distance between two feature vectors, such as WED, Neural Networks, SVM etc. may vary by input.

DOI: 10.1587/elex.6.623
Received April 05, 2009
Accepted April 13, 2009
Published May 25, 2009

In recent systems, in order to avoid this problem, the input data has been normalized, so that the system may be more independent to the words of the text [1, 2, 7].

In the proposed classifier, the features in the feature vector are sorted by magnitudes to make SOF (Sorted Order of Features). By employing the assumption that the magnitude order of directional features in independent-text writing by the same writer is tend to remain unchanged, the sequence similarity of the SOFs is employed to avoid data normalization. This assumption has been confirmed by experiments. Figure 2 shows examples of three feature vectors and their SOFs. If we suppose the Writer1-Data1 as test data and two other rows as train data, then distance between test data feature vector and two other ones are 20.9 and 3.74 respectively by WED, therefore, Writer-2 is selected as answer by this classifier that is a wrong selection. By using sequence similarity of SOFs as classifier, SOF1 is more similar to SOF2 and less similar to SOF3, therefore Writer-1 (second row) is selected that is a true selection. To determine the sequence similarity; LCS algorithm is used which is described in next sub-section. The LCS length of SOF1 and SOF2 is equal to 4 while LCS length of SOF1 and SOF3 is equal to 3.



	F1	F2	F3	F4	F5		
Writer 1 Data 1	1	3	2	4	5	⇒	SOF 1: F1-F3-F2-F4-F5
Writer 1 Data 2	3	19	6	12	15	⇒	SOF 2 F1-F3-F4-F5-F2
Writer 2 Data 1	1	3	4	5	2	⇒	SOF 3: F1-F5-F2-F3-F4

Fig. 2. Some Feature vectors with 5 elements and their corresponding SOF. SOF1 is more similar to SOF2 by using sequence similarity while their feature vectors are far from by WED.

3.2 LCS Algorithm

The LCS (Longest Common Subsequence) algorithm, finds the longest subsequence that is common in input sequences. When the number of sequences is constant, the problem is solvable in polynomial time by dynamic programming. The LCS recursive algorithm with time complexity $\theta(n^2)$ is as follows [9].

$$\begin{cases} M_{i,j} = \max(M_{i-1,j}, M_{i,j-1}) & A_i \neq B_j \\ M_{i,j} = M_{i-1,j-1} + 1 & A_i = B_j \end{cases} \tag{5}$$

A and B are the input sequences, and M is a two dimensional matrix which $M_{i,j}$ is the length of LCS of A from start to i and B from start to j.

© IEICE 2009
DOI: 10.1587/elex.6.623
Received April 05, 2009
Accepted April 13, 2009
Published May 25, 2009

3.3 Implementing the LCS as Classifier

After extracting the features from the documents, a feature vector with 56 features is generated. The features are sorted in increasing order in order to generate their SOF. Then the length of the LCS of test and train data's SOF is calculated, the train data with longer LCS is selected as result.

4 Experimental Results

To test the proposed method, two test benches were created. In the first one, 100 people's handwritings in Persian were gathered; each person wrote 5 different A5 pages with arbitrary texts and styles. Three pages of each person's handwriting were used to train the system, and the other two were used to test the system.

In the second one, 30 people's handwritings in English were selected from IAM database such that each write had been written 7 pages. 4 pages were used for the training and the 3 other ones where used to test.

Because there is more than one trained page, different LCS values are generated for each test data. To identify writer, 5 different approaches were used. Assuming there are 'n' training data per writer, at first the maximum value of the 'n' generated LCS was used to decide. The second approach used the Minimum value. Similarly, the Median, Sum, and Product of these values were used to decide about the writer of the document. After using the LCS classifier, in order to make the system more accurate, WED classifier was used. The WED was applied to the Top-5 outputs and the closer one to test data is selected as the writer. The results are summarized in Table I.

With regards to Table I, the proposed method had great results, especially when it is combined by WED.

Table I. The experimental results of applying proposed method on Persian and English database. WED column shows the results of using only WED on same features. Top-1 column shows the precision of proposed method to find writer directly. The column Top-5-A shows the method precision when the writer is in Top-5, and Top5-+WED column shows the method precision when WED is applied in Top 5 results in final decision.

	Persian Database (100 / 5)				IAM Database (30/ 7)			
	WED	Top-1	Top-5		WED	Top-1	Top-5	
			A+	WED			A	+WED
Maximum	77%	83%	91%	91%	80%	93.3%	94.4%	95%
Minimum		84%	91%	91%		92.2%	94.4%	95%
Median		86%	92%	92%		94.4%	95.5%	96%
Sum		89%	94%	94%		93.9%	95.5%	96%
Product		89%	95%	95%		94.4%	95.5%	96%

© IEICE 2009
DOI: 10.1587/elex.6.623
Received April 05, 2009
Accepted April 13, 2009
Published May 25, 2009

5 Conclusion

In this paper, we proposed a precise method to identify a writer. This method is text and language independent. It consists of two phases; feature extraction and classification. Gabor and XGabor filter in different directions and ratios have been used to extract features and an LCS based classifier has been proposed. The LCS based classifier, is a powerful classifier that may be used in any kind of features. The method was tested on two different database and experimental results were great. In future, we are going to apply this method on some other languages such as Japanese and Chinese.

DOI: 10.1587/elex.6.623
Received April 05, 2009
Accepted April 13, 2009
Published May 25, 2009