ISSN : 0974 - 7435 Volume 8 Issue 6

Trade Science Inc.

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 8(6), 2013 [855-860]

Local fragment distribution features for text-independent writer identification

Ding Hong[1]*, Yang Feng-ying[2], Zhang Xiao-feng[1]

[1]School of Computer Science and Technology, Nantong University, Nantong 226019, (CHINA)

[2]School of Information Engineering, Huanghuai University, Zhumadian 463000, (CHINA)

ABSTRACT

In this paper, an efficient method for text-independent writer identification using Local Fragment Distribution Feature (LFDF) is proposed. Local fragments, which are parts of the contour in sliding windows, contain the information of strokes. Our method uses the distributions of to create LFDF vector for each specific manuscript. In order to reduce the impact of stroke weight, the fragments which do not directly connect the center point of the sliding window are ignored. Then, the distributions of local fragments are counted and normalized into LFDF. At last, weighted Manhattan distance is used as similarity measurement. The proposed method offers state-of-art performance on ICDAR 2011 writer identification database with multi-languages and the experiments demonstrated that this method is suitable for text-independent writer identification.

KEYWORDS

Writer identification;
Stroke feature;
Local fragment distribution feature;
Weighted manhattan distance.

INTRODUCTION

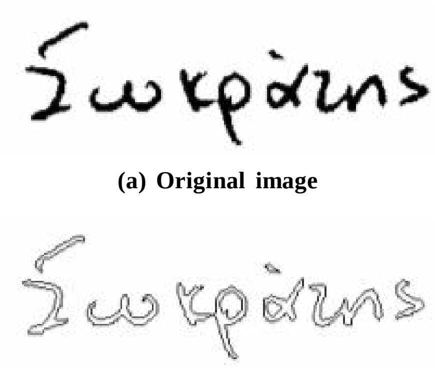
Writer identification is a behavioral biometric based on writing styles. It can provide an important clue for authentication and is widely used in security fields. Writer identification methods can be divided into two major categories: text-dependent and text-independent[1]. Text-dependent methods require the same fixed characters with training handwritings, such as signature verification. In text-independent methods, any handwriting documents with different text will be useful. These methods don’t concentrate on a whole character but on writing style features, such as texture, direction. So the text-independent methods have widely used in many applications.

In recent years, varies of methods have been proposed for text-independent writer identification. Bulacu et al.[2] proposed a serial features with direction, angle for writer identification. Li et al. proposed a micro-structure feature[3], and improved it[4]. Their methods obtained good performance on Chinese character identification. Ghiasi et al.[5] coded local structures into a length-angle form and used them to describe the direction of handwriting. Fiel et al.[6] used SIFT features to avoid the negative effects of binarization. Wen et al.[7] found features by counting the coding of local structure.

Learning from the idea of local structure distribution features, a method base on Local Fragment Distribution Feature (LFDF) is proposed in this paper. LFDF shows the writing style by counting the distribution of stroke in sliding windows. In order to reduce the impacts of stroke weights, this feature only counts the edge points directly connecting the center point in sliding windows. At last, the weighted Manhattan distance is used to measure the similarity between two LFDFs. The experiments show that the proposed method gets satisfactory performance on ICDAR 2011 writer identification database[8].

FEATURE ABSTRACTION AND SIMILARITY MEASUREMENT

The distribution of stroke feature is a hidden feature of handwriting. It is a writing style and can reflect the trend of stroke. The proposed method contains two main parts: feature abstraction and similarity measurement. The feature abstraction procedure counts the edge points in sliding windows and normalized the distribution into LFDF. The degree of similarity is identified by the weighted Manhattan distance. LFDF is not abstracted directly from the original image but from its edge for more valuable information and less redundant information. It is a reasonable method because handwritings can be recovered from edges of stroke. In experiments, Sobel detector is used. Figure 1 shows an example of contour detection. (a) is the original image, (b) is the detection result of (a) by Sobel detector.



(a) Original image

(b) The edge of (a)

Figure 1 : Edge detection.

The LFDF extraction

Everyone has his special writing styles and most of them can be extracted from stroke edges, such as directions, length and angles. These features have been successfully used for writer identification in previous literature. LFDF is also a feature of stroke edge which can reflect the above features. It is abstracted from local fragments which are parts of contour in sliding windows. Flow chart of feature abstraction is shown in Figure 2, which includes edge detection, loop counting and normalization. Loop counting is the main step, which contains local fragment extraction, distribution counting.

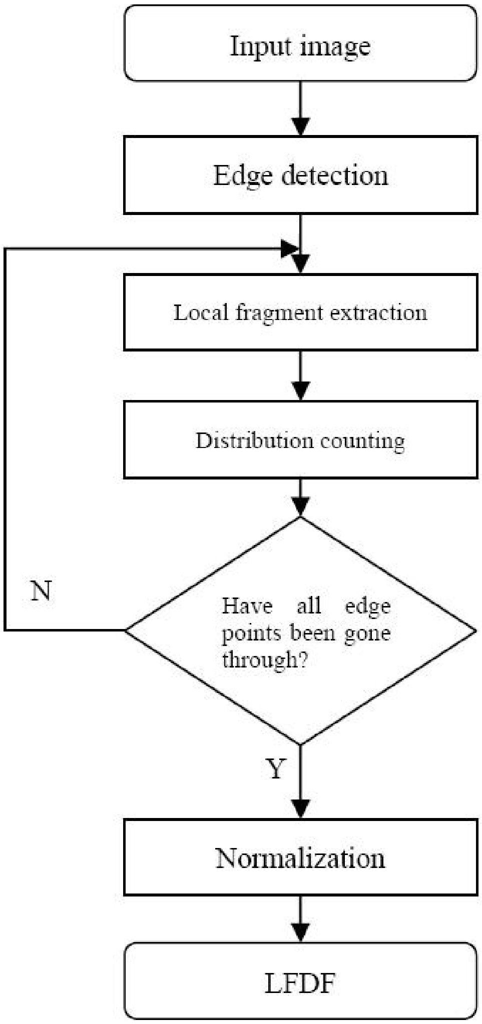


Figure 2 : Flowchart of LFDF extraction.

Fragment extraction

The rectangle in Figure 3 is a sliding window, whose center is an edge point marked with "+". The size of the window is $(2r+1)\times(2r+1)$, where r is the distance between the center and the rectangle border. Fragments are contour in sliding windows. As show in Figure 3, there are several fragments in the original window. In order to reduce the influence of stroke weight, the fragments not connecting the center point are ignored in the following steps. Figure 3 shows the local fragment extraction process. There are three fragments in the window and only the one connecting center point is used in next step.

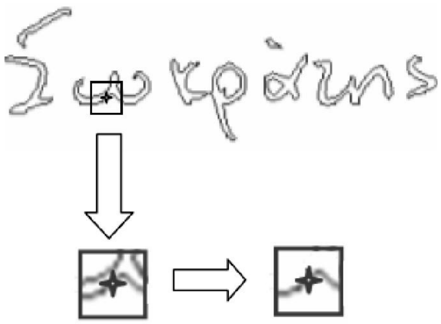


Figure 3 : Fragment extraction in a sliding window.

The distribution of obtained fragments was described in literature[10]. Its main contribution is reducing the influence of stroke weight. In the conditions of any writing instruments allowed, a writer will give handwritings with different weights. So, different stroke weights have negative influence for writer identification.

LFDF extraction

The probability distributions of local structures in sliding windows are used in literature[3,4,7,10]. These stroke distributions can reveal the hidden feature of the strokes and are counted in sliding windows which go through the image with edge points as their centers.

The existing local features only used a subset of related site pairs. A reasonable extension of these ideas is that considering more pairs may gain a more powerful feature. A 7×7 sliding window is shown in Figure 4. The subscript of every site is its group number. The proposed feature uses two kinds of edge point pairs. The first kind pairs are near the center. For every pair, its first group number is no less than its second number. The sites near center have high probability values. Even a little deviation of them will cause a negative influence. So, more pairs of this kind are counted for greater accuracy. The second kind pairs are far from the center and the first group number equals the second number. This kind is less important and counting all pairs will cause a lot of repeating computation.

So, LFDF can be extracted by the following steps:

1) Edge detection. It is an important preprocessing. In our experiments, Sobel detector is used.
2) Local fragment extraction. This step is shown in the previous section.

9_3	8_3	7_3	6_3	5_3	4_3	3_3
10_3	6_2	5_2	4_2	3_2	2_2	2_3
11_3	7_2	3_1	2_1	1_1	1_2	1_3
12_3	8_2	4_1	■	0_1	0_2	0_3
13_3	9_2	5_1	6_1	7_1	14_2	23_3
14_3	10_2	11_2	12_2	12_2	13_2	22_3
15_3	16_3	17_3	18_3	19_3	20_3	21_3

Figure 4 : Group numbers in a sliding window.

3) Counting the number of (I_{m1}, J_{m2}), where I and J are related pairs in a sliding window, $m1$ and $m2$ are their group number, $m1 \leqslant m2$when $m1 < m_t$ or $m1 = m2$ when $m1 \geqslant m_t$, where m_t is the parameter.
4) Go through all edge points and repeat step (2) and (3).
5) Normalization. Different images have different numbers of edge point. So, the distribution is normalized with $\sum_{I_m} N(I_m)$, where I_m is the site and $N()$ is the number. Then, the probability density of coding is

$$p(I_{m1}, J_{m2}) = \frac{N(I_{m1}, J_{m2})}{\sum_{I_m} N(I_m)} \qquad (1)$$

where $N(I_{m1}, J_{m2})$ is the number of pair .

The obtained LFDF is shown in Figure 5. The size of example window is 7×7, $m_t = 2$, the feature of every site contains the probability densities of the pairs between the current site and other sites.

The main part of feature extraction is repeat counting, which is easy to realize. As the size of sliding window increases, the feature dimension rapidly increases. But most sites far from center are nearly useless for their close to zero values. So, the size of sliding window is limited in a small range.

Similarity measurement

The proposed method directly computes the distance between two features. Several distance measurements and their weighted measurements have been tested in our experiments. Among these methods, the

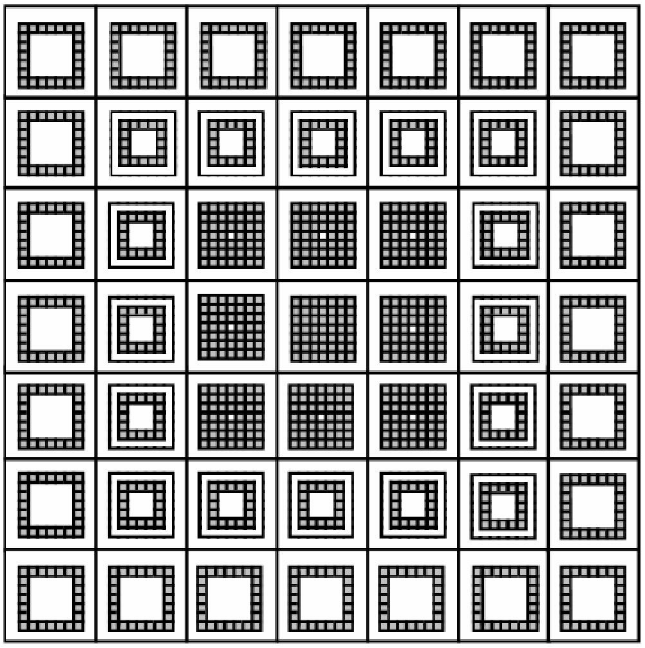


Figure 5 : An example of LFDF.

weighted Manhattan distance has obtained the best performance, whose definition is

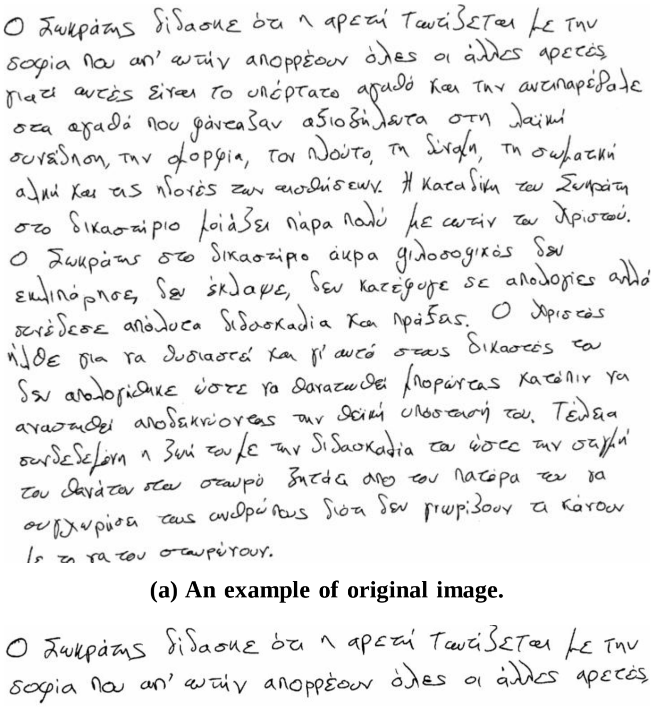
$$D = \sum_i \frac{|LFDF_{1i} - LFDF_{2i}|}{\sigma_i} \qquad (2)$$

where σ_i is standard deviation of the ith component of LFDFs, $LFDF_{1i}$ and $LFDF_{2i}$ are the ith components of two LFDF, respectively.

The similarity is measured by the nearest neighbor rule. The smaller value the LFDF is, the more similar two handwritings are.

EXPERIMENT

The proposed method has been applied on ICDAR 2011 writer identification database. ICDAR 2011 writer identification contest is the first contest in the field of writer identification. Its benchmarking dataset was created with the help of 26 writers that were asked to copy eight pages that contain text in several languages (English, French, German and Greek). They applied two different evaluation scenarios. In the first scenario, the whole images of the dataset were used. In the second scenario, the images were cropped and only two text lines were preserved. These two evaluation scenarios also applied in our experiments. Figure 6 shows two examples of these two scenarios. A cropped image has fewer characters than an original image, which increases the difficulty of feature abstraction.



Ο Σωκράτης δίδασκε ότι η αρετή ταυτίζεται με την
σοφία που απ' αυτήν απορρέουν όλες οι άλλες αρετές
γιατί αυτές είναι το υπέρτατο αγαθό και την αντιπαρέβαλε
στα αγαθά που φάνταζαν αξιοζήλευτα στη λαϊκή
συνείδηση, την ομορφιά, τον πλούτο, τη δόξα, τη σωματική
αλκή και τις ηδονές των αισθήσεων. Η καταδίκη του Σωκράτη
στο δικαστήριο μοιάζει πάρα πολύ με αυτήν του Χριστού.
Ο Σωκράτης στο δικαστήριο άκρα φιλοσοφικός δεν
εκλιπάρησε, δεν έκλαψε, δεν κατέφυγε σε απολογίες αλλά
συνέδεσε απόλυτα διδασκαλία και πράξεις. Ο Χριστός
ήλθε για να θυσιαστεί και γι' αυτό στους δικαστές του
δεν απολογήθηκε ώστε να θανατωθεί μπορώντας κατόπιν να
αναστηθεί αποδεικνύοντας την θεϊκή υπόστασή του. Τέλεια
συνδεδεμένη η ζωή του με την διδασκαλία του ώστε την στιγμή
του θανάτου στον σταυρό ζητάει από τον πατέρα του να
συγχωρήσει τους ανθρώπους διότι δεν γνωρίζουν τι κάνουν
με το να τον σταυρώνουν.

(a) An example of original image.

Ο Σωκράτης δίδασκε ότι η αρετή ταυτίζεται με την
σοφία που απ' αυτήν απορρέουν όλες οι άλλες αρετές

(b) An example of cropped image.

Figure 6. Examples of ICDAR2011 writer identification dataset.

We calculated the LFDF of each image and evaluated the similarity by the weighted Manhattan distance. In our experiments, $m_t = 3$ and three kinds of sliding window sizes are used: $11 \times 11, 13 \times 13$ and 15×15.

Two different measurements soft TOP-N and hard TOP-N criterion are used to evaluate the performance of the proposed method. Every document image of the database is calculated the distance to all other document images using the weighted Manhattan distance. The results are sorted from the most similar to the less similar image. The soft TOP-N criterion is the accuracy of at least one of the same writer is included in the N most similar document images. While the hard TOP-N criterion is the accuracy of all the N most similar document images are written by the same writer. It is a more strict criterion and difficult to get a high accuracy. In our experiments, the values of N used for the soft criterion are 1, 2, 5 and 10 and the values of N used for the hard criterion are 2, 5 and 7.

TABLE 1-4 show the performance of the proposed method. The performance slightly changes when different sliding window sizes are used. It shows the

TABLE 1 : Performance on original images (using soft evaluation).

Window sizes	TOP-1	TOP-2	TOP-5	TOP-10
11 × 11	98.1%	98.6%	99.0%	99.0%
13 × 13	98.1%	98.6%	99.0%	99.0%
15 × 15	98.6%	98.6%	99.0%	99.0%

TABLE 2 : Performance on original images (using hard evaluation).

Window sizes	TOP-2	TOP-5	TOP-7
11 × 11	92.8%	78.8%	42.8%
13 × 13	93.3%	80.8%	49.0%
15 × 15	93.3%	82.2%	45.7%

TABLE 3 : Performance on cropped images (using soft evaluation).

Window sizes	TOP-1	TOP-2	TOP-5	TOP-10
11 × 11	93.8%	98.6%	98.6%	98.6%
13 × 13	96.2%	97.1%	98.6%	98.6%
15 × 15	93.8%	96.6%	98.1%	98.6%

TABLE 4 : Performance on cropped images (using hard evaluation)

Window sizes	TOP-2	TOP-5	TOP-7
11 × 11	82.2%	50.5%	19.2%
13 × 13	83.7%	51.0%	18.3%
15 × 15	86.1%	52.3%	18.8%

TABLE 5 : Soft evaluation using ICDAR database of original images.

Methods	TOP-1	TOP-2	TOP-5	TOP-10
ECNU	84.6%	86.5%	88.0%	88.9%
QUQA-a	90.9%	94.2%	98.1%	99.0%
QUQA-b	98.1%	98.6%	99.5%	100.0%
TSINGHUA	99.5%	99.5%	100.0%	100.0%
GWU	93.8%	96.2%	98.1%	99.0%
CS-UMD	99.5%	99.5%	99.5%	99.5%
TEBESSA	98.6%	100.0%	100.0%	100.0%
MCS-NUST	99.0%	99.5%	99.5%	99.5%
The proposed	98.1%	98.6%	99.0%	99.0%

TABLE 6 : Hard evaluation using ICDAR database of original images.

Methods	TOP-2	TOP-5	TOP-7
ECNU	51.0%	2.9%	0.0%
QUQA-a	76.4%	42.3%	20.2%
QUQA-b	92.3%	77.4%	41.4%
TSINGHUA	95.2%	84.1%	41.4%
GWU	80.3%	44.2%	20.2%
CS-UMD	91.8%	77.9%	22.1%
TEBESSA	97.1%	81.3%	50.0%
MCS-NUST	93.3%	78.9%	38.9%
The proposed	93.3%	80.8%	49.0%

TABLE 7 : Soft evaluation using ICDAR database of cropped images.

Methods	TOP-1	TOP-2	TOP-5	TOP-10
ECNU	65.9%	71.6%	81.7%	86.5%
QUQA-a	74.0%	81.7%	91.8%	96.2%
QUQA-b	67.3%	79.8%	91.8%	94.7%
TSINGHUA	90.9%	93.8%	98.6%	99.5%
GWU	74.0%	81.7%	91.4%	95.2%
CS-UMD	66.8%	75.5%	83.7%	89.9%
TEBESSA	87.5%	92.8%	97.6%	99.5%
MCS-NUST	82.2%	91.8%	96.6%	99.5%
The proposed	96.2%	97.1%	98.6%	98.6%

TABLE 8 : Hard evaluation using ICDAR database of cropped images.

Methods	TOP-2	TOP-5	TOP-7
ECNU	39.4%	2.9%	0.0%
QUQA-a	52.4%	15.9%	3.4%
QUQA-b	47.6%	22.6%	6.3%
TSINGHUA	79.8%	48.6%	12.5%
GWU	51.4%	20.2%	6.3%
CS-UMD	51.9%	22.1%	3.4%
TEBESSA	76.0%	34.1%	14.4%
MCS-NUST	71.6%	35.6%	11.1%
The proposed	83.7%	51.0%	18.3%

good stability of our method. TABLE 5-8 show the comparisons of the proposed method with other methods mentioned in ICDAR 2011. The results corresponding to the highest accuracy are marked in bold. Though the performance of the proposed method in original scenario is slightly below the highest, its performance in cropped scenario exceeds the existing methods.

CONCLUSION

In this paper, a method based on LFDF is proposed. LFDF is extracted from the sliding windows by counting the edge point distribution within the fragments. In order to reduce the impact of the stroke weight, only the fragments connecting the centers of sliding windows are counted and others are ignored. The counting procedure is an easily implementation, which is mainly consisted of the repeat additions. Our feature is more powerful than the existing local structure features by counting more related pairs near the center of sliding windows. At last, the weighted Manhattan distance effectively measures the similarities of the LFDFs. The experiments on the ICDAR database show our method gets the state-of-art performance, especially using the cropped images. It means that the proposed method can abstract stable features and suit for writer identification in the conditions of fewer characters.

ACKNOWLEDGMENTS

This work was supported by the Program for Natural Science Foundations of China (61170171).

REFERENCES

[1] X.Li, X.Q.Ding, X.L.Wang; Semi-text-independent writer verification of Chinese handwriting.International Conference on Fountiers of Handwriting Recognition, **(2008)**.

[2] M.Bulacu, L.Schomaker; Text-independent writer identification and verification using textural and allographic features.IEEE Transactions on Pattern Analysis and Machine Intelligence, **29(4)**, 701-717 **(2007)**.

[3] X.Li, X.Q.Ding, L.R.Peng; A microstructure feature based text-independent method of writer identification for multilingual handwritings.Acta Automatica Sinica, **35(09)**, 1199-1208 **(2009)**.

[4] X.Li, X.Q.Ding; Writer identification based on improved microstructure features.Journal of Tsinghua University (Science and technology), **50(04)**, 595-600 **(2010)**.

[5] G.Ghiasi, R.Safabakhsh; Offline text-independent writer identification using codebook and efficient code extraction methods.Image and Vision Computing, **31(5)**, 379-391 **(2013)**.

[6] S.Fiel, R.Sablatnig; Writer Retrieval and Writer Identification using Local Features.The 10th IAPR International Workshop on Document Analysis Systems, 145-149 **(2012)**.

[7] J.Wen, B.Fang, J.L.Chen, Y.Y.Tang, H.X.Chen; Fragmented edge structure coding for Chinese writer identification.Neurocomputing, **86**, 45-51 **(2012)**.

[8] G.Louloudis, N.Stamatopoulos, B.Gatos; ICDAR 2011 Writer Identification Contest.International Conference on Document Analysis and Recognition, 1475-1479 **(2011)**.

[9] J.Bernsen; Dynamic thresholding of gray-level images.International Conference on Pattern Recognition, 1251-1255 **(1986)**.

[10] X.F.Zhang, Y.Lu; Handwritten and Machine Printed Text Discrimination Using an Edge Co-occurrence Matrix.International Conference on Audio, Language and Image Processing, **2**, 828-831 **(2012)**.