Writer Identification Using a Statistical And Model Based Approach

Diamantatos Paraskevas Gritzalis Stefanos Kavallieratou Ergina
Department of Information & Communication Systems Engineering
University Of Aegean
Karlovassi, Greece
{diamantatos,sgritz,kavallieratou}@aegean.gr

***Abstract*—The state-of-the-art writer identification systems use a variety of different features and techniques in order to identify the writer of the handwritten text. In this paper several statistical and model based features are presented. Specifically, an improvement of a statistical feature, the edge hinge distribution, is attempted. Furthermore, the combination of this feature with a model-based feature is explored, that is based on a codebook of graphemes. For the evaluation, the Firemaker DB was used, which consists of 250 writers, including 4 pages per writer. The best result for the statistical suggested approach, the skeleton hinge distribution, achieved accuracy of 90.8%, while the combination of this method with the codebook of graphemes reached 96%.**

Keywords- Writer Identification; Skeleton hinge distribution; Codebook of graphemes; Directional Features;

I. INTRODUCTION

This paper addresses the problem of offline, automatic writer identification, by the use of scanned handwritten document images. Writer identification is a behavioral handwriting-based recognition modality, which proceeds by matching unknown handwritings against a database of samples with known authorship and it is considered today as a hot and promising topic of research. Furthermore, identifying the author of a handwritten sample, using automatic image based methods, is an interesting pattern recognition problem with direct applicability in the forensic and historic document analysis fields.

The work presented here, can be considered as an improvement of previous works. The skeleton hinge distribution, a technique suggested in this paper, attempts to improve edge-hinge distribution [1] and edge-hinge combinations [2]. While a combination of this method with codebook of graphemes method [2,3] is explored. Works of the recent literature, presented next, mostly influence the work done in this paper.

Bensefia et al. [4] used graphemes generated by a handwriting segmentation method to encode the individual characteristics of handwriting. These graphemes are then clustered. Grapheme clustering is used to define a common feature space for all the documents in the dataset. The reported experiment results achieved accuracy of 90% on a dataset consistent of 88 writers (PSI), and 68% on a dataset of 150 writers (IAM-DB).

Bulacu et al. [1] computed edge-hinge distribution feature, an edge-direction feature. By traversing the image, all edge fragment directions are considered and stored in a histogram of directions. The nearest neighbor algorithm is used to match histograms of different images. Experimental results reported accuracy of 63% on the Firemaker DB [7] using 250 distinct writers.

Schomaker et al. [3] compute fragments of connected-component contours, which are classified to identify the writer. A codebook of graphemes is generated, by training a Kohonem SOFM on a large number of grapheme contours. Next, the graphemes are extracted from each document and matched to the graphemes of the codebook. A histogram of graphemes for every document is generated. Experimental results achieved accuracy of 95% on 10 writers, and 83% on 215 writers. When were combined with edge directional features 97% accuracy is achieved.

Laurens van der Maaten et al. [2] improved edge hinge directional features, by using various window sizes, while combining these features with a codebook of graphemes achieved identification accuracy of 97%. The proposed edge hinge based method, achieved 81% identification accuracy, on the Firemaker DB [7] which consists of 250 writers.

The contribution of the present work consists of:

- Introducing Skeleton hinge distribution, an improvement of previous edge-directional features.
- Experimenting with the combination of skeleton hinge distribution with codebook of graphemes.
- Suggesting that in writer identification, all stroke widths should be considered to be the same size.

In following section 2, the statistical features mentioned before are presented, along with the proposed feature, the skeleton hinge distribution. On section 3, the model-based approach is presented. Experimental data and results can be found on section 4, while in section 5 the conclusion is drawn.

II. STATISTICAL FEATURES

The statistical features have been explored extensively in automatic writer identification [1], like run length distribution, slant distribution, entropy, and edge-hinge distribution. In this section, the evolution of edge-direction distribution feature is described. Totally four methods for feature extraction, three existed and the proposed one, are presented that have similar characteristics. The edge-direction distribution [1], the edge-hinge distribution [1], edge-hinge combinations [2] and finally a novel feature proposed in this paper, the skeleton-hinge distribution.

2167-6445/14 $31.00 © 2014 IEEE
DOI 10.1109/ICFHR.2014.104

A. Edge-direction distribution

This feature extraction starts with edge detection. Edge detection generates a binary image in which only the edge pixels are kept. Next, each edge pixel is considered the center of a square neighborhood. All the pixels are checked, by the use of logical AND operators, to all directions, emerged from the central pixel and end on the periphery of the neighborhood, looking for the presence of another edge fragment. In fig 1, an example with 4-pixel length edge fragment quantized in 12 directions, is presented. All the verified instances are counted into a histogram that is normalized to a probability distribution p(φ). This distribution gives the possibility of finding in the image, an edge based fragment oriented at the angle φ to the horizontal. Moreover, the most dominant direction in p(φ) corresponds to the slant of the handwritten text.

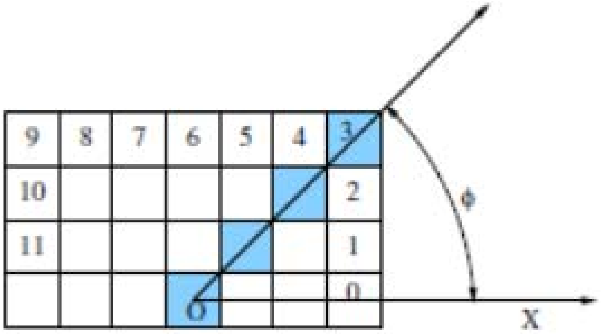

Figure 1. Extraction of edge-direction distribution.

B. Edge-hinge distribution

As reported by Bulacu et al. [1] the edge hinge distribution is a statistical feature, which outperforms all the other statistical approaches. The central idea in the edge hinge distribution is to consider, not one, but two edge fragments in the neighborhood, emerging from the central pixel, and subsequently compute the joint probability distribution of the orientations of the two fragments. This feature concerns the direction changes of a writing stroke in handwritten text. The edge-hinge distribution is extracted by the use of a window that scans an edge-detected binary handwriting image. Whenever the central pixel of the window is "on", the two edge fragments (i.e. connected sequences of pixels) emerging from this central pixel are considered only when φ1<φ2. In fig 2, an example with 4-pixel length edge fragment quantized in 24 directions. The directions are measured and stored in pairs. A joint probability distribution p(□1, □2) is obtained over a large sample of pairs.

C. Edge-hinge combinations

The edge-hinge combinations, proposed by Van der Maaten et al. [2], improved the edge hinge distribution by considering multiple pixel length edge fragments (i.e. window sizes), instead of just one. Experimenting with combinations of edge hinge distributions and using various fragment lengths, they improved the results of writer identification by up to 12% compared with edge-hinge distribution. The algorithm of this implementation is available at [5].

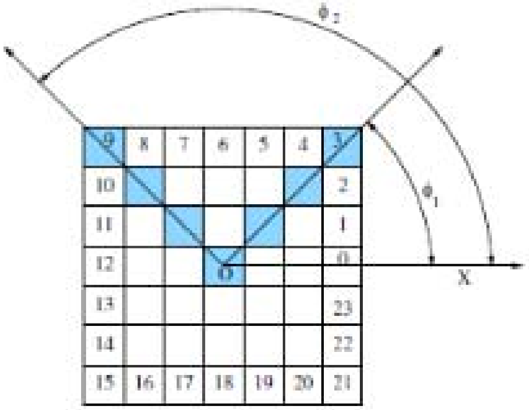

Figure 2. Edge Hinge Distribution Extraction

D. Skeleton-hinge distribution

The main problem with the current implementations is that the edges are usually close to each other, filling the feature matrix with unnecessary data. In order to overcome this problem, the same technique was used in combination with the skeleton of the image, instead of the edges. Henceforth, this technique will be referred as skeleton hinge distribution.

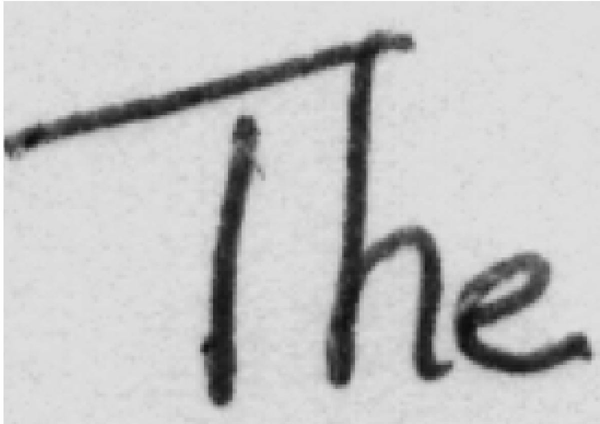

Figure 3. Hand written digitized text

Normally when something is written on a paper (fig.3), its thickness is considered to be a single line. When the image is digitized the same trace of ink is translated into several pixel lines. By considering the edge hinge distribution, on an edge image (fig. 4) a lot of unnecessary information, like the bottom or the side curves of the letters, is included in the feature vector.

Furthermore, differences in line thickness, from a variety of different pens, may produce significant variations in the extracted features, in both edge hinge distribution, and edge hinge combinations. The main suggestion in this paper is that all stroke widths, i.e. line thickness, should be considered to be the same size. This is achieved by skeletonizing the characters, to a single pixel width line.

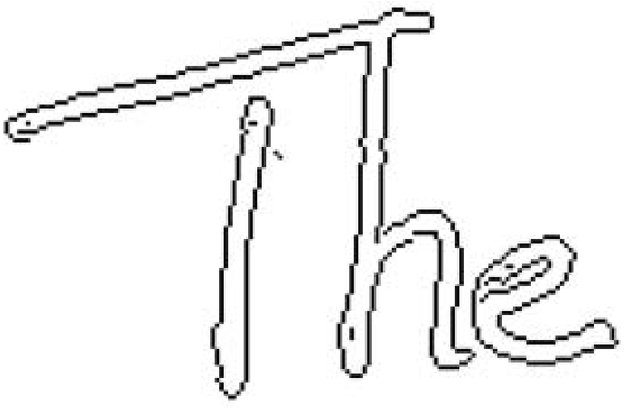

Figure 4. Edge image of hand written text

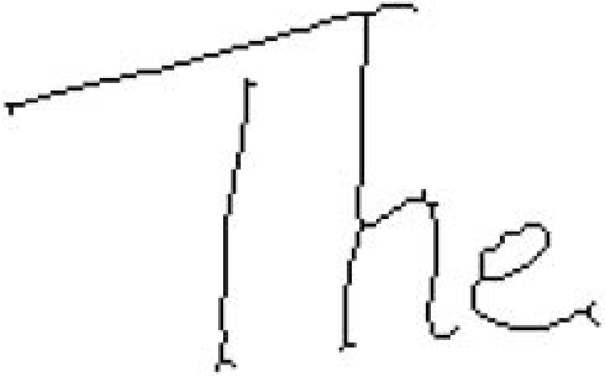

Figure 5. Skeleton image of hand written text

On the skeleton hinge distribution only the skeleton of the letters is considered (fig.5), a simple structure that takes into account the basic, only, required information, in order to match the features to already known ones.

In the proposed system, the first step is the image binarization, in order to clean the image from the unnecessary grey scale data (noise). Next, the skeleton of the image is extracted. Every pixel of the skeleton image that is "on" in the center of the window is considered. A combination of various windows of 3,5,7,9 pixels size [2] is checked for "on" pixels on the periphery of each window (see Fig.2). Only directional fragments with $\varphi1<\varphi2$ are counted and stored in pairs in a histogram. That histogram of directional fragments, is normalized into a joint probability distribution p(□1, □2).

The main ideas of edge hinge distribution, and edge hinge combinations, are present in the proposed technique. On the other hand, by applying this methodology to a skeleton image, a significant improvement on the results of writer identification task is observed (section IV).

It is important to mention that the resulting feature matrix includes more compact information and it is easier to compare two resulting matrices of test and train samples. Please check a successful application of the proposed system, in figure 6, where some text samples are provided over their results. On the left side, two train samples are presented, of different writers, and on the right, two test samples from the same two previous writers. Next, the surface of skeleton hinge distribution is presented. The left ones correspond to the train samples, while the right ones to the test samples. On the bottom part of the figure, the edge hinge combinations are shown. Again the left ones correspond to the train samples, and the right to the test samples.

III. Model-Based Features

In the model-based approach used in the works [2-3], it is assumed that each writer produces a recognizable set of writer specific character shapes, or allographs. This happens due to schooling and personal preferences. The core idea reflected in the above statement implies that a histogram of used allographs can characterize each writer. However, it is not possible to have a predefined list of allographs. Instead training is needed, in order to generate automatically a codebook, which sufficiently captures allograph information from samples of handwriting.

The approach used in this work actually relies on a codebook of models of graphemes. Graphemes are small strokes of handwriting, which are extracted by applying a robust segmentation algorithm on a handwritten image. It should be noted, that there is a distinction between graphemes and the fragments used in the statistical methods because of the different algorithms in use.

In Schomaker et al. [3], a codebook of graphemes is generated by training a Kohonen SOFM [6] on a large number of grapheme contours. The produced codebook is later used to construct feature vectors.

The process used to create feature vectors from the codebook is quite simple: From each text image, all graphemes are extracted and matched to the grapheme models of the codebook. Euclidean distance between the grapheme contours is used for the matching process. For each grapheme model in the codebook, every successful match is counted. The result is a histogram of graphemes, which characterizes the writer, and also identifies him.

A limitation in this approach is the long training time of Kohonen SOFM. As reported in [3] a training time of up to 122 hours can be required. Besides that, Kohonem SOFM may get stuck in local minima.

Van Der Maaten et al [2] proposed the use of random selection for the creation of graphemes rather than using Kohonem SOFM. In this method, no time consuming training is performed, overcoming the time limitation.

Instead of training a randomly number of graphemes are drawn from the large set of graphemes.

Both approaches, when were combined with the edge-hinge feature, achieved an identification performance of 97% on the Firemaker DB for 150 distinct writers and a codebook of 400 graphemes.

Here, an improvement was attempted, using different approach on the codebook generation, by only considering closed areas of the characters. Character closed areas, are the least affected by writer slant, very important as slant is a characteristic of the writer that can affect the skeleton hinge distribution.

By combining skeleton hinge distribution with a codebook of graphemes only generated by character closed areas, it was expected to be an ideal way, of securing skeleton hinge distribution, against forge attempts. A forge attempt can be made by simply changing the slant. But the results on this approach were not the expected ones.

IV. Results

A. Data

The accuracy of the technique presented on this paper, the skeleton hinge distribution, was evaluated by using the Firemaker DB [7]. This data set was used in order to be able to directly compare the achieved results with the reported ones by the other methods.

The Firemaker is a database of handwritten pages from 250 writers, including four pages per writer.

- Page 1 contains a copied text in natural writing style
- Page 2 contains a copied text in Upper-case text
- Page 3 contains copied forged text. The writers here try to impersonate another writer.
- Page 4 contains a self-generated description of a cartoon image in free writing style. In this last page, the text content and the amount of written ink varies considerably per writer.

All pages in Firemaker DB were scanned at 300-dpi gray scale. The text, that was asked to be copied, was specially designed in forensic praxis to cover a sufficient amount of different letters of the alphabet. In our experiments, only pages 1 and 4 were used. Page 1 was used as a train set. While page 4, was used as a test set.

B. Training

In order to train the system only page 1, from the Firemaker DB was used. Each page was binarized and the skeleton was extracted using the Matlab. The used procedure is the one described in the previous section for skeleton hinge distribution.

Train procedure was really fast, about 250 seconds on a laptop i7 2.5Ghz pc, and in comparison to the edge hinge distribution, about 35% faster. On the same machine edge-hinge distribution train took 384 seconds to complete.

C. Testing

In order to test the system only page 4 was used from the Firemaker DB. Testing process used the same procedure as the training process. Each page was binarized and skeletonized.

The test procedure was faster than training, due to the variations in the sizes of text, in page 4. Testing took around 200 seconds on a laptop i7 2.5 Ghz. Edge hinge distribution time was about 270 seconds. An improvement of about 35% can be observed here, too.

D. Matching

All the results reported on experiments section used a 1-nearest neighbor classifier and Manhattan distance. Euclidean and chi-square distances were also considered, but they performed worse.

E. Experiments

Various experiments were performed, using combinations of several parameters, e.g. window sizes, matching classifiers, etc. It is hard to compare our results, with results reported on other papers, because of the variation on the data sets. Our results will be only comparable with methods that used the same data set.

Furthermore, even on the same data set, results can have a significant variation. Some methodologies only used a fragment of the entire data set, without mentioning which one, exactly. Also there are differences in train and test sets. Even a slight change in these sets, can change the entire outcome.

Skeleton hinge distribution feature identification results are presented on table 1. These experiments used the entire data set of 250 writers. Like edge-hinge combinations method, a combination of fragment lengths i.e. window sizes, was used. Furthermore, for the nearest neighbor classifier Manhattan, Euclidian and chi-square distances were used. Our top result is identification accuracy of 90.8 % for a combination of fragment lengths of 5- and 9-pixel length window and Manhattan distance.

TABLE I. Skeleton Hinge Distribution

Fragment Length	**Skeleton Hinge Distribution Accuracy (Percentage)**		
	Manhattan Distance	***Euclidian Distance***	***Chi-square Distance***
3	80%	72%	53.2%
5	89,6%	77,2%	66%
7	90%	81,6%	69,6%
9	88%	85,2%	76%
3 , 5	85,2%	75,2%	58,4%
3 , 7	85,6%	75,6%	55,2%
3 , 9	86%	74,8%	53,2%
5 , 7	90%	78,8%	64,4%
5 , 9	**90.8%**	78,8%	67,2%
7 , 9	90%	83,2%	73,6%
3 , 5 , 7	86,8%	76,8%	60%
3 , 7 , 9	89,6%	76,8%	55,6%

Fragment Length	Skeleton Hinge Distribution Accuracy (Percentage)		
	Manhattan Distance	***Euclidian Distance***	***Chi-square Distance***
5 , 7 , 9	90%	79,2%	68,8%
3 , 5 , 7 , 9	89,6%	76,8%	60,4%

In addition, an attempt was made to combine skeleton hinge distribution, with codebook of graphemes method. The results of this experiment are presented on table II. The previous methods [2,3] reported accuracy of up to 97% on 150 writers, using a codebook of size 400, when the results where combined with edge-directional features. Unfortunately, it was impossible to train a codebook of 400 graphemes for 250 writers, due to memory issues.

Instead, a codebook of 225 graphemes was trained for 250 writers. A maximum accuracy of 95,6% was reached. It is important to note, that the other methods reported 97 % accuracy on 150 writers with a codebook of 400 graphemes. In our case, an experiment was also performed using 150 writers of the data set and a codebook of 225 graphemes. An accuracy of 96% was achieved.

TABLE II. SKELETON HINGE DISTRIBUTION COMBINED WITH CODEBOOK OF GRAPHEMES METHOD

Number Of Writers	CodeBook Size	Skeleton Hinge Distribution Combined With Codebook Of Graphemes Method		
		Manhattan Distance	***Euclidian Distance***	***Chi-square Distance***
250	225	95.6%	91.2%	78.8%
150	225	96%	94.7%	86.7%

V. CONCLUSION

In this paper a writer identification system were presented. Our experiments indicate that the use of a single feature, the skeleton hinge distribution, yields promising results.

The entire idea for skeleton hinge distribution came from the assumption that in writer identification, all stroke widths, i.e. line thickness, should be considered the same size. By applying skeletonization on characters, this criterion is met. All stroke widths are transformed to a single pixel line. The experimental results proved that the previous assumption is correct. We believe, that this assumption should be considered in other statistical methods as well, methods like run lenghts, entropy etc.

Further improvements can be achieved. A combination of features along with the skeleton hinge distribution like different statistical features can be used. Further research is needed in the area.

REFERENCES

[1] M. Bulacu, L. Schomaker, and L. Vuurpijl. Writer identification using edge-based directional features. In Proceedings of ICDAR 2003, pages 937–941, Edinburgh, UK, 2003

[2] Van der Maaten, L., Postma, E. (2005) Improving automatic writer identification, in:17thBelgium-Netherland Conference on Artificial Intelligence.

[3] Schomaker, L., Franke, K., Bulacu, M. (2007) Using codebooks of fragmented connected-component contours in forensic and historic writer identification, Pattern Recognition Letter28, 719–727.

[4] Bensefia, A., Paquet, T., Heutte, L. (2005) “A Writer Identification and Verification System,” Pattern Recognition Letters, vol. 26, no. 10, pp. 2080-2092.

[5] Electronic Source : http://homepage.tudelft.nl/19j49/Software.html

[6] T. Kohonen. Self-organization and associative memory: 3rd edition. Springer-Verlag New York,Inc., New York, NY, USA, 1989.

[7] L. Schomaker and L. Vuurpijl. Forensic writer identification: A benchmark data set and a comparison of two systems [internal report for the Netherlands Forensic Institute]. Technical report, Nijmegen: NICI, 2000.

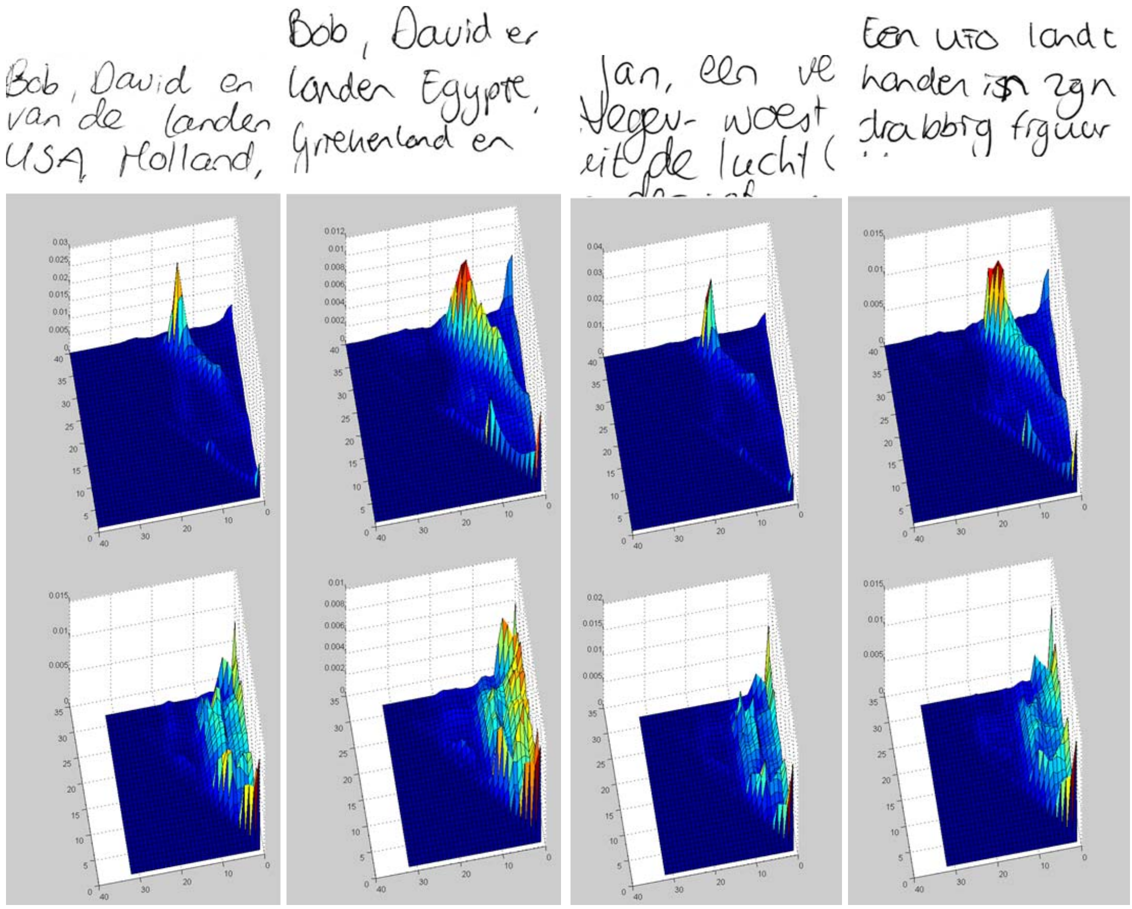

Figure 6. Text samples from the same two different writers (train set on left, test set on right) along with skeleton hinge distribution feature surface (middle) and edge hinge combinations feature surface (bottom). The text samples on the top are fragments of the text sample used, and are provided for illustrating the differences of handwriting between those texts.