2015 IEEE 2015 International Conference on Computer, Communication, and Control Technology (I4CT 2015), April 21 - 23 in
Imperial Kuching Hotel, Kuching, Sarawak, Malaysia

Framework for Human Identification through Offline Handwritten Documents

Shehzad Khalid, Uzma Naqvi
Department of Computer Engineering, Bahria University.
Islamabad, Pakistan
shehzad@bahria.edu.pk, uzmanaqvi@gmail.com

Imran Siddiqi
Department of Computer Science, Bahria University.
Islamabad, Pakistan
imran.siddiqi@gmail.com

Abstract— **Identification of individuals from handwritten documents using automated recognition systems has gained significant research interest due to the wide variety of applications it offers for forensic analysis, signature verification, classification of historical writings and other document analysis tasks. In this paper, we present a framework that combines different feature space representations of handwriting for an effective characterization of writers. Multiple distance functions are applied to each feature space which are then combined to enhance the overall recognition performance. The proposed identification framework evaluated on a standard database realizes significant performance improvements in terms of identification rate.**

Keywords-component; Handwritten documents; Writer identificatio;, Classification.

I. INTRODUCTION

Handwriting of an individual not only serves as a mode of communication but also carries specific traits of the writer. Today, handwriting is considered an effective behavioral biometric identifier[1]. A writer can be identified by his/her handwriting based on his unconscious practices and writing style. This identification of writers from writing samples finds a number of interesting applications in forensic and historical document examination. It is applied frequently in the court of law and banks for validating the authenticity of a document or for verifying signatures. Computerized systems for writer recognition made this tiresome task efficient as well as effective. The process of writer identification involves searching a handwritten sample of an unknown authorship in database of writing samples with known authors. Handwritten documents can be examined for their textual content as well as graphical/visual appearance to identify the writer. An effective writer recognition system retrieves a subset of similar writing samples of known writers on which human experts can further verify the identity of the writer of the document presented as query.

Writer identification techniques are generally classified into offline and online methods as a function of the writing acquisition method. In online writer identification, writing samples are provided directly on a digitizing device and are stored as trajectories represented as time series of 2D coordinates. Variety of information such as speed of writing, angles and pressure etc. can be extracted thus resulting in spatio-temporal feature space representation of handwriting. On the contrary, offline writer identification employs scanned images of handwritten documents. The temporal information is not available in offline writings and these approaches rely merely on spatial features associated with characters, words, lines or paragraphs.

Writer identification task can be carried out in text-dependent or text-independent modes. In text-dependent methods, writing samples with same textual content are compared. Text-independent methods compare writing samples of arbitrary text to find the identity of a questioned document and are more close to real world scenarios. These methods, however, realize lesser identification rates as compared to text-dependent methods [2].

Most of the writer identification methods generate global feature space representation of handwriting. Global features are extracted based on various statistical measurements obtained from a complete image of text. Global features can be broadly categorized into texture and structural features. Texture features are extracted by treating text as an image and applying techniques such Gabor filters, fractal analysis and co-occurrence matrices etc. On the other hand, approaches employing structural features [21] measure the structural properties such as average height, width and slope of handwritten characters. Codebook based structural features have also been investigated [7, 15-17] to identify the writer on the basis of localized features that are frequent in the writing style of a writer. Hybrid feature space representations by combining global and structural feature space representations have also been proposed in literature [4].

In this paper, we present a hybrid framework for offline writer identification that employs multiple feature spaces and distance functions. Similarity of handwriting samples is computed in multiple distance spaces which are further combined using a weighted approach to enhance the writer identification performance.

This paper is structured as follows. Section II provides a brief overview of the recent developments in offline writer identification. In section III, we introduce the text-independent features that have been used to represent handwritings. Section IV presents the proposed framework to combine the selected feature spaces and compute similarity in multiple distance spaces. The results of experimental evaluations to validate the proposed hybrid framework are presented in Section V while

978-1-4799-7952-3/15/$31.00 ©2015 IEEE

the last section concludes the paper with a discussion on possible extensions of this work.

II. RELATED WORK

Handwriting is accepted as a valid biometric modality and has been widely employed by Forensic or Questioned Document Examiners (FDEs/QDEs) [4, 22]. A wide variety of writer identification systems have been proposed in the recent years. The notable of these contributions are summarized in Fig. 1. Most of the existing work on this subject employs global feature space representations. Global features make use of the difference in ink trace and background and the textual content is not important. These features can be extracted at macro or micro levels [3,4]. Local features, on the other hand, are computed from small parts of the ink trace such as strokes or characters [5]. Extraction of local features requires segmentation of text into small units (characters, graphemes or strokes) and has been considered in a number of recent studies [1] [5-8].

Textural features have also proven to be effective in characterizing the writer of a given document. Texture based features consider each writing as a distinct texture and are independent of the content of document. Multi-channel Gabor filtering [9] and grey level co-occurrence matrices (GLCM) [10, 12] have been effectively used as texture descriptors for writer identification. Combination of different features to enhance the identification rates has also been investigated. Bulacu et al. [11], for example, combined texture based features with allograph based features and realized enhanced performances on writer identification task.

In the recent years, a number of codebook based writer identification approaches [7,15-17] have been proposed and have reported high identification rates. These approaches divide the handwriting into smaller units (graphemes or small fragments) and cluster these units into classes. These classes correspond to a representative set of strokes (the codebook) which can be combined to regenerate a given writing. Graphemes, which are under, over or well segmented characters, are obtained by dividing the text at the local minima of the upper contour of writing. Grapheme based codebook was introduced in [13] and the same idea was extended in [14]. Later studies [15-18] introduced a sub-grapheme level segmentation by dividing the writing into small fragments which are clustered to produce a codebook of small writing fragments.

Codebooks based writer identification methods either rely on a writer-specific or a universal codebook. In writer-specific codebooks, a separate codebook is generated for each writing and characterizes its writer. Methods based on a universal codebook generate a single global codebook of patterns and the probability of occurrence of these patterns in a particular writing is characteristics of its writer. Universal codebooks are known to be computationally less expensive and also outperform writer-specific codebooks in terms of identification rates [7]. In addition to codebook based features, orientation and curvature based features computed from a contours of writing have also shown promising performances in a number of studies [7,19].

After having discussed some recent advancements in offline writer identification, we discuss the features employed in our study in the next section.

III. FEATURES

In this section, we present the feature space representation that we have employed to represent the hand writing samples. We have employed two state-of-the-art feature space representations including codebook based probability distribution and contour based representation as presented in [7]. We briefly describe these features for the completeness of text.

Codebooks are aimed at grouping the frequent patterns in writing into clusters which characterize the writer. The codebook based method proposed in [7] extracts small writing fragments from a set of handwritten documents and applies a k-means clustering to group these patterns into a codebook of 100 clusters. This codebook serves as a common representation space. The small writing fragments in a given handwritten text are compared with the patterns in the codebook and the probability of occurrence of each pattern of the codebook in the writing is computed, the distribution being the characteristics of the writer.

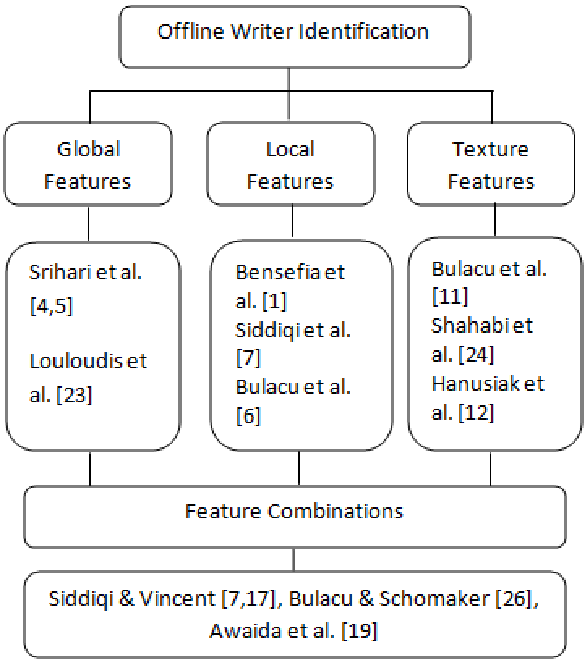

Fig.1: Taxonomy of recent writer identification techniques

In addition to codebook based features, we also employ the contour based features proposed in [7]. The interior and exterior contours in the writing are approximated by a set of polygons and a set of features capturing the orientation and curvature of writing is computed. These include histograms of angle between two segments and the slope of each segment in the polygonized contours. The complete implementation details of these features can be found in [7].

I. PROPOSED HYBRID FRAMEWORK

In this section, we present the proposed framework for writer identification using codebook and polygon features discussed earlier. Given the feature space representations of writing samples, we propose to compute multiple distance functions on each feature space separately and then combine the distances using the weighted average mechanism. An overview of the proposed framework is illustrated in Fig. 2. Two well-known metrics, Euclidean and chi-square distance, are applied to each of the feature space and the computed distances are combined in to a hybrid distance space. Finally, writer identification is carried out by using nearest neighbor classification.

Let *DB* be a labeled dataset containing *n* writing samples and $L = \{l_1, l_2, ..., l_n\}$ be the set containing the label information of corresponding instances in *DB*. For each sample in DB, the codebook and polygon based feature space representations are generated and stored in DB_{CB} and DB_P respectively. Given a query handwritten document, referred to as Q, the identification of writer using proposed framework comprises the following steps.

1. Generate the codebook and polygon based feature space representations of query sample referred as Q_{CB} and Q_P respectively.
2. Compute the Euclidean distance between codebook based feature space representation of query sample and the samples in *DB* as:

$$Dist^{i}_{CB\_ED} = \sqrt{\sum_{j=1}^{m} \left(Q^{j}_{CB} - DB^{ij}_{CB}\right)^2} \qquad (1)$$

where $Dist^{i}_{CB\_ED}$ is the Euclidean distance of ith sample in *DB* with query sample in the codebook feature space and *m* is the dimensionality of the feature vector.

3. Compute the chi-square distance between codebook based feature space representation of the query sample and samples in *DB* as:

$$Dist^{i}_{CB\_CH} = \sum_{j=1}^{m} \frac{\left(Q^{j}_{CB} - DB^{ij}_{CB}\right)^2}{Q^{j}_{CB} + DB^{ij}_{CB}} \qquad (2)$$

where $Dist^{i}_{CB\_CH}$ is the chi-square distance of ith sample in *DB* with the query sample in the codebook feature space.

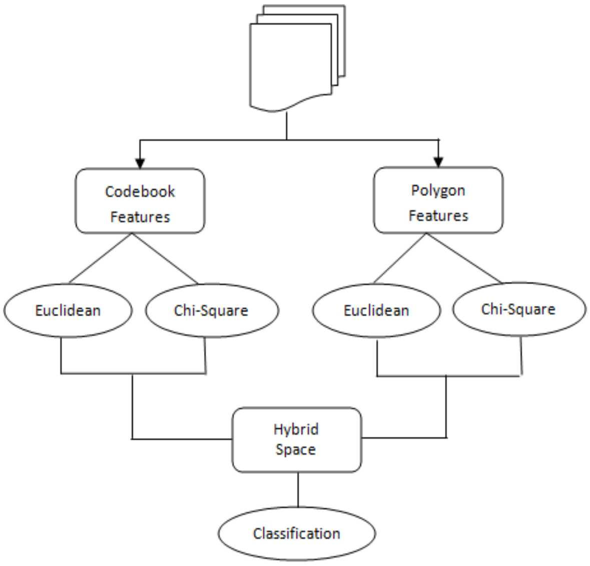

Fig.2: An overview of the proposed framework

4. Compute the Euclidean distance between polygon based feature space representation of the sample in question with those in *DB* as:

$$Dist^{i}_{P\_ED} = \sqrt{\sum_{j=1}^{m} \left(Q^{j}_{P} - DB^{ij}_{P}\right)^2} \qquad (3)$$

where $Dist^{i}_{P\_ED}$ is the Euclidean distance of ith sample in *DB* with the query image in the polygon feature space and *m* is the feature vector length.

5. Compute chi-square distance between polygon based feature space representations of query and reference image as:

$$Dist^{i}_{P\_CH} = \sum_{j=1}^{m} \frac{\left(Q^{j}_{P} - DB^{ij}_{P}\right)^2}{Q^{j}_{P} + DB^{ij}_{P}} \qquad (4)$$

where $Dist^{i}_{P\_CH}$ is the chi-square distance of the ith sample in the reference base with the query sample in the polygon feature space.

6. Normalize the computed distances as follows.

$$Dist^{i}_{CB\_ED} = \frac{Dist^{i}_{CB\_ED} - \mu_{Dist_{CB\_ED}}}{\sigma_{Dist_{CB\_ED}}} \forall i \qquad (5)$$

$$Dist^{i}_{CB\_CH} = \frac{Dist^{i}_{CB\_CH} - \mu_{Dist_{CB\_CH}}}{\sigma_{Dist_{CB\_CH}}} \forall i \qquad (6)$$

$$Dist^{i}_{P\_ED} = \frac{Dist^{i}_{P\_ED} - \mu_{Dist_{P\_ED}}}{\sigma_{Dist_{P\_ED}}} \forall i \qquad (7)$$

$$Dist^{i}_{P\_CH} = \frac{Dist^{i}_{P\_CH} - \mu_{Dist_{P\_CH}}}{\sigma_{Dist_{P\_CH}}} \forall i \quad (8)$$

where μ is the mean and σ is the standard deviation of the corresponding distances.

7. Compute hybrid distance between the writing in question and the samples in the training database using a weighted combination of the computed distances as:

$$Dist_{hybrid} = \omega_1 \times Dist^{i}_{CB\_ED} + \omega_2 \times Dist^{i}_{CB\_CH} + \omega_3 \times Dist^{i}_{P\_ED} + \omega_4 \times Dist^{i}_{P\_CH} \forall i \quad (9)$$

8. Identify the nearest neighbor of the query writing sample, indexed by *c,* from the dataset *DB* as:

$$c = \arg\min_{j}\left(Dist_{CB\_P\_hybrid}\right)\forall j \in DB \ (10)$$

The test sample is finally attributed to the class (l_c) of the nearest neighbor.

II. EXPERIMENTS

This section details the experiments carried out to evaluate the effectiveness of the proposed hybrid space for writer identification task. We first present the database employed in our study followed by the results of the experimental evaluations along with a comparison and discussion.

A. Dataset

We have employed the handwritten samples in of LAMIS-MSHD database [25] in our study. This database comprises writing samples of 87 different writers in French and Arabic. Each writer provided a total of 12 samples, 6 in French and 6 in Arabic. For our evaluations, we have employed only the French writings but the same approach can be applied to writings in Arabic or any other script as well.

B. Experiments and Results

We carry out a series of experiments using different combinations of training and test samples. The objective of these experiments is to compare the performance of the proposed hybrid framework with that of [7]. The codebook and polygon based features are extracted as described in Section III and the distances are combined (Eq. 9) using the following empirically determined weights.

$\omega_1 = 0.2$, $\omega_2 = 0.3$, $\omega_3 = 0.2$, and $\omega_4 = 0.3$

The different experiments and their accompanying results are summarized in Table 1. It can be observed from Table 1 that in all cases, the proposed framework outperforms the traditional feature combination used in [7]. This superiority is consistent in experiments involving 1, 2 and 3 test samples. It is interesting to note that the features in our study are the same as in [7], however, while [7] combines codebook and contour based features in a single feature vector, we keep the two representations separate and perform a combination in the distance space.

The average identification rates of experiments involving 1, 2 and 3 test samples are summarized in Table II. Naturally, the identification rates show a slight decrease when the number of training samples is reduced. But in all cases, the proposed hybrid framework reports higher identification rates as compared to those reported by the method in [7] highlighting the effectiveness of a hybrid representation space.

TABLE I: Writer identification rates on MSHD database

Training Samples	**Test Samples**	**Writer Recognition Rate**	
		Siddiqi et al.[7]	***Proposed Framework***
2,3,4,5,6	1	93.10%	**95.40%**
1,3,4,5,6	2	90.80%	**94.25%**
1,2,4,5,6	3	95.40%	**96.55%**
1,2,3,5,6	4	97.70%	**100.0%**
1,2,3,4,6	5	96.55%	**98.85%**
1,2,3,4,5	6	93.10%	**95.40%**
3,4,5,6	1,2	92.52%	**94.83%**
2,4,5,6	1,3	93.67%	**95.40%**
2,3,5,6	1,4	94.25%	**97.13%**
2,3,4,6	1,5	95.40%	**97.13%**
2,3,4,5	1,6	90.80%	**94.83%**
1,4,5,6	2,3	90.80%	**94.25%**
1,3,5,6	2,4	93.67%	**95.40%**
1,3,4,6	2,5	93.10%	**95.40%**
1,3,4,5	2,6	91.37%	**94.25%**
1,2,5,6	3,4	94.25%	**95.40%**
1,2,4,6	3,5	95.97%	**97.70%**
1,2,4,5	3,6	94.82%	**95.98%**
1,2,3,6	4,5	97.12%	**99.43%**
1,2,3,5	4,6	95.97%	**98.28%**
1,2,3,4	5,6	93.67%	**94.83%**
1,2,3	4,5,6	92.72%	**96.55%**
1,3,5	2,4,6	93.86%	**95.79%**
1,2,6	3,4,5	90.80%	**95.79%**
2,3,4	1,5,6	88.88%	**91.57%**

TABLE II: **Average identification rates using 1, 2 and 3 test samples**

Training Sample Size	**Test Sample Size**	**Writer Recognition Rate**	
		Siddiqi et al.[7]	***Proposed Framework***
5	1	94.44%	**97.01%**
4	2	93.83%	**95.86%**
3	3	93.29%	**94.93%**

III. CONCLUSION

This paper presented a framework for writer identification from offline handwritten documents. The objective of this study was not to present a novel set of features but to propose an effective combination of existing features which results in improved identification rates. Traditionally, multiple features are combined into a single large feature vector or the distances of multiple features are combined using a weighted average. The proposed framework, instead of considering the features individually and applying a single distance function, combines these features using multiple distance functions. The framework has been thoroughly evaluated by considering a number of combinations of training and test samples on MSHD database of French handwritings and performed consistently better than the traditional feature combination scheme. It should also be noted that the proposed hybrid representation is not limited to writer identification problem only and can be applied to a number of pattern classification problems where the existing feature spaces can be combined by multiple distance functions to enhance the overall classification rates. In our future work, we intend to extend the proposed framework by employing supervised dimensionality reduction techniques. Different distance metrics and their combinations can also be evaluated.

REFERENCES

[1] A. Bensefia, T. Paquet, L. Heutte. Handwritten Document Analysis for Automatic Writer Recognition. Electronic Letters on Computer Vision and Image Analysis.2005,pp. 72-86.

[2] M. Sreeraj, S. M. Idicula.A Survey on Writer Identification Schemes. International Journal of Computer Applications. 2011,pp. 23-33.

[3] S. M. Awaida, S. A. Mahmoud.State of the art in off-line writer identification of handwritten text and survey of writer identification of Arabic text.Educational Research and Reviews. 2012,pp. 445-463.

[4] S. Srihari, S.H.Cha, H. Arora, S. Lee.Individuality of Handwriting : A Validity Study. 6th International conference on Document Analysis and Recognition. 2001, pp. 106-109.

[5] S. N. Srihari, S. H. Cha, H. Arora S. Lee. Individuality of handwriting. Journal of Forensic Sciences. 2002

[6] M. Bulacu, L. Schomaker, L. Vuurpijl. Writer identification using edge-based directional features. 7th International Conference on Document Analysis and Recognition. 2003

[7] I. Siddiqi, N. Vincent. Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features. Journal of Pattern Recognition. 2010,p. 3853–3865.

[8] O. Kirli, M. Gulmezoglu. Automatic writer identification from text line images. International Journal of Document Analysis and Recognition. 2011, pp. 1-15.

[9] G. S. Peake, T. N. Tan. Script and language identification fromdocument images.British Machine Vision Conference, 1997.

[10] T. Tan. Written language recognition based on texture analysis.International Conference on Image Processing.1996.

[11] M. Bulacu. Statistical Pattern Recognition for Automatic Writer Identification and Verification. University of Groningen. 2007

[12] R. Hanusiak, L. S. Oliveira, E. Justino, R. Sabourin. Writer verification using texture-based features. International Journal on Document Analysis and Recognition. 2012

[13] A. Bensefia, A. Nosary, T. Paquet, L. Heutte. Writer identification by writer's invariants. IEEE Proceedings, Niagra. 2002

[14] A. Bensefia, L. T. Paquet , Heutte. Information retrieval based writer identification. ICDAR, UK. 2003

[15] L. Schomaker, M. Bulacu. Automatic writer identification using connected component contours and edgebased features of upper-case western script. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2004

[16] A. Seropian, M. Grimaldi, N. Vincent. Writer identification based on the fractal construction of a reference base. 7th International Conference on Document Analysis and Recognition. 2003

[17] I. Siddiqi, N. Vincent. Writer identification in handwritten documents. 9thInternational Conference on Document Analysis and Recognition. 2007

[18] G. Ghiasi, R. Safabakhsh. An efficient method for offline text independent writer identification. International Conference on Pattern Recognition. 2010

[19] S. M. Awaida, S. A. Mahmoud. Writer identification of Arabic text using statistical and structural features. An International Journal of Cybernetics and Systems. 2013,pp. 57-76

[20] H.E.S. Said, T.N Tan, K.D. Baker. Personal identification based on hand writing. Pattern Recognition. 2000,(33). pp 149-160,

[21] U.V. Marti, R. Messerli, H. Bunke, "Writer Identification Using Text Line BasedFeatures", Proc. ICDAR'01, Seattle (USA), pp 101-105, 2001

[22] A. K. Jain, F. D. Griess, S. D. Connell. On-line signature verification. Journal of Pattern Recognition. 2002,pp. 2963 – 2972.

[23] G. Louloudis, B.Gatos, N. Stamatopoulos. Competition on Writer Identification. International Conference on Frontiers in Handwriting Recognition. 2012

[24] F. Shahabi, M. Rahmati. A New Method for Writer Identification of Handwritten Farsi Documents. *ICDAR*, 2009

[1] Djeddi, C., Gattal, A., Souici-Meslati, L., Siddiqi, I., Chibani, Y., & El Abed, H. LAMIS-MSHD: A Multi-Script offline Handwriting Database, Proc. of International Conference on Frontiers in Handwriting Recognition, ICFHR 2014.