Arabic Writer Identification System Using the Histogram of Oriented Gradients (HOG) of Handwritten Fragments

Yaâcoub Hannad
MISC Laboratory, Ibn Tofail University,
Kenitra, Morocco
y.hannad@gmail.com

Imran Siddiqi
Department of Computer Science, Bahria University, Islamabad, Pakistan
imran.siddiqi@bahria.edu.pk

Youssef El Merabet
Department of Physics, Ibn Tofail University, Kenitra, Morocco
y.el-merabet@univ-ibntofail.ac.ma

Mohamed El Youssfi El Kettani
MISC Laboratory, Ibn Tofail University,
Kenitra, Morocco
elkettani@univ-ibntofail.ac.ma

ABSTRACT

This paper [1] presents an enhanced approach for writer identification from offline Arabic handwriting samples in text-independent mode. Based on the hypothesis that graphical fragments in handwriting are individual, we propose a technique based on texture analysis where the handwriting is divided into small fragments and each fragment is represented by the histogram of oriented gradients (HOG). The set of HOG descriptors for all the fragments in the writing is used to characterize its writer. The proposed system is evaluated using writing samples of the IFN/ENIT database realizing an identification rate of 86.62% on 411 writers.

KEYWORDS

Writer identification, Arabic handwriting, Histogram of oriented gradients (HOG), Handwritten fragments.

1 INTRODUCTION

Handwriting is known to be an effective biometric modality and identification of writers from handwritten documents finds applications in areas like forensic document analysis [1], classification of ancient manuscripts [2] and verification of signatures [3]. A validation study in [4] has demonstrated that handwriting of every individual is unique. The writer-specific characteristics can be extracted from a handwritten sample allowing recognizing the author of a given query document. This identification of writers from handwriting images has enjoyed significant research attention over the last two decades. The initial research focused on a limited number of writers and comparison of individual characters. The solutions matured over the years and techniques dealing with a large number of writers and unconstrained handwriting were developed.

Writer identification can be carried out at the time of writing (online) where in addition to the shapes of characters and words additional information in terms of writing speed, pressure and writing time etc. are also available. This naturally requires specialized hardware devices to capture the dynamic information in the writing. Offline writer identification relies on digitized images of handwriting and features extracted from writing images are exploited to characterize the writer. Traditionally, researchers also discriminate between text-dependent and text-independent writer identification. In text-dependent mode, each writer is required to produce the same pre-defined text. Hence, during identification, images comprising same textual content are compared. In text-independent mode, the textual content of images to be compared is different and more closely represents the real world scenarios.

The focus of the present study lies on offline, text-independent writer identification from handwritten documents in Arabic. Literature on Arabic writer identification is relatively limited once compared to that on handwritten text in the Roman script. Features typically employed to characterize writer from handwriting can be categorized into structural and statistical features which can be computed globally or locally. Structural features attempt to describe the structural characteristics of writing for instance, inclination, information on ascenders and descenders, loops etc. The work proposed in [5], for example, is based on structural features like line height and mean of the

MedPRAI-2016, November 22 - 23, 2016, Tebessa, Algeria

ACM 978-1-4503-4876-8/16/11 $15.00
DOI: http://dx.doi.org/10.1145/3038884.3038900

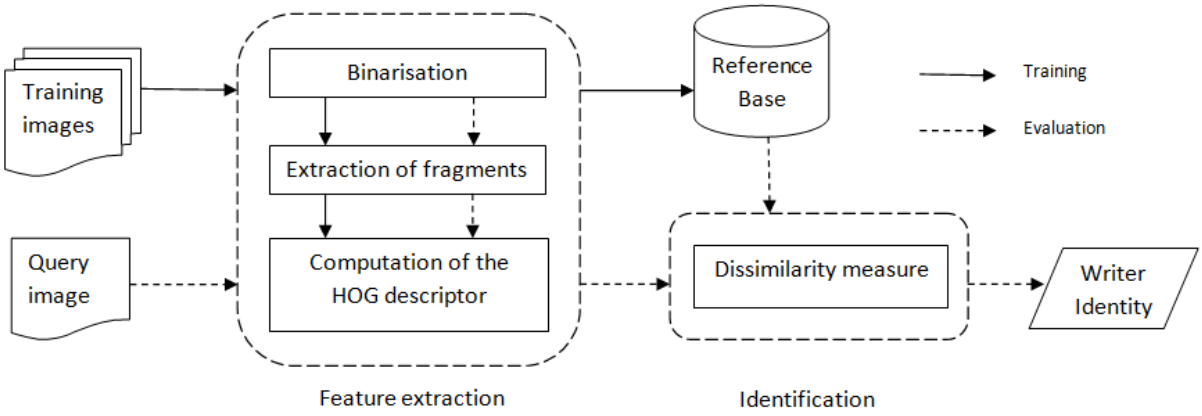

Figure 1. An overview of the proposed System

Figure 2. Extraction of the connected components: (a) A sample Arabic word. (b) The extracted connected components [14]

ascender's inclination. A direct comparison between handwriting fragments to identify authorship of a document is presented in [6]. Likewise, in [7], authors extract gradient distributions and chain code based features to improve writer identification. Textural features [8, 9] are also known to be effective for characterizing the author of a handwritten sample. Texture features extracted using Grey Level Run Length Matrix (GLRL) and Grey Level Co-occurrence Matrixes (GLCM) have been effectively employed for writer recognition in [8]. Likewise, authors in [9] exploit textural features LBP and LPQ for characterizing the writer from handwriting.

In some cases, different types of features are combined to enhance the overall identification rates. The work in [10], for instance, combines the textural features based on probability distributions of contour directions and contour hinges with allographic features to improve the identification rates. Likewise, a combination of different features is investigated in [11] to improve the identification performances.

An effective technique investigated in a number of studies [6, 10, 12, 13] is to exploit small writing fragments to characterize authorship from handwriting images. These methods divide the handwriting text into small fragments which are then employed to characterize the writer either through a direct comparison [6, 13] or by first generating a codebook [10, 12]. The written fragments are either compared directly (pixel values) or by first representing them in an appropriate feature space and then compared. Based on the same idea of small writing fragments, we presented in our previous work [14] a new approach that represents each writing fragment by a set of textural descriptors (LBP, LPQ and LTP). The technique realized promising results on writer identification from Arabic handwriting images. The present study extends this work [14] to investigate the effectiveness of histogram of oriented gradients (HOG) [15] to characterize writer style from small fragments of writing. HOG is a well-known descriptor in computer vision that has been mostly applied to object recognition problems. It would be interesting to study its effectiveness on small fragments of writing to identify the writer. The detailed methodology of the proposed technique is presented in the next section followed by the experimental results and analysis. The final section presents the concluding remarks and open research areas on this subject.

2 Proposed Methodology

The proposed methodology is based on pre-processing the handwriting image, feature extraction from small writing fragments and classification (identification). The key steps in the proposed system are shown in Figure 1 while each of these steps is discussed in detail in the following sections.

2.1 Pre-processing

The pre-processing step in our case involves converting the images to binary. Since we evaluate the system on contemporary handwriting documents, a global thresholding is defined using the Otsu threshold method [16] and employed to

binarize the images. Once an image is binarized, the connected components in the writing are extracted (Figure 2) for further processing.

2.2 Feature Extraction

Feature extraction involves dividing the writing into small fragments and representing each fragment by a vector. The division of writing into segments is based on the same technique as presented in our previous work [14]. Each connected component is divided into small blocks of size NxN where the window size is empirically fixed to 100x100. It should be noted that the window size considered in our study is relatively larger once compared to those employed in other similar studies characterizing writer identity from small handwriting fragments [6, 12]. The major reason is that these studies either work directly on pixel values or represent each pixel by simple shape descriptors. We, on the other hand, represent each fragment by textural descriptor(s) hence a significant proportion of handwritten text has to be present in the window to compute meaningful features. Figure 3 illustrates an example of handwritten fragments resulting from a connected component.

Figure 3. Examples of handwritten fragments extracted from a component [14]

Once the writing is divided into blocks, each fragment is represented by the histogram of oriented gradients (HOG). HOG [15] is a well-known computer vision descriptor that has been widely employed in a number of object recognition problems and is closely related to the Scale-Invariant Feature Transform (SIFT) descriptor [17]. For this study, we employ the same HOG parameters as defined in [18], i.e. 9 rectangular cells for each fragment and 9 bins per histogram per cell. These 9 histograms are then concatenated to produce an 81 dimensional feature vector that represents a fragment. This feature extraction allows to benefit from the high discriminatory power of handwritten fragments for writer identification as well as to exploit the efficiency of the HOG descriptor for an improved comparison between writing fragments during the identification step.
To represent a complete handwriting image H, the set of HOG descriptors hi computed from all the fragments is used to characterize the writer.

$$H = \{ h_i , i \le \text{card}(H) \} \quad (1)$$

The term card(H) refers to the total number of fragments in a sample. The process is repeated for writing samples of all writers under study and a reference base is produced.

2.3 Writer Identification (Classification)

During the identification stage, a questioned documented is presented to the system and the system needs to retrieve the identity of the writer of the document from the reference base. Like the reference samples, the query handwriting image is also divided into fragments and the HOG descriptor is computed for each fragment. The following dissimilarity measure is used to compare an unknown query sample U and a reference sample H.

$$\text{DIS(U,H)} = \frac{1}{\text{Card(U)}} \sum_{i=1}^{\text{Card(U)}} \underset{hj \in H}{\text{Min}}(\text{distance}(u_i - h_j)) \quad (2)$$

Where ui and hj are, respectively, the HOG descriptors of handwritten fragments of samples U and H. The distance is the Hamming distance between ui and hj and is defined as follows.

$$\text{distance}(u_i, h_j) = \sum_{n=1}^{N\text{dim}} |u_{in} - h_{jn}| \quad (3)$$

Where "Ndim" is the dimension of the HOG descriptor.
Finally, the writer of the queried sample U is identified as the least dissimilar writer in the reference database.

$$\text{Writer (U)} = \text{argmin} \left(\underset{Hi \in Base\ \text{Re}\ f}{\text{DIS(U, Hi)}} \right) \quad (4)$$

3 Experiments and Results

This section presents the details of the experiments carried out to validate the proposed identification technique. First, we present the details of the database used in our experiments. Later, we detail the different experimental results along with a discussion on the realized results.

Figure 4. A sample from the IFN/ENIT database.

3.1 Database

We have employed the widely used IFN/ENIT database [19] for experimental study of the system. The database comprises 2,200 handwritten forms with more than 26,000 handwritten Tunisian town/village names in Arabic collected from 411 different writers. A sample form from the database is illustrated in Figure 4.

3.2 Results and Discussion

The writer identification performance is evaluated on the complete set of 411 writers in the IFN/ENIT database. For each writer, 30 words are used as the reference (training) base while 20 words are used as the test set. The realized writer identification rates as a function of hit-list size are summarized in Figure 5 where can be seen that a Top-1 identification rate of 86.62% is realized.

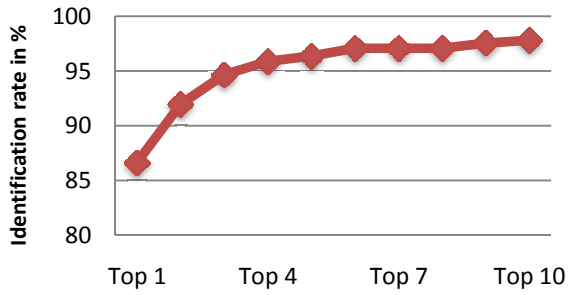

Figure 5. Writer identification rates as a function of hit-list size

Table 1: Comparison of identification rates for different texture descriptors

Systems	Database	Number of writers	Identification rate (Top1)
Hannad et al. [14] - LBP Descriptor	IFN/ENIT	411	73.48%
Hannad et al. [14] - LTP Descriptor			87.12%
Hannad et al. [14] - LPQ Descriptor			94.89%
HOG Descriptor			86.62%

We also compare the identification rate reported by the HOG descriptor with our previous work [14] that was based on LBP, LTP (concatenation of the positive and negative LTP) and LPQ descriptors. The experimental settings in [14] are exactly the same as those of the present study and the different identification rates are summarized in Table 1. It can be seen that the identification rate obtained by using the HOG descriptor outperforms that of the LBP descriptor and is very close to that of LTP. The highest identification rate, however, is realized by the LPQ descriptor [14]. Nevertheless, HOG represents an attractive choice due its relatively efficient computation and can be combined with other features to enhance the overall performance.

We also carried out a series of experiments to study the impact of the number of writers on the identification rates. There is a gradual decrease as the number of writers increases which is pretty much natural. The identification rates are relatively more stable after 200 writers reflecting the fact that writer characterization using HOG can be effectively employed on large databases.

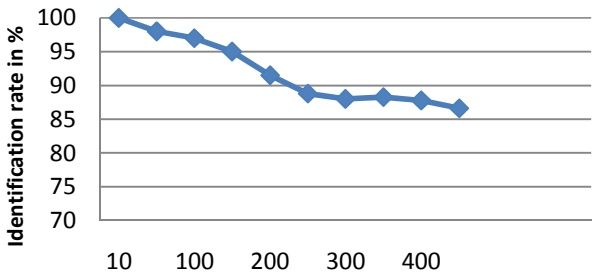

Figure 6. Writer identification rates as a function of number of writers.

4 CONCLUSIONS

We presented an offline text-independent writer identification system primarily targeting Arabic handwriting. The technique is based on local analysis of small writing fragments representing each fragment by the HOG descriptor. The set of descriptors for all the fragments in a given writing sample is exploited to characterize its writer. The system realized an identification rate of 86.62% when evaluated on the 411 writers of the IFN/ENIT database. In our further study on this subject, we intend to study the effectiveness of other textural descriptors to characterize the writing fragments. Combination of multiple features and feature selection to identify the best set of features for this problem is also planned to be carried out. Although the technique has been applied to Arabic handwriting, it is very much generic and can be applied to other scripts as well. We also plan to introduce a rejection threshold in the system so that the system is able to reject the writing samples for which the writer does not exist in the reference base.

REFERENCES

[1] Said, Huwida ES, Tienniu N. Tan, and Keith D. Baker."Personal Identification Based on Handwriting", Pattern Recognition, vol. 33, pp149-160, 2000.

[2] D.Arabadjis, F. Giannopoulos, C. Papaodysseus, S. Zannos, P. Rousopoulos, M. Panagopoulos, C. Blackwell. "New mathematical and algorithmic schemes for pattern classification with application to the identification of writers of important ancient documents." Pattern Recognition 46 (2013) 2278–2296.

[3] Rajesh Kumar, J.D. Sharma, Bhabatosh Chanda. "Writer-independent off-line signature verification using surroundedness feature". Pattern Recognition Letters 33 (2012) 301–308.
[4] Srihari, S. Cha, H. Arora, and S. Lee, “Individuality of Handwriting,” J. Forensic Sciences, vol. 47, no. 4, pp. 1-17, July 2002.
[5] S. Gazzah, N. Ben Amara, “Neural Networks and Support Vector Machines Classifiers for Writer Identification UsingArabic Script”, The second International Conference on Machine Intelligence (ACIDCA-ICMI 2005), pp. 1001-1005, Tozeur, Tunisia, 2005.
[6] D.Chawki, S. Labiba, "Une approche locale en mode indépendant du texte pour l’identification de scripteurs: Application à l’écriture arabe", CIFED'08, Rouen, France, pp. 151-156, Octobre 2008.
[7] Sameh M. Awaida & Sabri A. Mahmoud: “Writer identification of Arabic text using statistical and structural features”, Cybernetics and Systems: An International Journal, 44:1, 57-76, 2013.
[8] D. Chawki, S. Labiba, “A texture based approach for Arabic Writer Identification and Verification”, IEEE International Conference on Machine and Web Intelligence, pp 115 – 120, Octobre 2010.
[9] D Bertolini, LS Oliveira, E Justino, R Sabourin, “Texture-based descriptors for writer identification and verification”, Expert Systems with Applications, vol. 40, pp 2069-2080, 2013.
[10] M. Bulacu, L. Schomaker, A. Brink, “Text-Independent Writer Identification and Verification on Offline Arabic Handwriting”, Proc. of 9th Int. Conference on Document Analysis and Recognition (ICDAR 2007), IEEE Computer Society, pp. 769-773, vol. II, 23 - 26 September, Curitiba, Brazil, 2007.
[11] N. Abdi, M. Khemakhem, H. Ben-Abdallah, “An Effective Combination of MPP Contour-Based Features for Off-Line Text-Independent Arabic Writer Identification”, Communications in Computer and Information Science, Volume 61. ISBN 978-3-642-10545-6. Springer-Verlag Berlin Heidelberg, p. 209, 2009.
[12] Siddiqi, I and Vincent, N. "Writer identification in handwritten documents." Proc. Of the 9th International Conference on Document Analysis and Recognition, 2007.
[13] Bensefia, A., Paquet, T., and Heutte, L. "A writer identification and verification system." Pattern Recognition Letters, vol. 26 (13), pp. 2080-2092, 2005.
[14] Hannad, Y., Siddiqi, I., and El Kettani, M. E. Y. (2016). "Writer identification using texture descriptors of handwritten fragments". Expert Systems with Applications, 47, 14-22.
[15] Dalal, N. and Triggs, B., Histograms of oriented gradients for human detection, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, pp. 886-893.
[16] N. Otsu. "A Threshold Selection Method from gray Level histogram". IEEE Transactions on Systems, Man and Cybernetics, 1979, vol. 9, pp. 62-66.
[17] Lowe, D., "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, vol. 60, pp. 91110, 2004.
[18] O. Ludwig, D. Delgado, V. Goncalves, and U. Nunes, "Trainable Classifier-Fusion Schemes: An Application To Pedestrian Detection, " In: 12th International IEEE Conference On Intelligent Transportation Systems, St. Louis, 2009. V. 1. P. 432-437.
[19] M. Pechwitz, S. Maddouri, V. Märgner, N. Ellouze , H. Amiri, “IFN/ENIT - Database of handwritten arabic words”, in the 7th Colloque International Francophone sur l'Ecrit et le Document , CIFED 2002, Oct. 21-23, Hammamet, Tunis, 2002.