Expert Systems With Applications 47 (2016) 14–22

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Writer identification using texture descriptors of handwritten fragments

Yaacoub Hannad [a,*], Imran Siddiqi [b], Mohamed El Youssfi El Kettani [a]

[a] ACRIS Laboratory, Ibn Tofail University, Kenitra, Morocco
[b] Center of Computer Vision and Pattern Recognition, Bahria University, Islamabad, Pakistan

ARTICLE INFO

Keywords:
Handwritten documents
Writer identification
Texture descriptors
Handwriting fragments

ABSTRACT

This paper presents a texture based approach for identification of writers from offline images of handwriting. Contrary to the classical texture based techniques which extract texture information at page or block level, we exploit the texture at a very small observation scale. The proposed technique divides a given handwriting into small fragments and considers each fragment as a texture. Texture descriptors including histograms of Local Binary Patterns (LBP), Local Ternary Patterns (LTP) and Local Phase Quantization (LPQ) are then computed from these fragments. The writer of a document is characterized by the set of histograms calculated from all the fragments in the writing. Two writings are compared by computing the distance between the descriptors of their writing fragments. The technique evaluated on IFN/ENIT and IAM databases comprising handwritten text in Arabic and English, respectively, realized high identification rates.

1. Introduction

The analysis of handwriting and hand-drawn shapes has been an interesting research area for many centuries that has attracted a significant number of psychologists, graphologists, palaeographers and forensic experts to solve a wide variety of problems. With the theoretical and practical advancements in computers and the availability of appropriate tools and technologies, automated systems for analysis of handwriting are being researched and developed for more than three decades now. Although these computerized tools cannot completely replace the human analysis, they serve a great deal by assisting the human experts through features like reduction of search space, visualization, segmentation, efficient processing, etc. In addition to the classical task of handwriting recognition, other relevant problems include classification of writing styles, keyword spotting of handwritten words, studies on correlation between handwriting and different neurological disorders, prediction of writer demographics from handwriting and verification and identification of writers from handwritten samples. The focus of this study lies on the last of the aforementioned problems i.e. identification of individuals (writers) from samples of their handwritten text.

Formally, given a set of handwritten documents with known authorship, the writer identification task involves finding the writer of a query document comparing it with the samples in the reference base. Writer identification mainly finds applications in forensic document analysis (Said, Tan, & Baker, 2000), classification of historical archives (Arabadjis et al., 2013), verification of signatures (Kumar, Sharma, & Chanda, 2012), etc. Writer identification techniques are generally categorized into online and offline methods depending upon the type of writing samples considered in the study. Offline writer identification works on digitized images of writing and relies only on allographic variations to determine the identity of the author of a questioned document. Online writer identification is carried out at the time of writing on specialized devices and, in addition to the shape of characters, online methods also make use of dynamic features of writing including the writing trajectory, number and order of strokes and pen pressure etc.

As a function of textual content, writer identification techniques are categorized into text-dependent and text-independent methods. Text-dependent methods employ the same text in training and evaluation modes while text-independent methods are unconstrained in terms of the textual content of writing samples and hence represent more closely the scenarios encountered in the real world.

Among one of the pioneer works, a validation study (Srihari, Cha, Arora, & Lee, 2002) exploiting the handwritten samples of 1500 writers proved the hypothesis that handwriting is individualistic and can be used as an effective biometric modality. Later, a number of writer identification techniques were proposed based mainly on two broad groups of features, structural and textural. Structural features, computed at local or global levels, are aimed to capture the structural properties of writing like average line height, inter and intra-word distances and inclination etc. (Bulacu & Schomaker, 2007; Siddiqi & Vincent, 2010; Srihari et al., 2002). High identification rates have been reported by a number of writer identification schemes based

* Corresponding author. Tel.:+212 638093170 .
E-mail addresses: y.hannad@gmail.com (Y. Hannad), imran.siddiqi@bahria.edu.pk (I. Siddiqi), elkettani@univ-ibntofail.ac.ma (M.E.Y. El Kettani).

http://dx.doi.org/10.1016/j.eswa.2015.11.002
0957-4174/

Table 1
Performance of writer identification systems.

Study	Alphabet	Database	Number of writers	Identification rates
Srihari et al. (2002)	Latin	CEDAR	1500	98%
Kırlı and Gülmezoğlu (2012)	Latin	IAM	93	98%
Schomaker, Bulacu, and Franke (2004)	Latin	Firemaker	250	97%
Jain and Doermann (2011)	Latin	IAM	300	93%
Khalifa, Al-maadeed, Tahir, Bouridane, and Jamshed (2015)	Latin	IAM	650	92%
Siddiqi and Vincent (2010)	Latin	IAM	650	91%
Abdi and Khemakhem (2015)	Arabic	IFN/ENIT	411	90%
Abdi, Khemakhem, and Ben-Abdallah (2009)	Arabic	IFN/ENIT	82	90%
Bulacu et al. (2007)	Arabic	IFN/ENIT	350	88%
Hannad et al. (2015)	Arabic	IFN/ENIT	130	87%
Chawki and Labiba (2010)	Arabic	IFN/ENIT	130	82%

on structural features. These methods, however, are costly in terms of execution time mainly due to the complexity of segmentation and feature extraction steps.

Systems based on texture analysis of handwriting consider each writing as a different texture and rely on extracting a set of features from different regions of interest (blocks) or complete image (Bertolini, Oliveira, Justino, & Sabourin, 2013; Chawki & Labiba, 2010; Said et al., 2000) of writing. The main advantage of texture based methods is their efficiency in terms of execution time mainly thanks to the fast computation of texture features. Among different texture analysis systems proposed in the literature, an interesting work proposed by Bertolini et al. (2013) achieved high identification rates by applying two local texture descriptors on normalized image blocks. These include the Local Binary Patterns (LBP) (Ojala, Pietikäinen, & Mäenpää, 2002) and the Local Phase Quantization (LPQ) (Ojansivu & Heikkilä, 2008). Though these local texture descriptors have been widely employed in a number of texture classification (Ojala et al., 2002; Ojansivu & Heikkilä, 2008) and face recognition tasks (Lei & Li, 2012; Pujol & García, 2012), their only application to handwriting has been studied in (Bertolini et al., 2013) to the best of authors' knowledge.

Inspired by the effectiveness of the local texture descriptors in texture classification problems and the high discriminatory power of small writing fragments in characterizing the writer (Bensefia, Nosary, Paquet, & Heutte, 2002; Bulacu & Schomaker, 2007; Djeddi & Souci-Meslati, 2008; Siddiqi & Vincent, 2007), we proposed in our previous work (Hannad, Siddiqi, & El Kettani, 2015) a new approach for writer identification from Arabic handwritten samples. The technique in Hannad et al. (2015) is based on local analysis of texture using the LBP-histogram of small writing fragments as feature. The system realized promising identification rates with reduced execution time due to the fast computation of the LBP descriptor. The performance of the LBP descriptor, however, degrades when evaluated on large databases of handwritten samples. In an attempt to maintain high identification rates on relatively large databases, this study extends the work in Hannad et al. (2015) by introducing two more texture descriptors. These include the Local Ternary Patterns (LTP) (Tan & Triggs, 2007) and the Local Phase Quantization (LPQ) (Ojansivu & Heikkilä, 2008). Contrary to the approach presented in Bertolini et al. (2013) where the authors extract these descriptors by converting the writing into normalized texture blocks, we employ a much smaller scale of observation and extract these features from small fragments of writing.

An important issue regarding writer identification on Arabic documents has been raised by Bulacu, Schomaker, and Brink (2007) where the writer identification system, primarily developed for Latin scripts, was directly evaluated on Arabic handwritings and achieved lower identification rates. The authors concluded that writer identification on Arabic script is more challenging as compared to that on the Roman script. This hypothesis is supported by a performance comparison of some of the well-known works on writer identifications reported in the literature (Table 1). It can be seen that though Arabic writer identification systems have been evaluated on relatively smaller databases, the identification rates reported by these systems are not as high as those reported for text in scripts based on the Latin alphabet.

This study is aimed at developing an effective, text-independent writer identification technique for large databases of offline handwritten documents. The key objective is to improve the Arabic writer identification performance and attain identification rates of the same order as reported by studies on other scripts. This is realized by extending our work (Hannad et al., 2015) on texture analysis of small writing fragments. The texture information of writing is captured by using three texture descriptors namely LBP, LTP and LPQ. The effectiveness of these descriptors in characterizing the author of handwriting is evaluated on the complete IFN/ENIT (Pechwitz et al., 2002) and IAM (Marti & Bunke, 2002) databases. The stability of the system performance with respect to the size of the database and other system parameters is also studied.

This paper is organized as follows. In the next section, we describe the IFN/ENIT and IAM databases used in our experimental study. In Section 3, we introduce the proposed texture based features followed by a presentation of the dissimilarity measure used for classification. Section 5 details the experiments along with the realized results and analysis. The last section concludes the paper with a discussion on potential research directions on this problem.

2. Databases

The experimental study of the proposed system was carried out on two standard databases, the IFN/ENIT (Pechwitz et al., 2002) Arabic database and the IAM (Marti & Bunke, 2002) English database. Each of these databases is briefly discussed in the following.

2.1. IFN/ENIT database

The IFN/ENIT database is the most widely used Arabic handwritten database (Pechwitz et al., 2002). It comprises 2200 forms with more than 26,000 handwritten Arabic Tunisian town/village names collected from 411 different writers. All forms are scanned at a resolution of 300 dpi and are available as binary images. The database was mainly developed for training and evaluation of Arabic handwriting recognition systems. However, since the writer information is also stored with the forms, this database has been widely used for evaluation of writer identification systems. For the experimental study of our system, we kept the complete set of 411 writers but employed a reduced set of samples per writer to correspond to the real world problems where limited text is available for training and evaluation. These real world problems may include scenarios like those encountered by the forensic examiners where the authorship of a writing is to be concluded. The biggest challenge in such problems is the availability of a limited amount of text from which the authorship is to