Pattern Recognition Letters 59 (2015) 18–25

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

# Off-line writer identification using an ensemble of grapheme codebook features ☆

E. Khalifa [a], S. Al-maadeed [b,*], M.A. Tahir [a,c], A. Bouridane [a], A. Jamshed [c,d]

[a] Department of Computer Science and Digital Technologies, Northumbria University, Newcastle Upon Tyne, NE1 9ST, UK
[b] Department of Computer Science and Engineering, College of Engineering, Qatar University, Doha, 2713, Qatar
[c] College of Computer and Information Sciences, Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, 13318, KSA
[d] Predictify.me Inc., Raleigh, NC 27601, USA

ARTICLE INFO

*Article history:*
Received 27 March 2014
Available online 17 March 2015

*Keywords:*
Writer identification
Forensic document examination
Kernel discriminant analysis
Grapheme features

ABSTRACT

Off-line writer identification is the process of matching a handwritten sample with its author. Manual identification is very time-consuming because it requires a meticulous comparison of character shape details. Consequently the automation of writer identification has become an important area of research interest. The codebook (or bag of features) approach is a state-of-the-art computerized technique for writer identification. One way to achieve a high identification rate is to expose the personalized set of character shapes, or allographs, that a writer has adopted over the years. The main problem associated with this approach is the extremely large of number of points of interest that are generated. In this paper we extend the basic model to include an ensemble of codebooks. Additionally, Kernel discriminant analysis using spectral regression (SR-KDA) is used as a dimensionality reduction technique in order to avoid over-fitting. Fusion of multiple codebooks is shown to increase the identification rate by 11% compared with a single codebook approach.



## 1. Introduction

The identification of an individual from a sample of handwriting is a very important technique in forensic document analysis, with applications in the criminal justice system. The relationship between a character's shape and a writing style differs from person to person, with no two individuals writing in exactly the same way. To complicate matters an individual will not write a piece of text in exactly the same way twice. These two conflicting facts make writer identification in forensic document analysis a rather complex process. An automatic writer-identification system should be able to answer the question: "Who wrote this sample?" Typically an automated identification system performs a one-to-many recognition task, using a large database containing samples of known authorship, and returns a ranked list of candidates whose writing style is similar to the input sample. However, it remains challenging to devise effective procedures for identifying authors accurately from a sample of handwriting.

The bag of words approach is a state-of-the-art method for off-line writer identification and verification [4,5,7,8]. The approach involves the computation of a local descriptor using grapheme blobs, vector quantization via clustering, structured writer representation via localized histograms of vector codes, and similarity measures for kernel construction and classifier learning. Research is ongoing to develop better techniques for each of these methods.

Most current approaches to writer identification generate a codebook vector using only a single clustering algorithm. The main disadvantage with using only a single algorithm is the extent of memory usage because of the extremely large number of grapheme features (interest points). For example, approximately one million points of interest are generated for written samples of 100 authors. It is difficult to cluster effectively using only a single clustering technique. In this paper a novel approach, which utilizes an ensemble of codebook grapheme features, is proposed to address this problem. Ensemble techniques have been used repeatedly to improve the accuracy of single-classifier systems. The approach proposed here generates several diverse codebook vectors from randomly selected grapheme features. In order to reduce the high dimensionality of the resulting feature vectors, kernel discriminant analysis, using spectral regression, is employed. This approach avoids the over-fitting problem associated with combining multiple codebooks. The fusion of multiple codebook features can be applied to obtain a ranked candidate list.

The remainder of the paper is organized as follows: Section 2 reviews related feature extraction methods. The proposed ensemble of grapheme codebooks for feature extraction is discussed in Section 3. This section includes code generation and subsequent

☆ This paper has been recommended for acceptance by F. Tortorella.

* Corresponding author. Tel.: +974 4403 4262; fax: +974 4403 4241.
*E-mail address:* s_alali@qu.edu.qa (S. Al-maadeed).

http://dx.doi.org/10.1016/j.patrec.2015.03.004
0167-8655/

dimensionality reduction using spectral regression combined with Kernel discriminant analysis. It is shown how the technology can be applied to writer identification. Section 4 compares the experimental results obtained from both the single and multiple reference codebooks. An analysis of the results, including comparison with a number of existing similar techniques, is also given. Finally, some conclusions are drawn in Section 5.

## 2. Related work

This section points out some of the notable contributions to writer identification and the discussed approaches are not all inclusive. For interested readers, an exhaustive studies on the subject which informatively provide a comprehensive analysis and comparison of different approaches are presented in papers by Chen [10], Sreeraj and Idicula [27], Kore and Apte [20].

During the last decade, many researchers have contributed to the field of off-line writer identification. Srihari et al. [28] presented what is considered to be one of the most significant studies authenticating the individuality of handwriting. A dataset of handwritten samples of over 1500 writers was collected and examined. The authors identified a collection of macro and micro features that were extracted from each of the handwritten samples. While the achieved identification accuracy rates are highly notable, the dataset shares alike passages from all participant writers and manual segmentation is required, which could be considered as an obstacle in real world scenarios.

Schlapbach et al. [25] proposed a handwriting identification system based on shifting pixel-wide windows from left to right over each line of text to extract, at each position, nine basic geometric features. These features include the number of ink (black) pixels in the window, position, and contour directions to the upper and lower-most pixel, thus allowing the construction of a hidden Markov model (HMM) for each writer. An identification rate of 94% was achieved using only 50 writers from the Institut für Informatik und Angewandte Mathematik (IAM) dataset [22] by using a log likelihood output through the Viterbi algorithm to rank the writers. Schlapbach et al. [26] proposed a further improvement to the system by using a Gaussian mixture model instead of the hidden Markov model, where the system achieved an identification rate of 98.5% using 100 writers. However, perfect line segmentation and line normalization with respect to slant, skew, baseline location and height, was required in both techniques.

Bulacu and Schomaker [6] introduced a set of techniques that exploited two primary sources of information relating to the individuality of the handwriting samples. The first source includes the handwriting slant, curvature, and roundness of the handwriting sample captured by a joint directional probability distribution operating at the texture level. The second source of information includes character shapes, or allographs, that are captured by a grapheme-emission probability distribution operating at the character level. The authors achieved an identification accuracy rate of 89% using 650 writers from the IAM dataset, by combining textural level and allograph level analyses, which is also known as the grapheme codebook. Siddiqi and Vincent [17] presented a study that demonstrated ideas similar to the grapheme codebook technique. They shifted the focus to the sub-grapheme level to generate a reference base (codebook) using smaller stroke fragments. An 84% identification rate was achieved using 650 writers from the IAM dataset. In a later paper [18], Siddiqi and Vincent extended their work by introducing contour-based orientation and curvature features by combining them with the codebook features. The authors achieved an identification accuracy rate of 91% on 650 writers from the IAM dataset, thereby achieving the current benchmark for writer identification.

Said et al. [24] avoided the segmentation complexity stage and presented a work that was based on the overall look and feel of the writing (looking at handwriting as texture), where texture information is extracted by applying Gabor filter and co-occurrence matrices. From each handwriting sample, a total of 82 features are extracted by using multichannel Gabor filtering and gray-scale co-occurrence matrices. Two classifiers used the weighted Euclidean distance and the nearest neighbor to carry out the identification. Avoiding segmentation is a plus in this approach, however, it requires text line normalization, in which orientation and gaps between words and margins have a predefined size.

Amaral et al. [1] presented a baseline system using graphometric features which are extracted from different levels including documents, lines and words. These features provide information which includes relative placement habits (number of lines in forensic letter, right margin position, lower left margin position, upper margin position and bottom margin position) and relative relationship between individual word height (proportion of black pixels and height of the first word). A fusion of these features is submitted to support vector machine (SVM) classifier as a writer identification method. An identification rate of 80% is achieved on 20 writers.

Fiel and Sablatnig [12] presented a work based on the codebook method to generate clustering features extracted by using the scale invariant feature transform (SIFT) using various pages of handwriting. The advantage of using SIFT from the authors point view is to eliminate the negative effects of binarization. An identification rate of 90.8% using the IAM dataset of 650 writers is achieved. However, the dataset setup is notably different from the current dataset setup used in this paper. In [11] Daniels and Baird proposed a technique to investigate the performance of five highly discriminating features. These features include slant and slant energy, skew, pixel distribution, curvature, and entropy. The performance obtained by combining these features showed recognition rates competitive with other state of the art methods for writer identification.

Tang et al. [29] proposed two feature extraction methods: the stroke fragment histogram (SFH) and local contour pattern histogram (LCPH). Vasquez et al. [31] achieved writer identification based on extracting calligraphic features from each word in handwritten samples. An artificial neural network with a decision fusion block is used as classifier. An identification rate of 94.6% is achieved on 100 writers.

Previous contributions have shown bag of words feature extraction to deliver the highest off-line writer identification rate [4,5,7,8]. The theory of bag of words feature is built around the idea that each writer acts like a stochastic generator of ink blobs that he/she has learned over the years. These grapheme blobs are considered to be distinguishable between writers and similar for a given writer [4,5,7,8]. Another idea was to explore writer redundancy patterns by working on a much smaller scale of observation [18]. However, all previous contributors to off-line writer identification using the bag of words model were recorded at a maximum identification performance rate of 81% on the IAM benchmark dataset. The study detailed in the following section will extend the bag of words approach to "*bags*" of words approach, or what is called the fusion of multiple grapheme codebook features. This method achieves an 11% increase in identification rate performance when tested on different datasets. Table 1 shows advantages and disadvantages of some recent techniques.

## 3. Proposed ensemble of grapheme codebook features

In this section we discuss our proposed ensemble of grapheme codebook system in order to develop a writer identification system that is capable of carrying out a one-to-many examination in a large database of handwriting samples of known writers to produce a ranked list of potential candidates. The resulting list of likely candidates is intended to facilitate forensic document experts in reaching a decision concerning the identity of the writer of an unknown sample in a timely manner. In a practical forensic setup between four to six samples per writer can be enough to achieve acceptable results. The proposed approach is an extension of the traditional codebook approach. The traditional approach, as illustrated in Fig. 1, consists of