Pattern Recognition 78 (2018) 318–330

Contents lists available at ScienceDirect

Pattern Recognition

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

An improved online writer identification framework using codebook descriptors

Vivek Venugopal, Suresh Sundaram*

Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India

ARTICLE INFO

Article history:
Received 28 August 2017
Revised 3 November 2017
Accepted 24 January 2018
Available online 1 February 2018

Keywords:
Online writer identification
Codebook
Descriptor
Gap parameter,

ABSTRACT

This work proposes a text independent writer identification framework for online handwritten data. We derive a strategy that encodes the sequence of feature vectors extracted at sample points of the temporal trace with descriptors obtained from a codebook. The derived descriptors take into account, the scores of each of the attributes in a feature vector, that are computed with regards of the proximity to their corresponding values in the assigned codevector of the codebook. A codebook comprises a set of codevectors that are pre-learnt by a k-means algorithm applied on feature vectors of handwritten documents pooled from several writers. In addition, for constructing the codebook, we consider features that are derived by incorporating a so called ‘gap parameter’ that captures characteristics of sample points in the neighborhood of the point under consideration. We formulate our strategy in a way that, for a given codebook size k, we employ the descriptors of only $k-1$ codevectors to construct the final descriptor by concatenation. The usefulness of the descriptor is demonstrated by several experiments that are reported on publicly available databases.

1. Introduction

The problem of writer identification refers to the task of deciding on the authorship of a piece of handwritten document by comparing it against a set of samples saved in a database [1]. Based on the mode of data capture, such systems are categorized into either online or offline. The recent advances in technology has enabled the release of hand held devices, wherein the data entry is captured through an electronic pen / stylus. The tip of the stylus, as such, has the capability to capture the trajectory information such as (x, y) coordinates, time stamp and pen status from the handwriting. In the literature of writer identification, the analysis of such temporal data is referred to as ‘online’. The input to such a system consists of a set of strokes, each of which containing the sample points of the trace captured between a pen-down and pen-up signal. Off-line writer identification systems, on the other hand, capture the data as an image and subsequently establish the authorship by applying image processing techniques [2–5].

Another classification for online writer identification systems are those of text dependent and text independent approaches [6]. In the former, the handwriting samples of a writer are processed based on a specific transcript usually with the aid of a recognizer. The problem of signature recognition / verification is one such popular instance of text dependent writer identification [7–9]. In general, the use of the knowledge of the content of the data increases the accuracy of such systems. However, they fail in scenarios that require the textual contents of the documents to be different. Hence, as an alleviation to this, text independent writer identification systems capture the style information of handwriting and can identify the writer irrespective of the textual content. In this research, we focus our proposal towards such a system.

With regards to prior techniques being proposed for online writer identification, a popular one is that of the Gaussian Mixture Model-Universal Background Model (GMM-UBM) [10–12] - inspired from the domain of speaker identification [13]. The enrolled handwritten data from a set of writers is first used to learn the parameters of the UBM by employing the Expectation Maximization (EM) algorithm. Thereafter, individual GMMs are obtained for every writer from the UBM by applying the MAP adaptation on their corresponding training samples. The authorship of a test document is assigned to that writer, whose corresponding GMM outputs the highest log-likelihood score.

A number of strategies related to the domain of information retrieval have been explored in the literature of online writer identification. A score derived from the term-frequency inverse-document-frequency (tf-idf) weighing scheme is employed to represent the handwritten document of unknown authorship in [14–17]. Likewise, the works of [18–20] consider the concept of La-

* Corresponding author.
E-mail address: sureshsundaram@iitg.ernet.in (S. Sundaram).

https://doi.org/10.1016/j.patcog.2018.01.023

tent Dirichlet Allocation from the area of topic models. Here, each document is modelled with the assumption of being a combination of finite (shared) writing styles, that in turn is a combination of a set of text independent feature probabilities. The idea of subtractive clustering is explored to determine the unique writing styles / prototypes of a writer in [21]. Thereafter, for writer identification, a modified tf-idf approach and a nearest neighbor based method are considered. In the work [22], the idea of multi-fractals are used to model the segmented graphemes / substrokes. Subsequent to it, the authors utilize a weighting based on tf-idf framework. The scoring of the tf and idf term is based on a frequentist approach, that in a way, characterises the number of segmented grapheme patterns assigned to a given codevector in a codebook. In addition, based on the grouping of the graphemes for Persian / Arabic script, separate codebooks are constructed for each of them. Another work is that of Dwivedi et al. [23], where the exploration of a sparse coding framework has been proposed for learning the different prototypes of a given writer, which are then utilized for establishing his / her identity. The features considered for obtaining the sparse coefficients are extracted on the segmented substrokes and are motivated from the idea of Histogram of Gradients [24] popular in the area of computer vision. Finally, the tf-idf framework is incorporated on the sparse representation coefficients to characterise the author of the document.

Coming to other explorations, a Bayesian framework is utilized for identifying the writer of the handwritten text by considering the shape primitives that are segmented from the online trace [6]. The authors of [25] represent the dynamic features such as speed, pressure and shape of the temporal trace as a sequence of codes for writer identification. The probability density distribution of four dynamic features from each pre-identified stroke type are used to describe the individuality characteristics of a writer in [26]. Likewise, in another work [27], the distribution of the shape primitives in handwritten text are employed to characterize the orientation of writing trajectory. This is utilized in conjunction with the distribution of curvature and other dynamic features in a hierarchical matching scheme. In [28], a fusion of Dynamic Time Warping (DTW) and Support Vector Machine (SVM) approach has been presented for identifying the authorship of Arabic texts. The extraction of features in this work was done with regards to different levels such as the point, the stroke and the space between strokes.

In a recent work, a Beta-Elliptic model is proposed to characterize the velocity and spatial profiles of the substrokes – followed by classification using an ensemble of Multi Layer Perceptrons (MLP) [29–31]. Last but not the least, the utility of a Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) for learning the feature representation of an online handwritten document are explored in [32,33] , wherein promising results have been quoted over the prior works that used hand-crafted features.

Based on the preceding enumeration of works, we infer that the explorations being proposed for online writer identification have been inspired from areas such as speaker identification [10–12,21,28] and information retrieval [14–20,22,23].

2. A roadmap to the proposed strategy

Recent techniques (since year 2010) in the area of object retrieval have focussed on descriptors obtained from a codebook [34–38]. In this paper, we consider the notion from such methods to motivate the study of online writer identification. An important ingredient in developing a writer identification system is that of a codebook – that comprises a set of codevectors representing the frequently occurring writing patterns among writers in an average sense. The descriptors in our proposal are aimed to capture the relative location of the feature vectors corresponding to the handwriting samples of a writer with respect to their nearest codevector in the feature space. The idea of the same comes from the intuition that the relative location of feature vectors of the same writer are more or less aligned in near-by proximity in the feature space. However, such a trend may not be prevalent when considering the feature vectors across different writers.

Keeping the aforementioned discussion in perspective, we propose a strategy that encodes the sequence of feature vectors along the online trace of a document with descriptors from a codebook. Based on a distance criterion, each feature vector of the document is assigned to a specific codevector in the codebook. As we shall see in Section 5, the proposed descriptors take into consideration, the scores of each of the attributes[1] in a feature vector with regards of the proximity to their corresponding value in the assigned codevector. We show in Section 6 that such explicit scoring can provide useful cues for better discriminating handwritten samples of the different writers enrolled to the system, when compared to the Vector of Locally Aggregated (VLAD) descriptor.

The utility of codebook descriptors was first investigated by the authors in [39]. This article presents improvements with regards to the strategy, that provide higher writer identifications together with an extensive performance evaluation demonstrating its efficacy. These are outlined as follows:

1. The derivation in Section 5 is formulated in a way that, for a given codebook size k, we employ the descriptors of only $k-1$ codevectors to construct the final description (by concatenation) for the online handwritten document. This is different from the formulation in [39], wherein the descriptors from all the k codevectors had to be considered for describing the handwritten document. Likewise, while computing the descriptors of each of the $k-1$ codevectors in the present proposal, we take into regard the scores of the attributes of the feature vectors from the entire document.
2. For constructing the codebook, and subsequently the descriptor, we derive features / attributes (in Section 4) by incorporating a gap parameter that aids in capturing the characteristics of sample points in the neighborhood of the point under consideration. A study is also conducted on the variation of the writer identification rate in Section 8.2 with different values of the gap parameter.
3. One of the detailed steps described in [39] was that of the regression based normalization, that was applied to the features prior to generation of the codebook. The main premise to adopting such a normalization was to ensure that the codevectors obtained are not influenced by outliers. As an avoidance to this, we consider in Section 3 of this paper, a preprocessing step for removing the isolated sample points. Moreover, in lieu of the regression based normalization, we transform the derived point based features obtain across a document, so that they have zero mean and unit variance (z- score normalization). This indeed helps in saving computation load in the normalization step.
4. An empirical evaluation of the effectiveness of our proposal with several recent variants of VLAD [34–37] is presented in Section 8.3.
5. We describe a reduced $D \times (k-1)$ dimension variant of the descriptor in Section 8.4 and demonstrate its performance in writer identification.
6. Experiments of our proposal are conducted on several publicly available online handwritten databases. On the whole, we ob-

[1] We refer to the feature values in a feature vector as ‘attributes’. The attributes computed at each sample point of the online trace are stacked to form a feature vector. Accordingly, elsewhere in the paper, we interchangeably use ‘features’ or ‘attributes’.