Expert Systems With Applications 72 (2017) 196–206

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An online writer identification system using regression-based feature normalization and 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 20 September 2016
Revised 22 November 2016
Accepted 29 November 2016
Available online 15 December 2016

Keywords:
Online writer identification
Codebook descriptors
Feature normalization
IAM Online Handwriting Database
IBM UB1 database

ABSTRACT

This paper describes a strategy to identify the authorship of online handwritten documents. We regard our research framework to that of a retrieval problem and adapt the so called codebook based Vector of Local Aggregate descriptor (VLAD) that has been promising for the object retrieval application in image processing. The codebook comprises a set of code vectors with associated Voronoi cells computed from a clustering algorithm on a set of feature vectors along the online trace. However, we show that the VLAD formulation at times, cannot effectively discriminate between writers, when their respective feature vectors are not linearly separable in the Voronoi cell of the code vectors. To overcome this problem, we propose a novel descriptor that improves upon the VLAD formulation. Secondly, we explore a normalization for the feature vectors prior to the generation of the VLAD. Our method is different to the min–max and *z*-score in that it takes care in ensuring that the codevectors are not influenced by the presence of outliers in the data. The performance of our proposed descriptor with the new feature normalization are evaluated on two publicly available Online Handwriting Databases – the IAM and IBM-UB1. The results show a marked improvement over the VLAD.

1. Introduction

In a writer identification system, the main objective is to decide the authorship of a piece of handwritten document. Essentially we establish the identity of a handwritten sample by matching it against a set of enrolled samples with known authorship stored in a database. A prominent application of the same is in the area of forensic sciences (Jain, Ross, & Prabhakar, 2004). From the perspective of a biometric system, writer identification falls under the category of behavioral biometric and depending on the mode of data capture it can be classified into one of the following: online and offline.

The recent technological advancements has made available the use of hand held devices where in the data entry is obtained via an electronic pen/stylus. The tip of the stylus captures the dynamic information of the trace of the handwriting such as (x, y) coordinates and time stamp. The processing of data of this nature is referred to as 'online'. Contrast to it, offline writer identification systems obtain the handwriting data as an image and use image processing techniques for analyzing the same (Bensefia, Paquet, & Heutte, 2005; Bertolini, Oliveira, Justino, & Sabourin, 2013; Christlein, Bernecker, Hnig, Maier, & Angelopoulou, 2016; Nicolaou, Bagdanov, Liwicki, & Karatzas, 2015; Siddiqi & Vincent, 2010; Wen, Fang, Chen, Tang, & Chen, 2012).

Another categorization of writer identification system is with respect to the textual content – namely text dependent and text independent approaches (Namboodiri & Gupta, 2006). In the former, a specific piece of text is used for the generation of handwriting samples of a writer and the identification process usually involves the use of a recognizer. Though the use of the knowledge pertaining to the content of the data increases the accuracy of text dependent systems, they fail in scenarios where text documents comprising different contents need to be contrasted. For such applications, text independent writer identification systems become more applicable as they capture the style information of handwriting. Such systems are designed to identify the writer irrespective of the textual content.

1.1. Previous works on online writer identification

In this section, we outline the techniques proposed in literature for the problem of text independent online writer identification – the focus of our research. The authors in Liwicki et al. (2006), Liwicki, Schlapbach, and Bunke (2007), Schlapbach, Liwicki, and Bunke (2008) use a Gaussian Mixture Model–Universal Background Model (GMM–UBM) approach, a technique adapted from the area

* Corresponding author.
E-mail addresses: v.venugopal@iitg.ernet.in (V. Venugopal), sureshsundaram@iitg.ernet.in (S. Sundaram).

http://dx.doi.org/10.1016/j.eswa.2016.11.038
0957-4174/

of speaker identification (Reynolds, 1995) for modeling the writer's data. The training samples from a set of enrolled writers is first used to construct an Universal Background Model (UBM) via the Expectation–Maximization (EM) algorithm. Thereafter, the writer-specific GMMs are obtained from this model by a process of adaptation. During the testing phase, given a document of unknown authorship each of the writer-specific GMMs return a log-likelihood score. The identity of the text is established to the writer with the highest score. The GMMs being a generative model require a large number of parameters to be estimated for each writer.

Techniques adapted from the domain of information retrieval have also been attempted for online writer identification. The works in Chan, Viard-Gaudin, and Tay (2008) and Tan, Viard-Gaudin, and Kot (2008, 2009, 2010) provide a detailed description of the same. Here a measure based on the term frequency–inverse document frequency (tf–idf) framework is adopted to score the handwritten data of a writer. The various allographs of an alphabet are obtained by considering the segmentation output of the industrial MYScript SDK character segmentation and recognition engine.

The works of Ramaiah, Shivram, Porwal, and Govindaraju (2012) and Ramaiah, Shivram, and Govindaraju (2013); Ramaiah, Shivram, and Govindaraju, 2014) employ the concept of Latent Dirichlet Allocation (LDA) adapted from the area of topic models. The authors assume each document to be modeled as a random mixture over a set of finite (shared) writing styles which in turn is a mixture over an underlying set of text independent feature probabilities. In this way, this work departs from the conventional approach of modeling unique writing style for each writer. As such, the usage of LDA involves learning a large number of hyper parameters.

In Singh and Sundaram (2015), the use of subtractive clustering is adopted to explore the unique writing styles of a writer. Two strategies namely, a modified tf–idf approach and a nearest prototype method is suggested for writer identification. Whenever the distribution of data is multi-modal with closely spaced modes, the subtractive clustering scheme is not effective unlike the *k*-means or fuzzy *c*-means.

In addition to the above works, attempts have also been made with regards to identifying authorship of documents written in specific scripts – Chinese and Arabic. A shape primitive approach to the problem of Chinese writer identification is described in Li, Sun, and Tan (2009) with a hierarchical classifier structure used for the matching. The set of shape primitives represent the frequently occurring strokes in Chinese script. The distribution of features like azimuth and pressure from each stroke are used as features in Sun, Li, and Tan (2007) with a nearest neighbor classifier. This classifier being an instance based type can be highly susceptible to presence of outliers in the training data. In a recent work (Yang, Jin, & Liu, 2016), an exploration towards the usage of deep Convolutional Neural Networks (CNN) for online writer identification of documents written in Chinese is proposed. However, the authors comment on the scalability issue with respect to addition of new writers to the database.

With regards to Arabic online writer identification, multifractals were used to model the substrokes in Chaabouni, Boubaker, Kherallah, Alimi, and El Abed (2011). The authors categorized the substrokes into five groups on the basis of stroke patterns that occur in Arabic/Persian. In Gargouri, Kanoun, and Ogier (2013), a fusion of Dynamic Time Warping (DTW) and Support Vector Machine (SVM) approach has been presented for identifying the authorship of Arabic texts. Nevertheless, the DTW matching makes the identification system computationally expensive with the increase in number of writers. Recently, a Beta-Elliptic model was used to represent the velocity profile and spatial profile of the substrokes and classification was done by an ensemble of Multi Layer Perceptrons (MLP) (Dhieb, Ouarda, Boubaker, & Alimi, 2016a; 2016b; Dhieb, Ouarda, Boubaker, Halima, & Alimi, 2015).

2. Research framework

An overview of the works on online writer identification suggest that the research ideas being proposed have been motivated largely from areas of speaker identification (Gargouri et al., 2013; Liwicki et al., 2006; Schlapbach et al., 2008) and information retrieval (Chan et al., 2008; Ramaiah et al., 2013; Ramaiah et al., 2012; Tan et al., 2008, 2009, 2010). A contemporary area of research in image processing is the problem of object retrieval. Techniques proposed in the past four years have considered the use of descriptors obtained from a codebook – with the Vector of Local Aggregate descriptor (VLAD)[1] being considered as the state of art (Arandjelović & Zisserman, 2013; Jegou et al., 2012; Sanchez, Perronnin, Mensink, & Verbeek, 2013; Wang, Di, Bhardwaj, Jagadeesh, & Piramuthu, 2014). Being an image based approach, the codebook comprises a set of code vectors with the corresponding Voronoi cells computed from a clustering algorithm on a set of interest point feature vectors such as Scale Invariant Feature Transform (SIFT) (Lowe, 2004).

Analogizing to our problem, it is well-established that the codebook is an important ingredient in developing a text-independent online writer identification system (Chan et al., 2008; Li et al., 2009; Tan et al., 2008, 2009). Thus, this paper investigates on the merit of VLAD by adapting it for the online writer identification. Specifically, we replace the SIFT feature vectors with the set of point based feature vectors along the online trace of the handwritten data.

However, we demonstrate that at times, the formulation of the VLAD is confronted with an issue, that can reduce the discrimination between writers. This can occur when the point based feature vectors corresponding to the different writers are not linearly separable in the Voronoi cell of a codevector – thereby affecting the identification rate. To alleviate this drawback, we propose a novel descriptor with a modified formulation aimed at improving writer discrimination beyond VLAD. Our devised descriptor has a dimension twice that of the VLAD and therefore we consider reducing the number to half. Despite this reduction, we report improved results over the VLAD.

We also consider a normalization scheme for the point based feature vectors prior to the generation of the codebook in the training phase.[2] In contrast to the traditional techniques such as min–max and *z*-score, our method is different in that it takes care to ensure that the codevectors are not influenced by the presence of outliers in the data. Moreover, we show that our method normalizes the feature value in a way so that their corresponding histogram/distribution presents a reduced skew. Considering that our proposal relies on the descriptor from the codebook, this normalization scheme aids in enhancing the efficacy of the identification system.

The performance of the VLAD and our proposed descriptor with the new feature normalization are evaluated on two publicly available Online Handwriting Databases – the IAM and IBM-UB1. The results obtained demonstrate an improvement in writer identification rate over the system that uses the conventional VLAD formulation.

To summarize the preceding discussion, we highlight our present research contributions:

- Proposal of a feature normalization method.

[1] Further details regarding this method will be outlined in Section 5.1.

[2] We wish to clarify that in this phase, generation of the codebook is performed from a set of handwritten documents used as training examples in the writer identification system.