Forensic Science International 279 (2017) 41–52

Contents lists available at ScienceDirect

Forensic Science International

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

Review Article

Writer identification: A comparative study across three world major languages

Gloria Jennis Tan[a,b,*], Ghazali Sulong[a,b], Mohd Shafry Mohd Rahim[a,b]

[a] Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia
[b] School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, Terengganu, Malaysia

ARTICLE INFO

Article history:
Received 23 February 2017
Received in revised form 3 June 2017
Accepted 27 July 2017
Available online 4 August 2017

Keywords:
Handwriting identification
Text-independent writer recognition
Chinese handwriting writer recognition

ABSTRACT

This paper presents a review on the state of the art in offline text-independent writer identification methods for three major languages, namely English, Chinese and Arabic, which were published in literatures from 2011 till 2016. For ease of discussions, we grouped the techniques into three categories: texture-, structure-, and allograph-based. Results are analysed, compared and tabulated along with datasets used for fair and just comparisons. It is observed that during that period, there are significant progresses achieved on English and Arabic; however, the growth on Chinese is rather slow and far from satisfactory in comparison to its wide usage. This is due to its complex writing structure. Meanwhile, issues on datasets used by previous studies are also highlighted because the size matter – accuracy of the writer identification deteriorates as database size increases.

Contents

1. Introduction

Based on the writing method, writer identification is classified into online and offline writer identification. Speed of writing, angle or pressure is utilized in online writer identification, while, the features associated with words, characters, lines or paragraphs are utilized in offline case. Two-dimensional coordinates of successive points of the writing as a function of time are stored in order, i.e., the order to strokes made by the writer is readily available, in case of on-line. While, only the completed writing is available as an image in case of off-line. A spatio-temporal representation of the input is utilized in on-line case whereas, spatio-luminance analysis of an image is involved in the off-line case. In case of on-line information of the writing order and dynamics is available. While, only scanned images are available and much dynamic information is lost in an off-line way. On-line case has much higher performance rates reported as compared to off-line case. It is striking to consider developing off-line systems that primarily estimate the trajectory of writing from off-line data and then use on-line recognition algorithms due success of on-line systems [1].

* Corresponding author at: Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia.
E-mail addresses: jtgloria@gmail.com (G.J. Tan), ghazali.s@umt.edu.my (G. Sulong), shafry@utm.my (M.S.M. Rahim).

http://dx.doi.org/10.1016/j.forsciint.2017.07.034
0379-0738/

However, only few such feature extraction systems have been developed so far due to the difficulty of recreating the temporal data [2–4].

Further there can be text-dependent (script dependent) or text-independent (script independent) offline writer identification. Same content must be written by the writers to be compared in text dependent methods. However, content is not required to be same in text independent methods. There is broader applicability of text independent methods but they lack in obtaining the same high accuracy as obtained by text dependent methods [5]. Text-independent approach is not limited by text content, that, the writer must write exactly the same predefined or given text. Writers are required to write the same fixed text with the training handwritings in text-dependent methods. One of the text-dependent writer identifications is the signature verification, which is commonly used in credit cards. In text-dependent writer identifications we can utilize word position, font, distribution of strokes orientation and strokes arrangement as its characteristics because it is very relevant with the text contents. Orientation and curvature of contours, allographic features, higher-order moments, histogram and strokes matching etc. are some widely used text-dependent methods. The process of identification or verification of identity of the writer without any restrictions on the textual content of the samples being compared is the text-independent method. It will be useful to have any handwritten documents with different text as to extract writing styles any text with different characters is helpful, and it is more difficult to extract their features. There are wider applications of text-independent methods as having same texts in different documents is impossiblein most conditions. However, in many cases (e.g. in criminal investigation) where writers and characters are not determinable because no registration is available and off-line training is impossible, the problem becomes more difficult. The task of text-independent off-line handwriting identification is rather challenging due to unavailability of many valuable writing features such as shape features and dynastic writing information. The extraction of writing features can only be achieved from handwritten image; and it means a loss of valuable writing information.

Therefore, text-independent writer recognition has been researched for decades [6] and is drawing even more attention [7] in the current scenario of fast and efficient methods for investigative research. Extensive researches have been conducted in this field due to its importance in forensic analysis and documents authorization [8–14]. Also, a series of international writer identification contests [15–19] have been successfully organized. It has become an important problem to achieve secure and automatic personal identification and the hypothesis on which writer identification using handwriting is based upon is that people write uniquely and can be characterized based on the information present in their handwriting. The interest in it lies in the significance of its role in the criminal justice [20,21] system as well as a wide variety of fields ranging from security, forensics, financial activities to archeology (e.g. to identify ancient document writers). For forensic analysis, offline text-independent writer identification is very important and can be beneficial in various applications including banks, criminal justice systems, determining the authenticity of handwritten documents, and calligraphic relics identification, etc. A likely list of candidates is returned by a writer identification system after performing a one-to-many search in a large database with handwriting samples of known authorship. Therefore, the system must be learnt from a set of handwriting samples of each individual candidate.

In recent years, to confirm document authenticity in the financial district as well as revealing the identity of suspected criminals, etc. writer identification and verification have become commonly used applications. Qualification of handwriting analysis as expert testimony was decided by the United States v. Paul on May 13, 1999 and is therefore admissible. While coming to conclusion about the authenticity of a document in the court of justice the problem of writer identification and verification arises frequently. The same problem arises in banks regrading signature verification. It is a great contribution to forensics and critical criminal occurrences across three major languages around the world, larger databases based experiments on various languages, which is a real-world scenario.

Due to their particular characteristics in terms of [22]: frequently reduced number of handwriting samples, pencil or type of paper, variability of writing style, the presence of noise patterns, etc. or the unavailability of online information the forensic scenario present some problems. Consequently, current application domain still relies heavily on human-expert interaction. Therefore, it is very helpful to use semi-automatic recognition systems in which questioned handwriting sample is given and a list of possible candidates which are into a database of known identities is narrowed down, hence, the subsequent confrontation for the forensic expert is made easier [23,22].

In AI, biometrics—automatic personal recognition based on physiological or behavioral characteristics is one emerging technology which is spreading more widely in such organizations. A person is recognized by a biometric system which is essentially a pattern-recognition system based on a feature vector derived from a specific physiological or behavioral characteristic that the person possesses. A biometric system typically operates in verification or identification mode, depending on the application context. Manually intensive techniques are utilized by most handwriting identification experts nowadays. Although there is some literature available on prototype toolsets for document examination [24–26], but, no tool is available implying completely automated writer identification process. In a task of writer identification, it uses a query input and a database of identified writers returning the identity of the handwriting as an output. However, in forensic it is not a fully automatic task for practical reasons. Due to unique characteristics of each language scripts, the performance of writer identification from different languages strictly depends upon the threshold set and the selection window size, thus, posing a new challenge to the writer identification. Data heterogeneity and human interpretability problems are there, which is involved in most approaches and manipulation of a constant window size is performed ambiguously. The extent seamless integration of such methods in current forensic handwriting expertise is as yet unclear. The types of errors made by machine and human are quite different due to the fundamental differences between automatic and manual methods. Currently, it is not easy to find multiple researches on different languages using same dataset for benchmarking hence, making the comparison impossible on common grounds. Therefore, proposing reliable automatic and dynamic model methods which would be more appropriate to meet the robustness requirement on data heterogeneity and human interpretability of the results by forensic experts would be more valuable. It is the scenario of a real-world and can be effectively and efficiently fused in existing investigation procedures for all languages instead of being specific to a certain language.

There have been developed several types of writer identification techniques for various handwritten scripts of different languages. A large proportion of the research carried out in this field is based on the Latin script. The complications arising because of complicated Chinese [27–32] and unique stroke Arabic [33–35,18,36–38] script have motivated only a few studies on off-line text independent writer identification. The systems developed for Latin scripts have been tried on Arabic scripts in some studies that achieved variable degree of success [39] but Chinese handwritten text is comparably rare and not proportional to its widely usage.