Online Text Independent Writer Identification Using Character Prototypes Distribution

Siew Keng Chan¹, Yong Haur Tay¹
¹Computer Vision and Intelligent Systems (CVIS) Group, Faculty of Information and Communication Technology Universiti Tunku Abdul Rahman (UTAR), MALAYSIA
tayyh@mail.utar.edu.my

Christian Viard-Gaudin²
²IRCCyN -UMR CNRS 6597
Ecole Polytechnique de l'Université de Nantes FRANCE.
Christian.Viard-Gaudin@univ-nantes.fr

Abstract—Recently, with the advances in digital pen and paper technology, renewed attention was given on research for writer identification of online documents. This paper proposes a method to retrieve the writer of a document by comparing his handwriting with those stored in a reference database of documents. The query will consist of a testing online handwriting document, the output will be a ranked list of writers whose online handwriting documents are stored in the reference database. The proposed method is text independent and does not impose any constraint on the writing style; it is based on a distance measurement between the distributions of reference patterns defined at the character level. Experiment has been carried out on a database which consists of 82 subjects each providing two online handwritten documents, one stored as the reference and the second one is for retrieval purposes. The Top 1 rate reported is 95% accuracy.

Keywords— writer identification, online handwriting, allograph, information retrieval, term frequency

I. INTRODUCTION

Handwritings have been viewed as a personal characteristic due to its uniqueness among each individual. Thus, over the decades, it has been used in biometric identification applications such as the well-known signature verification applications and forensic document analysis through handwriting. Signature verification is a process of verifying whether a given signature is genuine or forgery against what it claims to be while forensic document analysis requires the identification and verification of the writer of a given handwritten document. In the field of forensic document analysis, it has drawn the attention to researches in the field of offline writer identification. Recently, with the advances of technology in digital pen and paper, researchers have began to set interests on online writer identification considering documents written using the digital pen and paper or digital tablet.

Digital pen and paper provides a cost-effective way for traditional paper processes to enter the digital world by fusing the traditional means of information-gathering with electronic communications. Digital pens are used like ordinary pens—only they are embedded with electronics capable of storing time-stamped content. Users write on paper printed with a faint irregular pattern of dots similar to map coordinates, enabling the pen to know which form is being used, what is being written, where on the form and when. This technology enlarges considerably the field of online handwriting, which is no longer restricted to PDA devices with very small input areas, and it is less physically intrusive than Tablet PCs in situations such as financial discussions or census interviews.

As a consequence the amount of online handwriting documents which are produced and stored is increasing very rapidly and it is necessary to provide additional functionalities so that these documents can be retrieved and processed in a smart way. We propose in this paper a method to retrieve the writer of a document by comparing the handwriting with those contained in a reference database of documents. The query will consist in an online handwriting document, the output will be a ranked list of writers whose online handwriting documents are stored in the reference database.

There are two problems to be considered for writer identification i.e. text independent and text dependent problems. The signature verification application would be a type of text dependent approach which considers the same content of text for each writer each time. However, the forensic document analysis which has been mentioned early would be the inverse whereby the identification process is carried out on different content of text, each time even for the same writer.

Many methodologies have been proposed over the years for writer identification. H. E. S. Said et al in [1] proposed a text-independent approach for offline writer identification and used the different textures of writer's handwriting for writer identification. The textures were extracted using multi-channel Gabor filtering technique and gray scale co-occurrence matrices. The identification accuracy of 96 percent was achieved on a 1000 testing documents from 40 subjects using Gabor features with nearest centroid classification using weighted Euclidean distance.

In [2], J. Chapran proposed a method for text-independent dynamic writer identification based on the segment analysis of a handwritten sample. Handwritten samples were grouped up based on the length and the direction of the segments of handwriting between two consecutive sample points and the corresponding dynamic features were computed for each group. The initial reported result of writer identification is of error 8.6% False Acceptance Rate (FAR).
A. Schlapbach et al in [8] proposed the use of Hidden Markov Model (HMM) based recognizers for text-independent offline writer identification and verification. A recognizer is built for each writer using text lines from the writer and nine features were extracted from these text lines. Unknown text lines are presented to each of the recognizers which then output log-likelihood scores for the text line. Experiment was carried out on 100 subjects, and accuracy of 96% was achieved on writer identification. On writer verification experiment using over 8600 text lines from 120 subjects including 20 impostors, an error of 2.5% was reported.

Zois et al in [9] proposed a methodology for writer identification using the feature vector which is morphologically transformed from the projection function of single words. The data set used in the system was composed of the word “characteristic” written 45 times by 50 writers each in Greek and English. Classification is performed using Bayesian classifier and multilayer perceptron. The accuracy reported was 95% for both English and Greek words.

M. Bulacu et al in [16] proposed a methodology for automatic writer identification and verification using probability distribution functions (PDFs) extracted from handwriting images to discriminate each writer. Two levels of analysis take place in the study namely texture level and allograph level. Slant and curvature were features extracted at the texture level and encoded into PDF using contour-based joint directional PDF. In the allograph level, graphemes were extracted and clustered to generate a graphemes codebook which then be used to produce the PDF. The study showed that with the combinations of features: directional, grapheme and run-length information yields greater accuracy in writer identification with best Top 1 rate of 85 – 87% and Top 10 with 96% and an equal error rate (EER) around 3% in verification. The experiments were tested using three data sets namely Firemaker, IAM and ImUnipen.

A. Bensefia et al in [5] [6] [7] proposed the use of graphemes as the features to discriminate writers and the use of Information Retrieval paradigm for the writer identification stage and mutual information between the grapheme distributions of two handwritings to be compared in the writer verification stage. Experiments were carried on three different data sets composed of 88 writers, 39 writers (historical documents) and 150 writers where each contributing two documents. The results reported for writer identification were around 90%. Our work is closely related to this approach, except that it is carried out with online handwriting whereas Bensefia works with offline handwriting, and another difference is that we explicitly use the character level instead of the grapheme level, characters are more difficult to extract than grapheme but we expect more consistency at this level.

In the following sections, explanations are given on the proposed methodology which explains the overall processes involved for online writer identification. There are three main stages namely character prototypes building stage, reference and testing document building stage and distance computation stage. The processes involved in each stage are explained further in its own section and a section is dedicated for the experiments carried out. A small section describes the database which has been used. Last section concludes on the methodology proposed as well as the future works.

II. PROPOSED METHODOLOGY

The system consists of three main stages: character prototypes building stage, reference and test document processing stage and distance computation stage. Character prototypes are first defined on an independent isolated word database. For each 26 letters, it produces a set of N prototypes (N=10, in all the following experiments), where each prototype is supposed modeling a specific allograph. Reference and test documents processing stage transform the online handwritten document in a fixed-size vector. These vectors are to be used in the later stage i.e. the distance computation stage which computes the distance between the vectors. The distances obtained is then be used in the ranking process to identify the writer of a test document. Processes involved in each stage are further discussed in the following sub-sections.

A simplified description of the methodology is presented on the example given in Table I. Five different allographs of the letter ‘f’ are displayed. On the next two rows, it can be seen the distribution of these different allographs on a text coming respectively from writers i and j, these vectors being stored in the reference database, while in the last row, this same vector has been extracted from a query handwriting document. With this example, we classify writer T as writer i in Top1 position, since the distance between the vectors of writer T and writer i is the smallest:

Euclidian distance computation (using the values in Table I):

\[ D \text{(writer } i, \text{ writer } T) = \sqrt{\sum_{k=1}^{5} (tf(k), f', i - tf(k), f', T)^2} = 0.03 \]  \tag{1} \]

while \( D \text{(writer } j, \text{ writer } T) = 0.47 \)

<table>
<thead>
<tr>
<th>Table I. Prototype Frequency Vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
</tr>
<tr>
<td>Prototype selected by k-means algorithm</td>
</tr>
<tr>
<td>tf \text{ writer } i</td>
</tr>
<tr>
<td>tf \text{ writer } j</td>
</tr>
<tr>
<td>tf \text{ Writer } T</td>
</tr>
</tbody>
</table>
The features considered in this communication are at the character level. One specific character exhibits different allographs. One writer might use the same allograph most of the time while another writer probably uses different allograph for the same character. Thus, it is feasible to consider this as a discriminative feature for writer identification.

A total of 16585 French words written by 373 subjects taken from IRONOFF database [17] are used in this stage. The words were undergone a process of segmentation where each word was segmented into character and each of the character was categorized accordingly in its alphabet database thus having 26 alphabet databases. The segmentation/recognition process is carried out using an industrial recognition software engine (MyScript SDK) [18] with its French language resources. Before segmentation process, the recognition result of each word was checked against its true label. Any word which was not correctly recognized has been discarded. When the recognition succeeds in finding the correct label then the path given by the Viterbi algorithm allows to define a segmentation at the character level. Each of the different segments of ink is stored in the corresponding database of characters. This stage produces 26 databases which are composed of a variable number of characters. Total number of characters being 87719, some sub-databases being more numerous than others: for instance 'e' letter which is very common in French language contains 11959 samples, while 'w' has only 139. Fig 1 shows an example of a French word, which is displayed using the ink tool provided by MyScript SDK and the process of recognition and segmentation carried out. The points of each character segmented are then put into graph to view to result of segmentation.

The characters in each alphabet database were further normalized in size and re-sampled with a fixed number of 30 points with each point exhibits seven features, namely x-coordinate, y-coordinate, direction of x-coordinates, direction of y-coordinates, curvature of x-coordinates and curvature of y-coordinates and the status of pen up or pen down. Then, each alphabet database is clustered into N character prototypes using k-means clustering algorithm.

The k-means clustering algorithm was modified to accommodate the presence of outliers. In the traditional k-means clustering algorithm, a point is assigned to a cluster by finding the minimum distance from the distances computed from the point to each centroid. However, this might cause a point to be grouped alone in a cluster which is known as outlier because of its large distances from other points. Thus, to eliminate this case, before a point is assigned its cluster, the maximum distance of the point with any member in each cluster is firstly computed. From these maximum distances, the minimum distance determines the cluster of which the point should be assigned to. The results at the end of this stage are clusters of character prototypes in each alphabet database.

### Table 1

<table>
<thead>
<tr>
<th>Actual image displayed using InkTool provided by Myscript SDK with each segmented character</th>
<th>Extracted characters’ points (normalized by maximum and minimum points) are mapped using graphs</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
</tr>
</tbody>
</table>

**Figure 1.** Segmentation process on French words with the extracted points graphed.

### IV. Reference and Testing Document Building Stage

This section describes the processes involved in computing the representative vector of the reference and test documents. Each document is firstly be segmented into characters using the same recognition engine mentioned in the previous stage. Thus, each document has a 26-character database labeled accordingly. Each character undergoes the same process of normalization and feature extraction as mentioned before. Then, each character from the document is assigned to the most similar prototype of this character by computing the minimum distance between this character and each character prototype of this same letter. Based on the assignments of the characters into the character prototype, the term frequency if can be computed using Eq. (2), where i is the writer number, a is the considered character where a ∈ \{'a'..'z'\}, k is the cluster number, where \(k \in \{1,2,\ldots,N\}\) with N as the number of clusters, \(TC(k)_{a,i}\) is the total number of characters a of writer i clustered in cluster k and \(TC_{a,i}\) is the total number of characters a from writer i.

\[
tf(k)_{a,i} = \frac{TC(k)_{a,i}}{TC_{a,i}}
\]

For the case of reference documents, the inverse document frequency idf is computed. This is to weight the importance of each character prototype for each subject. The computation is carried out using Eq. (3) where \(TRD_a\) is the total number of reference documents having the character a and \(TRD(k)\) is

\[
idf(a) = \log \left( \frac{|D|}{|D(a)|} \right)
\]
the total number of reference documents having the character prototype, \( a_k \). This \( idf \) value will be used during the distance computation between the reference documents and the test document, which is explained in the next section.

\[
idf(k)_a = \log \frac{TRD_a + 1}{TRD(k)_a + 1}
\]

(3)

V. DISTANCE COMPUTATION STAGE

This stage is actually the identification stage where the process is mainly to identify the writer of a given test document. The distance measure used is the Euclidean Distance and the computation is carried out between the term frequencies of each reference documents with the test document. The distances obtained are used as the key in ranking the writers of the reference documents starting with the minimum distance on top. A few considerations were taken place in the distance computations to eliminate the bias of computation such as a missing letter in either the reference document or the test document and the size of the documents due to the varying size and contents of the documents. The computation is carried out using Eq. (4) and (5). Weighting coefficients \( \alpha \), defined by Eq. (5) is to eliminate the bias been added to the total distance because either document might not contributed the specific alphabet. In the meantime, it normalizes the distance respective to the size of the documents.

\[
dist(writer_i, writer_T) = \sum_{a \in V} \alpha_a \times \left( \sum_{k \in C} idf(k)_a \left( tf(k)_a - tf(k)_T \right) \right)^2
\]

(3)

\[
\alpha_a = \sqrt{\frac{\#a_i \times \#a_T}{\#Ch_i \times \#Ch_T}}
\]

(4)

\( \#a \) is the total number of the particular alphabet from reference document, \( i \) or testing document, \( T \). \( \#Ch \) is the total number of characters extracted from the reference document, \( i \) or testing document, \( T \).

VI. DATABASES

There are two sets of data which have been used throughout the entire process. The first set of data is taken from the IRONOFF database for the character prototype building stage. In the data set, each subject is required to write pre-defined words. Only French words were extracted from the entire database. Fig 2 shows some of the example of words written by two subjects. The second set corresponds to the reference and testing documents. It is composed of a total of 164 online handwritten documents written by 82 subjects, each writer having written one reference document and one test document with different contents but all are in French language. Fig 3 shows examples of online handwritten documents of a writer from the reference document and testing document database.
VII. EXPERIMENTS AND DISCUSSIONS

Two experiments have been carried out. Experiment one is to test the efficiency of the proposed methodology for online writer identification. The 82 testing documents have been tested against the 82 reference documents to get an accuracy of writer identification. The initial accuracy achieved is of 95% for Top 1 rate (i.e. 4 out of the 82 test documents are not assigned the correct writer in first rank) and 98.78% for Top 6 rate (i.e. 1 out of the 82 test documents are not retrieved in the list of the first six candidates).

Second experiment is intended to check the discriminative power of each letter. However, not all letters can be tested because of the limited characters occurring in the documents for a particular alphabet. Table II shows that results drop severely when considering a single letter; the best results are obtained for letter ‘a’ which has an accuracy of 34% and letter ‘i’, both exhibit a 33% of Top1 identification rate, which is in any cases far below the 95% obtained when all letters are considered. Using only one letter is definitively not a precise approach but conversely using all the 26 letters might not be the best choice.

<table>
<thead>
<tr>
<th>TABLE II. PERCENTAGE OF ACCURACY WHEN USING A SINGLE LETTER WHEN COMPUTING THE DISTANCE BETWEEN TWO DOCUMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Char</td>
</tr>
<tr>
<td>%</td>
</tr>
<tr>
<td>Char</td>
</tr>
<tr>
<td>%</td>
</tr>
<tr>
<td>Char</td>
</tr>
<tr>
<td>%</td>
</tr>
<tr>
<td>Char</td>
</tr>
<tr>
<td>%</td>
</tr>
</tbody>
</table>

VIII. CONCLUSION AND FUTURE WORKS

The proposed methodology is proven to be applicable for writer identification based on the early stage of experimental result with 95 percent of accuracy on the Top 1 rate. However, further refinements could be investigated to improve the efficiency of such method and to achieve a better result. The initial planning is to use genetic algorithms to search for more discriminative alphabets using another set of database and to check the consistency between both databases on the discriminative alphabets. It is thought to be extended to the stage whereby each writer could have certain alphabets most descriptive one about oneself. However, a number of considerations have to be taken into account such as the absence of the descriptive alphabet in the documents and so forth.

REFERENCES