A Stochastic Nearest Neighbor Character Prototype Approach for Online Writer Identification

Guoxian Tan, Christian Viard-Gaudin, Alex Kot

A Stochastic Nearest Neighbor Character Prototype Approach for Online Writer Identification

Guo Xian Tan[1, 2], Christian Viard-Gaudin[2], Alex C. Kot[1]
[1]*Nanyang Technological University of Singapore*
[2]*IRCCyN, Ecole Polytechnique de l'université de Nantes*
tanguoxian@pmail.ntu.edu.sg, christian.viard-gaudin@univ-nantes.fr, eackot@ntu.edu.sg

Abstract

One novel technique for identifying the writer of an online handwritten document is proposed. This technique makes use of a character prototype distribution to model the specific allographs [1]used by a given writer. In this paper, we propose to extend and improve upon this newly established methodology [1] by making use of a stochastic nearest neighbor algorithm to estimate the character prototype distribution. The proposed method is text independent and relies on the automatic segmentation of the handwritten text at the character level. Our results show that this approach attained a writer identification rate of 99.2% when a reference database of 120 writers is used. Experiments related to the effect of the length of text of the document on the performance of the writer identification system are also reported.

1. Introduction

With the advent of pen-input devices such as Tablet PC and personal handheld devices, entering data into computers is now not just being limited to using the keyboard. Numerous initiatives have thus been funded to develop algorithms and computing platforms to handle the surge in demand for this seamless and natural interaction between human users and the machine [2]. All these led to the emergence and proliferation of handwritten on-line documents. Online handwritten digital documents are defined as those digital documents that not only provide information obtainable from offline digital documents, but also contain temporal information of the handwriting process [3]. Such additional information provides vital clues as to the identities of the writer. Writer identification has ubiquitous applications in digital rights management and forensic analysis in the prevention of fraud and identity theft cases. In addition, we can also perform writer-adaptation in tablet PCs and create, store or retrieve a profile of handwriting styles of the writers if we are able to automatically determine their identities.

Various techniques currently exist for writer identification. They can be generally classified into text-dependent or text-independent techniques, which can make use of both global as well as local features. Yashushi et al. proposed a HMM-based approach [4-6] that is robust to noise and small shape changes. However, this model is text-dependent. Jain et al. proposed to use dynamic time warping [3, 7] on the stroke direction, curvature and the height as features to do word matching. Such a method is language independent, but at the expense of a high computational complexity. Schomaker et al. [8, 9] used textural information such as the slant and curvature to transform them into probability density function as well as allographic information such as the graphemes to cluster into codebooks. This approach achieved significant improvements when compared to traditional techniques.

One newly-established technique by Chan et al. [1] makes use of character prototyping, where statistical information about the writer is being matched based on the prototype of that character. A top1 accuracy of 95% was achieved based on this text-independent approach. The advantage of this method is that it is more consistent at the character level as opposed to using the grapheme or word level. Our proposed work further improves upon the work done by Chan et al. by

[1] Allographs are different shapes and forms of the same alphabet.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

adopting a fuzzy-algorithmic approach, resulting in significant improvements in the performance.

The remainder of this paper is organized as follows: Section 2 describes the proposed methodology and experimental setup. Following that, section 3 then presents the experimental results. Finally, discussions and future areas to explore are described in section 4.

2. System Architecture

The writer identification system can be broadly divided into three main stages as shown in figure 1. The purpose of the prototype training stage is to build a set of character prototypes to model the different allographs of the 26 Latin alphabets ('a' to 'z'). The character prototypes are built and trained using the IRONOFF database [10] of 16585 isolated French words written by 373 subjects. The segmentation and extraction of the characters to build the set of character prototypes is done by an industrial text recognition engine, MyScript [11]. Thereafter, in the document labeling stage, the same recognition engine is used to segment and label the online handwritings of a writer document at the character level. Each of the extracted characters is then mapped to the set of character prototypes corresponding to its alphabet and transformed into a distribution of frequency vectors. Finally, in the last classification stage, frequency vectors are then classified to identify the writers.

In this paper, we propose a stochastic nearest neighbor algorithm at the document labeling stage. After an automatic segmentation and preprocessing process, the extracted characters are then assigned using a Mahalanobis distance to the respective character prototypes built in the previous prototype training stage. This distance is used to take care of the specificities of the feature space and the resulting shape of the clusters in this space. Hence, the variances among the different feature vectors are factored into the proposed methodology. The assignment to the respective prototypes, termed as stochastic nearest neighbor assignment, is derived from a fuzzy c-means algorithm [12, 13] which uses an exponential kernel function as described in Eq. 1 to determine a partial membership to the respective character prototypes.

$$P(p_k \mid x_\alpha) = \frac{\exp\left(-\beta \times dist(p_k, x_\alpha)\right)}{\sum_{k'=1}^{N} \exp\left(-\beta \times dist(p_{k'}, x_\alpha)\right)} \quad (1)$$

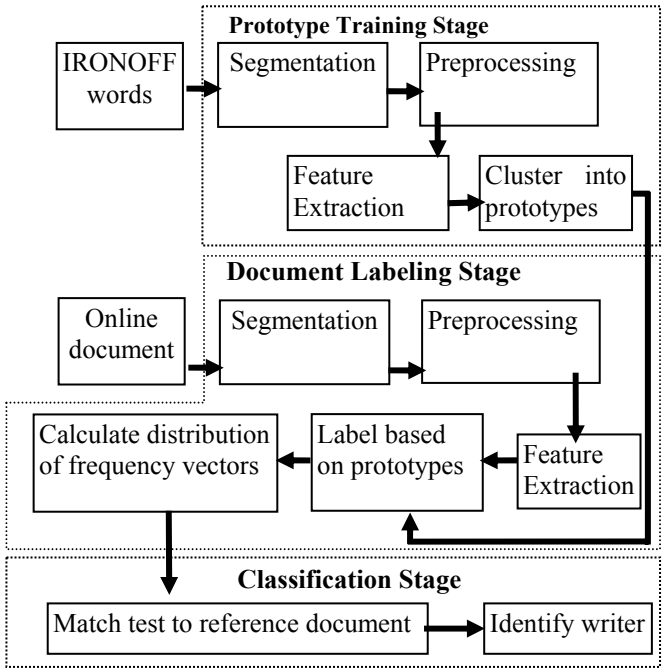

Figure 1. Block diagram of character prototype method

For each writer *i*, $P(p_k \mid x_\alpha)$ is the probability that a given segmented character x_α, which has been recognized as of the alphabet α, $\alpha \in \{'a', 'b', ..., 'z'\}$, is assigned to prototype p_k, $k \in [1, N]$. This represents the partial membership of p_k as discussed earlier. $dist(p_k, x_\alpha)$ represents the Mahalanobis distance. In Eq. 1, β is a tuning parameter which is set to be 0.01 and *N* is the number of prototypes used, which is set to 10 based on our past results.

$$tf_k = \frac{1}{M} \sum_{x=1}^{M} P(p_k \mid x_\alpha) \quad (2)$$

$$idf_k = \log \frac{\sum_{k'=1}^{N} \sum_{i=1}^{R} tf_{k',i} + \varepsilon}{\sum_{i=1}^{R} tf_{k,i} + \varepsilon} \quad (3)$$

$P(p_k \mid x_\alpha)$ is then used to calculate the distribution of frequency vectors, the term frequency (*tf*), as shown in Eq. 2. Each component of the term frequency vector will be weighted with the inverse document frequency (*idf)*, as shown in Eq.3. The usage of *tf* and *idf* is commonly employed in information theory for document retrieval [14] and will be used for classification later. In Eq. 2, *M* is the number of characters corresponding to the alphabet α. In Eq. 3, *R* is the number of reference writers and ε is a small value to prevent any numerical problems.

3. Experimental Results

We have collected online handwritten documents from 120 writers. Each writer wrote 2 documents, with a digital pen and paper technology: one is considered as a reference document and the other is taken as a test document. As presented in table 1, using Euclidean distance between the frequency vectors during classification resulted in our stochastic nearest neighbor approach attaining a high accuracy of 98.3% for writers that are ranked correctly in the top 1 position. This translates into a misclassification error of 2 misclassified writers, with both of them being misclassified in the top 2 position. This indicates that the writer identification system has confused the misclassified documents with only 1 other document from a different writer out of 120 writers. Comparisons with the method by Chan et al. [1] show a significant improvement. An accuracy of 96.7% (4 misclassified writers out of 120) was achieved using their technique.

Table 1. Performance of writer identification using different distance measures for classification

1NN[2]	Stochastic Nearest Neighbor Approach	
	Euclidean distance	Chi^2 distance
96.7%	98.3%	99.2%

This improvement over Chan et al.'s results can be explained as follows. Their methodology hinges on the concept that each character can only be assigned to one particular prototype. This is flawed in reality because there often exist overlapping handwriting styles for different writers. Our observations reveal that there are numerous instances when the characters are close to more than one prototype in the vector space. For such instances, the discrete allocation of prototypes used by Chan et al. does not yield good results. A writer whose dominant handwriting style does not fit well into an existing prototype will be weakly modeled using their approach. Therefore, in our proposed methodology, each character is assigned a certain degree of all the prototypes depending on how close they are to that prototype, allowing us to realize a higher accuracy.

3.1 Classification using different distance measure

Experiments were also conducted using different distance measures for classification. It is observed that the Chi^2 distance measure, as described in Eq. 4, outperforms the Euclidean measure in our writer identification system, achieving a top 1 writer identification rate of 99.2%. This is equivalent to a misclassification error of only 1 misclassified writer, with the misclassified writer being in the top 2 position. The rationale, which can explain the better result obtained with the Chi^2 distance, is that the Chi^2 distance considers a relative difference between the two components of the distributions instead of an absolute difference as with the Euclidean distance and this is more meaningful with respect to the style of writings that we would like to distinguish.

$$\text{Dist}(\text{writer}_i, \text{writer}_T) = \sum_{k=1}^{N} \frac{idf_k \left(tf_{k,i} - tf_{k,T}\right)^2}{tf_{k,i} + tf_{k,T}} \quad (4)$$

3.2 Effect of length of text in test documents

A series of experiments have been conducted to investigate the amount of text required for sufficient accuracy of the writer identification system. Figure 2 shows the achievable of the identification rate by varying the number of characters in the test documents, while keeping the reference documents unchanged. A random reduction of characters from all the alphabets in the original test document is carried out until the required number of characters per test document is attained. As illustrated in figure 2, the misclassification error remains almost constant when at least 160 characters are present in each test document and then suffers a drastic loss in performance once it falls below this threshold. This can be explained by the fact that the handwriting styles of different writers cannot be effectively modeled without sufficient allographic information. A minimum length of text is necessary to perform a reasonably accurate statistical representation of the handwriting styles. Figure 2 also suggests that beyond this minimum threshold, any further increase in the amount of allographic information does not serve to improve the accuracy of the system. Hence, our methodology requires a minimum threshold of 160 characters in each test document for sufficient performance to be achieved. This is equivalent to approximately 30 words, or 3 lines.

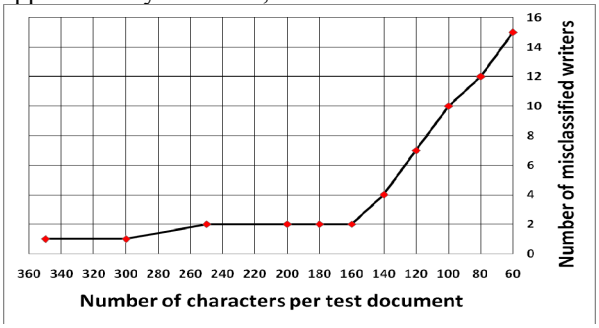

Figure 2. Number of misclassifications against the number of characters in each test document.

[2] 1-Nearest Neighbor algorithm adopted by [1]

4. Discussions

From the experimental results, the proposed methodology is able to generate high accuracies of 99.2% (one misclassified writer out of 120) with the misclassified writer being identified correctly in the rank 2nd position. This is a remarkable improvement over Chan et al.'s results [1], where they reported an accuracy of 95.12% (four misclassified writers out of 82), with all the misclassified writers only being correctly identified in the rank 11th position. Furthermore, we have also determined that the length of text for each document has to be at least 160 characters in order to achieve sufficient performance when using this stochastic nearest neighbor character prototype approach for writer identification.

In this paper, the segmentation of the characters from the handwritten documents was done automatically using a text recognition engine based on the French language, where a corresponding French lexicon was attached to increase segmentation accuracies. This tool attained an average character recognition accuracy of 91% on the document datasets used in our experiments. With such a level of accuracy, characters which are not correctly recognized or segmented do not play a pivotal role on the identification writer rate. Hence, it will be interesting to investigate the performance of the writer identification system where the language domain is unknown or in an environment which contains a mixture of multi-lingual documents. Our current work focuses on this aspect of developing a general framework for writer identification based on multiple languages.

5. Acknowledgements

This research is jointly supported by Nanyang Technological University of Singapore, the French Merlion Scholarship, and the ANR grant CIEL 06-TLOG-009.

References

[1] S.K Chan, C. Viard-Gaudin and Y.H Tay "Online Text Independent Writer Identification Using Character Prototypes Distribution", *Proc. of SPIE-IS&T Electronic Imaging: Document Recognition and Retrieval XV*, 2008, vol. 6815, pp.1-9

[2] S. Oviatt, P. Cohen, L. Wu, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson and D. Ferro, "Designing the User Interface for Multimodal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future Research Directions", *Human-Computer Interaction*, 2000, vol. 15, No. 4, pp. 263-322.

[3] A.K. Jain and A. M. Namboodiri, "Indexing and Retrieval of On-line Handwritten Documents", *Proc. of the 7th Int. Conf. on Document Analysis & Recognition*, 2003, pp.655-659.

[4] Y. Yasushi, T. Nagao and N. Komatsu, "Text-indicated Writer Verification Using Hidden Markov Models", *Proc. of the 7th Int. Conf. on Document Analysis & Recognition*, 2003, pp.329-332.

[5] Z. He, X. You and Y.Y Tang, "Writer identification of Chinese handwriting documents using hidden Markov tree model", *Pattern Recognition 41*, 2008, pp. 1295-1307.

[6] E. Caillault and C. Viard-Gaudin, "Mixed Discriminant Training of Hybrid ANN/HMM Systems for Online Handwriting Word Recognition", *Int. J. of Pattern Recognition and Artificial Intelligence*, vol.21, no.1, 2007, pp.117-134.

[7] R. Niels, L. Vuurpijl and L. Schomaker, "Automatic Allograph Matching in Forensic Writer Identification", *Int. J. of Pattern Recognition and Artificial Intelligence*,vol.21, no.1, 2007, pp.61-81.

[8] L. Schomaker and M. Bulacu, "Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Uppercase Western Script", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, No. 6, June 2004, pp. 787-798.

[9] M. Bulacu and L. Schomaker, "Text-Independent Writer Identification and Verification Using Textural and Allographic Features", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no.4, Apr 2007, pp. 701-717.

[10] C. Viard-Gaudin, P-M Lallican, S. Knerr and P. Binter, "The IRESTE On/Off (IRONOFF) Dual Handwriting Database", *Proc. of 5th Int. Conf. on Document Analysis and Recognition*, Sep1999, pp.455-458.

[11] Vision Objects Industrial Text Recogniser SDK, "MyScript Builder Help", documentation, http://www.visionobjects.com/about-us/download-center/_263/myscript-products-datasheets.html, 2008.

[12] J. Han and M. Kamber, "*Data Mining: Concepts and Techniques*", Elsevier, 2006, pp.383-460.

[13] W. Pedrycz and P. Rai, "Collaborative Clustering with the use of Fuzzy C-means and its Quantification", *Journal of Fuzzy Sets and Systems*, 2008, pp. 1-29.

[14] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", *Journal of Information Processing & Management, 1988* 24(5), pp. 513–523.