On-line cursive handwriting characterization using TF-IDF scores of graphemes

Muriel VISANI, Quang Anh BUI and Sophea PRUM
Laboratory L3i, Department of Computer Science, University of La Rochelle
La Rochelle, France
Email: mvisani@univ-lr.fr, qbui01@univ-lr.fr, sprum@univ-lr.fr

***Abstract*—In this paper, we present an approach for characterizing the on-line cursive handwriting of different writers, which may consist in identifying the writer or his handwriting style. This method is inspired from information retrieval methods and is designed to be embedded in an adaptive word recognizer. We perform experiments assessing the effectiveness of the proposed method for writer identification. Additional preliminary experiments also show that the handwriting style can be used to personalize our cursive word recognizer, enabling the word recognition rates to be increased significantly even with a basic adaptive scheme, which is very encouraging.**

***Keywords*-Writer identification; handwriting style classification; adaptive cursive word recognition; on-line features; classification using TF-IDF.**

I. Introduction

The general application context of this work is the recognition of cursive words handwritten in a form using an electronic tablet. In this context, we developed a word recognizer using a bi-character model [1] based on HMM and SVM. Although its overall accuracy is good, it varies a lot dependending on the writer's handwriting style. The final objective of this work is to build a preliminary module that identifies the writer (or more likely his handwriting style) in order to build multiple instances of our word recognizer, each instance being adapted to a given handwriting style.

Because of our application context, we consider Latin script in a closed-world context (writers are typically supposed to be employees of a company); also, our writer characterizer must rely on on-line features, be text-independent, and remain efficient even when there is not much handwriting for each writer (we use handwriting contained in forms).

Among the large amount of work on writer identification [2], we can cite the methods in [3], [4], which are also based on TF-IDF. While the former uses off-line features only, the latter requires a minimum length of text for each writer. As far as we know, none of them has been applied to handwriting style classification nor integrated in an adaptive word recognizer. In [5], we mainly focused on a method using TF-IDF computed from letter clusters, which requires a preliminary step of character segmentation and recognition. As shown in [5], this introduces a systematic bias due to character segmentation/recognition errors.

In this paper, we present a TF-IDF based method using grapheme clusters (*c.f.* section II), and preliminary experiments showing its accuracy for writer identification as well as for handwriting style characterization, when embedded in an adaptive cursive word recognizer (*c.f.* section III).

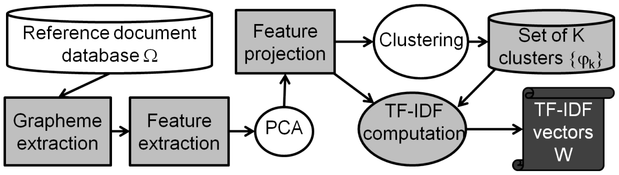

Figure 1. Proposed approach (learning stage)

II. Overview of the proposed approach

A. Writer identification

Figure 1 gives an overview of the **learning stage**. For every grapheme in the reference dataset, we extract 45 local on-line and off-line features (the latter being extracted from an approximate reconstruction of the off-line signal). We use local features because 1) they are less affected by local noise/distorsion than global features and 2) in our application, we often do not have enough handwriting to compute global features. We use the same grapheme and feature extractors as our word recognizer; more details are given in [1]. In order to reduce the redundancy between features as well as the dimensionality of the problem, Principal Component Analysis (PCA) is applied in the feature space. Clustering is applied on the feature vectors in the PCA subspace, resulting in K clusters φ_k. Each document D in the reference database Ω is then described by its TF-IDF vector $W = (w_1, \dots, w_K)^T$, where:

$$w_k = TF_{\varphi_k}(D)IDF_{\varphi_k}(\Omega) \qquad (1)$$

where $TF_{\varphi_k}(D)$ is the *Term Frequency* of graphemes from cluster φ_k in document D. The main idea behind the *Inverse Document Frequency* $IDF_{\varphi_k}(\Omega)$ is that the more the cluster φ_k is rare in the reference database, the higher is the value of $IDF_{\varphi_k}(\Omega)$ (for more details please refer to [5]).

During the **recognition stage**, for each grapheme x in the query document T, we first compute its feature vector and project it onto the PCA subspace. Second, we use an exponential kernel function to estimate the probability that x has been generated by each cluster (see [5]) and determine the most likely cluster for x. We then compute the TF-IDF vector of T using Eq (1). Finally, the writer of T is identified using the nearest neighbour rule between T and the documents D in Ω based on their TF-IDF vectors and any distance (normalized cosine, χ^2, Euclidean…).

B. Handwriting style and adaptive word recognition

Our final objective is to ameliorate our cursive word recognizer [1] by personalizing it. However, in our application context we do not have enough handwriting to train a separate recognizer for each writer; as a result we personalize our word recognizer towards the handwriting style (and not the identity) of the writer. Therefore, we apply clustering on the documents from Ω using the TF-IDF weights of their graphemes. As an output, we obtain S document clusters, each cluster being representative of a different handwriting style (handwriting slant, "roundness"…). We use k-means and settle parameter S using $V-$measure. A separate word recognizer is trained for each cluster (*ie.* for each handwriting style) using a very basic adaptive scheme: for training, we use the whole reference dataset Ω, where the documents from the corresponding cluster are increased of a factor α compared to documents from other clusters.

When a new document is presented to the system, the handwriting style classifier determines which is the nearest cluster and then the corresponding word recognizer is used.

III. Experimental results

First, for writer identification, we use cursive words extracted from the IRONOFF dataset [6]. We consider words written by 10 to 300 different writers, with 30 words per writer (20 words as reference, 10 words as test) and $K = 80$ clusters. For comparing the TF-IDF vectors, we use the χ^2 distance measure as it outperforms the normalized cosine and Euclidean distances. Figure 2.a) gives the writer identification rates when increasing the number of writers. We can see that the identification rates drop dramatically when increasing the number of scriptors, which may be partly explained by the fact that we do not have much handwriting for each writer (nor for training nor for testing).

Second, we perform preliminary experiments for the basic adaptive version of our cursive word recognizer, using the same dataset and 300 writers. We compute S=9 clusters of handwriting styles, and the number of writers per cluster varies from 7 to 49. Figure 2.b) shows the word recognition rates when varying the weight factor α. We can see that our adaptive word recognizer using weights 7 times superior for the writers of the corresponding handwriting style outperforms the generic (non-adaptive) word recognizer (weight=1) of about 1,65%; the confidence intervals as well as a paired T-test show that this difference is significant.

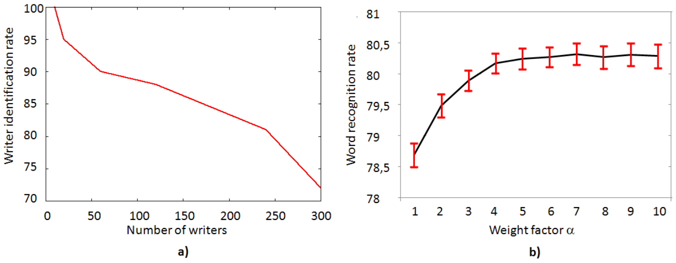

Figure 2. a) Writer identification rate when increasing the number of writers from 10 to 300. b) Cursive word recognition rates, when varying the weight factor α from 1 to 10. When α=1, the word recognizer is generic (not personalized). The higher is α, the more personalized is the word recognizer. Red intervals correspond to 95% confidence intervals.

IV. Conclusion

In this paper, we present a method for characterizing on-line cursive handwriting, *ie.* recognizing writers or handwriting styles. This method is based on TF-IDF computed directly from graphemes (avoiding bias due to character segmentation/recognition). It performs quite well for writer identification, even when there is not much handwriting for each writer (as long as the number of writers is limited). When embedded in our word recognizer to personalize it towards the handwriting style of the documents, word recognition rates are significantly increased, even when using a basic adaptive scheme. We are currently integrating more effective clustering methods and more elaborate adaptive schemes in order to ameliorate word recognition rates.

Acknowledgment

Grateful acknowledgement is made for financial support by the Poitou-Charentes Region (France).

References

[1] S. Prum, M. Visani, and J.-M. Ogier, "Cursive on-line handwriting word recognition using a bi-character model for large lexicon applications," in *Intl Conf on Frontiers in Handwriting Recognition (ICFHR)*, 2010, pp. 194–199.

[2] A. Schlapbach and H. Bunke, "A writer identification and verification system using hmm based recognizers," *Pattern Analysis and Applications*, vol. 10, no. 1, pp. 33–43, 2007.

[3] A. Bensefia, T. Paquet, and L. Heutte, "A writer identification and verification system," *Pattern Recognition Letters*, vol. 26, no. 13, pp. 2080–2092, 2005.

[4] G. X. Tan, C. Viard-Gaudin, and A. C. Kot, "Automatic writer identification framework for online handwritten documents using character prototypes," *Pattern Recognition*, vol. 42, no. 12, pp. 3313–3323, 2009.

[5] Q. A. Bui, M. Visani, S. Prum, and J.-M. Ogier, "Writer identification using TF-IDF for cursive handwritten word recognition," in *Intl Conf. on Document Analysis and Recognition (ICDAR)*, 2011, pp. 844–848.

[6] C. Viard-Gaudin, P.-M. Lallican, S. Knerr, and P. Binter, "The IRESTE On/OFF (IRONOFF) Dual Handwriting Database," in *ICDAR*, 1999, pp. 455–458.