Feature Evaluation for Discriminating Handwriting Fragments

Claudio De Stefano, Francesco Fontanella, Angelo Marcelli, Antonio Parziale, Alessandra Scotto Di Freca

Feature Evaluation for Discriminating Handwriting Fragments

Claudio DE STEFANO [a], Francesco FONTANELLA [a], Angelo MARCELLI [b], Antonio PARZIALE [b] and Alessandra SCOTTO di FRECA [a]

[a] *Dipartimento di Ingegneria Elettrica e dell'Informazione*
University of Cassino and Southern Lazio
Via Di Biasio, 43
04303, Cassino (FR), ITALY

[b] *Dipartimento di Ing. dell'Informazione, Ing. Elettrica e Matematica Applicata (DIEM)*
University of Salerno
Via Ponte don Melillo, 1
84084, Fisciano (SA), ITALY

(destefano, fontanella)@unicas.it, (amarcelli, anparziale)@unisa.it, a.scotto@unicas.it

Abstract. The large majority of methods proposed in literature for handwriting recognition assume that words are produced drawing large parts of the ink without lifting the pen, other than horizontal bars and dots. This fundamental assumption, however, does not always hold: while some educational systems provide explicit training for producing continuous handwriting, minimizing the number of pen-up during the production of a word, others do not. As a consequence, whenever the handwriting presents pen-up within a word, the recognition performance can drop significantly. In a preliminary study, we presented an algorithm for discriminating among different types of ink appearing in handwriting, namely isolated characters, cursive, dots, horizontal and vertical bars, based on the use of a suitable set of features. In this paper, we have characterized the discriminative power of each considered feature according to different measures and we have proposed a method for combining the different feature rankings. We have also used the Fischer's Linear Discriminant Analysis (LDA) for exhaustively selecting the best feature subsets with increasing number of features. Finally, we have compared the results obtained by using the feature subsets provided by LDA with those obtained with the feature subsets selected according to our feature ranking. The experimental results, on different datasets of handwritten words, showed that our approach successfully achieves its aim allowing to reduce the computational cost without affecting the overall performance of the recognition process.

1. Introduction

Handwriting generation studies, and more in general studies on motor control and trajectory planning, show that the complex movements involved in handwriting are composition of elementary movements, each corresponding to an elementary shape or stroke. Such strokes are drawn one after the other during handwriting and the fluency emerges from the time superimposition of them (Plamondon 1995, Grossberg & Paine 2000). Following this line of thought, we have conjectured that handwriting recognition can be achieved by providing the system with a *reference set*, i.e. a set of words whose transcripts are given, decomposing each of the reference word into strokes, and matching the strokes with the transcript so as to associate to each of them the ASCII code corresponding to the character the stroke belongs to. Once the reference set has been provided, handwriting recognition can be achieved by looking within the unknown word for sequences of strokes whose shape resembles that of sequences of strokes found in the reference set, labeling the sequence of the unknown as the matching ones in the reference set, and then combining the labels according to the writing order (De Stefano & al., 2010).

There are cases, however, when our conjecture does not hold. Those are the cases when the word is not produced by keeping the pen-tip in constant contact with the paper, so to have a continuous ink, but lifting the pen here and there while drawing. While such a habit is still within the domain of handwriting generation models, that can explain why and under which circumstances such a behavior appears, it may produce undesired effects in our prototype. Because of the pen lift, in fact, some of the movements do not produce an ink trace on the paper, and therefore some of the strokes are missed. So the sequence of strokes cannot be reconstructed completely, and some of the invariants may disappear, compromising the results of whole process.

To deal with those cases, we proposed in a preliminary study (De Stefano & al., 2011) a method for extracting from a word image the sub-images corresponding to pieces of ink produced without lifting the pen. Each sub-image was described by a suitable set of features and then classified as cursive, isolated character, vertical line, horizontal line, dot or noise. According to this approach, sub-images corresponding to cursive fragments can be processed as described before, while those containing characters can be passed to an OCR module. Thus, the recognition of the whole word can be obtained by composing the results of each module according to the position of the corresponding sub-images in the word image.

To better understand the effectiveness of the above approach, in this study we have characterized the discriminative power of each considered feature in classifying the pieces of ink produced without lifting the pen

as isolated characters or cursive. The basic motivation of our work is to answer this main question: "is it possible to describe handwriting movements just analyzing static images?" We will show that with a suitable set of features extracted from the original images it's often possible to associate each pieces of ink to one of the above two classes.

The remainder of the paper is organized as follows: Section 2 describes the set of considered features, Section 3 illustrates the feature evaluation measures, while the analysis and the discussion of the experimental results, together with some concluding remarks, are eventually left to Section 4.

2. Feature description

The aim of the feature extraction process is that of allowing the classification of connected components of ink traces, possibly produced by writers without lifting the pen, in two main classes: isolated characters and cursive. The basic idea is that a simple shape is generated by a simple motor program. The simpler the motor program, the smaller the quantity of ink the connected component contains. However, in order to improve the fluency of handwriting, a writer may introduce extra strokes, or ligatures, to connect the last stroke of a character and the first of the following one, instead of lifting the pen between the final point of the former and the initial point of the latter. Accordingly, we expect that images of isolated characters will contain less ink (and less strokes) than those of cursive, and that the ink will not span prevalently along the writing direction (De Stefano & al., 2011)

In order to estimate the features of connected components of ink traces, we proceed as follows: The word image is processed for extracting the bounding box of each connected component (see Figure 1a). Then, each component is analyzed by considering its size, the number and the distribution of its black pixels and the size of the word it belongs to (see Figure 1b). In particular, we consider the coordinates of the top-left and bottom right vertices of the bounding box (X_{min} , Y_{min} , X_{max} , Y_{max}), the width and the height of the bounding box (W_{comp} , H_{comp}), the total number of pixels and the number of black pixels included in the bounding box (P_{comp} , BP_{comp}), the width and the height of the bounding box of the word (W_{word} , H_{word}).

Starting from these basic features, an additional set of features is computed, whose description is reported in Table 1. The features *HR*, *AR* and *PAR* are meant to capture the spatual, and hence the temporal, extension of the handwriting, while *FF* is meant to capture the spatial density of ink.

In order to evaluate the shape complexity of the ink trace, we have considered the number of transitions between white and black pixels along consecutive rows/columns of the component. These values have been arranged in two histograms, namely ink-mark on the horizontal (IM_x) and vertical (IM_y) axis, where each bin represents the above number of transitions for a group Δ along a row or a column, respectively (see Figure 1b). These features can be seen as a measurement of the complexity of the ink: an empty or flat ink-mark on both horizontal and vertical axis suggests that the component presents scattered black pixels and is likely to be noise, whereas higher values correspond to more complex shapes.

Finally, we have estimated the center-zone of the word and we have considered as features the y-coordinate of the upper side of the center-zone (say CZ_{Ymin}). Table 2 summarizes the whole set of considered features.

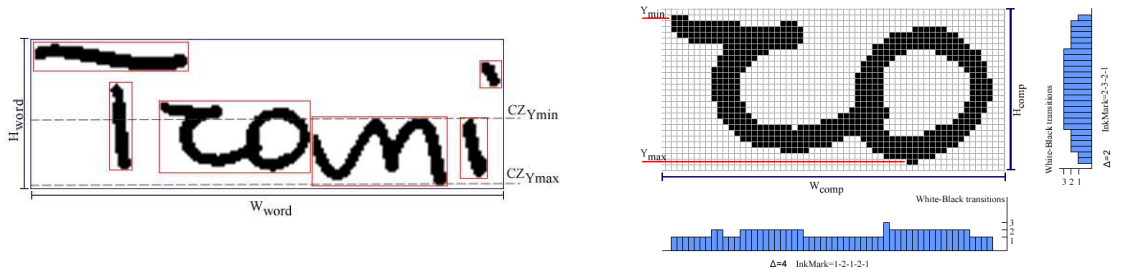

Figure 1: the image of word "Trani" with the bounding box of each connected component and the center zone; a connected component extracted from the word image (right).

Table 1: description of additional features

height ratio (HR)	aspect ratio (AR)	proportional aspect ratio (PAR)	fill factor (FF)
$HR = \frac{H_{comp}}{H_{word}}$	$AR = \frac{W_{comp}}{H_{comp}}$	$PAR = \frac{W_{comp}}{H_{word}}$	$FF = \frac{P_{comp}}{BP_{comp}}$

Table 2: the set of adopted features

F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
IM_x	IM_y	X_{min}	Y_{max}	BP_{comp}	FF	AR	W_{word}	H_{word}	HR	PAR	CZ_{Ymin}

3. Feature evaluation

Two different approaches have been followed for evaluating the effectiveness of each feature and for identifying the subset of them having the highest discriminative power. The first approach is based on the use of standard univariate measures, while the second one uses the Fischer's Linear Discriminant Analysis (LDA).

In the first case, we have considered five standard univariate measures, where each of them ranks the available features depending on their ability in discriminating pieces of ink belonging to either isolated characters or cursive. In our study, we have considered the following univariate measures: Chi-square (*CS*) (Liu & Setiono, 1995), Relief (*R*) (Kononenko, 1994), Gain Ratio (*GR*), Information Gain (*IG*) and Symmetrical Uncertainty (*SU*) (Hall, 1999). The final ranking of all the features is computed by using the Borda Count rule, according to which, a feature receives a score that depends on its position in the rankings provided by each univariate measure. Once the final ranking has been obtained, subsets including increasing number of features (top1; top1 and top2; etc.) are used by a Support Vector Machine (SVM) classifier for testing their discriminating power.

The second approach for evaluating the behavior of subsets including increasing number of features is based on the use of the Fischer's Linear Discriminant Analysis (LDA). In this case we have exhaustively generated from the 12 available features, all the possible subsets of *k* distinct features, without repetitions, varying k from 1 to 12. Thus we created 4095 feature subsets, including 12 sets with only 1 feature, 66 sets with 2 features, 220 sets with 3 features, and so on up to the only set of 12 features. For each subset, the separation index *S* between the two classes has been computed. Denoting with *0* and *1* the two classes to be discriminated, *S* is defined as the ratio of the variance between classes to the variance within classes, using the mean vectors μ_0, μ_1 and the covariance matrices Σ_0, Σ_1 of class 0 and 1, respectively, and $\vec{\omega}$ is described in (De Stefano & al., 2014)

$$S = \frac{\sigma^2_{betwwen}}{\sigma^2_{within}} = \frac{\left(\vec{\omega}(\vec{\mu}_1 - \vec{\mu}_0)\right)^2}{\vec{\omega}^T(\Sigma_1 + \Sigma_0)\,\vec{\omega}}$$

The parameter *S* is a measure of how well the feature subset is able to discriminate between the two classes. It is worth noticing that *S* is a non-decreasing function with respect to the number of features included in a subset. This is the reason why we used S for ranking subsets including the same number of features. Once the best subset including *k* distinct features has been determined using the parameter *S* (with *k* ranging from 1 to 12), we used once again the SVM classifier for testing the discriminating power of that subset.

4. Experimental results

In order to ascertain the effectiveness of the proposed approach, two real world datasets involving handwritten words have been taken into account, namely RIMES and ELSAG database.

The RIMES database is a publicly available dataset used for performance evaluation of handwriting recognition systems (Grosicki, & 2008). It is composed of French words written by more than 1300 volunteers. To validate our algorithm, we extracted 4047 words from the test set and we showed them to 6 human experts. For each word, an expert had to classify manually each connected component and provide its transcript. At the end of this process, 9869 components were manually classified and transcribed, 5101 of them were cursive and 4768 isolated characters.

In the ELSAG database, a set of images representing postal addresses, acquired at 200/300 dpi, was processed in order to segment single words. Then, from each word, the connected components of ink traces, corresponding to cursive or isolated character, were extracted and described by using the above mentioned features. Moreover, in order to evaluate the classification results, each fragmented word image has been shown to 10 experts, and they were asked to label each fragment, to produce the ground truth. At the end of this process, a dataset of 26143 labeled samples has been obtained, containing 15838 isolated characters and 10305 cursive.

Feature evaluation based on the univariate measures has been applied to both databases, producing the results summarized in Table 3. Similarly, LDA approach produced the results reported in Figure 2, where the occurrence of each feature in the optimal subsets selected by LDA is shown. On the basis of these results and applying the previously discussed criteria, we obtained for both evaluation approaches, 12 subsets with increasing number of features, starting from the one including just 1 feature to that including all the 12 features. The effectiveness of each feature subset has been evaluated by implementing a SVM classifier using those features and measuring the recognition performance. In particular, we used for the SVM's the standard algorithm of regularized Support Vector Classification (C-SVC) with a Radial Basis Function kernel. The classification results reported in Figure 3 refer to the application of the 10-fold validation approach and show the plot of the recognition rate as a function of the number of features.

The analysis of these results confirms the effectiveness of the considered features, allowing us to obtain a maximum recognition rate equal to 92.55% and 93.65% for RIMES and for ELSAG database, respectively. The data in the plot show that satisfactory results can be obtained even considering only the top 3 features according

to the Borda Count overall ranking: in this case, in fact, a recognition rate of about 90% is obtained for RIMES database, while a recognition rate higher than 92% is obtained for ELSAG database. It is worth noticing that the results of the Borda Count are comparable, or in some cases better, than those obtained by the LDA. This aspect is particularly meaningful since the univariate measures combined by the Borda Count perform the feature ranking considering one feature at a time, while LDA performs an exhaustive search considering all the possible feature combination, thus implying a very high computational cost.

Future work will include exploiting the information about the classification reliability. Such kind of information would allow the designer of the system the implementation of a reject option for accepting only the high reliable classification on the basis of few features, thus limiting the use of more complex and computationally expensive feature only to the confused cases.

Table 3: Feature ranking according to the Borda Count overall measure. For each row, the leftmost value indicates the best feature, while the rightmost value denotes the worst one.

RIMES	F2	F8	F11	F7	F5	F1	F6	F9	F12	F10	F3	F4
ELSAG	F2	F8	F11	F5	F1	F7	F6	F10	F9	F4	F3	F12

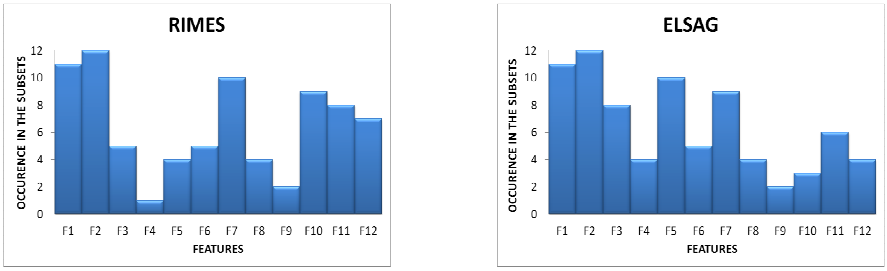

Figure 2: Occurrence of each feature in the optimal subsets selected by LDA for RIMES database (left) and ELSAG database (right).

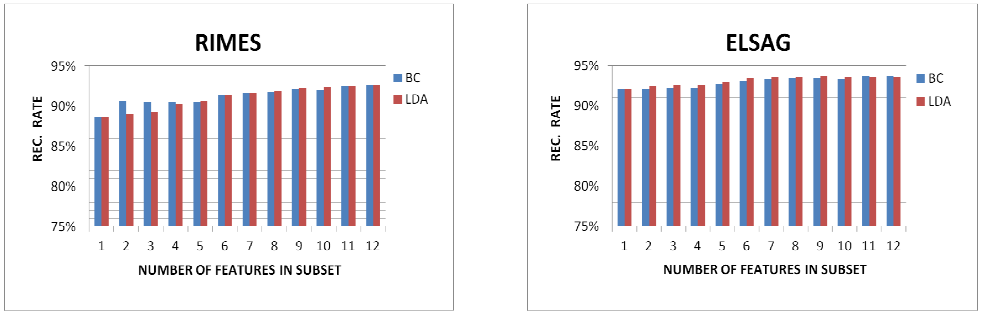

Figure 3: SVM classification results with 10 fold validation on features subsets for RIMES database (left) and for ELSAG database (right).

References

Plamondon, R. (1995). A kinematic theory of rapid human movements. Part I: Movement representation and generation, *Biological Cybernetics,* 72, 297-307.

Grossberg, S., & Paine, R.W. (2000). A neural model of corticocerebellar interactions during attentive imitation and predictive learning of sequential handwriting movements, *Neural Networks,* 13: 999-1046.

De Stefano, C., Marcelli, A., Parziale, A., Senatore, R. (2010). Reading Cursive Handwriting, Proc. Int. Conf. on Frontiers in Handwriting Recognition - ICFHR 2010, Kolkata (INDIA), November 16-18, pp. 95-100.

De Stefano, C., Marcelli, A. & Parziale, A. (2011). Segmenting Isolated Characters Within Cursive Words", Proc. of the 15th International Graphonomics Society Conference (IGS 2011), Cancun, MEXICO, IGS press, pp. 156-159.

De Stefano, C., Fontanella, F., Marrocco, C., Scotto di Freca, A., "GA-based feature selection approach with an application to handwritten character recognition", Pattern Recognition Letters, Vol. 35, pp. 130-141.

Liu, H. & Setiono, R. (1995). Chi2: Feature Selection and Discretization of Numeric Attributes. In: ICTAI, IEEE Computer Society, pp 88–91.

Kononenko, I. (1994). Estimating Attributes: Analysis and Extensions of RELIEF. In: European Conference on Machine Learning, pp 171–182.

Hall, M. (1999). Correlation-based Feature Selection for Machine Learning. PhD thesis, University of Waikato.

Grosicki, E. & al. (2008). RIMES evaluation campaign for handwritten mail processing. Proc. of the Int. Conf. on Frontiers in Handwriting Recognition, Montreal, Canada, pp. 1- 6.