A novel verification system for handwritten words recognition

Laurent Guichard
INRIA Bretagne Atlantique
IRISA, Rennes, France
lguichar@irisa.fr

Alejandro H. Toselli
Instituto Tecnológico de Informática
UPV, Valencia, Spain
ahector@iti.upv.es

Bertrand Couasnon
INSA de Rennes
IRISA, Rennes, France
couasnon@irisa.fr

Abstract

In the field of isolated handwritten word recognition, the development of highly effective verification systems to reject words presenting ambiguities is still an active research topic. In this paper, a novel verification system based on support vector machine scoring and multiple reject class-dependent thresholds is presented. In essence, a set of support vector machines appended to a standard HMM-based recognition system provides class-dependent confidence measures employed by the verification mechanism to accept or reject the recognized hypotheses. Experimental results on RIMES database show that this approach outperforms other state-of-the-art approaches.

1. Introduction

The interest for developing effective verification systems (VSs) for handwritten word recognition applications (HWR) that can distinguish when their outputs are not recognized with enough certainty (and consequently rejected) is still an active research topic. Such VSs are crucial and vital for several security-sensitive applications, as for example the case of the recognition of handwritten postal-address, legal amounts handwritten in bank checks, etc.

Commonly, VSs involve two parts: the confidence measures computation (CMs), which gives an idea of the achieved recognition quality of each word image, and the thresholding-based procedure, which stands for trading off between errors and rejections.

In the literature we can find a wide diversity of VSs for HWR. On one hand are the VSs directly applying a rejection rule to the HWR hypotheses scores [7, 6, 9]. For HWRs based on Hidden Markov Models (HMMs), by far the most successfully employed statistical tool according to the state-of-the-art, the VS rejection mechanism relies on the HMM decoding scores. Those approaches are limited by the intrinsic nature of the HWR, aimed at maximizing the recognition but not the rejection. On the other hand, some VSs, independent from the HWR, re-score the HWR hypotheses before performing the accept/reject action. [8] employs a multilayer perceptron (MLP) to reevaluate the hypotheses, although this kind of classifiers are not designed for the rejection task. We propose to use the latter approach with support vector machines (SVM) to re-score the HWR hypotheses as they already proved their ability to verify isolated handwritten digits [1, 2].

As mentioned above, VS approaches rely on thresholding methods, which intend to adjust threshold values to decide whether accept or reject given recognized hypotheses. The formulation of the best error-reject tradeoff and the related optimal reject rule is given in [3]. According to this, the optimal error-reject trade-off is achieved only if the *a posteriori probabilities* of the classes are known exactly. As they are always affected by errors, [4] suggests the use of multiple reject thresholds to obtain the optimal decision and reject regions. Nevertheless, most VSs employ a single threshold to accept/reject the selected hypothesis. Therefore, the VS we detailed here includes a method to generate artificial classes, each related to a threshold, in order to absorb the problem raised by inexact *a posteriori probabilities*.

In this paper, we present a new independent VS which aims at improving both rejection and recognition capabilities of the verified HWR. Our approach employs an alternative SVM-based confidence measures relying on the HWR grapheme segmentation information, and applies multiple thresholds to optimize the error-rejection trade-off.

This work is organized in the following way. Section 2 details our above-mentioned VS. Experimental results and conclusion are presented in sections 3 and 4.

2. Proposed verification system approach

The proposed VS is suitable for HWRs based on grapheme/character-segmentation (explicit or implicit). For a given word image input s, the HWR outputs the N-best recognized hypotheses along with their corresponding grapheme segmentations and recognition scores. This list of N-best hypotheses serves as input of our VS approach. To represent this list, we employ the following notation: $\langle h_1 = (w_1, r_1), \dots, h_N = (w_N, r_N)\rangle$, where w_i and r_i denote respectively the transcription and grapheme segmentation of the ith recognized hypothesis h_i of word image s. In turn, each hypothesis $h_i = (w_i, r_i)$ is associated with a sequence of grapheme-label and sub-image pairs: $\langle(c_{i,1}, g_{i,1}), \dots, (c_{i,n_i}, g_{i,n_i})\rangle$, where n_i is the number of recognized (grapheme/character) labels of the corresponding hypothesis transcription w_i. Furthermore, each h_i has an associated probability $P_{HWR}(h_i)$ emitted by the HWR.

Our VS approach is compounded by three different modules: *grapheme feature extraction*, *N-best hypotheses re-scoring* and *hypothesis selection and verification*.

The first module makes use of the segmentation information provided by HWR to split input word image into the corresponding grapheme sub-images (i.e. character images in our case). Then, a feature extraction process transforms each of these sub-images into a 95-dimensional real-value vector composed of the following set of features:

- 8th order Zernike moments (45 components);
- 8-contour directions histogram using Freeman chain code representation (48 components);
- Normalized grapheme pixels distributions within area above word upper line and area between base and upper lines (2 components).

The second module performs a re-scoring of each N-best recognized hypotheses by using SVM classifiers, each of which modeling a specific grapheme class c from the whole grapheme classes set considered in the recognition. In this way, given a pair $(c_{i,j}, g_{i,j})$ with $i \in [1, N]$ and $j \in [1, n_i]$, the corresponding SVM assigns it a new score $P_{SVM}(c = c_{i,j}|g_{i,j})$. The SVM output score is approximated to a *posterior probability* by using the *softmax* function, as described in [10]. Once all individual grapheme probabilities have been computed, a global SVM score of hypothesis h_i is calculated as the geometric mean of their respective grapheme scores:

$$P_{SVM}(h_i) = \sqrt[n_i]{\prod_{j=1}^{n_i} P_{SVM}(c = c_{i,j}|g_{i,j})} \quad (1)$$

We realized after some informal experiments that this way of computing the SVM global score works properly well for this case. Moreover, this makes the SVM score independent from hypothesis length (number of graphemes) and thereby comparable across different length hypotheses.

The final confidence measure (CM) of hypothesis h_i is then computed by linearly combining their respective global HMM and SVM scores:

$$P(h_i) = \alpha P_{SVM}(h_i) + (1-\alpha) P_{HWR}(h_i) \quad \forall i \in [1, N] \quad (2)$$

This linear combination of classifier scores aims at balancing the weakness of each of them by the empirically tuned coefficient α.

Once all hypotheses of the N-best list have been rescored, the third and last module is in charge to select the best one (i.e. with the maximal CM score) and to perform the accept/reject action on it. In order to do this, the hypotheses are first re-ordered according to their new CM scores, defining a new list: $\langle \hat{h}_1, \dots, \hat{h}_N\rangle$, such that $P(\hat{h}_i) \geq P(\hat{h}_j) \quad \forall 1 \leq i < j \leq N$. Then, the reject/accept action decision is conducted by the thresholding mechanism using the computed difference of the two best re-scored hypotheses

$$d_{12} = P(\hat{h}_1) - P(\hat{h}_2)$$

as a value to be compared with the corresponding threshold. Experiments conducted by other works [8] have shown that this strategy gives the best results.

As was mentioned in section 1, the proposed verification mechanism is based on multiple class-dependent thresholds. To define these classes, we have clustered into different length-classes all word transcriptions from the HWR lexicon according to their length. It is worth mentioning that the use of length-class-dependent thresholds serves somewhat to mitigate the problem related to the fact that it is not comparable, for example, rejection of 10-characters words with one character error respect to rejection of 2-characters words with one character error.

Formally, the set of length-classes is defined as:

$$\Omega = \{length(w) : w \in \text{Lex}\}$$

where $length$ is a function returning the number of graphemes of word transcription w. We also employ $\omega_j \in \Omega$ with $j \in [1, |\Omega|]$ to denote an element belonging to Ω. Thus, each of the length-classes:

$\omega_1, \omega_2, \ldots, \omega_{|\Omega|}$ has been linked to a respective threshold: $t_1, t_2, \ldots, t_{|\Omega|}$, whose values are set up during the tuning phase. The detailed description of this tuning phase is, for the moment, out of the scope of the present paper.

The verification process performs for a given selected hypothesis $\hat{h}_1$ and its associate threshold t_{j} ($t_{\mathrm{j}} \rightarrow \omega_{\mathrm{j}} = length(\hat{h}_1)$) the accept/reject action of word image s, according to:

$$\textbf{if } d_{12} \geq t_{\mathrm{j}} \textbf{ then } \text{accept } \hat{h}_1 \textbf{ else } \text{reject } \hat{h}_1$$

3 Experiments

3.1 Experimental setup

Experiments have been carried out on the RIMES database used at the ICDAR 2009 competition [5]. The database contains a total of 59 202 running words with their transcriptions and a vocabulary-size of 1 612 different words. Table 1 presents basic statistical information of the corpus along with the partition definition employed to carry out the experiments.

Table 1. Basic statistics of the RIMES-DB words corpus and its standard partition.

Num. of:	Training	Valid.	Test	Total	Lex.
words	44 196	7 542	7 464	59 202	1 612
charact.	230 259	39 174	38 906	308 339	65

The HWR used here is a standard HMMs-based recognizer which extracts feature vectors using a sliding window, models lexicon words by a concatenation of continuous left-to-right grapheme HMMs and employs the Viterbi algorithm to look for the HMM-concatenated models that maximize the probability to produce the given feature vector sequence. We participated to the ICDAR 2009 competition with this HWR (IRISA system), details and results can be found in [5].

To assess our VS, comparisons have been made between our approach and others already published:

SVM-ST: VS presented in section 2 using SVM-rescoring and just a global single reject threshold.

MLP-ST: VS employing MLP classifier-based grapheme re-scoring (see [8]). As **SVM-ST**, it uses just a global single reject threshold.

HMM-ST: as described in [7], a global single reject threshold is applied to the difference between the CMs of the first and second HWR hypotheses.

SVM-MT: our VS explained in section 2 using SVM-rescoring and multiple reject thresholds.

The SVM classifiers employed to re-score graphemes use a Gaussian kernel and were trained with the one-against-all strategy for multi-class SVM classification. In this sense, grapheme samples to train SVM and MLP classifiers were obtained through segmenting the word images of the training set with our HMMs-based HWR in forced alignment mode.

The RIMES-DB partition sets employed in the experiments are highlighted in table 1. While HMMs, SVMs and MLPs parameters learning is carried out on the training set, multiple thresholds tuning is performed on the validation set using an algorithm derived from [11]. Finally, reported results of the comparisons among the different approaches have been obtained on the test set.

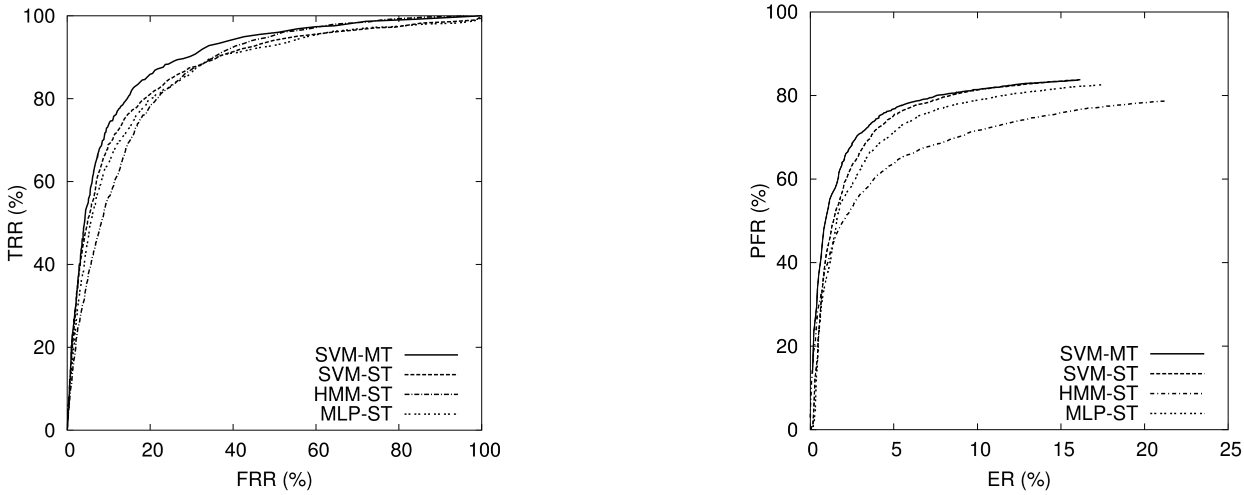
For the VS using multiple reject thresholds, a number of 17 thresholds were set according to the number of classes produced by regrouping the RIMES lexicon words with the same lengths, (i.e. RIMES lexicon contains words varying from 1 to 17 characters). The number of hypotheses generated by the HWR for each recognized word-image was set to 10.

To compare the performance of the different VS approaches, the *Receiver Operating Characteristic* (ROC) curve which plots the True Rejection Rate (TRR) versus the False Rejection Rate (FRR) was used. The TRR (resp. FRR) is defined as the number of wrong (resp. well) recognized words that are rejected divided by the number of well (resp. wrong) recognized words. In addition, the area under a ROC curve provides an adequate overall estimation of the rejection capabilities. This area is denoted as AROC. The *Performance* (PFR) versus *Error Rate* curve is also plotted to demonstrate the increase of well recognized words brought by the VS. The PFR (resp. ER) is defined as the number of well (resp. wrong) recognized words divided by the total number of words.

3.2 Evaluation of the proposed VS

The following results were all obtained on the test set partition. Figure 1-(a) presents the ROC curves obtained through the four different VS approaches: **SVM-MT**, **SVM-ST**, **HMM-ST** and **MLP-ST**. It can be observed that **SVM-MT** and **SVM-ST** are the best performing approaches in the FRR range of 0% to 30%. Clearly in that range, **SVM-ST** outperforms **HMM-ST** and **MLP-ST**, corroborating in this way the CM quality of the approach. Similarly, **SVM-MT** outperforms all of the others, including **SVM-ST**, confirming that multiple-thresholds-based VSs generally performs better than single-threshold one.

Additionally, figure 1-(b) plots the VS performance



(a) (b)

Figure 1. (a) ROC curve for each VS. (b) performance (PFR %) versus error-rate (ER %) for the different VS.

versus error rate for each of the proposed approaches. Once again, it is specially notable for the ER range of 0% to 2.5%, the good performance achieved by **SVM-MT** and **SVM-ST** with respect to the others.

Those experiments demonstrate the superiority of our VS **SVM-MT**. One important feature to notice is the improvement in term of performance even without rejection. Indeed, the performance of the HWR (**HMM-ST**) increases from 78.6% to 83.7% when adding our VS (**SVM-MT**)

For each VS, table 2 gives the AROC values, the TRR values for a FRR set to 10% and the PFR values without rejection and for an ER set to 2.5%.

Table 2. AROC values, TRR values for a constant FRR set to 10%, PFR values without rejection (PFR_1) and PFR values for a constant ER set to 2.5% (PFR_2)

Approach	AROC	TRR(%)	PFR_1(%)	PFR_2(%)
SVM-MT	0.899	73.3	83.7	68.4
SVM-ST	0.874	68.9	83.7	63.1
MLP-ST	0.864	64.5	82.3	58.4
HMM-ST	0.822	56.3	78.6	53.6

4 Conclusion

This paper introduces an alternative independent verification system using a confidence measure based on SVMs rescoring and multiple rejection thresholds to verify handwritten word recognized hypotheses. The experimental results obtained show that the proposed approach boosts the rejection capabilities of the HWR as, for example, the performance increases from 53.6% to 68.4% for an error rate set to 2.5%. It also improves the global recognition performance which rises from 78.6% to 83.7% when rejection is disabled.

References

[1] A. Bellili, M. Gilloux, and P. Gallinari. An mlp-svm combination architecture for offline handwritten digit recognition. *IJDAR*, 5(4):244–252, July 2003.

[2] C. Chatelain, L. Heutte, and T. Paquet. A two-stage outlier rejection strategy for numerical field extraction in handwritten documents. In *ICPR*, volume 3, pages 224–227, 2006.

[3] C. K. Chow. On optimum error and reject tradeoff. *Information Theory Society*, 16(1):41–46, Jan. 1970.

[4] G. Fumera, F. Roli, and G. Giacinto. Reject option with multiple thresholds. *Pattern Recognition*, 33:2099–2101, 2000.

[5] E. Grosicki and H. E. Abed. Icdar 2009 handwriting recognition competition. In *ICDAR*, 2009.

[6] M. N. Kapp, C. Freitas, and R. Sabourin. Handwritten brazilian month recognition: An analysis of two nn architectures and a rejection mechanism. *IWFHR*, 0:209–214, 2004.

[7] A. L. Koerich. Rejection strategies for handwritten word recognition. In *IWFHR*, pages 479–484, 2004.

[8] A. L. Koerich, R. Sabourin, and C. Y. Suen. Recognition and verification of unconstrained handwritten words. *TPAMI*, 27(10):1509–1522, 2005.

[9] S. Madhvanath, E. Kleinberg, and V. Govindaraju. Holistic verification of handwritten phrases. *TPAMI*, 21(12):1344–1356, 1999.

[10] J. Milgram, M. Cheriet, and R. Sabourin. Estimating accurate multi-class probabilities with support vector machines. In *IJCNN*, volume 3, pages 1906–1911, 2005.

[11] H. Mouchere and E. Anquetil. A unified strategy to deal with different natures of reject. In *ICPR*, volume 2, pages 792–795, 2006.