A two-stage outlier rejection strategy for numerical field extraction in handwritten documents

Abstract

In this article, we propose a segmentation-driven recognition system which aims at extracting numerical fields from handwritten documents. We show that a crucial point of the system is the rejection ability of the handwritten numeral classifier. Therefore, we propose a simple two-stage outlier rejection strategy, and we show the benefit of this strategy on the numerical field extraction results.

1 Introduction

Nowadays, no system is able to read automatically a whole page of cursive handwriting without any *a priori* knowledge. This is due to the extreme complexity of the task when dealing with free layout documents, unconstrained cursive handwriting, and unknown textual content of the document. Nevertheless, it is now possible to consider restricted applications of handwritten text processing which may correspond to a real industrial need. The extraction of numerical data (file number, customer reference, phone number, ...) in an incoming mail document (see figure 1) is one particular example of such a realistic task.

The main idea of our approach is to exploit the known syntax of a numerical field to locate it in a text line [2]. For example, a french phone number is always made of ten digits, with optional separators between each pair of digits. Thus, the extraction of a phone number in a text line consists in the detection of a sequence of ten digits with optional separators in the whole line sequence. This is performed by a numeral component recognition stage followed by a syntactical analysis of the recognition hypotheses, which filters the syntactically correct sequences with respect to a particular syntax known by the system. Thus, a crucial point of this system is the ability of a classifier to discriminate numeral patterns from the rest of the document: word, fragment of word, noise, etc. that one can call *outliers*. In this article, we propose a simple two-stage outlier rejection strategy which improves the final system performance.

This paper is organized as follows. In section 2 we present an overview of the numerical field extraction system with a brief description of each processing stage. Section 3 deals with the outlier rejection strategy embedded in the system. We present in section 4 our experimental results on a database of real handwritten incoming mail documents. Conclusion and future works are drawn in section 5.

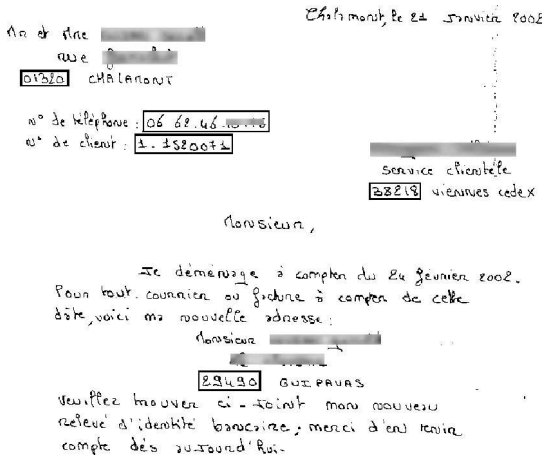

Figure 1. Incoming mail document

Figure 2. Examples of numerical fields.

2 Numerical field extraction

The numerical field extraction strategy relies on a syntactical analysis of the lines of text in order to filter the syntactically correct sequences with respect to a particular syntax known by the system. Hence, a recognition stage is required to distinguish numerical components (isolated and touching digits) and separators (point or dash) from the rest of the document (outliers). This is performed thanks to a segmentation-driven recognition described below, which provides a three-level recognition trellis with confidence

values for each component, concatenated over all the line (see figure 3, where 'X' denotes a confidence value).

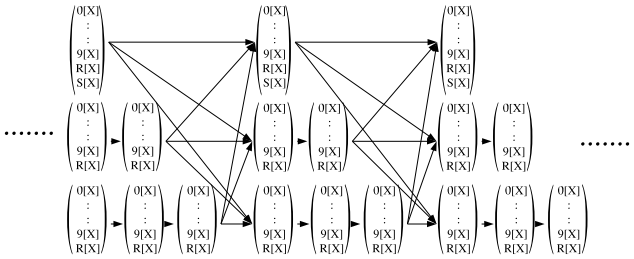

Figure 3. Line trellis obtained by concatenation of the component trellis.

Text line model: a line model is defined for each kind of numerical field, which provides the syntactical constraints of a text line that may contain a numerical field. Models are made of the 12 states described above: 10 classes of digit + separator (S) + outlier (Reject: R) . As an example, the phone number text line model is presented in figure 4. Authorized transition between states have been learned on a handwritten document database containing numerical fields. The exploration of the trellis is performed according to the confidence values of the recognition hypotheses by dynamic programming [9] under the constraints of the model.

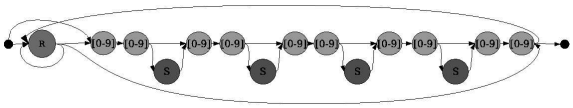

Figure 4. Phone number line model.

Segmentation-driven recognition: the aim of this recognition stage is to detect the components which belong to a numerical field: single or touching digits, and separators. All the remaining components are outliers and must be rejected. Hence, components are successively considered as: numerical components, separators and outliers, according to three different strategies :

Digits: the single and touching digit recognition is performed thanks to a segmentation-driven recognition which successively considers a component as a single, double and triple digit. Figure 5 gives an example of the segmentation-driven principle for the recognition of a component as a double digit: several cutting paths are generated and are submitted to a digit classifier. The path which maximizes the confidence product is retained (in this example, the first path). This stage is re-itered for the recognition of triple digits.

Drop fall	ascending left	ascending right	descending left	descending right
cutting path				
digit classifier output	0[98] 8[82]	2[27] 8[35]	0[73] 8[36]	0[92] 8[34]
confidence product	81	09	26	32

Figure 5. Double digit recognition example

Separators: the separator recognition is performed thanks to a small classifier based on contextual features [2].

Reject: since most of the components are outliers, the numeral and separator recognition hypotheses must be submitted to an outlier rejection system which provides a confidence value for the reject class. This outlier rejection system is described in section 3.

Hence the outputs of the recognition stage performed on each component are concatenated over all the line to produce finally a 3-level recognition hypothesis trellis (see figure 3).

3 Outlier rejection strategy applied on a digit classifier

In this section, we focus on the design of an outlier rejection strategy, based on a standard 10 class digit classifier. As seen in the previous section, the digit classifier should be able to output 11 confidence values: ten for digit classes and one for the outlier class.

If the discrimination between handwritten digits is now a quite well-solved problem, the outlier rejection is still a tough problem due to the extrem variability of outlier patterns. The analysis of a database of outliers leads us to consider roughly two kinds of outliers (see figure 6): **(i) Obvious outliers** which have a very different shape from isolated digits like noise, fragment or entire words, stroke, points or dash, etc. **(ii) Ambiguous outliers** which have a similar shape with single digits: letter, group of letters or fragment of word mainly. These outliers are more difficult to distinguish from digits. This observation leads us to consider a two-stage strategy to reject outliers (see figure 7):

Figure 6. obvious (first line) and ambiguous (second line) outlier examples.

The first stage is used to reject obvious outliers. As it seems easy to distinguish obvious outliers from digits, we propose to design a 2-class classifier based on a restricted number of features. The aim is to reject as many outliers as possible, *while accepting all the digits*. Thus, this stage provides a binary decision (accept/reject).

The second stage aims at discriminating ambiguous outliers from digits among the patterns accepted by the first stage. As it seems to be a tough problem, we propose a soft decision, based on the analysis of the confidences values of a 10-class numeral classifier. We now detail these two stages.

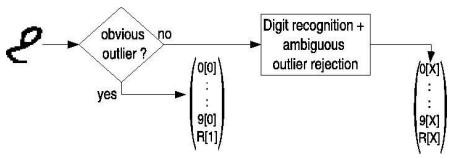

Figure 7. Two-stage outlier rejection strategy

3.1 Obvious outlier rejection

This first stage acts as a filter which decides whether a pattern is an obvious outlier or not. For that, we have designed a 8-feature set and a Support Vector Machine classifier (RBF kernel) [12], known to be very accurate for two-class discrimination problems.

The 8 features are: (f_1) height/width ratio, (f_2) black pixel density, (f_3) number of water reservoirs (metaphor to illustrate a valley in a component, see [7] for more details), (f_{4-6}) number of intersections with two horizontal and one vertical straight lines, (f_7) number of end points, (f_8) number of holes. The SVM has been trained on a database of 7,500 patterns (5,000 outliers, 2,500 digits) with a cost matrix penalizing the false rejection (FR) with respect to the false acceptance (FA). The parameters C and σ of the SVM and the cost parameter have been fixed empirically to minimize the Area Under ROC Curve (AUC). The results of the training shows on the test database that 54% of the outliers can be rejected efficiently while rejecting only 0,2% of the true digits.

3.2 Ambiguous outlier rejection

The second stage of our approach deals with the discrimination between ambiguous outliers and digits among the patterns accepted by the previous stage.

Several techniques have been designed for the rejection of outliers: training a classifier with outlier data [5], modeling the target classes and perform a distance rejection strategy [6], use of one class classifiers [11], reject outliers with respect to the outputs of a classical digit classifier [8]. We have chosen this latter solution, applied on a MultiLayer Perceptron (MLP), for the following reasons:

- As the classifier has to process a whole page of handwriting, we cannot use a large time consuming classifier during the decision stage (this constraint prohibits for example multiclass SVM and one-class SVM). MLPs well suit this condition because they have an extremely fast decision processing.
- If the use of model-based classifiers (RBF, one class classifier, etc.) allows a distance-based rejection strategy, these classifiers suffer from a poor discrimination ability, and the modelisation of classes in high dimensional spaces is still a difficult problem. Oppositely, MLPs have very good discrimination performance and are well adapted to high dimensional spaces [1, 5].

We have thus designed a combination of two MLPs, trained on 130,000 digits, with a 117-structural/statistical feature set developped in our previous work [3], and a 128-feature set extracted from the chaincode [4]. A product rule combination is performed between the two MLPs. On a test database containing 60,000 digits, the classifier combination provides a recognition rate of 98.44%, 99.48% and 99.75% in TOP 1,2,3 respectively.

From this point, the rejection rule is the following : a confidence value for the reject class is estimated with a Look Up Table (LUT) according to the confidence value of the first proposition of the digit recognizer. The LUT has been generated by considering the behaviour of the digit classifier on a database of 2,300 digits and 4,000 outliers. Thus, the classifier provides 11 confidence values (10 digits + reject). The softmax function is applied on the 11 confidence values to output *a posteriori* probability estimates.

The outlier rejection ability is evaluated with the Receiver-Operating Characteristic (ROC) curve which is a graphical representation of the trade-off between the false rejection (true digit rejection) and false acceptance (outlier acceptation) rates for every possible cut off (confidence value of the first proposition of the MLP). We show on figure 8 the ROC curve obtained when we only use the digit classifier, and when the digit classifier is preceded by the SVM classifier described above. The trade-off between FR and FA is clearly better with the use of the SVM classifier as prior filter to the digit classifier.

4 Results

This section presents the results of the numerical field recognition system. We have evaluated our approach on a database of 293 handwritten incoming mail documents containing ZIP codes, phone numbers and customer codes. As

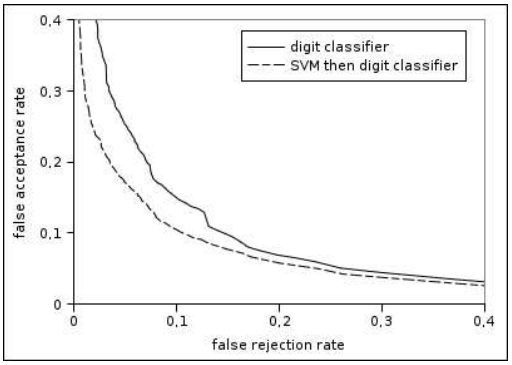

Figure 8. ROC curve

we propose an information extraction system, the performance criterion is the trade-off between recall and precision rates. The recall and precision rates are defined as:

$$recall = \frac{\text{nb of fields well recognized}}{\text{nb of fields to extract}}$$

$$precision = \frac{\text{nb of fields well recognized}}{\text{nb of fields proposed by the system}}$$

The syntactical analysis is performed thanks to the forward algorithm, which provides the n best alignment paths. A field well detected in *TOP n* means that the right recognition hypothesis for a field stands in the n best propositions of the syntactical analyser. Figure 9 shows the recall-precision trade-off for different values of n, while using or not the SVM classifier as a prior outlier rejection stage.

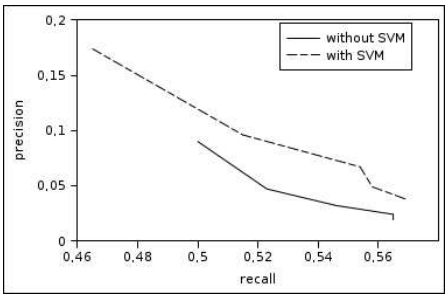

Figure 9. Recall-precision trade-off

We notice the improvement of the recall-precison trade-off while using the SVM classifier. Indeed, the use of a binary classifier as a prior outlier rejection stage bans obvious outlier and helps the syntactical analyser to find the best path.

5 Conclusion and future works

Thanks to a simple feature vector and a SVM classifier used as prior to a digit classifier, we have improved the outlier rejection ability of the digit recognizer. We have shown the influence of such an improvement when the classifier is embedded in a segmentation-driven recognition process. Our future works will focus on the integration of contextual features in order to take into account the location of the components related to the mean line. Another perspective is to replace the recognition rate criterion of the two classifiers (SVM and MLP) by a criterion which minimizes the Area Under ROC Curve [10] in order to improve the outlier rejection capacity of the system.

References

[1] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[2] C. Chatelain, L. Heutte, and T. Paquet. Segmentation-driven recognition applied to numerical field extraction from handwritten incoming mail documents. *Document Analysis System, Nelson, New Zealand*, pages 564–575, 2006.

[3] L. Heutte, T. Paquet, J. Moreau, Y. Lecourtier, and C. Olivier. A structural/statistical feature based vector for handwritten character recognition. *Pattern Recognition Letters*, 19:629–641, 1998.

[4] F. Kimura, S. Tsuruoka, Y. Miyake, and M. Shridhar. A lexicon directed algorithm for recognition of unconstrained handwritten words. *IEICE Trans. on Information & Syst.*, E77-D(7):785–793, 1994.

[5] J. Liu and P. Gader. Neural networks with enhanced outlier rejection ability for off-line handwritten word recognition pattern recognition. *Pattern Recognition*, 35:2061–2071, 2002.

[6] J. Milgram, R. Sabourin, and M. Cheriet. An hybrid classification system which combines model-mased and discriminative approaches. *ICPR'04*, pages 155–162, 2004.

[7] U. Pal, A. Belaïd, and C. Choisy. Water reservoir based approach for touching numeral segmentation. *ICDAR*, 2001.

[8] J. Pitrelli and M. Perrone. Confidence-scoring post-processing for off-line handwritten-character recognition verification. *ICDAR'03*, 1:278–282, 2003.

[9] L. R. Rabiner. A tutorial on hidden markov models and selected apllications in speech recognition. In *Readings in Speech Recognition*, pages 267–296. Kaufmann, 1990.

[10] A. Rakotomamonjy. Optimizing auc with support vector machine. *European Conference on Artificial Intelligence Workshop on ROC Curve and AI*, 2004.

[11] D. Tax and R. P. W. Duin. Combining one-class classifiers. In *MCS '01*, pages 299–308, 2001.

[12] V. Vapnik. The nature of statistical learning theory. *Springer*, 1995.