Recognition-based *Vs* Syntax-directed Models for Numerical Field Extraction in Handwritten Documents

Clement Chatelain
LITIS lab., Rouen, FRANCE
clement.chatelain@insa-rouen.fr

Laurent Heutte
LITIS lab., Rouen, FRANCE
laurent.heutte@univ-rouen.fr

Thierry Paquet
LITIS lab., Rouen, FRANCE
thierry.paquet@univ-rouen.fr

Abstract

In this article, two different strategies are proposed for numerical field extraction in weakly constrained handwritten documents. The first extends classical handwriting recognition methods, while the second is inspired from approaches usually chosen in the field of information extraction from electronic documents. The models and the implementation of these two opposed strategies are described, and experimental results on a real handwritten mail database are presented.

Keywords: Handwriting recognition, document analysis, information extraction, numerical fields

1 Introduction

Nowadays, a huge number of both electronic and paper documents, including machine printed and handwritten documents is daily exchanged between administrations, industries and end users. The automatic processing of these document flows can significantly improve the end user satisfaction, and therefore has been identified as the bottleneck of client services dedicated to the incoming mail documents analysis. While information extraction has been intensively studied on electronic documents [BSW99, FM99, Lee97] and machine-printed documents, few studies have dealt with information extraction in handwritten documents. We can mention the extraction of particular fields in forms, the extraction of relevant information on bank checks, such as the legal amount [HBPB^{+}97], the date [MSBS02] or the signature [MYHK03].

In this paper, we focus on the extraction of numerical fields in unconstrained handwritten documents. As opposed to the extraction of handwritten numerical fields in forms, such as bank checks for example, the system cannot rely on the strongly constrained layout to locate the fields of interest before their recognition. Consequently, the system has to manage both the localisation task and the recognition task. Such problem falls into the general framework of information extraction, for which shallow sentence models have been proposed and proved to be effective, particularly on electronical documents. These models rely on the modeling of the positive (relevant) and the negative (irrelevant) information in two different ways:

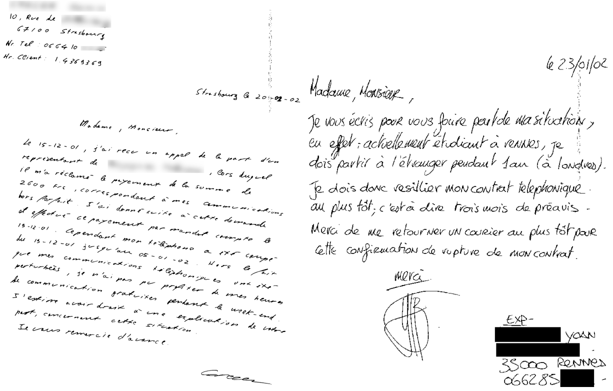
- A fine modeling of the relevant information (the handwritten numerical information in this study).
- A shallow modeling of the irrelevant information (the remaining of the document: text, graphics, etc.)

Using such a framework, shallow model of the negative information serves as a model of a rejection class and therefore does not require a deep modeling of the sentence using complex language models. Such shallow models have often been applied within the context of the Message Understanding Conferences [MUC98] in the nineties, for example to extract proper nouns [BSW99], document title and abstract [FM99], or fact [Lee97] from electronical documents.

The paper is organized as follows: First, we present the context of the study (section 2). In section 3, we analyse the strategies and the models involved by the numerical field extraction and recognition task. Two general strategies are retained, and their implementation is roughly described. In section 4, we present the performance of each implementation on a real incoming mail document database, and compare the two approaches by means of the recall/precision trade-off, a common measure for evaluating Information Extraction systems. Conclusion and future works are drawn in section 5.

2 Context of the study

The documents treated in this study are the incoming mails documents, which are daily massively received by the big organizations. Today, the automatic reading of the incoming documents is limited to certain kinds of constrained documents: mainly forms, bank checks, invoices, etc. The free-layout handwritten mails (see figure 1) are still extremely difficult to treat.



[redacted]
10, Rue de [redacted]
67100 Strasbourg
Nr. Tel : 066410[redacted]
Nr. Client : 1.4369369

Strasbourg le 20-02-02

Madame, Monsieur,

Le 15-12-01, j'ai recu un appel de la part d'un representant de [redacted], lors duquel il m'a réclamé le payement de la somme de 2600 Frs, correspondant à mes communications hors forfait. J'ai donné suite à cette demande et effectué ce payement par mandat compte le 19-12-01. Cependant mon téléphone a été coupé du 15-12-01 jusqu'au 05-01-02. Hors le fait que mes communications téléphoniques ont été perturbées, je n'ai pas pu profiter de mes heures de communication gratuites pendant le week-end. J'estime avoir droit à une explication de votre part, concernant cette situation. Je vous remercie d'avance.

le 23/01/02

Madame, Monsieur,

Je vous écris pour vous faire part de ma situation, en effet : actuellement étudiant à rennes, je dois partir à l'étranger pendant 1an (à londres).

Je dois donc resilier mon contrat telephonique au plus tôt; c'est à dire trois mois de préavis -

Merci de me retourner un courier au plus tôt pour cette confirmation de rupture de mon contrat.

Merci

EXP-
[redacted] YOAN
[redacted]
35000 RENNES
066285[redacted] -

Figure 1. Handwritten incoming mail documents.

This is due to the extreme complexity of the automatic reading task when dealing with free layout, unconstrained cursive handwriting, and unknown textual content documents. Indeed, let us remark on figure 1 that there is no stable document model since the content and the structure may strongly differ between two documents: the header may take place at the top or at the bottom of the document. Let us also note the various handwriting styles: cursive, script or mixed, which is to add to the intrinsic handwriting variability.

Within the framework of the automatic processing of incoming mail documents, the numerical information constitutes relevant information as they generally contain information relative to the customer, which allow to determine its identification (thanks to a telephone number or a customer code), his kind of contract (customer code) or its geographical localization (postal code) via a customer database.

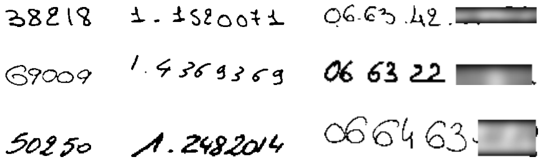


Figure 2. Numerical fields examples: Zip codes, customer codes, phone number.

In this paper, we focus on the strictly numerical sequences such as Zip codes, customer codes, or phone number (see figure 2). Hence, the mixed numerical/textual fields such as date are not considered.

3 Handwritten numerical fields extraction strategies

In this section, we address the problematic of numerical field extraction in handwritten documents, and we consider the different strategies to tackle this tough problem. Then, the problem of information modelisation is discussed in section 3.1, which leads to two opposed strategies described in section 3.2.

3.1 Information modelisation

While the recognition of handwritten isolated numerical sequences can now be considered as a solved problem (see the recognition of numerical bank check amounts [KG97], ZIP codes [LSF04] or dates [MSBS02]), challenges in building reading systems now consist in dealing with weakly constrained handwritten documents in which the location of the numerical information is unknown. Therefore, following the Sayre's paradox, the localization of the entities must be coupled with their recognition to be reliable. As numerical fields might appear anywhere in the document, then two strategies can be envisaged to tackle the problem:

- The first one would consist in building a complete reading system taking into account numerical information as well as textual information.
- The second strategy would consist in building a shallow reading system dedicated to the detection and recognition of numerical entities only.

As mentionned in [PS00], the first strategy appears to be unrealistic at present, considering the state-of-the-art of off-Line handwriting recognition systems. Indeed, such strategy would require the use of textual knowledge (lexical, linguistic knowledge) so as to cope with textual entities, in addition to the use of sophisticated models of handwritten numerical information. Besides, such complex modelisation, even if achievable, would probably lead to poor performance due to the low performance of the current recognition systems of unconstrained handwritten words and sentences. From these considerations, the second strategy consisting in building a shallow reading system appears to be more realistic and promising. Such a system dedicated to the extraction and recognition of a specific subpart of the written message only fall into the category of the Information Extraction Systems.

Such systems have been intensively studied among the Message Understanding community [MUC98] for the extraction of relevant textual information in electronic documents. In opposition, few approaches have considered Information Extraction systems in handwritten documents [Koc06, EYGB02]. The general strategy adopted by any of these systems is based on a text line model or a sentence model, integrating both the relevant information (key words, street names, proper names, etc), and the non-relevant information, which is the remaining of the document (out-of-lexicon words, street number, non-relevant words, punctuation, etc). Coupling a precise model of the expected entities with a coarser model of the non-relevant information provides a shallow model of the information dedicated to the extraction task. Then, the optimal information labelling can be obtained by aligning the model on the data sequence using a dynamical programming scheme. Thus, the shallow model of a line of text which may contain a numerical field can be represented on figure 3.

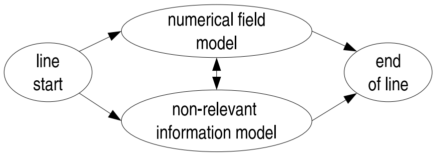


Figure 3. Shallow text line model with potential numerical field.

Such a model allows the extraction of numerical fields, but raises multiple questions:

How to model the numerical fields?

As opposed to handwritten word modeling, which exploits lexicon knowledge using lexicon driven recognition strategies so as to overcome character segmentation of cursive handwriting, numerical field models cannot exploit such reliable constraints. However, it is possible to exploit the known syntax of each kind of numerical field in order to constraint the solutions. Indeed, digit sequences generally obey some syntactical rules depending on the kind of field considered. The number of digits, the presence and position of possible separators are generally known, although uncertain. For example, a french telephone number always consists of ten digits generally gathered by pairs and possibly separated by points or dash (see figure 2). These syntactical rules concerning the fields are the only a priori knowledge available for such problem, and therefore has to be exploited by the numerical fields models.

How to model nonrelevant information?

The modeling of the non-relevant information is a difficult problem due to the diversity and the variability of the non-numerical shapes: textual information (words), punctuation, noise, digits that do not belong to a field of interest, symbols, etc. Following Information extraction strategies proposed in previous studies, and as we do not aim at recognizing the non-relevant entities, it is possible to apply a coarse or shallow model of the non relevant entities. For example, the 26 character classes can be gathered together into one single character class; similarly the 10 digit classes can be put together into one single meta class too. Doing this way, the simplest model of the non relevant information consists in a single rejection class, including all the non relevant shapes. This approach has been used for example by Bikel [BSW99] to model all the non-relevant information for the proper names extraction task in electronic texts.

Which processing stage are implied and how to connect them?

Finally, in addition to the information modeling, the third question raised concerns the implementation strategy of the models. Traditionally, one of the most difficult problem when recognizing cursive handwritten information is the segmentation of the entities into characters or digits. Here, the question is to know when (where?) a segmentation process should take place in the whole process. Indeed, while segmentation appears necessary to recognize the relevant information (sequence of digits), its behaviour when applied on the non relevant information is quite uncertain. Faced to a complex Segmentation/Recognition/Rejection problem, many strategies can be envisaged, all based on the 4 following processing stages: (i) Segmentation of handwritten components, (ii) Recognition of the numerical components, (iii) Identification of non-numerical components, (iv) Detection of numerical fields of interest using syntactical analysis. Among all the possible strategies, we have particularly focussed our attention on two important opposed strategies. Namely a ”recognition-based” strategy and a ”syntax-directed” strategy (see figure 4). We now describe these two strategies in more details, and for each of them we also derive the involved models.

3.2 Recognition-based *vs* syntax-directed strategies

The most intuitive strategy consists in sequentially applying the 4 stages in the order in which they are mentioned above. This is probably the most natural strategy for off-Line handwriting recognition. This strategy starts with a digit segmentation/recognition stage, then the information of interest is modeled and detected at the final level of the whole process. A first segmentation stage provides handwritten elementary components that are potential handwritten digits. A second stage recognizes the numerical components thanks to a digit classifier. Non numerical components can be detected during a third stage by analysing the digit classifier outputs. The fourth and final stage is the localization of the fields that is performed using a syntactical line model fed with the classification/Rejection hypotheses. In this strategy, the extraction stage is based on the recognition of the numerical entities to build the relevant numerical fields which verify the line model syntax. For this reason, one can call this approach a recognition-based strategy.

The second strategy aims at detecting the numerical fields at the early stage of the whole process. For this reason it is a syntax-directed strategy. Following this strategy, the segmentation/recognition stage is applied at the end of the process, while the rejection of the non relevant entities is carried out prior to the digit segmentation/recognition stage by exploiting syntactical properties of the fields of interest.

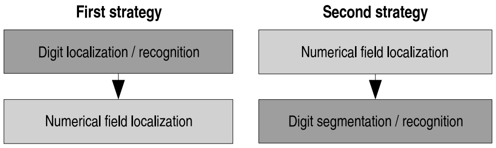

First strategy
Digit localization / recognition
Numerical field localization
Second strategy
Numerical field localization
Digit segmentation / recognition

Figure 4. General strategies for numerical field extraction.

Figure 4 gives a brief outline of the two strategies. From this figure, it appears obvious that the two main processing stage are inverted in these two strategies. In this work, these two strategies have been implemented in order to compare them. However, there are many implementation variations possibilities. Thus, we now briefly study the various choices related to the strategies implementation, and discuss in particular the crucial segmentation problem. For more details concerning the implementation, please refer to the description of each systems in the two previous IWFHR [CHP04, CHP06].

3.2.1 First approach implementation

This first extraction strategy is based on a digit recognition method applied to the whole document. This stage is followed by a discrimination stage between the numerical and non-numerical components. The field localization is then performed by the alignment of the digit localization/recognition hypotheses on the models which integrate the *a priori* knowledge related to the fields.

The central point for the implementation of this strategy is the component segmentation. It can be implemented according to two approaches depending on whether an explicit or implicit segmentation is considered. In this work we have implemented an explicit segmentation strategy that has already been successfully used in previous work at LITIS laboratory [HBPB$^+$97, Koc06]. Starting from the whole components, a descending segmentation method based on both the *drop fall* and the *water reservoir* algorithms is performed (please refer to [CHP06] for more details). Coupled with a digit recognition classifier based on neural networks and efficient feature sets, it provides a strictly numeral 3-level digit segmentation/recognition trellis.

Given this numeral trellis, a rejection stage must also be considered to reject all the components that are not digits (digit fragments, words, word fragments, punctuation, noise, etc). This appears to be a very complex task: indeed, the Sayre's paradox becomes the following: it is necessary to segment to recognize and recognize to segment, but it is also necessary to recognize to reject and reject to recognize, segment to reject and to reject to segment! In this study, the rejection stage is implemented based on the analysis of the scores of the digit classifier outputs.

The last stage of the system is the global segmentation/recognition/rejection decision overall the text lines, i.e. the filtering of the valid sequences among the complete trellis with respect to the known *a priori* constraints. The *a priori* knowledge such as the number of digits of a researched field and the potential position of the separators are merged into some text line models, for each kind of fields of interest. This stage has been performed using our incoming mail document database, where the position the nature and the value of the numerical field are labelled. The analysis of these fields has allowed us to identify the syntactic rules governing these fields. For example, a french postal code is made of five digits, which allow us to build the associated line model: it is made of an undefined number of reject components, followed by the five digits, followed again by an undefined number of reject components. The figure 5 presents the line models for a telephone number.

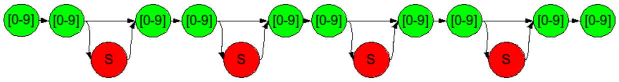


Figure 5. Phone number model.

The research of the best recognition path under the constraints of the models is performed using the well-known dynamic programming forward algorithm [Rab90].

3.2.2 Second approach implementation

In opposition to the first approach where the numerical field localization occurs at the end of the processing flow, this second strategy aims at performing the numerical field localization as soon as possible. For that, a two-stage strategy is proposed :

1) This first stage aims at both extracting the fields of interest, and rejecting the remainder of the document, without any digit recognition.

2) The second stage is dedicated to the recognition of the localised numerical fields. The numerical field localisation hypotheses are submitted to a numerical sequence recognition process inspired from the early works of the literature such as those originally proposed for ZIP code recognition or bank checks processing.

The central point of this second strategy lies in a low level (before recognition) line model. This model has to integrate the specific syntax of the searched numerical fields (number of digits, presence and position of separators) together with a low level description of the handwritten entites based on the definition of syntactical classes (textual components, digit components, etc...). This low level description of the component s has been chosen segmentation-free. Consequently, a numerical component belongs to one the following syntactical classes: D (Digit), DD (Double Digits), S (Separator). Let us note that touching digits containing more than two digits could also be considered, but these components are very rare, and we did not consider these classes for the current implementation. Concerning the textual components, only one single generic class, called Reject class, has been considered so as to describe all the following non-relevant component: isolated character, fragment of word, word, diacritic, punctuation mark, symbol. These 4 syntactical classes allow the definition of a Markovian model that models the syntactical constraints of the numerical fields. Thanks to the definition of these 4 low level syntactical classes, the localization of a numerical field within a text line is performed by searching for the best low level label sequence (see figure 6). Once the fields localized, a numerical field recognizer is applied in order to both determine their numerical value and validate the localisation hypotheses. Thus, the implementation requires three processing stages (see [CHP04] for more details):

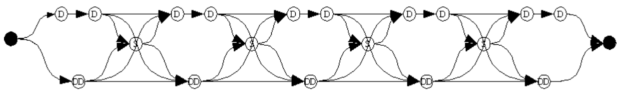


Figure 6. Phone number markovian model for a segmentationless strategy.

Syntactical classes discrimination: the 4-class classifier is built on the same architecture as the digit recognizer of the first approach (combination of two MLP).

Syntactical analysis (numerical field localisation): this stage achieve the fields extraction thanks to the syntactic analysis of the lines of text. The parser corrects the potential classification errors of the previous stage by aligning the recognition hypotheses on the markovian model. This is performed by the Viterbi algorithm [Rab90].

Numerical field recognition: as the field has been localized, the recognition can benefit from the syntactical information provided by the localisation stage. Indeed, the components of the extracted field has been labelled as “Digit”, “Separator” or “Double digit”. Therefore, the recognition stage only has to determine find the numerical values of the “Digit” and “Double Digit” component. Thus, this stage is simply based on the digit recognizer and the descending segmentation of the first approach.

This extraction method is an interesting alternative to the use of a digit segmentation/recognition applied on the whole document, since only the extracted fields are submitted to the recognition stage. The recognition stage is then reduced to a much more constrained recognition problem.

4 Results and comparison

The systems are evaluated on a handwritten incoming mail documents database divided into learning and test databases (293 documents each). These documents contain ZIP codes, phone numbers, and customer codes with labeled position and numerical values[1]. As we propose information extraction systems for handwritten documents, the results are presented by means of the recall/precision trade-off.

As both systems provide the n best alignment paths for each line, a well detected field in ”TOPn” means that the right recognition hypothesis for a field stands in the n best propositions of the syntactical analyser. It is obvious that the larger n, the more the recall increases, and the more the precision decreases. Figure 7 shows the recall-precision trade-off of both approaches for different n values. On can remark that the syntax-directed approach provides significantly better results that the recognition-based approach. Indeed, the first approach reach 57% recall in TOP1 with a 20% precision, whereas the second is beneath 50% recall with a lower precision rate. As these results are multiobjectives, one can say that all the trade-off of the first approach are dominated by those of the second method.

The main reason to explain these results is that the intensive and systematic segmentation of the first approach yields too much propositions with too few contraints. to filter the numerous hypotheses. In particular, the segmentation of textual components often produces entities which shapes are very similar to digits as ’0’ or ’1’. When the global decision has to be taken overall the line, the number of hypotheses is so important that errors often occur. On the contrary, the second approach limits the number of hypotheses thanks to the syntactical classification. Finally, one can say that the digit recognition yield potential errors and does not serve the localization process.

5 Conclusion

In this article, two opposed strategies have been proposed for the extraction of numerical field extraction in handwritten documents. The two strategies have been

[1] As this database contains private information, it is unfortunately unavailable for public diffusion.

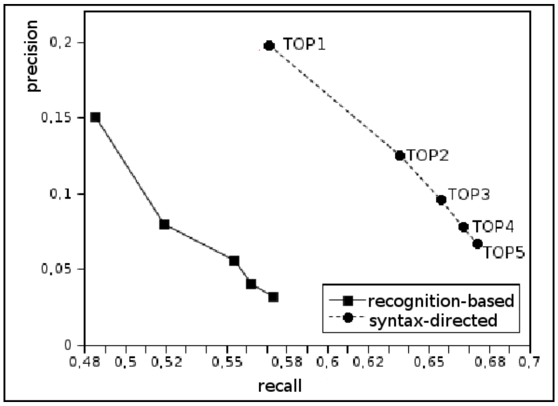


Figure 7. Recall-precision trade-off of both approaches.

implemented, leading to two generic and industrialisable systems. Indeed, the strategies can be applied on any kind of numerical fields provoded that the syntax is constrained enough. They also can be applied on any document written on any language.

In the following of our work, the syntax-directed method will be exploited and improved. The main evolution will be the combination of of this work with the keyword spotting system developped by Koch [Koc06]. Although numerous questions and technical problems are still open (segmentation and recognition of both numerical and textual components, conception of efficient models), this should make the system more reliable.

References

[BSW99] D. M. Bikel, R. L. Schwartz, and R.M. Weischedel. An algorithm that learns what's in a name. *Machine Learning*, 34(1-3):211–231, 1999.

[CHP04] C. Chatelain, L. Heutte, and T. Paquet. A syntax-directed method for numerical field extraction using classifier combination. *IWFHR, Tokyo, Japan*, pages 93–98, 2004.

[CHP06] C. Chatelain, L. Heutte, and T. Paquet. Discrimination between digits and outliers in handwritten documents applied to the extraction of numerical fields. *IWFHR, La baule, France*,page 475-480, 2006.

[EYGB02] A. El-Yacoubi, M. Gilloux, and J.-M. Bertille. A statistical approach for phrase location and recognition within a text line: An application to street name recognition. *IEEE Trans. on PAMI*, 24(2):172–188, 2002.

[FM99] D. Freitag and A.K. McCallum. Information extraction with hmms and shrinkage. *AAAI-99 Workshop on Machine Learning for Information Extraction*, pages 31–36, 1999.

[HBPB$^+$97] L. Heutte, P. Barbosa-Perreira, O. Bougeois, J.V. Moreau, B. Plessis, P. Courtellemont, and Y. Lecourtier. Multi-bank check recognition system : consideration on the numeral amount recognition module. *IJPRAI*, 11:595–618, 1997.

[KG97] G. Kim and V. Govindaraju. *Banckcheck Recognition Using Cross Validation Between Legal and Courtesy Amount*, pages 195–212. Automatic Bank Check Processing,. World Scientific, 1997.

[Koc06] G. Koch. *Catgorisation automatique de documents manuscrits : application aux courriers entrants*. PhD Thesis, Universit de Rouen, 2006.

[Lee97] T. R. Leek. Information extraction using hidden Markov models. Master's thesis, UC San Diego, 1997.

[LSF04] C.L. Liu, H. Sako, and H. Fujisawa. Effects of classifier structures and training regimes on integrated segmentation and recognition of handwritten numeral strings. *IEEE Trans. on PAMI*, 26:1395–1407, 2004.

[MSBS02] M. Morita, R. Sabourin, F. Bortolozzi, and C.Y. Suen. Segmentation and recognition of handwritten dates. *IWFHR*, pages 105–110, 2002.

[MUC98] *Proceedings Message Understanding Conference (DARPA)*. San Francisco, Morgan Kaufmann Publishers, 1991-1998.

[MYHK03] V.K. Madasu, M.H.M. Yusof, M. Hanmandlu, and K. Kubik. Automatic extraction of signatures from bank cheques and other documents. *Biennial Australian Pattern Recognition Conference*, 2:591–600, 2003.

[PS00] R. Plamondon and S.N. Srihari. On-line and off-line handwriting recognition : A comprehensive survey. *IEEE Trans. on PAMI*, 22(1):63–84, 2000.

[Rab90] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Readings in Speech Recognition*, pages 267–296. Kaufmann, 1990.