Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 26 (2005) 1118–1127
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Automatic extraction of numerical sequences in handwritten incoming mail documents

G. Koch, L. Heutte *, T. Paquet

Laboratoire PSI—FRE CNRS 2645, UFR des Sciences, Université de Rouen, F-76821 Mont-Saint-Aignan Cedex, France

Received 10 July 2003; received in revised form 28 September 2004
Available online 18 November 2004

Abstract

In this paper, we propose a method for the automatic extraction of numerical fields in handwritten documents. The approach exploits the known syntactic structure of the numerical field to extract, combined with a set of contextual morphological features to find the best label for each connected component. Applying a Markov model based syntactic analyzer on the overall document allows to localize/extract fields of interest. Reported results on the extraction of zip codes, phone numbers and customer codes from handwritten incoming mail documents demonstrate the interest of the proposed approach.

Keywords: Document analysis; Image processing; Handwritten document; Knowledge extraction; Numerical sequences

1. Introduction

Today, firms are faced with the problem of processing incoming mail documents: mail reception, envelope opening, document type recognition (form, invoice, letter,...), mail object identification (address change, complaint, termination,...), dispatching towards the competent service and finally mail processing. Whereas part of the overall process can be fully automated (envelope opening with specific equipment, mail scanning for easy dispatching, printed form automatic reading), a large amount of handwritten documents cannot yet be automatically processed. Indeed, no system is currently 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, unknown textual content of the document (Lorette, 1999; Plamondon and Srihari, 2000). Nevertheless, it is now possible to consider restricted applications

* Corresponding author. Tel.: +33 2 35 14 68 77; fax: +33 2 35 14 66 18.
E-mail address: Laurent.Heutte@univ-rouen.fr (L. Heutte).

0167-8655/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2004.10.006

of handwritten text processing which may correspond to a real industrial need. The extraction of numerical data (file number, customer reference, phone number, zip code in an address,...) in a handwritten document whose content is expected (incoming mail document) is one particular example of such a realistic problem.

In this paper, we propose a method for the automatic extraction of numerical fields in handwritten incoming mail documents. For this purpose we postulate that the spatial organization of the connected components belonging to a numerical field obeys a specific structure which can be exploited during the extraction phase. Although this hypothesis cannot be considered as a fully realistic one, the proposed method is an interesting alternative to the use of a digit recognizer which would be applied on the whole document. It mainly aims at detecting potential numerical fields. Indeed, numerical fields often contain broken and/or touched digits. Therefore, the recognition of a numerical field needs to implement a segmentation/recognition strategy which would be quite time consuming if applied on the whole handwritten document. Moreover, it would also require to implement postprocessing steps in order to reject non numerical elements. The proposed method will serve as a syntactical filter prior to recognition. Two steps are required for this extraction. The first one is dedicated to the automatic labelling of the connected components. It consists in providing hypothesis on each connected component (hypothesis such as digit, touched digit, separator, word,...). The second step consists in applying a syntactical analyser on each line of text in order to determine the best label sequence using the known syntax of the numerical field we want to detect.

The paper is organized as follows. Section 2 is devoted to the justification and motivation of the proposed method. Section 3 describes the intrinsic and contextual features used to characterize the connected components; results of the connected component labeling module are also given. We present in Section 4 the syntactical analysis stage aimed at extracting specified numerical fields. Experimental results on a real database of handwritten incoming mail documents are provided in Section 5. Finally, some conclusions and future works are drawn in Section 6.

2. Overview of the proposed approach

The proposed approach aims at extracting numerical fields such as zip codes, phone numbers or customer codes, in unconstrained handwritten incoming mail documents. In most applications of handwritten document automatic reading (bank-check processing, form reading, postal address interpretation), domain-specific knowledge is used to localize the field of interest (Nagy, 2000; Plamondon and Srihari, 2000): for example, in bank-check processing, it is often assumed that the courtesy amount field is located in the top right corner of the check image or that special boxes or icons may help locate this field (Impedovo et al., 1997). Unlike these applications in which domain-specific constraints may be applied to perform a reliable localization, the extraction of numerical fields in a full page of handwriting (incoming mail document) is not straightfoward: as one can see in Fig. 1, the fields of interest we are looking for can occur anywhere in the document (heading, body of text,...) or can sometimes be absent.

At first sight, a naive solution to our problem would be to use a common handwriting recognition module in order to recognize all patterns in the document (digits, letters, words) and then to select the only information of interest (digits). However, the use of a recognition module on a full page of handwriting is expensive in computer time and obviously not reliable when dealing with an open vocabulary (Heutte et al., 2004; Kim et al., 1999; Koerich et al., 2003; Vinciarelli et al., 2003). Besides, as the information to be extracted represent only a small part of the document, the complete reading of the document is not necessary. A second approach, probably more realistic, would be to use a single digit recognizer on all the connected components of the document. When looking forward to the consequences of such an approach, it appears that due to the potential presence of connected and/or broken digits, a segmentation driven recognition strategy would be