Measurement 61 (2015) 88–99

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

An automatic document processing system for medical data extraction

Francesco Adamo, Filippo Attivissimo *, Attilio Di Nisio, Maurizio Spadavecchia

Electrical and Electronic Measurements Laboratory, Department of Electrical and Information Engineering (DEI), Politecnico di Bari, Via E. Orabona 4, 70125 Bari, Italy

ARTICLE INFO

Article history:
Received 5 February 2014
Received in revised form 8 October 2014
Accepted 21 October 2014
Available online 29 October 2014

Keywords:
Medical data
Document image processing
Medical services
Performance evaluation

ABSTRACT

This paper illustrates an automatic document processing system for the extraction of data contained in medical laboratory results printed on paper. The final goal of the research is to automate the collection of medical data and to enable an efficient management and dissemination of the information. The following processing steps of the system are described in detail: image preprocessing; layout analysis for the identification of the tables contained in the document; extraction and classification of the laboratory results. Among the many features of the system there are the use of an open source OCR engine, as a basis of further processing, and the storage in XML format of the data retrieved, for ease of sharing. The knowledge base used to guide the data extraction is also explained. The proposed approach has been tested on several document formats and performance analyzed.

1. Introduction

Medical investigation is conducted to extend the life of people and to improve quality of life for patients with severe diseases; today it is common to run into doctors having different specializations which work together towards the common goal of curing a disease. This requires a multidisciplinary research organization utilizing advanced medical technologies and medical research institutes of different universities. Besides, in order to guarantee the statistical significance of the studies, a sufficient amount of data from clinical trials and medical examinations should be collected. Thus, the scientific communities of several medical fields are working on electronic databases containing clinical analyses and laboratory results useful for researches, medical investigations, epidemiological studies, quality control, and so on [1–3]. It should be stressed that an efficient and undistorted communication of medical research results and hospital data is one of the most important heritages of the medical scientific community.

As a matter of fact, however, a complete transition towards paperless practices has not been accomplished or, in some cases, is not possible at all, and paper continues to be used for diagnoses, laboratory results, and prescriptions. This constitutes and obstacle for the creation of electronic databases and electronic medical records [4]. Indeed, it has been noticed that the manual entry of data into medical records takes a long time and often produces errors [5–7]. Several causes of data errors involving human intervention have been individuated in the literature, such as typing errors, calculation errors, incomplete transcriptions, non-adherence to guidelines and non-adherence to data definitions. In some cases, lack of motivation and absence of training may negatively affect the entire set up and organization of electronic medical records. Moreover, the absence of common practices among various medical centers produces discrepancies and nonuniformity of data.

* Corresponding author.
E-mail addresses: adamo@misure.poliba.it (F. Adamo), attivissimo@misure.poliba.it (F. Attivissimo), dinisio@misure.poliba.it (A. Di Nisio), spadavecchia@misure.poliba.it (M. Spadavecchia).

http://dx.doi.org/10.1016/j.measurement.2014.10.032
0263-2241/© 2014 Elsevier Ltd. All rights reserved.

On the contrary, the adoption of automatic systems not only avoids errors in calculations, such as conversions of measurement units and computation of derived quantities, but produces improvements also in the enforcement of guidelines and, more in general, it incentivizes the adoption of clear and unique data definitions. Indeed, any lack of uniformity between data produced by different laboratories, such as differences in naming conventions, measurement units and missing values can be readily put in evidence.

Therefore, the automatic conversion of paper documents into digital resources is an important and nontrivial task that greatly contributes to the preservation and dissemination of medical archives. In this paper an automatic system able to extract the data contained in tabular-like form in printed medical laboratory results, converting them into an electronic form which can be stored in databases and further processed is proposed.

A review of algorithms for the automatic analysis of printed documents is presented in Section 2, followed in Section 3 by a thorough description of the implemented system. Performance is analyzed in Section 4 and final remarks are given in Section 5.

2. Related work

Extracting information from printed medical laboratory results in an automated way is not a simple task, and requires several processing steps [3,8–10]. Medical image archiving and management allows a fast and objective diagnosis even from remote locations [11]. There are many successful applications in this field [12–15]. Document automation systems are available for other purposes, such as generating special printed forms to be compiled by hand and processed by OCR [16]. In [17] a method is proposed for the automatic text classification in biomedical research documents, based on the use of a support vector machine. However, a complete system tailored to laboratory results is not available. In this work several components have been integrated with the aim of building such a system. In order to better understand the features of the proposed method, a short premise should be done about the components of the system and the processing flow.

The main components required for processing the document are: digitization, pre-processing, layout analysis, OCR, correction of the OCR results and document understanding. We consider these ones as components rather than consecutive steps, because they can be used several times with different working parameters in order to advance in the data extraction. For example, in our proposed approach, layout analysis and OCR are iterated two times in order to find column headers in the printed table of laboratory results. A third OCR run is performed with parameters tailored to the contents expected in each table cell, which are predicted on the basis of the column and the row where each cell is located.

Layout analysis allows to retrieve the structure of the document by using, essentially, graphical features such as position, distance, orientation and size of the components being analyzed, which can be connected components, characters, words, text lines, paragraphs, and so on. Layout analysis has its roots in image segmentation algorithms, and is a fundamental step towards document understanding, in which the logical relations between document components are fully exploited. Image segmentation [18,19] and text region extraction [20–22] are one of the most debated issues in the document images analysis [17] and many problems are currently unresolved. Over the last two decades, several techniques have been proposed, all referable to three classes; bottom-up algorithms, top-down algorithms and hybrid algorithms [23]. In the bottom-up approaches text components are identified starting at the character level, then characters are aggregated into words, and finally text lines, paragraphs and higher level components are built to reassemble the whole pages; examples are the use of the Voronoi diagram [24], the Docstrum algorithm [25], the Kruskal algorithm [26] and the probabilistic approach [27]. Alternatively, in the top-down approaches, the pages are split into columns, then into paragraphs and finally in the text lines and words. Examples are the XYcut [28] and whitespace analysis [29]. Finally, hybrid approaches can be regarded as a mix of the above two approaches in an attempt to overcome the limitations of these algorithms. Neural techniques have been applied not only to OCR and word recognition, but also to layout analysis [30]. Due to the importance of layout analysis in document image understanding, considerable effort has been dedicated to the performance evaluation of these algorithms [17,31]. No single algorithm can be considered optimal and different approaches should be chosen depending on the specific application.

In this work layout analysis is dedicated essentially to the analysis of tables, because this is the form in which laboratory results are generally reported. Table detection can be performed by analyzing gaps between words and rows, as described in [32]. However laboratory results often are not tabulated in a strict manner, for example some cells can span more than one column, and no cell box delimiters are printed. For this reason in our system a knowledge base is used, in order to extract table data.

We have explored the possibility of integrating in our system an open source OCR software. Really, any OCR could have been used because there are not special requirements about advanced functions and the layout analysis is performed mainly by our system. The only

Table 1
Row classification according to the cell types it contains.

	Test name cell	Test result cell	Test measurement unit cell	Test reference range cell
Test row	Present	Present	Do not care	Do not care
Result row	Absent	Present	Absent	Absent
Name row	Present	Absent	Do not care	Do not care