MedLab: Medical Labratoary Test Document Analysis Using HoG and SVM

Ali Samei, Alireza Tvakoli Talrghi, Mohammad Mahdi Dehshibi
1ISPR Lab., Department of Computer Science Faculty of Mathematics
2Intelligent Systems and Pattern Recognition Laboratory (ISPR Lab)
Shahid Behshti University
Iran
a.samei@isprlab.com, a_tavakoli@sbu.ac.ir, dehshibi@iranprc.org

ABSTRACT: *This paper presents an automatic document processing system for the extraction of data which are illustrated in medical laboratory results printed on a paper. The final goal of the research is to make the collection of medical data automatic and to enable an efficient management and description of the information in a way that a patient or a senior medicine student can understand the document just like an expert physician. In order to reach the mentioned goals, following the forthcoming steps was necessary in the proposed method. (i) Image pre-preprocessing, (ii) layout analysis for the identification of the tables' data contained in the document' (iii) extraction and classification of the laboratory results using template matching and HoG features in combination with a Support Vector Machine (SVM). Providing information for the application user need to constructing a knowledge base in which the relevant information of Wikipedia is used. The proposed approach has been tested on several document formats and performance analysis shows its superiority as well as simplicity.*

Keywords: Medical Data, Document Image Processing, Medical Services, HoG, Support Vector Machine

Received: 15 June 2015, Revised 18 July 2015, Accepted 21 July 2015

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], [2]. 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. Indeed, it has been noticed that the manual entry of data into medical records takes a long time and often produces errors [3]. 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 inconsistency of data.

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. The main components required for processing the document are: digitization, pre-processing, layout analysis, OCR, correction of the OCR results and document understanding. 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 [4] and text region extraction [5] are one of the most debated issues in the document images analysis [6], [7] 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. 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 [8], the Docstrum algorithm [9], the Kruskal algorithm [10] and the probabilistic approach. 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 and whitespace analysis. 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 [11]. Because of the importance of layout analysis in document image understanding, considerable effort has been dedicated to the performance evaluation of these algorithms. No single algorithm can be considered optimal and different approaches should be chosen depending on the specific application.

In this paper an automatic system is able to extract the data contained in tabular-like form in printed medical laboratory results and to convert them into an electronic form which can be stored in databases and further processed is proposed. The stricter of this paper is as follows: Section II dedicates to describe the proposed method, performance of the proposed method is analyzed in Section III, and final remarks are given in Section IV.

2. Proposed Method

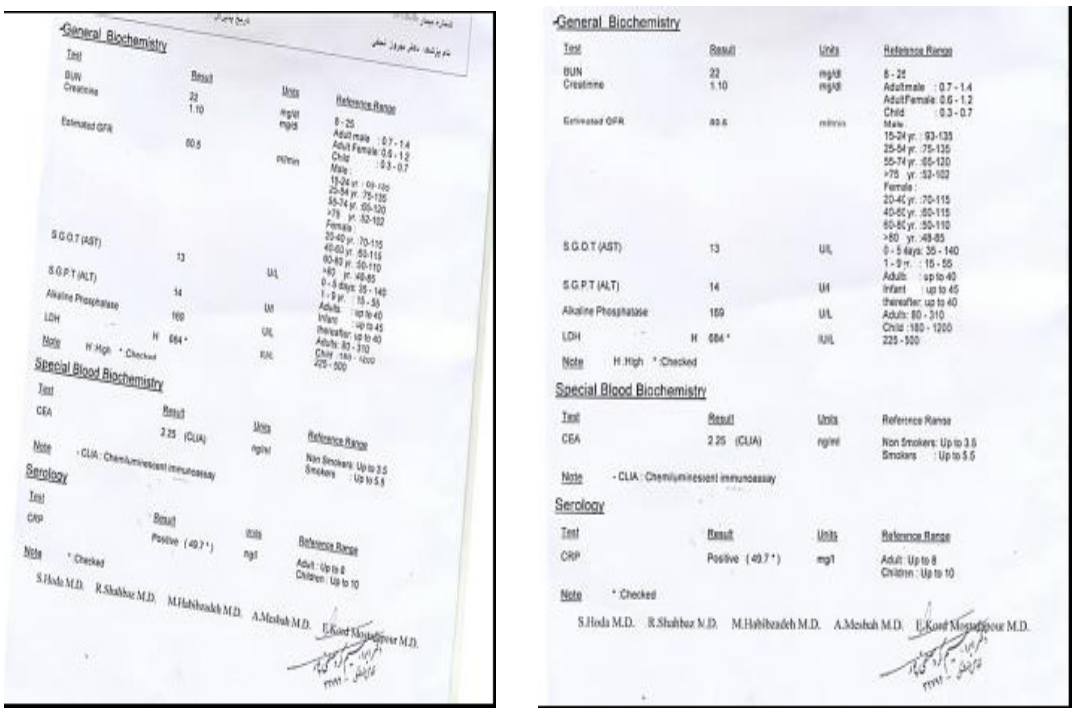
The conversion of paper and electronic documents into standard electronic forms is a key step in medical research. There are many advantages in using standard electronic records such as (1) better utilization of the storage, (2) fast retrieval and transmission, (3) ease of data analysis, and (4) the possibility of comprehensive statistical studies. Unfortunately, medical documents are very different in terms of format, and medical lexicon is large; this leads to difficulties in an automatic conversion of records. In the forthcoming subsections, we will describe the proposed method, which aims at automatically converting paper-based medical reports into an Android document.

2.1 Pre-processing

The first step in document normalization is to detect whether the region of interest (ROI) is skewed. To achieve this, the center point of image is calculated by equation (1).

$$C_x = ROI_{width}/2 \, , C_y = ROI_{height}/2 \qquad (1)$$

Then, two vectors from the right and two vectors from the left vertices of the plate's corner and extracted region's corner are created, respectively. Finally, the left and right angles, θ_{left} and θ_{right}, are calculated, and the ROI is rotated on the greater angle (refer to the equation 2). Figure 1 shows the alignment steps.



(a) (b)

Figure 1. (a) Skewed ROI (b) Aligned document

$$\theta_{left} = \arccos\left(\frac{v^1_{left} \cdot v^2_{left}}{\left\|v^1_{left}\right\|_2 \times \left\|v^2_{left}\right\|_2}\right) \times \frac{180}{\pi}$$

$$\theta_{right} = \arccos\left(\frac{v^1_{right} \cdot v^2_{right}}{\left\|v^1_{right}\right\|_2 \times \left\|v^2_{right}\right\|_2}\right) \times \frac{180}{\pi} \qquad (2)$$

$$\theta = \max\left\{\theta_{left}, \theta_{right}\right\}$$

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \times \begin{bmatrix} x \\ y \end{bmatrix}$$

where θ, is rotation angle, x and y are coordinates of the original image, x' and y' are coordinates of the transformed image, v^1 is the vector which lies across the centroid to vertices of the plate's corner, and v^2 is the vector which lies across the centroid to vertices of the region's corner, $|.|$ is the inner product of two vectors and $||.||_2$ is the norm of a vector.

The final stage of normalization is to reduce the sensitivity of the system to the noise, illumination variations, and image scale. To achieve this, Median filter is applied to the aligned doc. Median filter is a nonlinear operation which is more effective than convolution when the goal is to simultaneously reduce noise and preserve edges. After filtering the ROI, this region is scaled to 600×400 pixels.

2.2 Character Recognition

In a laboratory test document, there are two sort of data, the former is an abbreviation for symptoms and the latter is a numerical indicator which helps physician to judge about the health's condition of subjects. In order to extract these elements, two approaches are taken into account, which will be described.

2.2.1 Symptom Recognition

In this stage, several templates were constructed with different font style and size for each symptom. Providing such data may seem to be wasteful; nevertheless, as the number of symptoms are limited, such data would help increasing the speed of extraction process and would ease the inference stage for a real time smart phone application.

To reach the aim of template matching, the score of cross correlation for each pair of template and the document is calculated and the highest score is considered as the symptom.

The use of cross-correlation for template matching is motivated by the equation (3).

$$\gamma(u,v) = \frac{\sum_{x,y}\left[f(x,y) - \bar{f}_{u,v}\right]\left[t(x-u,y-v) - \bar{t}\right]}{\sqrt{\sum_{x,y}\left[f(x,y) - \bar{f}_{u,v}\right]^2 \sum_{x,y}\left[t(x-u,y-v) - \bar{t}\right]^2}} \qquad (2)$$

where $f(x, y)$ and $t(u, v)$ indicate the image and template, respectively. Moreover, is the mean of the template and $\bar{f}_{u,v}$ is the mean of $f(x, y)$ in the region under the template.

2.2.2 Number Recognition

To this aim, HoG feature descriptor [12] in combination with SVM is used. The essential thought behind the histogram of oriented gradients (HoG) descriptor is that local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions. The image is divided into small connected regions called cells, and for the pixels within each cell, a histogram of gradient directions is compiled. The descriptor is then the concatenation of these histograms. This descriptor has several key advantages over other descriptors. Since it operates on local cells, it is invariant to geometric and photometric transformations, except for object orientation. The HoG descriptor algorithm is as follows:

i. Divide the image into small connected regions called cells, and for each cell compute a histogram of gradient directions or edge orientations for the pixels within the cell.

ii. Discretize each cell into angular bins according to the gradient orientation.

iii. Each cell's pixel contributes weighted gradient to its corresponding angular bin.

iv. Groups of adjacent cells are considered as spatial regions called blocks. The grouping of cells into a block is the basis for grouping and normalization of histograms.

v. Normalized group of histograms represents the block histogram. The set of these block histograms represents the descriptor.

After the generation of descriptors, a recognition based on SVM step is performed. SVM looks for an optimal hyperplane as a decision function in a multi-dimensional space to separate between classes. Given a training set of instance label pairs (x_k, y_k); where $x_k \in R^n$ are the training examples, and $y_k \in \{-1, 1\}$ the class label. The SVM consists in mapping xk into a high dimensional space by the function ψ. After that, it finds a linear separating hyperplane of the form: $w.\ \psi(x) + b = 0$ with the maximal margin in this higher dimensional space. In the case of L_2 soft-margin SVM classifier, the optimization problem is given as follows [13]:

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{k=1}^{m}\xi_k$$
$$\text{s.t.} \qquad (3)$$
$$y_k(w.\psi(x_k) + b) \geq 1 - \xi_k, \xi_k \geq 0, \forall k$$

where $C > 0$ is the penalty parameter of the error term. The solution of this problem is obtained using the Lagrangian theory.

3. Experimental Results

In this Section experimental results are illustrated. The system performance has been analyzed in detail by applying the proposed method to the recognition of printed laboratory tests, and quantitative results are reported. The database used in this experimentation included about 120 definitions, tailored for recognizing four different kinds of documents coming from different laboratories. These documents contained laboratory test prescribed regularly to patients by nephrologists.

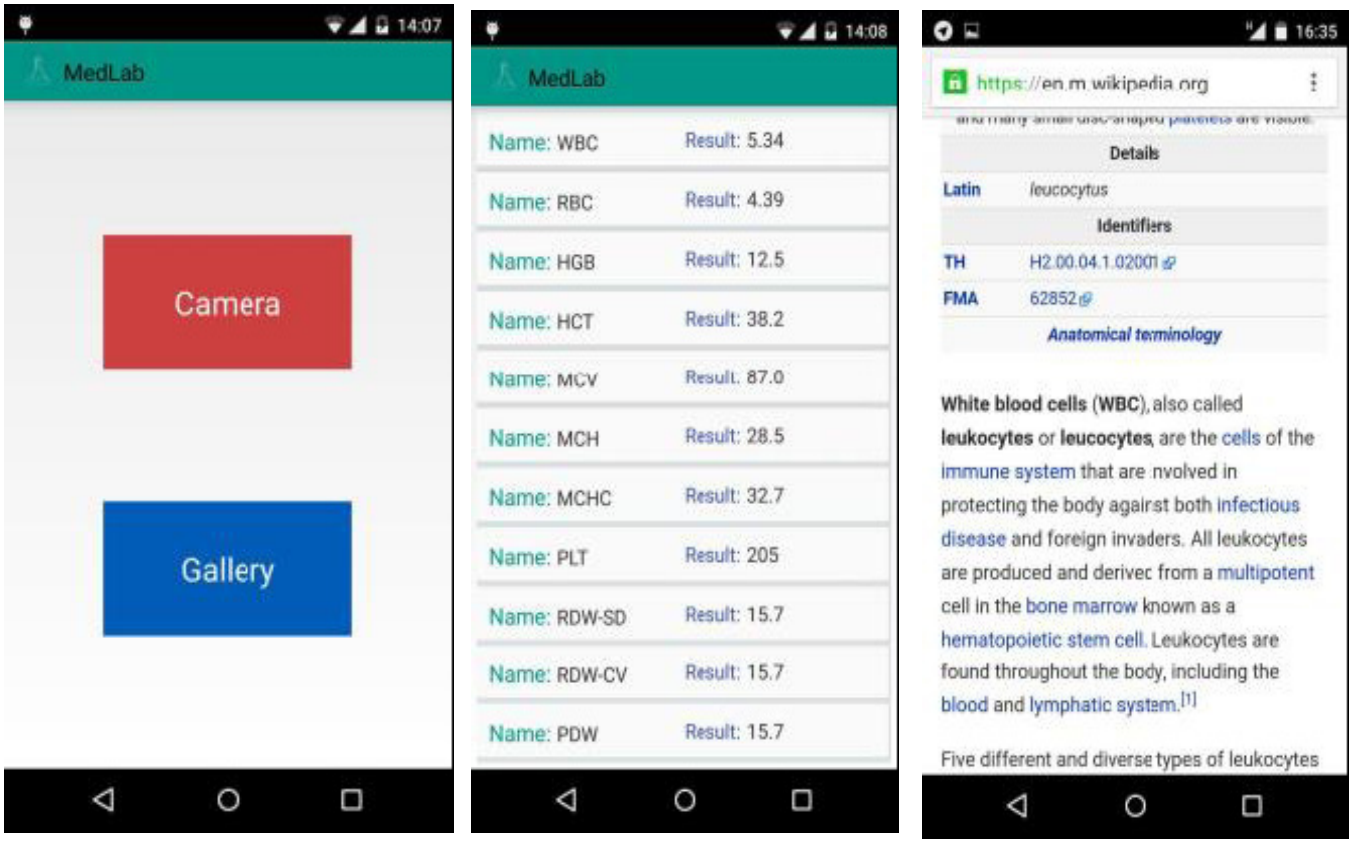
The final evaluation procedure consisted in the following steps:

• Selection of 20 pages of laboratory tests, in English language. Each page, whose size is A4, has been captured by a smart phone camera, giving image files in JPG format.

• Creation of a ground-truth database containing the medical tests present on each page. Each record, which will be referred later as true test, contains the following fields: testId, normalized test name, test result, normalized measurement unit, page identifier pageId. The ground-truth database contained 480 true tests.

• Processing of the images with the proposed system, and creation of output page containing the extracted data, as is illustrated in Fig. 2.

• Comparison between true tests and estimated tests, and automatic creation of a report.



(a) (b) (c)

Figure 2. (a) Application Interface (b) Result of OCR (c) Provided information on Wikipedia in case of tabbing on a symptom

In order to provide a report for the efficiency of the proposed method, a set of experiments is conducted. These experiments involve all sub algorithms which support the final goal, i.e., Character Localization/Segmentation, and OCR. The overall accuracy of the proposed method is calculated as follows:

$$\text{accuracy\%} = (\text{Localization}) \times (\text{Segmentation}) \times (\text{Recognition})^{(\text{number of document's component})} \quad (3)$$

The definition of each event counted by the statistics and their relative frequency of occurrence (obtained dividing by the number of true tests) are reported. The test mismatch error rate, which counts true tests that does not match any estimated test on the same page, is about 5%.

Statistic Name	Relative Frequency	Absolute Frequency	Description
Test matches	95.4%	458	The true test matches an estimated test on the same page
Test mismatches	4.6%	22	The true test does not match any estimated test on the same page
Numeric mismatches	2.1%	10	The true test name matches an estimated test name on the same page, but the result, which is expressed in numeric form, differs.
String mismatches	0.4%	2	The true test name matches an estimated test name on the same page, but the result, which cannot be expressed in numeric form, differs.
Aggregated errors	4.6%	22	Any event that contributes to test mismatches or spurious test names statistics.

Table 1. Performance Evaluation

Test names unrecognized were almost due to excessively bad character recognition, with only one case (0.2%) of wrong segmentation in which the test name and the test result were merged in a single cell. Out of the 10 unrecognized test names, 6 occurred on the same page. It should be noted that the spurious test names statistic is zero, meaning that the system has never interpreted some characters as tests when they are not. This is true also for a page, included in the experiment, which contained only the laboratory template and no tests to extract. Results of the experiments are tabulated in Table I.

4. Conclusion

The manual insertion of laboratory results by the medical or nursing staff is time consuming and produces errors. However, the use of electronic medical records and databases has many benefits, among which the possibility of improving treatments of diseases more readily and accurately, and the availability of clinical data for research purposes. The algorithm presented here, which has been experimented on laboratory results of patients, overcomes the limitations of manual entering and is able to extract and interpret data originated from different laboratories. A first advantage of the proposed method consists in the reduction of time expenditure when creating digital records from printed documents. In general, extra time burdens for using clinical information systems and lack of flexibility have slowed the adoption of electronic medical records. It should be noted that, in the intended use of our system, the documents can be processed automatically in batches whereas human intervention, for checking and eventual correction of the results, can be postponed to a second time.

The second important advantage is that the reported error rate of 5%, which aggregates all kinds of errors encountered in our experimentation, is acceptable if compared with procedures that include manual intervention. Moreover, mobile apps are created to make everyday life easier to be rolled out and we can claim that this aim was reached in our application, especially for non-expert people.

References

[1] Black, N., Tan, S. (2013). Use of national clinical databases for informing and for evaluating health care policies, Publisher, City.

[2] Ramin, M., Ahmadvand, P., Sepas-Moghaddam, A., Dehshibi, M. M. (2012). Counting the Number of Cells in Immunocytochemical Images Using Genetic Algorithm, in: Hybrid Intelligent Systems (HIS), 12th International Conference on, IEEE, 2012, 185-190.

[3] Goldberg, S.I., Niemierko, A., Turchin, A. (2008). Analysis of data errors in clinical research databases, in: AMIA annual symposium proceedings, *American Medical Informatics Association*, 242.

[4] Dehshibi, M.M., Fazlali, M., Shanbehzadeh, J. (2014). Linear principal transformation: toward locating features in N-dimensional image space, Publisher, City.

[5] Yazdani, D., Arabshahi, A., Sepas-Moghaddam, A., Dehshibi, M. M. (2012). A multilevel thresholding method for image segmentation using a novel hybrid intelligent approach, in: Hybrid Intelligent Systems (HIS), 2012 12th International Conference on, IEEE, 2012, 137-142.

[6] Belan, P., Araujo, S., Librantz, A. (2013). Segmentation-free approaches of computer vision for automatic calibration of digital and analog instruments, Publisher, City.

[7] Dehshibi, M. M., Allahverdi, R. (2012). Persian Vehicle License Plate Recognition Using Multiclass Adaboost, Publisher, City.

[8] Kise, K., Sato, A., Iwata, M. (1998). Segmentation of Page Images Using the Area Voronoi Diagram, Publisher, City.

[9] O'Gorman, L. (1993). The document spectrum for page layout analysis, Publisher, City.

[10] Simon, A., Pret, J. C. (1997). A fast algorithm for bottom-up document layout analysis, Publisher, City.

[11] Singh, S. (2013). Optical character recognition techniques: a survey, Publisher, City.

[12] Dalal, N., Triggs, B. (2005). Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, IEEE, 2005, 886-893.

[13] Vapnik, V., Golowich, S. E., Smola, A. (1996). Support vector method for function approximation, regression estimation, and signal processing, in: Advances in Neural Information Processing Systems 9.