Exploratory Data Analysis of Eus Elastography Sample Movies for Cancer Detection

Florin GORUNESCU[1], Elia EL-DARZI[2]

[1] University of Medicine and Pharmacy of Craiova
fgorun@rdslink.ro
[2] Harrow School of Computer Science,
University of Westminster, London, UK
eldarze@westminster.ac.uk

Abstract. Endoscopic ultrasound elastography (EUS elastography) is a recent elasticity imaging technique allowing assesing the difference between malignant and benign tumors. The method characterizes the difference of hardness between diseased tissue and normal tissue. This information can recently be obtained during real-time scanning, the results being shown in a color sample movie. The aim of this paper is to perform an exploratory data analysis of the digitalized sample movie, in order to provide an optimum input for a subsequent application of the intelligent systems in the computer-aided diagnosis process. The technique has been applied in a concrete case of malignant and benign patients, with a high performance using neural networks as clasifier.
Keywords: noninvasive cancer detection, endoscopic ultrasound elastography, digitalized sample movie, exploratory data analysis
Math. Subjects Classification 2000: 62-07; 68T05

1 Introduction

Palpation of the body is the classical method used by physicians to detect the presence of abnormalities that might indicate pathological lesions, usually because the mechanical properties of diseased tissue are typically different from those of the normal tissue that surrounds it. Recently developed methods of management in cancer diseases to replace palpation include a routine use of biopsy of the affected organ. However, biopsy is an invasive method, with inherent complications that may cause even the death of the patient. Consequently, the use of non-invasive alternatives is highly necessary, moreover since competitive computational technologies, which can be successfully employed towards this purpose, are currently available.

Because the mechanical properties of normal and diseased tissues are of pathological relevance, the development of a direct measure of tissue elasticity would be very important for the characterization of lesions, in addition to the information already obtained by conventional imaging methods. Endoscopic ultrasonographic elastography (EUSE) is a recent elasticity imaging technique

that reveals directly the physical properties of tissues. The method characterizes the difference of hardness between diseased tissue and normal tissue. This information can recently be obtained during real-time scanning, the results being shown in color superimposed on the conventional grey-scale image. Therefore, in the images resulted after scanning, colors express the difference of elasticity between healthy and diseased tissue.

An efficient noninvasive diagnosis procedure, based on machine learning techniques, needs an accurate digitalized correspondent of the EUSE output, represented by a sample movie.

In order to assess the issue regarding the way in which the sample movies are interpreted and that finally leads to major consequences as concerns the given diagnosis, we propose the employment of exploratory data analysis techniques in order to preprocess and transpose the corresponding digitalized sample movies into a vector (string) pattern that will be used in the subsequent machine learning process.

2 Exploratory data analysis

Exploratory Data Analysis (EDA) is closely related to the concept of Data Mining (DM). DM is becoming increasingly popular as a business information management tool, where it is expected to reveal knowledge structures that can guide decisions in conditions of limited certainty. An important difference in the focus and purpose between DM and EDA is that DM is more oriented towards applications than the basic nature of the underlying phenomena.

As opposed to traditional hypothesis testing designed to verify *a priori* hypotheses about relations between variables, EDA is used to identify systematic relations between variables when there are no (or not complete) *a priori* expectations as to the nature of those relations.

Computational EDA methods include both simple basic statistics and more advanced multivariate exploratory techniques, designed to identify patterns in multivariate data sets. A large selection of powerful exploratory data analytic techniques is also offered by graphical data visualization methods that can identify relations, trends, outliers and biases "hidden" in unstructured data sets.

Perhaps the most common and historically first widely used technique explicitly identified as graphical EDA is *brushing*, an interactive method allowing the (on-screen) selection of specific data points or subsets of data and the identification of their characteristics, or of their effects on relations between relevant variables. One of many applications of the brushing technique is to select (highlight) in a matrix scatterplot all data points that belong to a certain category (e.g. the black and white frame within the framework of a color sample movie).

3 Transforming EUSE sample movies into numerical database

EUSE was recently reported to offer supplemental information that appears to obtain a better characterization of tissue and that might enhance conventional EUS imaging [1]. It was developed to analyze structures in real time, with the information being represented in transparent color sample movies. Basically, the EUSE is performed during the EUS examinations with one sample movie of 25 seconds, recorded on the hard disk drive embedded in the EUS sono-elastography module, used in conjunction with a linear endoscope, in order to minimize the variability and to increase the reliability of the image acquisition. Each acquired EUSE sample movie is subjected to a computer-enhanced dynamic analysis using a public domain Java-based image processing tool (developed at the National Institutes of Health, Bethesda, Maryland [2]). Different elasticity values are marked with different (hue) colors (on a scale of 1 to 255) and the EUSE information is shown as color sample movie. Technically, a EUSE sample movie (dynamic image) consists in a sequence of 125 frames (static images). The system uses by default a rainbow color-coded map red-green-blue (RGB), where hard tissue areas are marked with dark blue, medium hard tissue areas with cyan, intermediate tissue areas with green, medium soft tissue areas with yellow, and soft tissue areas with red. Moreover, the system provides the corresponding color histograms of each frame of the sample movie. Recall that in computer graphics and photography a (hue) color histogram is a representation of the distribution of colors in an image, derived by counting the number of pixels of each of given set of color ranges in a typically two-dimensional (2D) color space. We illustrate a sample movie frame and its corresponding (hue) histogram in the case of a malignant tumor in Figure 1.

Fig. 1 EUSE sample movie frame and corresponding (hue) histogram

In order to apply the machine learning methodology to differentiate between the sample movies, characterizing either benign or malignant tumors, we firstly need to digitalize them (Java-based image processing technique). Since the corresponding EUSE sample movie (dynamic image) consists in a sequence of 125 frames (static images) displaying 255 colors, then, from mathematical point of view, to each sample movie corresponds a 125×255 matrix (a_{ij}), each row representing a certain frame of the sample movie and each column representing a pixel color.

In previous pilot studies [3], [4], [5], we tried to assess the EUSE characteristics of tumors using the dynamic image analysis performed by physicians, enhanced by a statistical analysis of the (hue) histogram corresponding to the (near) best image of each sample movie.

4 Exploratory data analysis of the digitalized sample movies

Summarizing, to each patient corresponds a 25 seconds EUSE sample movie, converted into a 125×255 matrix, representing the distribution of colors, corresponding to the tumor elasticity. By analyzing the distribution of colors, it is possible to detect between benign and malignant tumors. So far, the EUSE image analysis has been performed using the human perception of color hues (enhanced by statistical tools) only. Computer-enhanced dynamic analysis of the sample movies is the objective of our current study, with a consequent classification of the benign and malignant tumors based on the digitalized image analysis performed using the machine learning methodology.

Firstly, since the EUSE database consists in raw data, it is compulsory to previously perform an exploratory data analysis in order to detect outliers and unusual patterns. Unfortunately, due to technical features of the image acquisition, most of the EUSE sample movies contain a certain number of black and white frames, which represent outliers of the database, making difficult the decision process. Such frames are recognized by using a brushing technique (both visualization and analytical approaches) to highlight in the matrix plot all data points that belong to this category. Concretely, such a black and white frame is characterized by a very few number of different values, since the frame does not contain color hues, but a grey scale only. Accordingly, the corresponding histogram contains extreme values (outliers) equaling either zero or very large values (for hues of black and white). Afterwards, these frames (rows in the corresponding matrix) will be deleted. Next, since the color frames contain marginal colors hues of the RGB spectrum, making difficult the postprocessing stage, by a brushing method of filtering the extreme hues the corresponding matrix will contain only the main information.

Secondly, since the natural input of the intelligent systems is represented by vectors (strings), we need a way of summarizing the main characteristics embedded in the matrix corresponding to EUSE sample movie in a vector

pattern. Since a_{ij} represents the frequency of the (hue) color j in the i-th frame, then $a_j = \frac{1}{125}\sum_{i=1}^{125} a_{ij}$ represents the mean frequency of the (hue) color j in the sample movie. Consequently, the vector $(a_1, a_2, ..., a_{255})$ -in the general case of all colors hues presented- represents an average (hue) histogram summarizing the information provided by a EUSE sample movie.

In Figure 2 and Figure 3 the (hue) histograms of both black and white and preprocessed color frames are displayed.

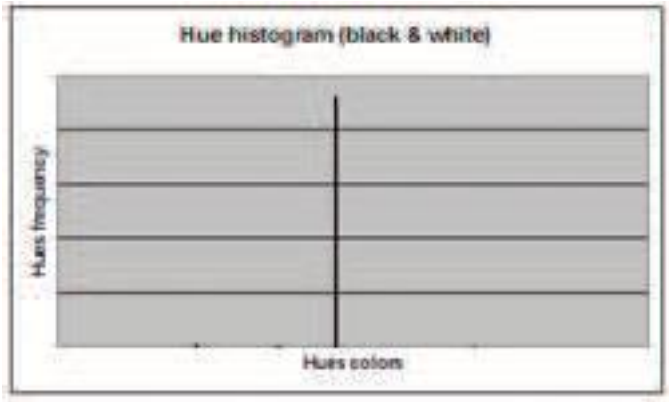

Fig. 2 The (hue) histogram corresponding to a black and white frame

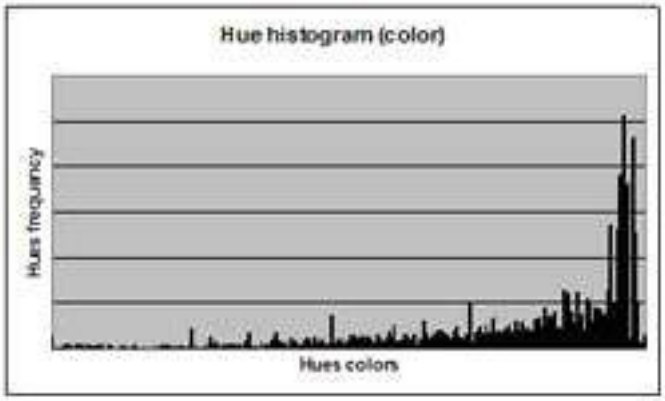

Fig. 3 The (hue) histogram corresponding to a color frame

5 Material and results

A number of 66 individuals were examined by EUSE during a 6-month period at the Department of Gastrointestinal Surgery, Gentofte University Hospital, Hellerup, Denmark and the Endoscopy Laboratory Department of Gastroenterology, University of Medicine and Pharmacy of Craiova, Romania. Cancer patients with various primary tumor locations staged by EUSE were also included in the group. The medical examination has been performed by two experienced EUS examiners in a typical clinical setting with previous knowledge of the patient's underlying disease.

EUS, EUS-FNA and EUSE of the tumors were performed during the same EUS examination with a Hitachi 8500 US system with an embedded SonoElastography module (Hitachi Medical Systems Europe Holding AG, Zug, Switzerland), used in conjunction with a EG 3830 Pentax linear endoscope (Pentax, Hamburg, Germany).

Once the numerical database constructed using the above EDA techniques, it can be processed by machine learning techniques in order to obtain an optimum decision concerning the tumor types.

Basically, departing from the numerical database containing the average digitalized form of scanned images of different tumoral tissues, together with the corresponding diagnosis that was established, without any doubt by doctors, the intelligent systems are trained to learn to associate a certain color pattern to the corresponding diagnosis (benign/malignant).

Concretely, in a neural network approach, an accuracy ranging between 96% and 99.99% has been obtained, proving the efficiency of both this image exploratory/processing technique and the neural network classification power.

6 Discussion and conclusions

The EUSE was recently reported to offer supplemental information that appears to obtain a better characterization of tissue and that might enhance conventional elasticity imaging procedures. The methodology we have developed, based on processing and analyzing images, enables exploration by Artificial Intelligence means of the EUSE digitalized sample movies, in order to obtain an optimal prediction of cancer, by using a noninvasive methodology.

Accordingly, the purpose of this paper is to demonstrate the suitability and ability of EDA techniques in computer-aided diagnosis problems, seen as a reliable tool of the dynamic tumors images analysis.

Acknowledgments. The EUSE sample movies were obtained from the Department of Gastrointestinal Surgery, Gentofte University Hospital, Hellerup, Denmark and the Endoscopy Laboratory Department of Gastroenterology, University of Medicine and Pharmacy of Craiova, Romania. Thanks go to Prof. Peter Vilmann, Denmark, and Assoc. Prof. Adrian Saftoiu, Romania, for providing the data and the necessary technical support.

References

[1] **M. Giovannini, L. Hookey, E. Bories et al.**:- Endoscopic ultrasound elastography: the first step towards virtual biopsy? Preliminary results in 49 patients. Endoscopy, 38, 2006, 344-348.

[2] **W. Rasband**:- ImageJ: image processing and analysis in JAVA, National Institutes of Health (available from: http://rsb.info.nih.gov/ij/)

[3] **A. Saftoiu, P. Vilmann, H. Hassan, F. Gorunescu**:- Analysis of endoscopic ultrasound elastography used for characterization and differentiation of benign and malignant lymph nodes, Ultraschall in der Medizin (European Journal of Ultrasound), 27(6), 2007, 535-542

[4] **A. Saftoiu, C. Popescu, S. Cazacu, D. Dumitrescu, C. V. Georgescu, M. Popescu, T. Ciurea, F. Gorunescu**:- Power Doppler Endoscopic Ultrasound for the Differential Diagnosis between Pancreatic Cancer and Pseudotumoral Chronic Pancreatitis, Journal of Ultrasound in Medicine, 25(3), 2006, 363-372

[5] **A. Saftoiu, P. Vilmann, T. Ciurea, G. L. Popescu, A. Iordache, H. Hassan, F. Gorunescu, S. Iordache**:- Dynamic analysis of endoscopic ultrasound (EUS) elastography used for the differentiation of benign and malignant lymph nodes, Gastrointestinal Endoscopy, 66(2), 2007, 291-300