COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 113 (2014) 284–289

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Raw data extraction from electrocardiograms with Portable Document Format

*Nuria Ortigosa**, *Vicente M. Giménez*

I.U. Matemática Pura y aplicada, Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, Spain

ARTICLE INFO

Article history:
Received 25 June 2013
Received in revised form
2 September 2013
Accepted 19 September 2013

Keywords:
ECG storage formats
Portable Document Format
Scalable Vector Graphics

ABSTRACT

During the last two decades there has been a thorough research and development of standards and protocols in order to cope with different electrocardiogram formats from heterogeneous acquisition systems. Despite the efforts of public and private consortiums on creating a standardized electrocardiogram (ECG) storage format, there is still not a single one. Indeed, there is also the necessity of access to raw data of the ECGs previously acquired. Most of these documents have been saved as Adobe PDF files, since for medical staff it is an easy format for later visualization. However, this format presents difficulties when trying to access original raw data for subsequent studies and signal analysis. In this manner, this paper presents an application that obtains plain numerical data from ECG files stored with PDF format. Data can also be exported to one of the most common file formats in existence, to be easily accessed thereafter.

1. Introduction

The cost of an electrocardiograph is low, specially when compared to other medical equipment (such as, for example, the radiological one). This is probably why different ECG recording companies have been working with such a large number of different data formats. During the last 20 years, the definition and adoption of a single standardized format for ECG recordings has been promoted for public and private consortiums [1,2] due to the fact that interoperability between them could save around 77.8 billion of dollars per year [3], just in the United States.

Therefore, a European funded project called OpenECG [4] started in 2002 to promote the adoption of the SCP-ECG ("Standard Communications Protocol for computed assisted ECG") by implementing visualization tools and interoperability standards to help manufacturers on their implementations [5].

Similarly, in 2001 the United States Food and Drug Administration (FDA) requested a standard which eased the storage and exchange of ECGs information, since it received a large number of annotated ECGs collected in a wide variety of formats. Thus, the HL7 (a not-for-profit international organization for sharing health information) developed the Health Level 7-annotated ECG (HL7-aECG) standard [6], a new XML-based format.

Another format to store medical data is DICOM [7], which initially was developed for medical image storage and, in 1995, finally became a European standard also used for cardiac and vascular information.

Surprisingly, efforts towards a single standardized ECG format are not only supported by the Standard Development Organizations. In 2003, Philips Medical Systems published the XML (extensible markup language) schema that they used for their entire line of ECG products [8], and began to deliver this information to its costumers and to European OpenECG project members [9,10].

Nevertheless, although interoperability farther than at regional or national level is a desirable goal, it may still take another 20 years to be fully achieved [11]. In this sense,

* *Corresponding author*. Tel.: +34 963877000x88398.
E-mail address: nuorar@upvnet.upv.es (N. Ortigosa).
0169-2607/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2013.09.014

integrated applications for standardized data formats exchange have been developed to cope with the numerous different formats and to facilitate their visualization [12]. For example, van Ettinger et al. [13] have implemented a conversion library and an ECG viewer to work with HL7-aECG, SCP-ECG and DICOM files. In order to facilitate interoperability for records obtained in Italy, Marcheschi et al. [14] designed a network infrastructure to manage with different standards. Similarly, [15,16] use XML language as a central platform to exchange data between these formats. Moreover, in [17] Trigo et al. have recently developed a modular application to exchange ECGs information of different data formats across healthcare information systems.

Unfortunately, ECGs previously stored often present data formats different from those commented above. In particular, many electrophysiologists and medical staff usually save only the plots of the ECG recordings as Adobe PDF ("Portable Document Format") due to the fact that, when ECG acquisition equipment is replaced in hospitals, ECG recordings stored with proprietary formats hamper the free access to previous ECGs in the medical history. Therefore, PDF format assures them an easy visualization in any computer and allows subsequent revision for clinical studies and patients evolution analysis. Thus, in this paper we present an application developed to recover raw data from ECG recordings saved in PDF format, in order to facilitate later signal processing of ECGs previously stored by medical staff in this format, and to integrate them with the existing information system of the hospital.

The rest of the paper is organized as follows. Features of ECGs stored with PDF format, a description of their structure, and the application interface are described in Section 2. Section 3 reports the application computational requirements, some particularities and discussion of results. Finally, conclusions and future work are drawn in Section 4.

2. Materials and methods

2.1. SVG files

The ECGs recorded with PDF format whose raw data we wanted to extract were acquired using the Philips PageWriter TC50 [18]. When revising the technical sheets of this electrocardiograph we noticed that, when Philips Medical Systems began to develop their ECG data format based on XML, they turned to the Scalable Vector Graphics (SVG) application language [9], which is ideal to display easily two-dimensional graphics.

SVG [19] is a markup language which is able to formalize a set of graphical elements such as rectangles, lines and polygons. The most relevant feature of SVG documents is that they can be considered simultaneously as both text and images. This ambivalence points out to SVG as an ideal format to connect textual and graphical information in just one file [20].

Therefore, we first tried to convert electrocardiograms to SVG data format, to check whether the Philips electrocardiograph stored PDF files as vector graphics. In order to do so, we used the free software tool Inkscape for conversion [21]. We confirmed that the converted SVG file was formed by a set of graphical elements located with their absolute coordinates, and not by a bitmap data set. As a result, we decided to extract raw data by implementing an application whose input were a PDF file and the generated output contained the corresponding true-value numerical data of each lead, in order to make easier the subsequent signal analysis.

2.2. Electrocardiogram leads

An electrocardiogram is the main instrument for the diagnosis of cardiovascular diseases. It can be defined as the graphical representation of the electrical activity of the heart.

In electrocardiography, the term "lead" refers to the measurement of voltage between two electrodes, which are placed on the patient's body. To perform a standard 12-lead ECG, it takes 10 electrodes: 6 electrodes for the chest leads (V1, V2, V3, V4, V5, V6), and 4 electrodes to acquire the limb leads (I, II, III) and the augmented limb leads (aVL, aVF, aVR).

Depending on the duration of the leads that the electrophysiologist prefers to visualize, ECGs can be stored using different printout formats. Some of the most popular printout formats are: 3×4, $3 \times 4+1$, and 6×2. Details of these printout lead organization will be provided in Fig. 2 and Section 3.

2.3. Electrocardiogram structure

As aforementioned, the SVG file obtained by Inkscape from ECG PDF conversion presents a structure similar to an XML file. The horizontal and vertical coordinates (in pixels) of each sample from the ECG lead is represented under a label called ⟨*path*⟩. Coordinates can be defined as absolute or relative to the last coordinate, depending on whether after the label ⟨*path*⟩ the token is d = "M or d = "m, respectively. In our case, we have configured Inkscape options to generate SVG files with absolute coordinates, being the abscissa and the ordinate coordinates separated by a comma.

We must remark that, under the label ⟨*path*⟩ there appear not only the polylines which correspond to the ECG leads, but also the lines that define the background grid and the reference pulses. They can be differentiated by their corresponding line widths and colours, as well as by their order of appearance along the SVG file and the number of samples for which they are defined.

2.4. Application interface

The presented application has been programmed using GUIDE, the MATLAB graphical user interface development environment. Fig. 1 shows the main window of the ECG data extraction application.

The interface is divided in three subsections. In the left top half the input ECG parameters are shown, such as the paper speed and the amplitude scale of the limb and the chest leads. It is standard to represent each microVolt (μV) of amplitude as 10 mm, and each second as 25 mm. However, as faster paper speeds and different scales can be used, these parameters can be modified by the user.

Then, the user must proceed to load the desired PDF file whose raw data is going to be extracted. Once it has been selected, the application calls the Inkscape program as a background task, in order to perform the PDF conversion to SVG