Catarina Alexandra da Quinta Silva Cavaco

Licenciada em Ciências de Engenharia Biomédica

New Visualization Model for Large Scale Biosignals Analysis

Dissertação para obtenção do Grau de Mestre em
Engenharia Biomédica

Orientador : Doutor Hugo Filipe Silveira Gamboa,
Prof. Auxiliar, Faculdade de Ciências e Tecnologias - Universidade Nova de Lisboa

Co-orientador : Doutor Ricardo da Costa Branco Ribeiro Matias,
Prof. Adjunto, Escola Superior de Saúde - Instituto Politécnico de Setúbal

	Júri:
Presidente:	Doutora Carla Maria Quintão Pereira
Arguente:	Doutor Vasco Miguel Moreira do Amaral
Vogais:	Doutor Hugo Filipe Silveira Gamboa Doutor Ricardo da Costa Branco Ribeiro Matias

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Setembro, 2014

New Visualization Model for Large Scale Biosignals Analysis

To my parents and grandparents

Acknowledgements

This chapter in my life would not have been possible without all the emotional and scientific support of many people who have accompanied my route through these last years.

First, I would like to demonstrate my sincere gratitude to my advisor, Professor Hugo Gamboa, for the opportunity he gave me. I am very grateful for his encouragement, guidance and for the opportunity of working with his research team. I am also very thankful to my co-adviser, Professor Ricardo Matias, for all the knowledge and enthusiasm which allowed me to achieve my goals.

The opportunity to participate in this project helped me growing and it is rewarding to know that I could contribute for the research in such an important area.

I would like to thank all the PLUX - Wireless Biosignals, S. A. team members for sharing the last months with me. Learning in such a dynamic and creative environment definitely motivated me and contributed to the achievements of this work. A special thanks to Ângela Pimentel for her support. I also would like to express my gratitude to Ricardo Gomes for his constant support, dedication, patience and interest which resulted in a great contribution to make this work move forward.

To my university colleagues Maria Inês Silva, Ricardo Eleutério, Pedro Ferreira, Ana Carolina Pádua, Inês Vale and David Branha, I want to thank you for sharing these years with me. Finally, a special thanks goes to my colleagues Marta Santos e Ana Luísa Gomes for all the support, motivation and knowledge exchanged over these last few months.

I would also like to show my gratitude to my friends Marta, Andreia, Ricardo, Raquel, Lara e Sara. Be assured that the meaning and space you fill in my life is quite bigger than this single paragraph. To Inês Oliveira and Daniela Azul I also thank for your support during this work.

Last but not least, I am very thankful to my whole family, specially to my parents and grandparents. To my mum who continuously supported, believed in me and pulled me up when I needed most, I will always be grateful. To Pedro who always stood by me with love and friendship giving me his unconditional support I thank you also.

Having the data is not enough. I have to show it in ways people both enjoy and understand.

Hans Rosling

Abstract

Benefits of long-term monitoring have drawn considerable attention in healthcare. Since the acquired data provides an important source of information to clinicians and researchers, the choice for long-term monitoring studies has become frequent.

However, long-term monitoring can result in massive datasets, which makes the analysis of the acquired biosignals a challenge. In this case, visualization, which is a key point in signal analysis, presents several limitations and the annotations handling in which some machine learning algorithms depend on, turn out to be a complex task.

In order to overcome these problems a novel web-based application for biosignals visualization and annotation in a fast and user friendly way was developed. This was possible through the study and implementation of a visualization model. The main process of this model, the visualization process, comprised the constitution of the domain problem, the abstraction design, the development of a multilevel visualization and the study and choice of the visualization techniques that better communicate the information carried by the data. In a second process, the visual encoding variables were the study target. Finally, the improved interaction exploration techniques were implemented where the annotation handling stands out.

Three case studies are presented and discussed and a usability study supports the reliability of the implemented work.

Keywords: Long-term Biosignals, Big Data, Biosignals Visualization, Biosignals Annotation, Medical Monitoring.

Resumo

Os benefícios da monitorização de longa duração têm recebido uma atenção considerável na área da saúde. Uma vez que os dados recolhidos constituem uma importante fonte de informação para médicos e investigadores, a escolha deste tipo de estudos tem-se tornado frequente.

No entanto, este tipo de monitorização pode resultar em conjuntos de dados de grandes dimensões o que torna num desafio a análise dos biosinais adquiridos. Neste caso, a visualização que é um ponto-chave na análise de sinais, apresenta muitas limitações e a manipulação de anotações da qual dependem alguns algoritmos de *machine learning*, torna-se uma tarefa complexa.

Por forma a superar estes problemas uma inovadora aplicação baseada nas tecnologias *Web* para a visualização e anotação de biosinais de um modo rápido e *user friendly* foi desenvolvida. Tal foi possível através do estudo e implementação de um modelo de visualização. O principal processo deste modelo, o processo de visualização, compreendeu a determinação do domínio do problema, o desenho da abstração utilizada, o desenvolvimento de uma visualização em multiníveis, e o estudo e escolha das técnicas de visualização que melhor transmitem a informação transportada pelos dados. Num segundo processo, a escolha das variáveis de codificação visual foi alvo de estudo. Finalmente, as melhores técnicas de exploração interativa foram implementadas das quais se destaca a manipulação de anotações.

Neste trabalho são ainda apresentados e discutidos três estudos efetuados com a aplicação desenvolvida. Por fim, é apresentado um estudo de usabilidade, o qual sustenta a confiança no trabalho implementado.

Palavras-chave: Biosinais de longa duração, *Big Data*, Visualização de Biosinais, Anotação de Biosinais, Monitorização Médica.

Contents

List of Figures

List of Tables

Listings

Acronyms

ACC Accelerometry

API Application Programming Interface

CSS Cascading Style Sheets

D3 Data-Driven Documents

DOM Document Object Model

ECG Electrocardiography

EDA Electrodermal Activity

EEG Electroencephalography

EMG Electromyography

HDF5 Hierarchical Data Format 5

HRV Heart Rate Variability

HTML Hypertext Markup Language

JSON JavaScript Object Notation

OCD Obsessive-Compulsive Disorder

OOD Object-Oriented Design

OpenGL Open Graphics Library

RESP Respiratory

SAX Symbolic Aggregate Approximation

SMA Signal Magnitude Area

SUS System Usability Scale

SVG Scalable Vector Graphics

WebGL Web Graphics Library

1

Introduction

1.1 Motivation

In the last years, the technological innovation and the interest in quality of life have been motivating the research and development of novel resources in the clinical field. However, it is still evident the demand for the improvement in diagnosis and monitoring of some pathologies. This is the case, for example, of different neurological disorders, such as Obsessive-Compulsive Disorder (OCD) and Parkinson disease.

Therefore, the study, development and implementation of new solutions to monitor healthy and pathological conditions became increasingly imperative in biomedical research.

The data acquired will be an important source of information to clinicians, who will be able to make their clinical diagnosis and treatment plans relying on the extracted information. It also supports researchers decision-making, as well as, inform patients, helping in their recovery. For this reason, the analysis of the acquired biosignals has a significant part either in research or clinic applications. To perform this analysis a variety of concepts and methods are involved, from which data visualization and machine learning algorithms stand out in this work.

Data visualization consists in the integration of the human visual perception with the processing power of computers. Its basic idea is examining, understanding and transmitting information. Since humans main input sense is visual, data visualization is considered essential and highly promising for signal analysis [32], [39]. However, data visualization presents some limitations such as the display capacity and the human perceptual and cognitive capacity [47].

Machine learning is the application of algorithms to computers in order to optimize a performance criterion. These algorithms can follow supervised, semi-supervised and unsupervised approaches [4]. However, the first two depend on the use of labelled data what implies that this task has to be the most accurately possible. Nevertheless, the annotation of the biosignals is a demanding task that depends on human subjective intervention and requires specific knowledge.

In some clinical cases, such as neuromuscular and sleep disorders, the required monitoring of the patients results in large amounts of data. Although the analysis of biosignals is a key feature in many medical applications, in these cases difficulties become evident. This is due to the fact that we are dealing with big data - massive data sets that present problems when processed and analysed by standard tools [68]. In other words, both difficulties previously described become more complex and the data processing will present several limitations [30], [50].

As a result, there is a high demand for novel visual exploration tools that achieve the better conjugation between processing and storage capabilities of computers and the visual, creative and knowledge capabilities of humans.

Considering the current needs described, this work introduces an innovating solution that allows the visualization and annotation of large scale biosignals. In order to overcome the problems introduced by the demand of big data acquisition, this work aimed to present a novel biosignals analysis platform which provides a useful experience to its users through an innovative inspection of the data. Consequently, the solution presented represents an effective biosignals visualization and interaction approach that can be adapted to any type of biosignal.

In this work, the signal acquisition was done with the OpenSignals software [24]. The signal processing algorithms and the visualization code were respectively developed in Python [8] and JavaScript [17] Programming Languages. The layout of the platform was achieved with the use of Hypertext Markup Language (HTML)5 [16] and Cascading Style Sheets (CSS)3 [16].

This dissertation was developed at PLUX - wireless biosignals S.A.[1].

1.2 State of the Art

The emergence of a variety of visualization tools, during the last years, has aimed to counter the gap verified between our ability to collect and store data and to process it. However, engineers and technicians who work with time series data daily (even in the medical field) face the difficulty of work with tools hard to learn, implement and manipulate [2].

[1] http://www.plux.info/

1.2.1 Visualization Techniques

Over 200 years later from the invention of the first and most common visualization technique, the line graph, by William Playfair [33], the development of new and better visualization techniques continues to face the challenges of our days. The large number of well-known data visualization techniques existing today, despite useful for simple data and basic analysis tasks, present huge limitations in respect to analysis of large scale data sets.

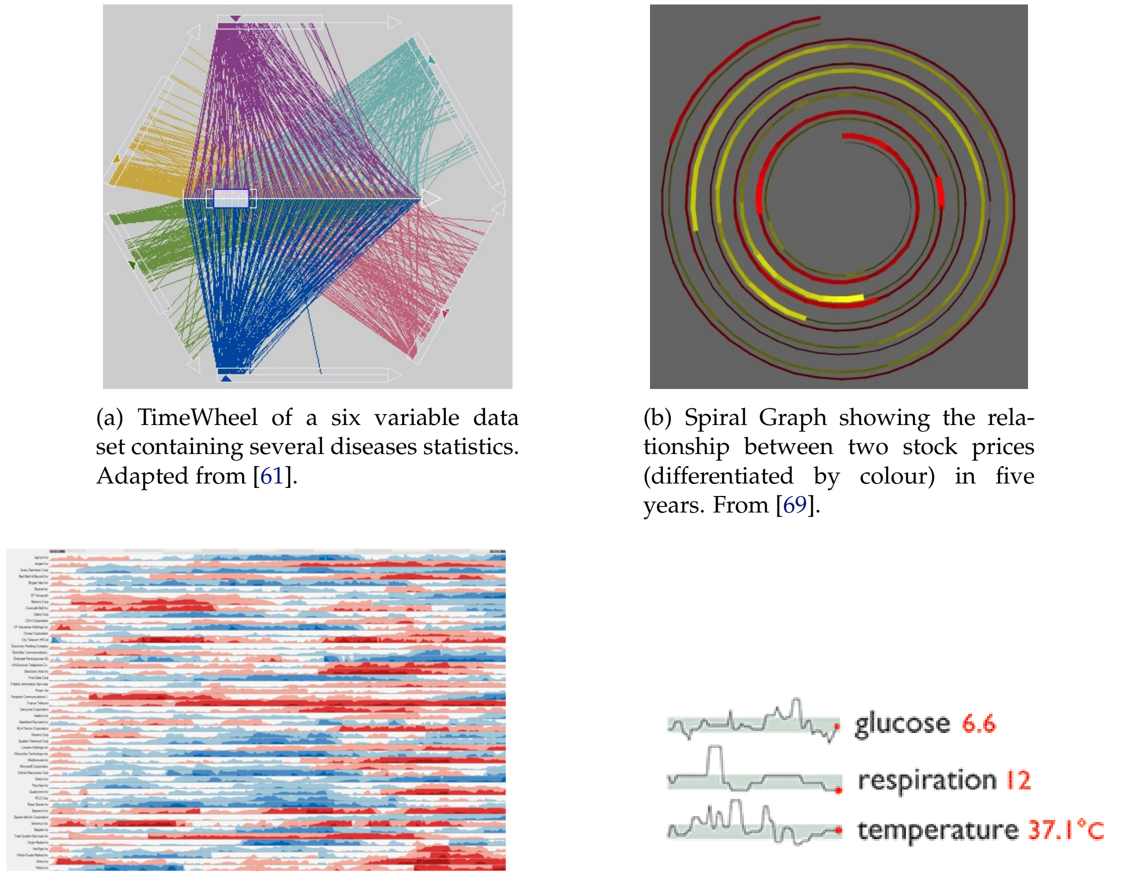
Besides the standard techniques described in [69], more sophisticated visualization techniques can be found, which may be combined with the aim of developing a specific visualization system.

The axes-based visualizations techniques such as Parallel Coordinates [35] and Polar Parallel Coordinates [13] allow the analysis of multidimensional data sets. In [61] it was proposed the TimeWheel technique as an alternative to Parallel Coordinates for visualizing time series with multiple dimensions. The basic idea of this representation is to present the time axis at centre and the encoding time-dependent attributes axes around it in a circular arrangement. Similar to Parallel Coordinates, lines make a connection between each time point in the considered data and the corresponding points on each of the attribute axes. Figure 1.1(a) presents an example of a TimeWheel representation. Nevertheless intuitive to express multidimensional data, axes-based visualizations are limited by the length or width of the screen and the number of axes to be considered.

In [69] it was presented a different approach based on a spirally shaped time axis with the purpose of visualizing large data sets and detect patterns and periodic behaviours of data. The Spiral Graph also allows the comparison of different time series through the use of intertwined graphs and different visual encodings for each. However, there is a limited number of comparisons for each graph and this technique requires an appropriate parameterization in each use [2]. Figure 1.1(b) represents an example of a Spiral Graph.

One of the most recent and promising technique is the Horizon Graph, a stacked graph based visualization technique which enables the performance comparison of a large number of time-dependent variables and the detection of outliers and predominant patterns. It is built by adding colour bands to a line graph and mirroring the negative values respecting the x-axis, then, through a technique called two-tone pseudo colouring the colour bands are overlaid. Finally, the use of small multiples (a series of small graphs stacked one above the other) enables to display a large number of data in a single view [56]. Figure 1.1(c) shows an example of a Horizon Graph. The performance of this novel representation was investigated in different studies. The study presented in [27] shows that Horizon Graphs are more efficient than line charts in the studied tasks and it also provides some recommendations about the number of bands to be used and the best chart size. However, the number of simultaneous time series and bands studied were limited only to two and four respectively. In [33] it was done an experiment with more techniques and higher numbers of time series, considering the previous recommendations. In this investigation the authors concluded that for comparisons in time series with a large visual span Horizon Graphs are generally more efficient.

A different approach arises with the use of Sparklines. A Sparkline is a word-sized graphic usually without axis or other scale. This technique enables, therefore, the visual display of a large amount of data in condensed charts in line with text [62]. In [54], it was suggested that Sparklines can help to reduce clinical diagnostic errors. Figure 1.1(d) represents an example of Sparklines.



(a) TimeWheel of a six variable data set containing several diseases statistics. Adapted from [61].

(b) Spiral Graph showing the relationship between two stock prices (differentiated by colour) in five years. From [69].

(c) Horizon Graph representing the daily performance of 50 different stocks over a year. From [56].

(d) Sparklines usage in clinical analysis. From [62].

Figure 1.1: Examples of data visualization techniques.

The choice of the appropriate technique to represent data sets should weight the different strengths and weaknesses that each technique presents for diverse tasks, and evaluate the proper parameterization to use. Table 1.1 summarises the presented visualization techniques.

Table 1.1: Summary of the presented visualization techniques.

Visualization Technique	Type	Purpose	Limitations
TimeWheel	Geometrically transformed display	Multidimensional data sets analysis.	Limited by the screen's length/width and by the number of axes to consider.
Spiral Graph	Geometrically transformed display	Patterns and periodic behaviours detection in different time series.	Requires appropriate parameterizations in each use.
Horizon Graph	Stacked display	Comparison of a large number of time dependent variables with detection of outliers.	The better results are limited to three bands.
Sparklines	Icon-based display	Visual display of data in condensed charts in line with text.	Presents few details about the data.

1.2.2 Visualization Toolboxes

Nowadays, a variety of toolboxes are available for creating data visualizations taking into account diverse techniques. With the aim of bringing data to life, there are different tools for different programming languages and most of them are free and open source. In [14] a considerable list of tools is presented, however, only some of them are suitable to time series visualization.

Data-Driven Documents (D3) is a promising JavaScript library developed by Michael Bostock for efficient manipulation of documents based on data in order to custom web based visualizations. D3 operates by binding data to a Document Object Model (DOM) and then applies data-driven transformations to the document. In the end, the concept and rules are provided by the user for D3 to execute. This fast library supports large data sets and dynamic behaviours and uses the full capabilities of HTML, Scalable Vector Graphics (SVG) and CSS. Still, it does not support older browsers. A diverse collection of toolkits and plugins built on D3 can be used and improved in different conditions. From this list stands out the Cubism plugin used to visualize time series on horizon graphs and the Rickshaw toolkit used to create interactive time series graphs [48].

Flot [2] is a well-known plugin used in data visualization and built on JQuery [17], the most popular JavaScript library. This plugin is focused on attractive and interactive visualization.

Two more JavaScript libraries excel in research, the dygraphs [3] and the Envision [4].

[2] http://www.flotcharts.org/

[3] http://dygraphs.com/

[4] http://www.humblesoftware.com/envision

They produce fast, interactive and dynamic charts for time series data sets, where dygraphs also supports massive data sets and has the ability to display error bars around the data.

One of the most recent libraries for data visualization was created by Cyrille Rossant. Galry [58] is a Python library for 2D plotting like matplotlib [31] (more popular). However, by contrast to matplotlib, Galry provides an efficient way to visualize large data sets (up to 100 million data points) by using the Open Graphics Library (OpenGL).

As well as Galry, bokeh [5], from Continuum Analytics, is a python library which enables a high-performance interactive visualization over very large data sets. However, at this moment, it uses HTML5 canvas, instead of Web Graphics Library (WebGL) to display the data in the browser.

With the aim of bind the best features of two popular libraries, matplotlib and D3, arises the mpld3 project [6]. From this conjunction results an improved visualization of the data in the browser.

Table 1.2 summarises the presented visualization toolboxes.

Table 1.2: Summary of the presented visualization toolboxes.

Toolbox	Language	Rendering	Developed for massive data sets?
D3	JavaScript	SVG	yes
Flot	JavaScript	HTML Canvas	no
Dygraphs	JavaScript	HTML Canvas	yes
Envision	JavaScript	HTML Canvas	no
Galry	Python	OpenGL	yes
bokeh	Python	HTML Canvas	yes
mpld3	Python	SVG	no

1.2.3 Visualization Applications

Usually, visualization alone is not enough, particularly when dealing with large data sets and multiple entities. Therefore, applications for visualization and exploratory viewing of time series data sets have been arisen in the last years in order to enable an appropriate user interface and convenient results analysis. Today, several applications, considering different analytical methods, are available to commercial, academic or research use. In [2] is provided a survey of applications for different purposes.

Cluster and Calendar based Visualization tool [64] is probably one of the first and most popular systems for time series data visualization. The application clusters similar daily data patterns and then displays the average patterns as line plots as well as a calendar with each day coloured according to the cluster that it represents - this is also known as the temporal clustering method. Even though providing a good overview of multiple

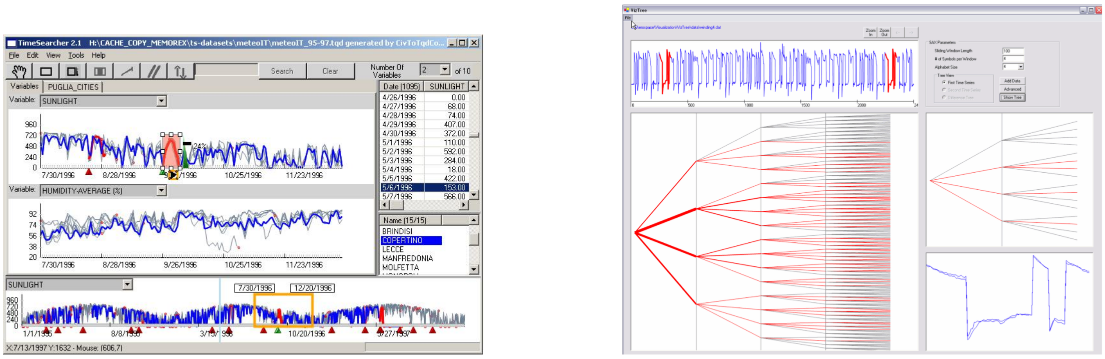
[5] http://bokeh.pydata.org/

[6] http://mpld3.github.io/

time series data, the application is limited to calendar-based data and requires previous knowledge about patterns to identify.

Another common tool for time series visual exploration is TimeSearcher2 [10], an extension of one previous work that allows users to visualize long time series of multiple heterogeneous variables by a combination of filter and pattern search capabilities. Here, the concept of timebox, a rectangular query region, already introduced in previous work is enhanced. While the original timebox is used to filter the data and reduce the scope of the search, an improved timebox, the searchbox, enables to perform a specific pattern search anywhere in the remaining data. Besides flexible, intuitive and interactive, this application presents some issues, particularly when dealing with extensive data sets. In this case the process becomes cumbersome, the application has limited scalability and the query model may be considered too simplistic. Figure 1.2(a) presents an example of the TimeSearcher2 interface.

By contrast to the previously described concept of query-by-example, it was developed the VizTree [42], a visualization system for massive time series data sets that discovers interesting patterns (frequently occurring or surprising patterns) without the need of a first example. This application is based on Symbolic Aggregate Approximation (SAX) [43]. The data is discretized into a fixed length of subsequences, converted to symbols and then the symbols are concatenated to form symbol strings. Finally, a modified suffix tree is built, where frequency and other properties of patterns can be differentiated by colours and other visual properties. VizTree benefits from the ability to scale very large databases and to discover non-trivial patterns. However, it is not very intuitive. Figure 1.2(b) illustrates an example of the VizTree interface.


TimeSearcher 2.1 H:\CACHE_COPY_MEMOREX\ts-datasets\meteoIT\meteoIT_95-97.tqd generated by CivToTqdCo...

Variables: PUGLIA_CITIES

Variable: SUNLIGHT

Variable: HUMIDITY-AVERAGE (%)

Number Of Variables: 2 of 10

Date (1095)	SUNLIGHT
4/26/1996	0.00
4/27/1996	68.00
4/28/1996	74.00
4/29/1996	407.00
4/30/1996	372.00
5/1/1996	110.00
5/2/1996	592.00
5/3/1996	284.00
5/4/1996	18.00
5/5/1996	422.00
5/6/1996	153.00
5/7/1996	566.00

Name (15/15)
BRINDISI
COPERTINO
LECCE
MANFREDONIA
MOLFETTA

X:7/13/1997 Y:1632 - Mouse: (606,7)

(a) TimeSearcher2 interface. The box in red selects the pattern of interest. In the graph below the result of the search is displayed with red triangle markers under the horizontal axes. On the right side the result is shown in detail. From [10].

(b) VizTree interface. At top is shown the the input time series. The bottom left panel shows the subsequence tree for the time series. The parameter setting area is on the top right. The bottom panel plots the actual subsequences when the user clicks on a branch. From [42].

Figure 1.2: Interfaces of applications for time series visualization.

With the intention of stimulating research in the study of complex biomedical signals Goldberger et al. (2000) [21] developed a resource that comprises three interrelated components. The PhysioNet offers free web access to a large biosignals database (the PhysioBank) and to a wide collection of software for viewing and analysing the signals (the PhysioToolkit). A particularity of the PhysioBank data sets is that these can include ground-truth information and, as stated before, the presence of annotations in data sets is a key feature in signal analysis. Besides the ground-truth information and the presentation of a wide variety of classic signal-processing functions, it works by command-line tools, which makes it lowly user friendly.

Notwithstanding the increasing in number of novel visualization applications for time series, it is verified that many of them are thought only for a specific domain [15] or focus on a limited number of tasks. Besides, some of the literature found presents tools that seem promising but are underexplored.

1.2.4 Visualization Applications in Clinical Cases

The high concern in clinical systems for improving the physical and psychological wellness has resulted in the emergence of a variety of new and crucial systems to record, visualize and analyse different types of biosignals.

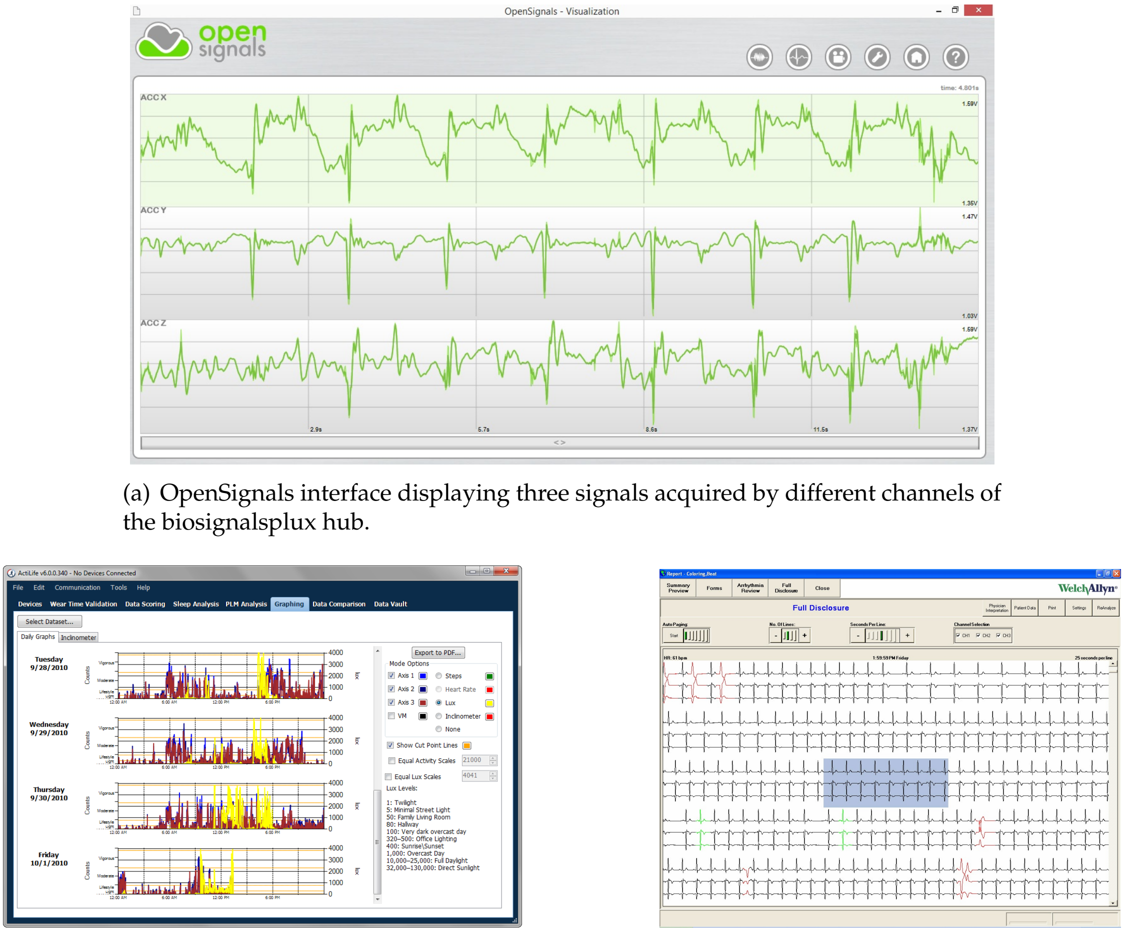
OpenSignals, developed by PLUX, is a software application which enables biosignals visualization in real-time or offline from a previously recorded data set, even large. In a user friendly interface, the entire signals are displayed. Then, an efficient navigation through the signals can be performed by zooming and panning. This tool also provides an EMG or Heart Rate Variability (HRV) analysis and has the possibility of open multiple channels, in other words, different biosignals from a recording [24]. Figure 1.3(a) shows the OpenSignals interface. This application can integrate the biosignalsplux system [7], an advanced biosignal monitoring platform for sports and biomedical research. Besides the OpenSignals, this system is also composed by a wearable hub and body sensors that provide the acquisition of ECG, Electrodermal Activity (EDA) or ACC signals, among others.

ActiLife 6 [1] is a software application that enables the visualization of data acquired from an accelerometer, the ActiGraph monitor. The ActiLife 6 contains a selection of analysis tools, including the calculation of sleep statistics or the identification of low activity periods, in order to provide a better and faster analysis of the data. Figure 1.3(b) shows the ActiLife 6 interface. This software integrates the ActiGraph system, the most used actigraphy monitoring system in research and clinical trials involving physical activity and sleep assessment, even in long-term and real-time studies.

In cardiology, one of the most common examinations is the Holter, an ambulatory electrocardiography for a minimum 24-hour period, conducted with the purpose of screening for ECG disturbances. The Welch Allyn Holter System Application, from Welch Allyn [70], is a software application used to analyse ECG signals in Holter examinations. The

[7] http://www.biosignalsplux.com/

application interface, shown in Figure 1.3(c), allows the selection of specific tasks enabling, for example, the review of the entire record with color-coded abnormal events or the conducting of specific assessments such as ST or HRV measurements. In the end, this software creates a report that includes selected ECG strips, some measurements and technician's comments. This application together with the Welch Allyn Holter Recorder constitute the Welch Allyn Holter System, one of many systems available to perform Holter examinations.



Mode Options

☑ Axis 1 ○ Steps
☑ Axis 2 ○ Heart Rate
☑ Axis 3 ● Lux
☐ VM ○ Inclinometer
○ None

☑ Show Cut Point Lines

☐ Equal Activity Scales 21000

☐ Equal Lux Scales 4041

Lux Levels:

1: Twilight
5: Minimal Street Light
50: Family Living Room
80: Hallway
100: Very dark overcast day
320–500: Office Lighting
400: Sunrise\Sunset
1,000: Overcast Day
10,000–25,000: Full Daylight
32,000–130,000: Direct Sunlight

☑ CH1 ☑ CH2 ☑ CH3

(a) OpenSignals interface displaying three signals acquired by different channels of the biosignalsplux hub.

(b) ActiLife 6 interface. From [1].

(c) Allyn Holter System Application interface. From [70].

Figure 1.3: Interfaces of clinical applications.

In spite of the high development of techniques, tools and applications to enable signals visualization, the visualization and processing of time series data sets faces yet some issues, particularly when dealing with big data. The possible integration of applications in real life platforms to monitoring diseases increases the demand for novel solutions. Therefore, the tools described in this section contribute to the evolution of signals visualization in general and also to the work that was developed in this thesis.

1.3 Objectives

The main purpose of this work is to develop a novel application which enables the visualization and exploration of biosignals in a big data context. Its demand is sustained by the need of enhance the monitoring of health and pathological conditions and improve the biomedical research.

Therefore, investigation for algorithms that allow the visualization of biosignal features, in a fast and user friendly way for further analysis by users, will be performed. This will be a powerful tool which will enable hidden information and knowledge discovery from the data. The developed application will also allow the annotation of biosignals.

Different biosignals will be collected (ECG, EMG and ACC signals). Then these biosignals will be tested and analysed in the developed visualization system. The application will help finding patterns in the data, thereby improving, for example, the diagnosis of several diseases.

In spite of this work only be applied to ECG, EMG and ACC signals, it is also desired that this novel tool may be applied to another type of needs that wrap the visualization and analysis of big data in biological context.

In order to achieve these objectives the following steps were planed:

1. Identify the requirement for an interface and the challenges that will be faced.
2. Study about the end users and its goals.
3. Develop a technique that allows overcoming the stated problems.
4. Test the interface with different biosignals.
5. Present the interface developed to possible users in order to evaluate it.

1.4 Thesis Overview

The remainder of this thesis is organised as follows and schematically represented in Figure 1.4.

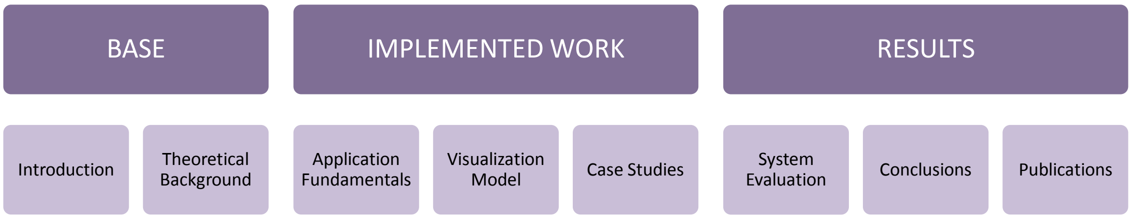


Figure 1.4: Thesis Structure.

The first two chapters comprise the thesis basis. While Chapter 1 presents the motivation and objectives of this work and also revises the related work, Chapter 2 provides the main theoretical concepts that fundament this work.

The following Chapters present the application development. In Chapter 3 the fundamentals of the application are discussed. Chapter 4 exposes the followed visualization model and Chapter 5 presents the case studies considered.

Finally, the results achieved are considered. Chapter 6 discusses the usability test performed and Chapter 7 draws the main conclusions and provides the directions for further research. Appendix A contains the publications resulting from this work.

2

Theoretical Background

This chapter provides the most relevant base concepts for this work. A review on time series, focusing on biosignals, and its acquisition and processing is the first subject to approach. Then, data visualization is introduced. In the end, it is presented a brief review about convenient graphics concepts.

2.1 Time Series

A time series can be defined as a set of observations made in a chronological order and is a common form of recorded data.

The nature of time series can be assigned to a variety of fields, ranging from finance to science. In order to understand physiological mechanisms in humans, the analysis of time series in medicine, sports or research is one of utmost importance fields where these can be applied.

2.1.1 Biosignals

Biomedical signal, or simply biosignal, is a term used to characterize signals that can be measured from a biologic system. To address the extraction and storage of these signals, an acquisition system has to be applied to the human body.

Biosignals result from electrical, magnetic, chemical or mechanical activity during biological events such as heart beat or muscle contraction and are classified considering their nature, application or their characteristics [37].

Electrodermal Activity (EDA) signals which reflect the changes in the electrical properties of the skin, Electroencephalography (EEG) signals which monitor the cerebral activity and Respiratory (RESP) signals which measure respiration rate and amplitude are some examples of biosignals that can be measured from the human body.

The following biosignals were applied in the developed visualization tool and will be later discussed. Therefore, they are here highlighted.

Electrocardiography (ECG)

The ECG signal is the most accessible and the most widely acquired bioelectric signal of the human body. It consists in the recording of the electrical activity of the heart. This is a cyclic signal and its basic waveform consists on a P wave, a QRS complex, a T wave and a U wave [37] (Figure 2.1).

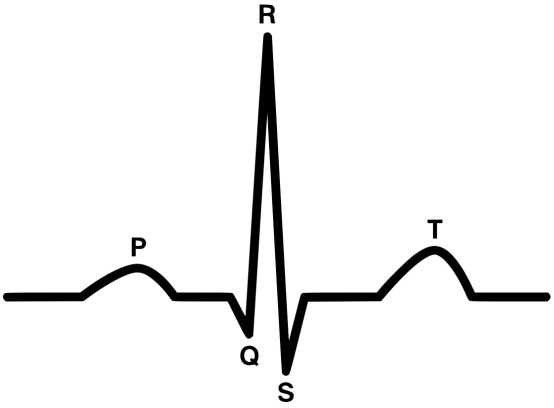


Figure 2.1: ECG basic waveform. This consists on a P wave, a QRS complex, a T wave and a U wave (not represented). From [24].

Therefore, the ECG analysis provides a fundamental way of cardiac monitoring that allows the detection of cardiac abnormalities, called arrhythmias. Arrhythmias may occur in healthy hearts with minimal consequences, but they may also indicate serious heart diseases [46].

Electromyography (EMG)

An EMG signal is an electrical signal that represents the neuromuscular activity. It is generated in a muscle during its contraction, which enables the quantification of the neuromuscular function. Although the resulting electrical potentials can be recorded through an invasive technique called intramuscular EMG, in this work the recording was performed from the skin surface through a technique called surface EMG [37], [57], [59]. Figure 2.2 represents an EMG signal where is possible to observe an activation pattern of the gluteus maximus.

The EMG signals are mostly used to measure the degree of muscle activation and to access the neurophysiologic mechanisms of fatigue [6], [11]. For this reason, they can be

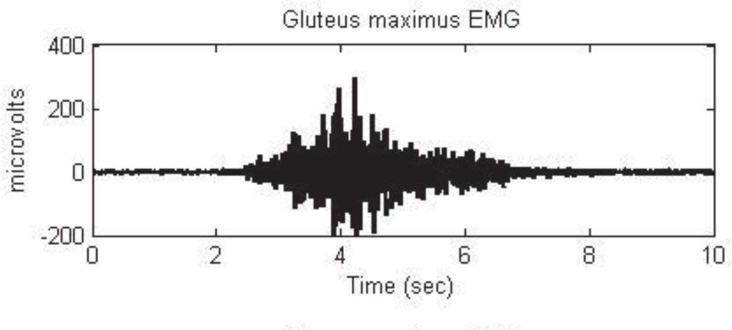


Figure 2.2: EMG signal. In this figure is possible to observe an activation pattern of the gluteus maximus assessed during a hip extension with knee extension on prone position movement. Adapted from [60].

applied in diagnosis of neuromuscular diseases, muscle therapy and rehabilitation and in the sports field through the analysis of the human movement biomechanics [59]. These signals can also be analysed along with other biosignals to achieve better results in the study being performed [57].

Accelerometry (ACC)

The ACC signal enables the performance of kinematic studies providing the measurement of the applied acceleration acting along a reference axis. Therefore, its analysis provides crucial information that can be used in functional status and monitor falling studies [34], [38], [40], sleep analysis [1] and in neuromuscular diseases diagnosis [44].

The study of the human movement is not recent, leading to the development of motion sensors, like accelerometers. These can provide an objective method of measuring physical activity through the access of the rate and intensity of body movement in up to three plans. Their low cost when compared with other motion capturing systems, small size and the fact that it is not necessary to be used in laboratorial environment made them widely accepted as wearable devices to assess the motor behaviour [20]. The operating principle of accelerometers is based on a proof mass (a mechanical sensing element) attached to a mechanical suspension system. According to Newton's Second Law, the inertial force resulting from acceleration or gravity will displace the proof mass. Acceleration may then be measured electrically, but its quantification depends on the accelerometer class and calibration is always needed in order to set the relationship between the output and the reference value [74]. The output of an accelerometer worn on the body will depend on important factors such as position and orientation of the accelerometer, the posture of the subject and the performed activity [44].

Figure 2.3 represents a recording from an accelerometer showing a subject's motion.

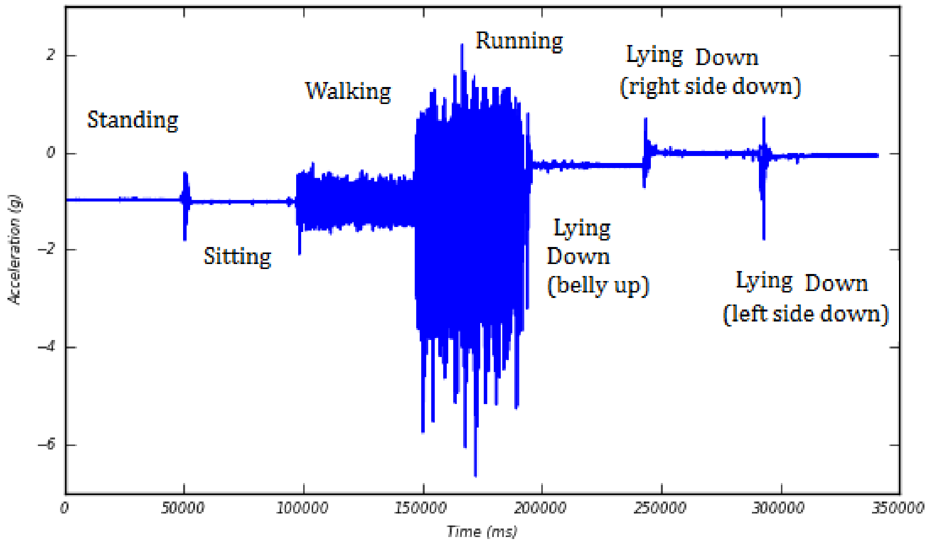


Figure 2.3: Recording from an accelerometer showing a subject's motion. In this case the subject was asked to perform specific tasks for about one minute. From [44].

2.1.2 Time Series Processing and Data Mining

Usually in many biomedical applications, as in other fields, the acquisition itself is not enough. So it is required signal processing in order to get the relevant information hidden on it. Therefore, a variety of processing methods and algorithms may be applied to the data, always considering the goal that user wants to achieve [47].

Data mining consists in the application of algorithms in order to extract useful information from large and complex volumes of data [4].

The fundamental problem in data mining is how to represent time series data in order to reduce its dimension. This reduction is achieved by several methods such as Principal Component Analysis, Discrete Fourier Transforms or SAX, that transform the time series data to another domain, principal component space, frequency and symbolic domain, respectively [19].

This is the primary step to efficiently deal with the following data mining tasks. These tasks facilitate the exploration and analysis of massive time series data sets. These are the classic time series data mining tasks and may be combined to obtain more sophisticated data mining applications.

- **Indexing or Query by Content.** Search for the most similar time series considering a specified query and a similarity/dissimilarity measure.
- **Clustering.** Group data into clusters based on a given similarity/dissimilarity measure.
- **Classification.** Determine which class a data set, sequence, or subsequence belongs to, in an unlabelled time series.

- **Prediction or Forecasting.** Given a time series containing n data points, predict the future behaviour at time $n+1$. This task can be regarded as a type of clustering or classification, however in this case the focus is a future state whereas in clustering and classification is a current state.

- **Segmentation.** Construct a model with k segments from a given time series containing n data points ($k << n$). Segmentation may be performed to create a high level representation of time series that supports indexing, clustering and classification.

- **Summarization.** Create an approximation of a given time series (extremely large) retaining its essential features to fit on a single page or a computer screen for example. Anomaly detection is a special case of summarization where only surprising, interesting, unexpected or novel behaviours are reported [55].

In addition to being considered an important mechanism to present processed time series for further analysis, visualization is also a powerful tool to facilitate the mining tasks [19]. In fact, the approaches above are the basis of some visualization applications already presented in the state of the art [10], [42], [64].

2.2 Data Visualization

The human vision is highly sophisticated and specifically suited for supporting visual patterns. In fact, it has been frequently nominated as the ultimate data-mining tool.

From the integration of this human visual perception with the processing power of computers arises a novel concept of visualization. The concept of data visualization can be seen as a relationship between technology, art and science where the results on one view can stimulate the work in the others [63]. These three different views are schematically depicted in Figure 2.4.

2.2.1 Objectives

The integration of the human's flexibility, creativity, and general knowledge with the massive computational capacities enables an improved data exploration.

The aim of data visualization is to support the analysis, understanding, and communication of data, models and concepts. It allows generating and confirming hypotheses and it is a quick manner to be aware of large amounts of information.

The benefit of visual data exploration is allowing users to interact directly with the data mining process by getting insight into the data and making conclusions. [39].

2.2.2 Limitations

Besides the benefits of data visualization, there are three types of limitations to consider when designing a visualization system:

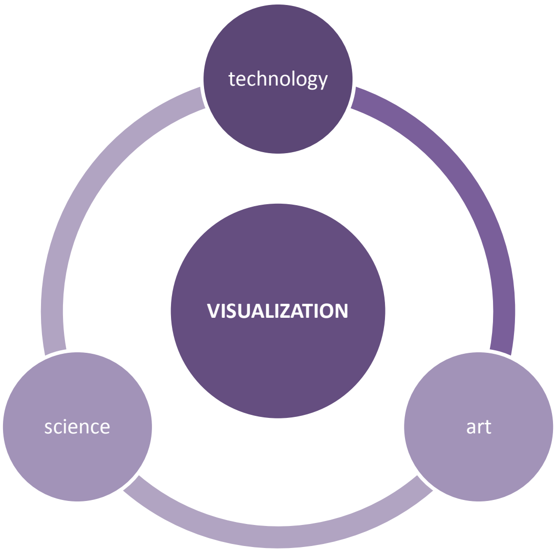


Figure 2.4: Views on visualization. Data visualization can be seen as a relationship between technology, art and science where the results on one view can stimulate the work in the others.

- **Computation capacity:** the memory and the time taken to run an algorithm are limited resources in the computer side.
- **Display capacity:** the resolution of the screen is also a limitation.
- **Human perceptual and cognitive capacity:** on the man side, memory and attention are finite resources. Humans are also vulnerable to a phenomenon known by change blindness that occurs when a change, even if big, is not noticed by the observer [47].

2.2.3 Basic Model of Visualization

In [63], a generic model for visualization is proposed. The central process in this model is visualization, where data is transformed into a time varying image according to a specification (this includes a specification of hardware, algorithms and parameters). In the following process, the image is perceived by a user resulting in an increase in knowledge (the gain depends on the image, the current knowledge of the user and the particular proprieties of the perception and cognition of the user). Then, the interactive exploration enables to adapt the specification of the visualization based on the current knowledge in order to explore the data further. Figure 2.5 represents schematically this model.

The whole model should have as a starting point the Visual Analytics Mantra proposed by Keim (2010) "Analyse first, show the important, zoom, filter and analyse further, details on demand." Indeed, the challenge in visualization research is to develop methods completely guided by this mantra with the aim of tightly integrating time series mining and visualization [2].

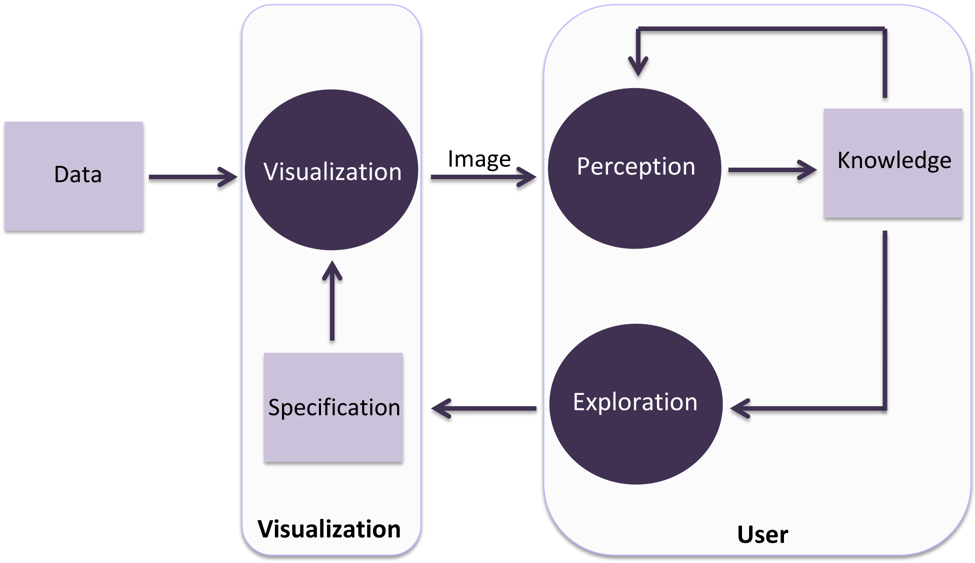


Figure 2.5: Simple model of visualization. The circles denote processes, while boxes denote containers. In visualization process, data is transformed into a time varying image according to a specification. In the following process, the image is perceived by a user resulting in an increase in knowledge. Then, the interactive exploration enables to adapt the specification of the visualization based on the current knowledge in order to explore the data further. Adapted from [63].

2.2.4 The Visualization Process

What, why and how are three questions that have to be answered when developing or choosing a visualization method [2].

Time series data is the answer this work gives to the first question and the need for a new visualization tool that enables the visualization of long term biosignals to help clinical diagnosis is the answer given to the second. The answer to the second question also encompasses the study about what specific tasks the user seeks to accomplish in order to get the better analysis of the biosignals. Allowing the user to make annotations in the biosignals visualization is one of the tasks to be taken in account and to be developed in this thesis.

After a specific domain problem has been identified, it has to be abstracted into a more generic representation. This process has become even more indispensable with the increasing on volumes of data to analyse [2]. The generic operations used in abstraction include sorting, filtering or even finding anomalies. Often, abstraction also involves transforming data from the original raw data into derived dimensions that can be different from the original data type [47].

The answer to the third question, how, arises now and includes the choice of the visualization design and algorithm. It will also be investigated in this thesis.

Figure 2.6 represents schematically the visualization process discussed.

The consideration of these questions will help fulfilling the next three major criteria

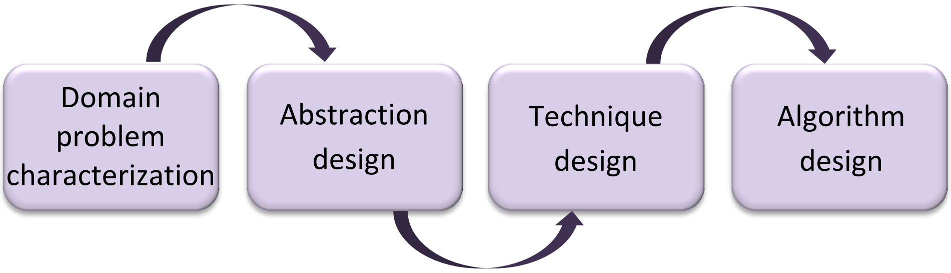


Figure 2.6: The visualization Process. The visualization process can be split into four levels where the output of one is the input of the other. Adapted from [47].

that any visualization should satisfy:

- **Expressiveness:** Showing exactly the information contained in the data.
- **Effectiveness:** Obtaining intuitively recognizable and interpretable visual representations considering the capabilities of the human visual system and the application background.
- **Appropriateness:** Considering the benefit of the visualization process with respect to achieving a given task, a cost-value ratio. The cost is assessed by the time efficiency (computation time spent) and the space efficiency (exploited screen space). In the display space the benefit of showing as much as possible at once minimizing the need for navigation should be balanced with the cost of overwhelming the user with visual clutter [2].

2.2.5 The Perception and Cognition Process

The graphical perception, already discussed in the state of the art, refers to the ability to decode the information encoded on graphs. Therefore, in order to facilitate the compression of data sets, the visual encoding variables used in visualization, such as position, shape or colour, have to be assessed [27].

2.2.6 The Interactive Exploration Process

Data exploration enables the direct interaction of the user with the data in order to change the visualizations according to his exploration objectives. The following interaction and distortion techniques represent some examples that can be applied to the visualization process, allowing the user to focus on some details in the visualization:

- **Projections:** Changing the projections of the visualization is possible to explore a multidimensional data set in a different way.
- **Filtering:** Selecting a desired subset enables the focus on a defined segment or pattern of the data. This can be done by browsing or querying.

- **Zooming:** Presenting the data in highly compressed form provides an overview of it, and presenting the data in higher zoom levels enables to display more details of one part of the data [39].

2.3 Graphics Fundamentals

2.3.1 Object-Oriented Design

Object-Oriented Design (OOD) is a popular approach to analyse and design a system through a variety of techniques and principles. According to [71] OOD is the best way to work with graphics. Actually, graphics are objects, so this is a natural framework for them.

OOD enables the development of flexible and reusable software inducing to abstraction and hierarchy. Another advantage of this approach is that its components are relatively modular, so the system can continue to function even when parts of it are discarded or malfunction.

In OOD, objects are the basic components of the system. They are relatively simple, useful for a variety of purposes and often polymorphous. The objects can communicate with each other through simple encapsulated messages and also inherit attributes and behaviour from other objects [71].

2.3.2 Graphics Rendering

Nowadays, the best way to generate dynamic graphics involves modern browsers. They can load and render data incrementally and allow the following of the visual analytics mantra. Therefore, knowledge of web-standard technologies can also be convenient in this work.

On a web browser, a complex language is used to structure the content. This language is the HTML and its hierarchical structure is called DOM. CSS is the mechanism that defines how to style the visual presentation of DOM elements. After parsing the HTML and generating the DOM, the browser can apply visual rules to the DOM contents and generate an image from a model. This process is called rendering.

Although HTML enables specifying semantic structure or assigning hierarchy, relationships, and meaning to content, it does not address the visual representation of the document. In HTML the Canvas element enables rendering graphics on a resolution-dependent bitmap canvas. However, it is necessary to use a script to actually draw the graphics. Usually the script used is JavaScript.

Another way of generating and manipulating visuals is the SVG elements which are more reliable and flexible than the canvas element [48].

OpenGL is an alternative to HTML. Using the interaction with a graphics processing unit enables rendering 2D and 3D computer graphics in a faster way [58]. WebGL derives from OpenGL and provides a similar rendering functionality, but in an HTML context [72].

Therefore, while Canvas and SVG provide bitmapped and vector capabilities, respectively, WebGL provides 3D drawing context [72]. All these elements have been introduced in HTML5.

3

Application Framework

The architecture of the ideal tool to visualize big data in the context of biosignals was rigorously studied. This chapter presents the fundamentals of the developed visualization tool.

Firstly, the main requirements of the application are enumerated. Then, the system architecture is presented, where the relationship between the data access and its visualization is explained. Finally, the choice of the Hierarchical Data Format 5 (HDF5) to store the data is justified and the file structure created for this work is presented.

3.1 System Requirements

The demand for a tool that enables the visualization of long-term biosignals (more than 24 hours of records) has been exposed. The limitations of the existing approaches were studied and the development of the novel visualization tool took into account some base requirements. Therefore, the developed tool has to:

- be applicable to all types of biosignals
- enable the possibility to explore up to 10 days of continuous acquisitions
- show the time lapses where the signal acquisition was interrupted
- allow the handling of annotations
- present a fast user friendly interface

Lastly, the proposed model has to represent a commitment between usability and performance, allowing the user (a clinician or a researcher) to analyse a long-term biosignal without having to deal with signal processing algorithms directly.

3.2 System Architecture

As previously mentioned and exemplified, the web standards currently available provide some of the best tools needed to the creation of a rich graphical user interface. In addition, a web-based platform also eliminates complex installation and configuration procedures. Therefore, the implemented visualization tool is a web-based platform.

Although it is possible to visualize local HTML files directly in the web browser, for security reasons the browser cannot load local files via JavaScript [48]. For this reason, a client-server model was implemented. Since in the prototyping phase it is much faster to store and host everything locally, it was used a local web server instead of a remote one.

3.2.1 Web Server

The server was implemented in Python language and the communication between the visualization platform and the server is done with WebSockets [67].

As a result, the visualization platform, on the client side, sends request messages to the server, which responds by serving a message containing the requested information. This flow of information is schematically represented in Figure 3.1. The request and response messages are on JavaScript Object Notation (JSON) structure, a lightweight data-interchange format.

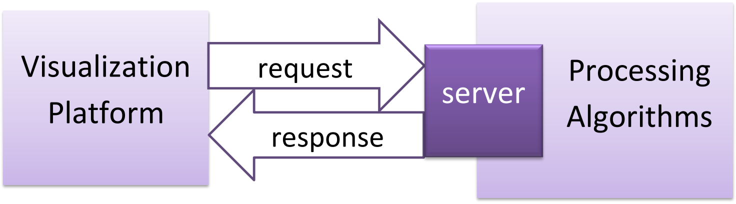


Figure 3.1: The visualization platform, on the client side, sends request messages to the server program, which responds by serving a message containing the requested information.

Through the use of Python language to access and manipulate the data and JavaScript language to present the data and deal with the user-interface tasks, a higher performance is reached.

3.2.2 Application Programming Interfaces (APIs)

The requests sent by the browser intend to:

- get the signal parameters

- get specific data intervals
- handle annotations
- access a detailed analysis of a specific interval

Considering these requests, it was implemented different python APIs with specific algorithms. According to the request done, a particular API is run and the processed data is returned in a string format. Then, this string is parsed as a JSON object.

Some of the implemented algorithms need to load a specific JSON file previously created. This file contains the base variables needed for the algorithms.

The files organisation is presented in Figure 3.2.

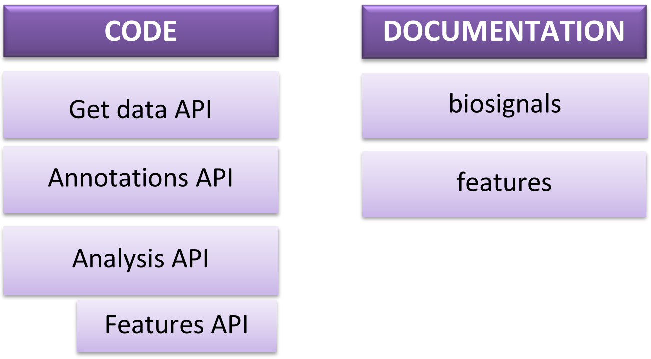


Figure 3.2: Files Organisation. While the code folder contains all the processing algorithms, the documentation folder contains the JSON files with the base variables needed for the algorithms.

3.2.3 Visualization Platform

The JSON objects, sent as responses to the previous requests, are interpreted by a series of JavaScript libraries. Then, the results are dynamically displayed on the page.

While HTML5 and CSS3 allow the definition of the overall layout of the platform, the jQuery library enables to handle user interface events.

3.3 Data Architecture

Since the focus of this work involves the manipulation of very large amounts of data, a suitable data format for storage and accessibility had to be considered. Previous studies have already showed that although presenting some limitations, the HDF5 format is the best option for the intended task [24].

The HDF5 was specially designed to store and access large data sets, allowing a fast random access to any point of the data. This is a portable and extensible format that supports a variety of datatypes. Its hierarchical structure comprises two types of objects:

- **Datasets:** multidimensional arrays of homogenous data type;
- **Groups:** container structures that hold an arbitrary number of other groups and datasets.

This file format also includes file and/or group attributes.

To handle the HDF5 files in the created APIs it was used the h5py python library [5].

The processing API to which each record file is submitted, before its visualization in the developed application, involves general processing that includes filtering, subsampling and event detection. The event detection algorithms are specific for each biosignal type and will be discussed later in Chapter 5. However, files from different biosignal types will present the same groups. These groups are presented in Table 3.1.

In each one of these groups are created n datasets, one for each day present in the data.

Table 3.1: Groups in the processed HDF5 file.

Group	Description
Processed	Filtered data.
Sub5max	Maximums in a subsampled data for 5 minutes of biosignal.
Sub5mean	Means in a subsampled data for 5 minutes of biosignal.
Sub5min	Minimums in a subsampled data for 5 minutes of biosignal.
Sub1max	Maximums in a subsampled data for 1 minute of biosignal.
Sub1mean	Means in a subsampled data for 1 minute of biosignal.
Sub1min	Minimums in a subsampled data for 1 minute of biosignal.
Events1	Data that represents a specific event in a biosignal.
Events2	Data that represents a specific event in a biosignal.
AnnotMessages	Annotations messages.
AnnotStart	Start positions of the annotations messages.
AnnotEnd	End positions of the annotations messages.

4

Visualization Model

The visualization model followed in this work is presented in this chapter.

Firstly, it is described the core idea of the visualization process. The domain problem is characterized and the approaches followed, such as the abstraction implemented and the chosen multi-level visualization method, are here explained. Finally, the visualization techniques selected for this model are presented.

After the central process of the visualization model had been discussed, the visual encoding variables used in the model are exposed.

The perception of the image resulting from the visualization process results in a gain in knowledge that can be increased with its interactive exploration. Therefore, the interactive visualization tools selected for the visualization model are also presented.

4.1 Visualization Process

4.1.1 Domain Problem Characterization

Benefits of long-term monitoring have drawn considerable attention in medicine and healthcare. Nowadays, biosignal monitoring in living environment with the aim of enable emergency response and long-term health management has been attracting special importance. In addition, the choice for long-term monitoring studies has become frequent in cases such as, sleep and movement monitoring, control of cardiac problems and falls detection.

In response, the number of systems which enable the recording of large data sets for different biosignals has been increasing and several studies have reported the achievements in long-term signal acquisition [25], [28], [36], [53].

The issue of long-term monitoring arises when dealing with these massive sources of information, making the analysis of the signals acquired a big challenge.

Visualization, as already stated, is a key point in signal analysis. However, consider, for example, the case of modern Holter recorders, which use sampling rates of 1000Hz or higher to acquire ECG signals over more than 24 hours [28]. Displaying a complete signal (24 hours) would result in drawing more than 80 million data points.

Since the standard computer's screen has only some thousands of available pixels [51], representing the previous signals, not only exceeds the capabilities of the visualization device, but also results in a massive time and memory consuming rate. Even if, the visualization is possible, the representation of excessive data points will surpass the human perceptual capacity, resulting in incorrect interpretations of the data both in time and space. Figure 4.1 represents this last case.

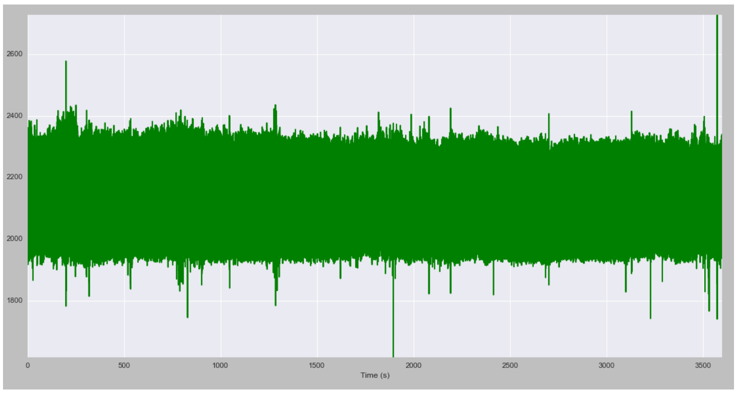


Figure 4.1: Overcrowded Display. ECG signal acquired during 1 hour with a sampling rate of 1000Hz (more than 3 million data points). It is possible to observe the overcrowded and cluttered display that does not give the proper information to the user.

4.1.2 Abstraction Design

After establishing the domain problem, this was abstracted into a more generic representation. In fact, abstraction is an essential process in big data visualization.

The created abstraction took into account that during most of the time, the biosignals collected have no information of interest. The crucial information is contained in special situations – in cyclic or sporadic events. Hence, different types of events were defined considering different biosignal types. These events are described later in Chapter 5.

The created abstraction allows to highlight the key information, contained in the events, and to suppress the irrelevant details. As a result, the visualization model developed does not consider the whole signal, which considerably reduces the number of data points to plot.

Therefore, categorical abstractions of the numerical data, which incorporate knowledge about the analysed biosignal, were performed in order to achieve the intended results.

4.1.3 Visualization in Layers

The model proposed is also composed by a multi-level visualization architecture. This type of architecture enables a simple navigation through the biosignal in order to provide an easy search and focus in the regions of interest.

The developed visualization architecture presents seven standard information layers divided into a defined number of intervals. The choice of the standard information layers had to obey the following conditions:

i) The maximum interval of sequential acquisition considered was 10 days. This is a significant value considering the studies already referenced.

ii) Each layer should be divided into an integer number of intervals with the size of the layer immediately below.

The chosen layers and their number of intervals are described in Table 4.1.

While the first layer displayed gives the user a global overview of the whole biosignal, the others provide a more detailed visualization of the selected interval of the above layer (by default the selected interval is the first one).

Figure 4.2 schematically represents the first three layers that can be displayed. Each layer is divided into a defined number of intervals, each one with the size of the next layer. In this figure is displayed the first fifteen minutes of the eighth hour of the first day.

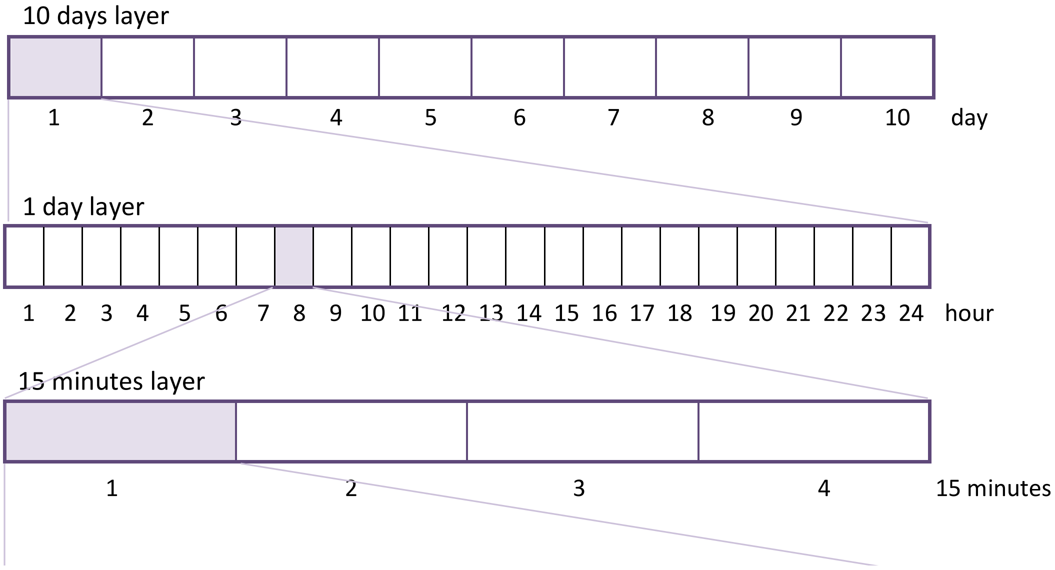


Figure 4.2: Layers Scheme. The three layers represented are the first ones that can be displayed. Each layer is divided into a defined number of intervals, each one with the size of the next layer. Here, the first fifteen minutes of the eighth hour of the first day are displayed.

The higher layers (10 days, 1 day, 1 hour and 15 minutes) have into account the abstraction already discussed. The 5 minutes and 1 minute layers represent the selected interval by the subsampling technique discussed in [24]. The last layer represents the smaller interval with raw data. The type of data that each layer represents is also presented in Table 4.1.

Table 4.1: Possible Layers in the Visualization Model.

Layer	Number of Intervals	Data Representation
10 days	10	Events
1 day	24	Events
1 hour	4	Events
15 minutes	3	Events
5 minutes	5	Subsampled
1 minute	60	Subsampled
1 second	1	Raw

The total number of layers displayed in the platform depends on the total duration of the biosignal that is being analysed. The first layer to be displayed corresponds to the lowest layer that can represent the whole signal. All the layers below this one will also be displayed. For example, if the biosignal has 20 hours of total duration, the model will present only 6 layers where the 1 day layer (the lowest one that can represent the whole signal) will be the first to be displayed (Figure 5.4).

In order to provide a better organisation of the variables needed to implement the different layers, it was created a *layers* dictionary in a pre-existing JSON file. This file contains the dictionaries of all biosignals considered in this work. Consequently, each biosignal contains a *layers* dictionary and other signal information dictionaries that will later be discussed. Each *layers* dictionary is composed of 7 different dictionaries, one per standard layer. The keys of these dictionaries are presented in Table 4.2.

Table 4.2: Layer keys per *layers* dictionary in the biosignals JSON file.

Key	Description
Intervals Number	Maximum number of intervals in which the layer can be divided into.
Total points	Number of points that can be drawn in the layer.
Total Seconds	Seconds represented by the layer.
Total in Day	Number of complete layers that exist in one day.
Function	Name of the function used for getting the data that should be represented here.
Key	Key to use for getting the data from the HDF5 file.
Chart	Type of chart to be used in this layer.
Axis	Axis labels of the chart represented in this layer.
Legend	Legend of the chart represented in this layer.

As example, an excerpt of the biosignals JSON file is presented. Listing 4.1 represents the *10days* layer in the *layers* dictionary for the ECG dictionary.

Listing 4.1: Excerpt of the *layers* dictionary in biosignals JSON file.

```
{
"ECG": {
  "layers": {
    "10days": {
      "Intervals Number": 10,
      "Total points": 960,
      "Total Seconds": 864000,
      "Total in Day": 0.1,
      "Function": "events_10days_data",
      "Key": "events2",
      "Chart": "3line_chart",
      "Axis": {"x": "time (d)", "y": "Heart Rate (bpm)"},
      "Legend": ["HR/15 min (Max)" , "HR/15 min (Mean)", "HR/15 min (Min)"]
    },
    "1day": {
```

4.1.4 Visualization Techniques

After designing the abstractions and the layout for the visualization process model, the natural next step is the choice of the visualization techniques that most efficiently communicate the data information.

Three different visualization techniques were considered in this work. Besides line and bar graphs, two standard techniques in data visualization, horizon graphs, already described in Section 1.2, were also considered. Other techniques, such as dots plots and heat maps, were also tested [23], however, it was selected only those that fitted better in the devised final result.

The main interface combines only the line and bar plots to show the data selected. While the bar graphs are used for the 1 hour layers the line graphs are used for the remaining ones (Figures 5.2 and 5.3). The presence of grey bars in any of the layers of the main interface indicates the time lapses where the acquisition was interrupted. This situation can be observed in the first layer of the Figure 5.4.

In order to help the comprehension of the represented signals, in each layer there is a legend. With exception of the 1 hour layer, the presence of three tones of green represents, from the darker to the lighter green, the maximum, the mean and the minimum of the represented event. In the 1 hour layer the presence of different tones of green allows the distinction of different parts in the signal, for example, in the ACC signal, while the dark green represents the activity moments, the lighter represents the rest moments.

The horizon plots are used in the analysis page, later discussed, enabling the comparison of a variety of features in a restricted space (Figure 4.7).

The visualization customization was done through the creation of SVG elements by the D3 JavaScript library and the visualization techniques described were rendered with the two D3 plugins already presented in section 1.2 (Cubism and Rickshaw). Since

WebGL provides a similar rendering functionality in a faster way, it could be used instead of the SVG. However, the toolboxes currently available to perform this type of rendering are still too complex or not stable enough to provide a suitable visualization.

4.2 Perception and Cognition Process

To better support the analysis tasks, the visual encoding of the developed platform had a crucial part in the process. In order to enhance the graphical perception, some variables, such as positions, shapes and colours, were carefully studied.

A fast and responsive design was assumed and enabled by the Foundation Framework developed by ZURB [1]. Figure 4.3 represents the response to a simulation of two different screen resolutions.

The platform also includes thoughtful default colours [26] that avoid the colour-blind effect.

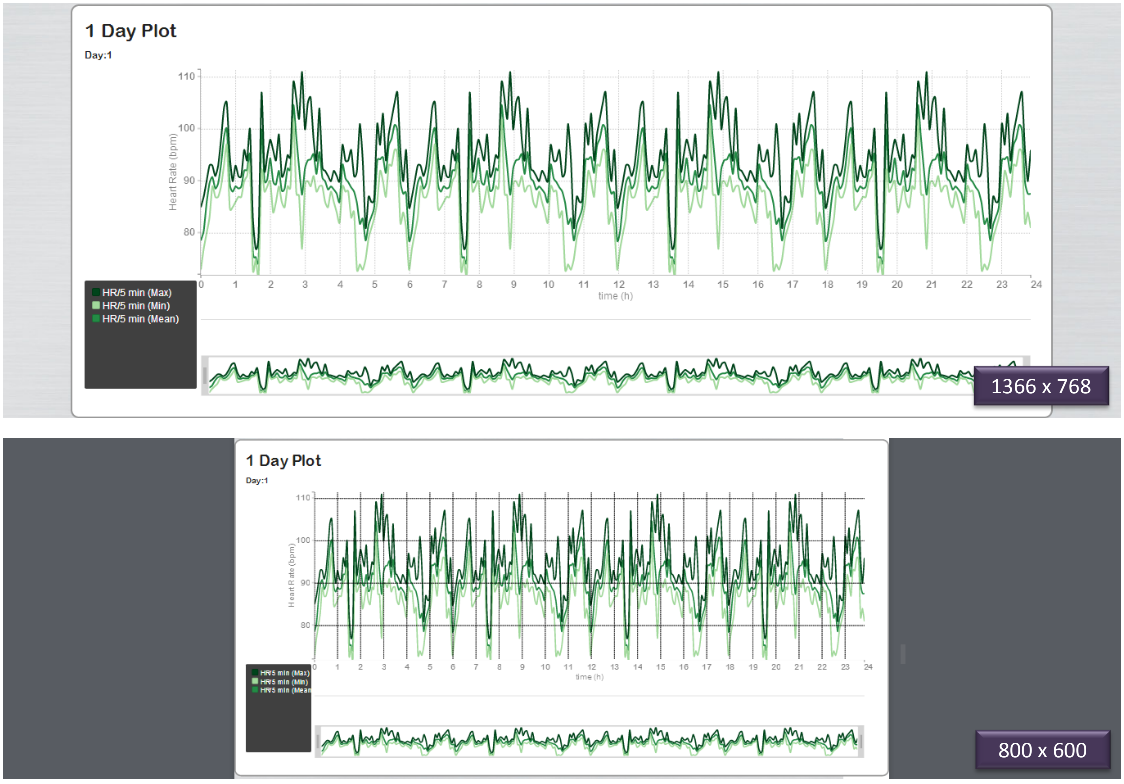


Figure 4.3: Response to a simulation of two different screen resolutions. While the first image represents the response to a screen resolution of 1366 x 768, the second one represents the response to a screen resolution of 800x600. It is possible to observe the resizing of the chart.

[1] http://foundation.zurb.com

4.3 Interactive Exploration Processes

4.3.1 Basic Interaction

The basic interaction techniques selected for the interactive exploration process are now presented. Since they improve the visualization exploration and can be adapted to different types of platform, these techniques are often used. For these reasons and for the fact that the majority of users are used to these tools and know how to use them, they were also selected for the developed platform.

- **Zoom and Pan:** The dynamic change and view of the mapped data allows the focus on some parts of interest. Figure 4.4 shows an example of a zoomed and panned graph in the platform.

- **Tooltips:** Support the visualization showing the instant of time and the amplitude value that the hovered point represents. Figure 4.4 shows an example of a tooltip in the platform.

- **Modal Dialogs:** Enabling the dialog between the interface and the user allows the user guidance in some tasks, alerts of possible errors and presents same useful information. Figure 4.5 represents an example of a modal dialog used in the platform.

- **Save:** Saving the exploration carried out on a pdf format represents an usual way of presenting results.

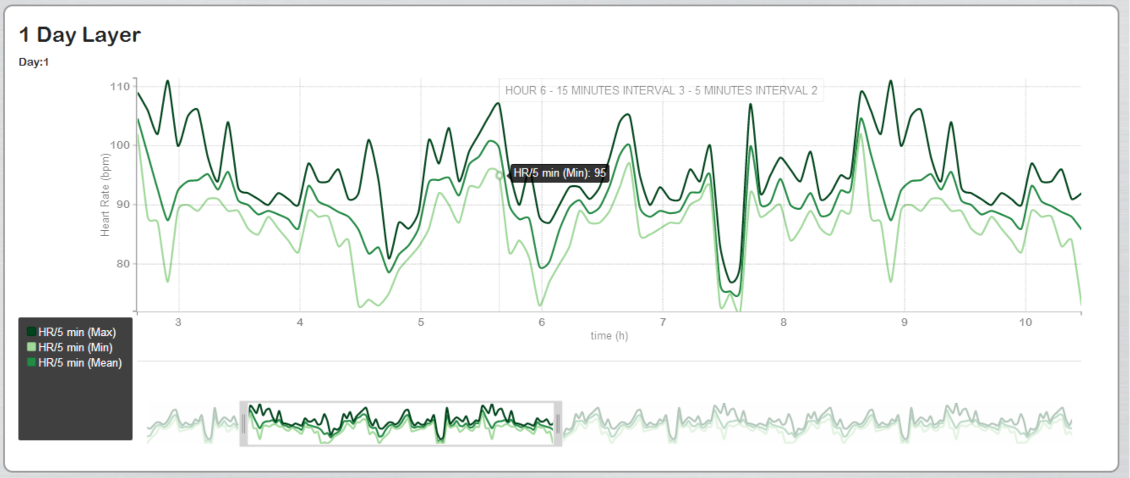
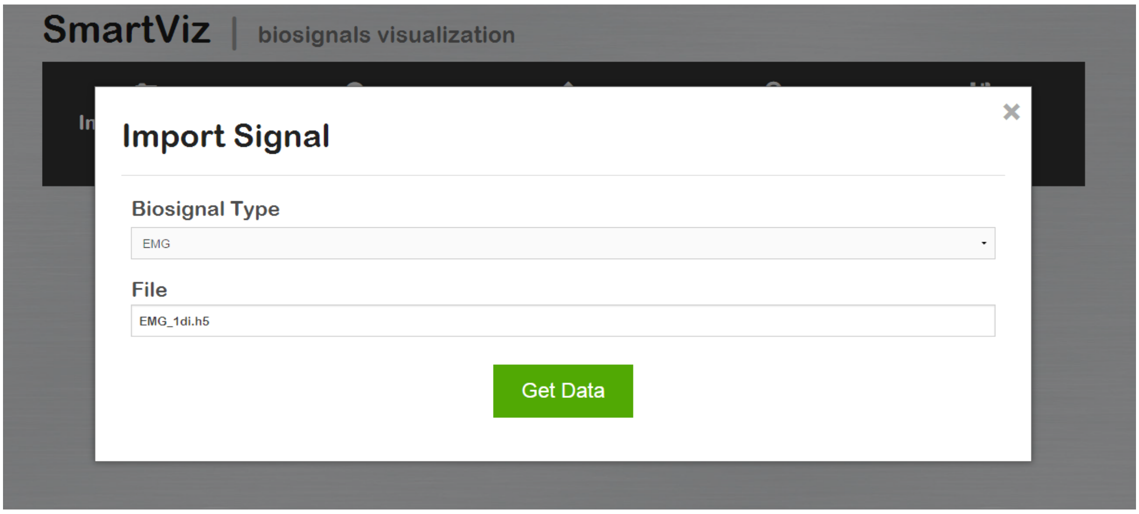


Figure 4.4: Zoom and pan of the previous graph. A tooltip is also present, displaying the information related with this specific graph.



Import Signal

Biosignal Type

EMG

File

EMG_1di.h5

Figure 4.5: Example of a modal dialog present in the developed platform.

4.3.2 Navigation

Considering the core of the visualization process in this work, an interactive visualization technique to support the navigation between the different layers was implemented.

As explained before, the default data to be displayed in the different layers (with exception of the first one) corresponds to the data of the first interval in the layer above. Hence, an appropriate navigation technique was required.

The technique implemented enables the switch of the displayed data through a double click inside an interval of one layer. This will switch the layers below with data from the selected interval.

This innovative type of browsing through the time axis is exemplified in Figure 4.6. The selection of the sixth interval in the 1 day layer, i.e., the sixth hour of that day will switch the data in the 1 hour layer in order to correspond to the data of the sixth hour. This switch continues until the last layer.

4.3.3 Annotations

The demand and importance of annotations has already been exposed. Therefore, the developed platform enables the annotations handling, thereby improving the biosignal interpretation through the notes made by the user in the relevant biosignal regions.

Annotations are visually represented in the platform by a timestamp placed at the beginning of the annotation time frame, in a division below the x axis. When the timestamp is clicked, it shows its message and a rectangular grey region highlighting the time frame of the annotation. Figure 4.6 illustrates the presence of annotations in the platform. It is also possible to have more than one annotation at the same timestamp.

A new annotation can be inserted by clicking the “Add Annotation” item in the platform. This will open a dialog box where it is possible to write the message of the annotation. Then, the user only has to click at the beginning and at the end of the selected time frame and a timestamp appears at the respective division. The layers below also will present the new annotation (if it was placed in their range). This is the case represented in Figure 4.6.

The platform also enables the removal of annotations. This is possible through a click at the annotation, which activates it, and another click with the control key pressed.

When a layer is loaded, either for the first time or if it is updated, every annotation that belongs to the selected time interval is displayed.

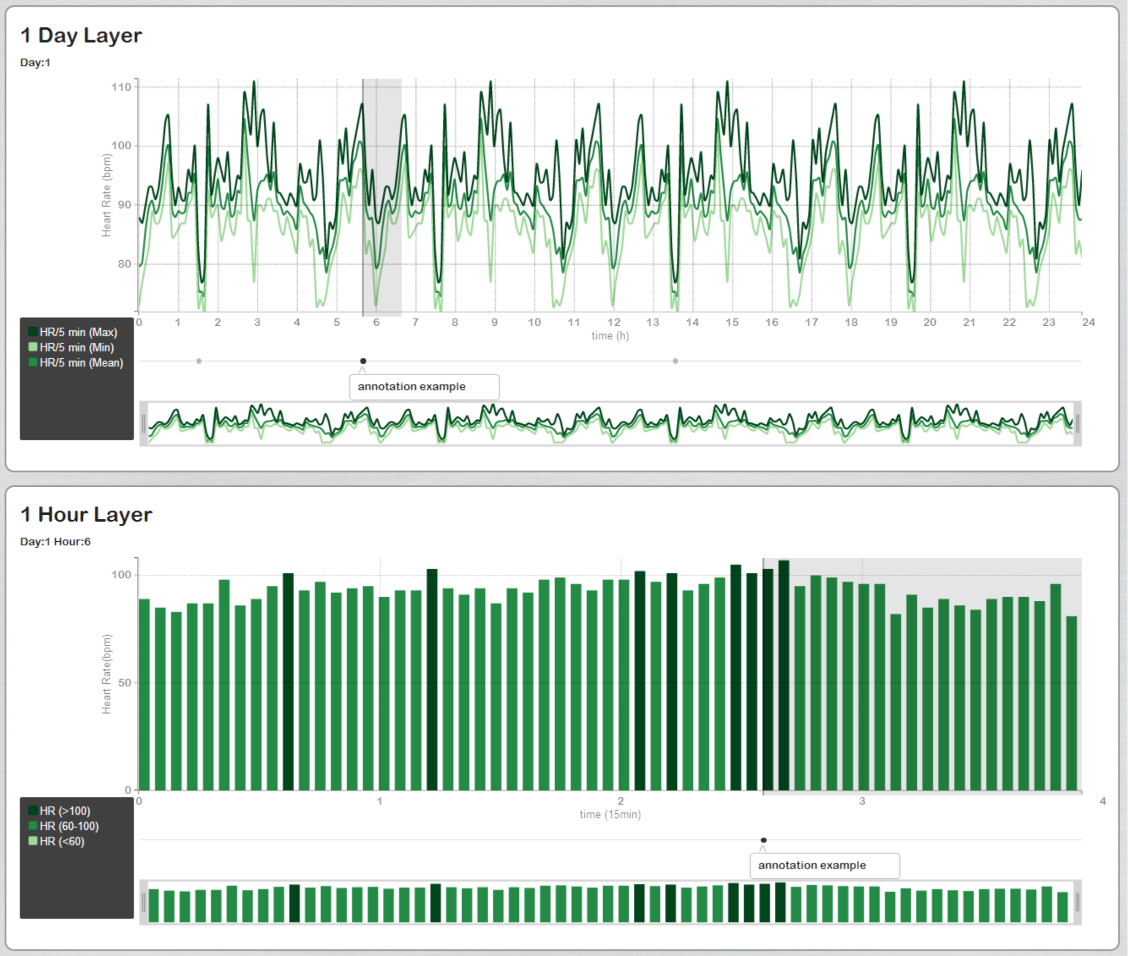


Figure 4.6: The selection of the sixth interval in the 1 day layer, i.e. the sixth hour of that day, will switch the data in the 1 hour layer in order to correspond to the data of the sixth hour. The 1 day layer presents three annotations, however only the second one belongs to the interval displayed at the 1 hour layer.

Each annotation is composed by a message and its positions of start and end (in seconds). The annotations are processed by a specific API where different algorithms were developed in order to handle the needed tasks. The approach followed sort the annotations by its start values, what enables an easiest access to the data during the reading task.

4.3.4 Further Analysis

The choice of key signal features is a fundamental method to discover important information hidden on the signal. Therefore, it is possible in the developed application requesting for a more detailed analysis. By clicking the "Analysis" item in the platform it is possible to select the layer from which the user needs more details. After the choice is made, a dialog box arises presenting the results. Figure 4.7 illustrates an example.

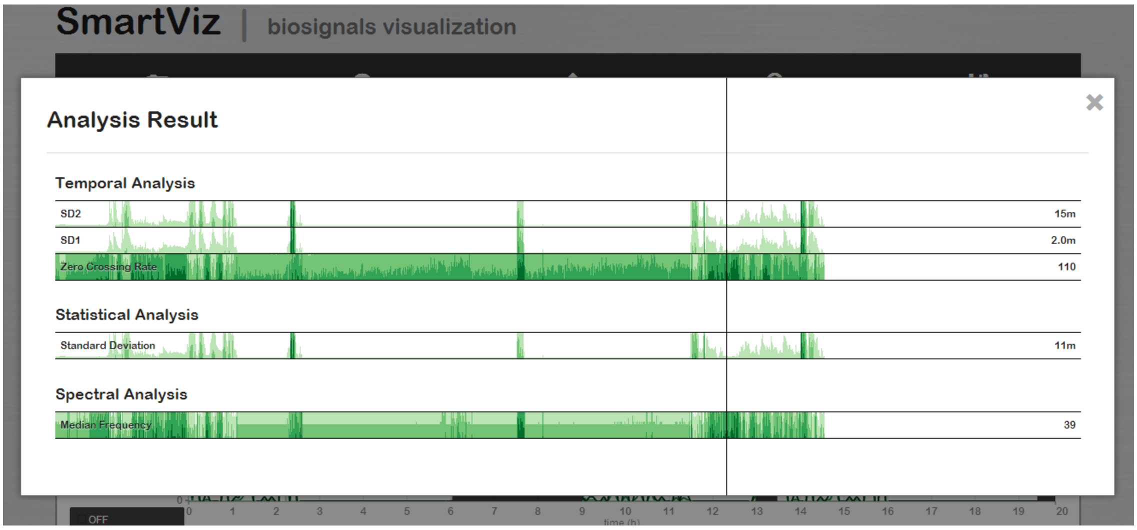


Figure 4.7: Further analysis result of the 1 hour layer from an EMG signal.

The features selected to analyse a specific layer depend on the biosignal that is being analysed and are discussed in Chapter 5. However, all the implemented features are grouped in the following domains:

- **Temporal domain:** Its features are computed through simple mathematical and statistical metrics, thus containing the basic information about the raw data. They are often used as pre-processing steps.
- **Statistical domain:** Although belonging to the temporal domain, it is often described separately. This domain representation presents an overall analysis of the signal magnitude.
- **Spectral domain:** Involves the frequency spectrum analysis and the examination of its dominant frequencies [18], [20].

As well as the *layers* dictionaries in the biosignals JSON file, a *features* dictionary with the selected features per biosignal was also created in the biosignals JSON file. Each one of these dictionaries is composed by three dictionaries, one per domain.

In order to improve the performance of the analysis API, the different features are also organised in a features JSON file. In this file the features are grouped by domain and each feature contains a dictionary with the keys presented in Table 4.3.

Table 4.3: Feature dictionary in features JSON file.

Key	**Description**
Description	Brief description of the feature
Imports	Module needed to execute the function
Function	Function that computes the feature
Parameters	Function inputs

As an example it is presented an excerpt of the features JSON file. Listing 4.2 represents the dictionary of the *Triangular Index* feature. This feature is a key in a temporal domain dictionary.

Listing 4.2: Excerpt of Triangular Index feature dictionary in features JSON file.

```
{
"Temporal": {
  "Triangular Index": {
    "description": "Computes the total number of all NN intervals divided by
                    the height of the histogram of all NN intervals measured on
                    a discrete scale with bins of 1/128 seconds.",
    "imports": "from temporalFeatures import triangular",
    "function": "triangular",
    "parameters": ["signal", "fs"],
  },
```

5

Case Studies

An adaptive model for different biosignals was one of the requirements of the developed application. Therefore, in this chapter the developed application is applied to three different biosignals. ECG, EMG and ACC were the three biosignals selected.

Firstly, the acquisition system used for recording these signals is described. The three biosignals are then described. The events selected to perform the needed abstractions are revealed and the features used to achieve a more detailed analysis are also presented. Finally, the results of an analysis with the developed application are discussed and illustrated.

5.1 Data Acquisition and Processing

The acquisition system consisted in a set of biomedical sensors, a wireless acquisition unit and a smartphone. A surface EMG sensor, a triaxial ACC sensor, and a 3 lead ECG sensor were used. The positions where the biomedical sensors were placed in the body are represented in Figure 5.1. The wireless acquisition unit was the bioPLUX research system and, as well as the sensors, was developed by PLUX.

The biosignals were acquired with a sampling frequency of 1000 Hz and with a 12 bit resolution. This data was sent via Bluetooth to the smartphone, saved in a text file and then converted to an HDF5 file.

The performed acquisitions are described in Table 5.1. All the acquisitions were carried out in a home atmosphere performing daily living activities.

After the acquisitions, the signals were processed.

In order to provide longer signals to visualize in the developed tool, the last signals were replicated or combined. The ECG and the ACC signals were replicated until they

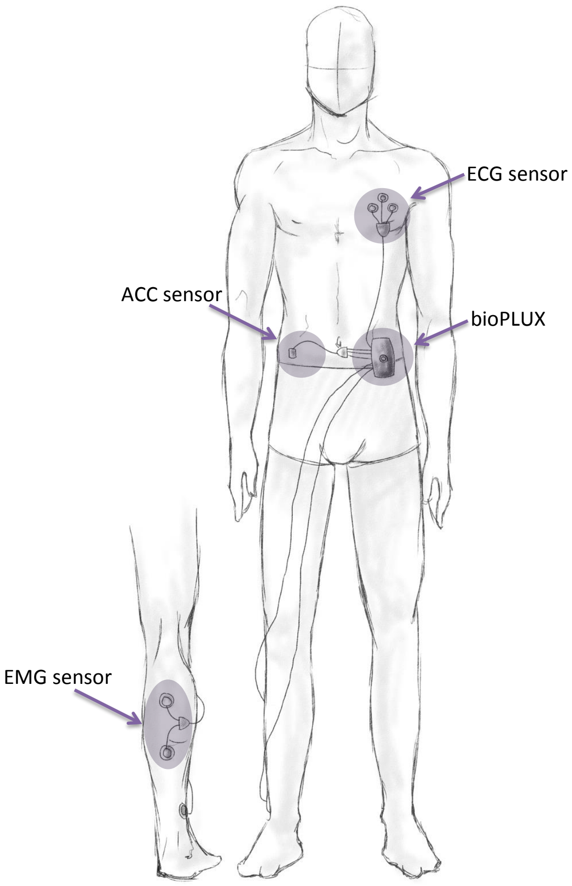


Figure 5.1: Placement of the biomedical sensors on the subject. While the EMG sensor was placed at the gastrocnemius muscles of the right leg, the ACC sensor was worn at the right hip and the ECG sensor was placed at the left side of the chest. The data acquired by the bioPLUX was sent via Bluetooth to a smartphone.

Table 5.1: Performed Acquisitions.

Acquisition	Biosignals Acquired	Acquisition Duration
1	ECG and ACC	6 hours
2	EMG	6 hours
3	EMG	4 hours

achieved the duration of 10 days, in the ECG case, and 2 days 9 hours 32 minutes and 55 seconds, in the ACC case. The EMG signal consisted in the combination of the two EMG acquisitions with time lapses where the acquisition was interrupted. Thus, the EMG signal was composed by the following sequence: acquisition 2, 3 hours of interrupted acquisition, acquisition 3, 30 minutes of interrupted acquisition and acquisition 2.

These files were finally visualized and analysed at the visualization tool.

5.2 Electrocardiography (ECG)

5.2.1 Events

In order to perform the abstraction described in section 4.1.2 it was considered one of the most important events in an ECG waveform. The R peak in the QRS complex is one of the most important to analyse [49] in an ECG waveform. For this reason the R peaks were the main events to consider in the developed tool.

Therefore, while the first group of events in the HDF5 file resulting from the processing contains the instants when these peaks occur, the second group of events in the HDF5 file contains the number of peaks per minute (the heart rate). The second group could easily be computed through the access of data in the first group. However, its implementation during the processing workflow allows a fast access to the data during the use of the platform, since it has to access a lower number of data samples and does not need to perform any other calculation.

The automated R peak detection was performed with the Pan and Tompkins detection algorithm [52].

Hence, from the highest to the lowest layer the represented events at the interface are: heart rate per 15 minutes, heart rate per 5 minutes, heart rate per minute, heart rate per second, subsampled signal, subsampled signal and raw signal.

5.2.2 Features Selection

The selection of features to provide a further analysis of the ECG signal was done considering the HRV whose standard features provide a vital source of information regarding the heart's activity. The HRV represents the variation of the intervals between heartbeats (RR intervals). Consequently, it is considered a derivation of the main event above considered.

The HRV features have been widely studied and are well documented in the literature. Actually, the literature indicates two major trends in them. While statistical features are used to characterize the distribution of heart periods, frequency features are usually used to relate mechanisms of the autonomic nervous system [3].

Table 5.2 presents the selected features for the analysis of ECG signals in the proposed visualization model [3], [29], [45]. These features are computed considering signal slices with 5 minute of duration [45] and can be computed for the 1 day, 1 hour and 15 minutes layers.

5.2.3 Results

Figures 5.2 and 5.3 present the result of one possible analysis of the ECG signal considered. Through the zoom, pan, the creation of annotations and the use of tooltips, it was possible to find a probable artefact caused by the chest muscles contraction. Double

Table 5.2: ECG Features.

Feature	Domain	Description
Triangular Index	Temporal	Total number of all RR intervals divided by the height of the histogram of all RR intervals measured on a discrete scale with bins of 1/128 seconds
SD	Temporal	Short and Long diameter of a Poincaré plot which describes the fast and slower HRV components
SDNN	Statistical	Standard deviation of all RR intervals
pNN50	Statistical	Percentage of pairs of adjacent RR intervals that differ by more than 50 ms
AVNN	Statistical	Average of all RR intervals
rMSSD	Statistical	Square root of the mean of the sum of the squares of differences between adjacent RR intervals
LF	Spectral	Total spectral power of all RR intervals between 0.04 and 0.15 Hz
HF	Spectral	Total spectral power of all RR intervals between 0.15 and 0.4 Hz
LFHF	Spectral	Ratio of low to high frequency power
VLF	Spectral	Total spectral power of all RR intervals between 0 to 0.04 Hz

clicking in the graphs allowed to switch the data in the layers below until the desired occurrence is displayed.

5.3 Electromyography (EMG)

5.3.1 Events

The event with most importance in the EMG analysis is therefore the signal activation (onset and offset times). However, the EMG signal is also often analysed by the root mean square value [60], [65].

Consequently, while the first group of events in the HDF5 file resulting from the processing contains the number of onsets per minute, the second group of events in the HDF5 file contains the root mean square values. The detection of the onset time of muscle activity is performed with the Teager-Kaiser energy operation method [73] and the root mean square values are computed with 900 samples windows.

Hence, from the highest to the lowest layer the represented events at the interface are: number of contractions per 15 minutes, number of contractions per 5 minutes, number of contractions per minute, root mean square, subsampled signal, subsampled signal and raw signal.

5.3.2 Features Selection

In order to perform the further analysis of the EMG signals, several features were selected taking into account its performance and usability. Table 5.3 presents the selected features. These features were computed with 3 seconds, 1 second and 0.5 seconds windows for the 1 hour, 15 minutes and 5 minutes layers respectively.

Table 5.3: EMG Features.

Feature	Domain	Description	Reference
Zero Crossing Rate	Temporal	Number of times that the signal changes from positive to negative or vice versa	[41], [57]
SD	Temporal	Short and Long diameter of a Poincarré plot which describes the faster and slower EMG components	[22]
Standard Deviation	Statistical	Dispersion from mean value	[11], [57]
Median Frequency	Spectral	Frequency where power spectral density reaches 50% of its distribution	[6], [11], [41]

5.3.3 Results

Figure 5.4 presents the result of one possible analysis of the EMG signal considered. Since the signal only has 20 hours of record, the first layer displayed is the 1 day layer. The last one, the 1 second layer, is not showed in Figure 5.4 since it does not present relevant information in this case.

The analysis of the first layer displayed enables to get the general intervals where there was or not acquisition and the activity and inactivity intervals where the muscle in study did or did not perform any contraction. With the analysis of the 1 hour layer it is possible to observe that the number of contractions is higher in the first 15 minutes of the hour in study. The remaining layers show the EMG signal contractions with more detail.

5.4 Accelerometry (ACC)

5.4.1 Events

In order to settle the abstraction for the ACC signals it was considered that these signals represent different activities performed by the subject. These activities can be classified as static states (lying, sitting or standing) and dynamic states (walking or running) [1], [18]. One of the most widely used and accepted feature to distinguish between the last two states is the Signal Magnitude Area (SMA). This feature is described by Equation

5.1 and was computed for each second of data. When the SMA value exceeded a pre-set threshold then the subject is classified as being in a dynamic state [12], [18], [34], [38], [40].

$$SMA = \frac{1}{t}(\int_0^t |x(t)|dt + \int_0^t |y(t)|dt + \int_0^t |z(t)|dt) \tag{5.1}$$

As a result, the first group of events in the HDF5 file resulting from the processing contains the SMA values per second.

Hence, from the highest to the lowest layer the represented events at the interface are: SMA per 15 minutes, SMA per 5 minutes, SMA per minute, SMA per second, subsampled signal, subsampled signal and raw signal.

5.4.2 Features Selection

In order to further analyse the ACC signal, several features were selected taking into account their performance and usability. Table 5.4 presents the selected features. These features were computed with the total acceleration signal [44] and with a 3 seconds, 1 second and 0.5 seconds windows for the 1 hour, 15 minutes and 5 minutes layers respectively.

Table 5.4: ACC Features.

Feature	Domain	Description	Reference
Autocorrelation	Temporal	Degree of correlation of a signal with itself	[44]
Root Mean Square	Statistical	Square root of the mean of the squares of the original values	[44]
Mean	Statistical	Signal central tendency	[18], [44]
Standard Deviation	Statistical	Dispersion from mean value	[18]
Median Frequency	Spectral	Frequency where power spectral density reaches 50% of its distribution	[44]
Fundamental Frequency	Spectral	The lowest frequency harmonic	[44]

5.4.3 Results

The result of the ACC signal analysis is similar to the EMG signal, however instead of activations, here the idea of static or dynamic state is presented. Figure 5.5 illustrates a result of one analysis of the ACC signal. This enables to quantify the amount of time that the subject remained in each state. Through the navigation in the lowest layers it is possible to distinguish different activities.

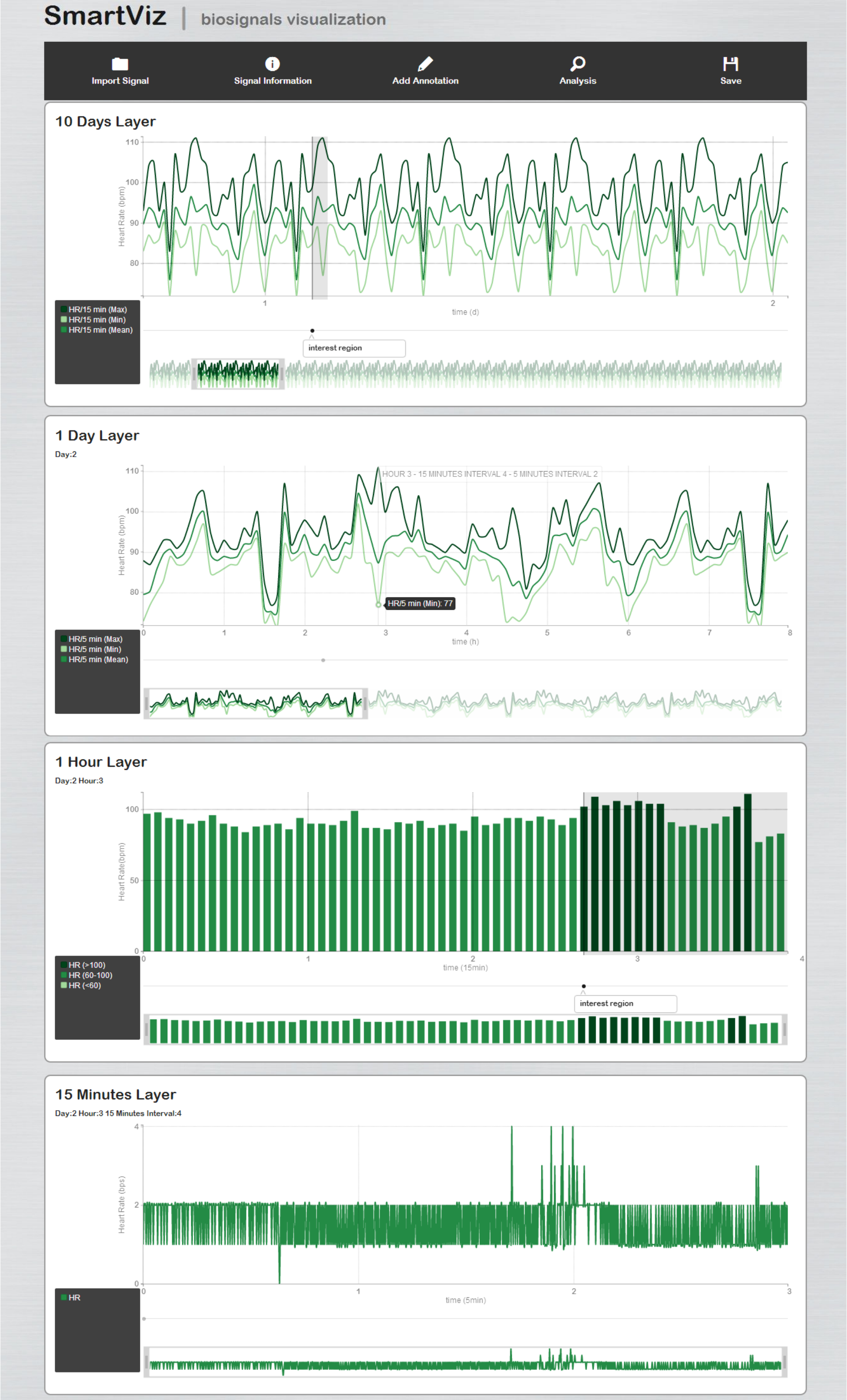


Figure 5.2: ECG signal analysis result. Through the zoom, pan, the creation of annotations, the use of tooltips and the navigation trough the layers it was possible to find a probable artefact caused by the chest muscles contraction.

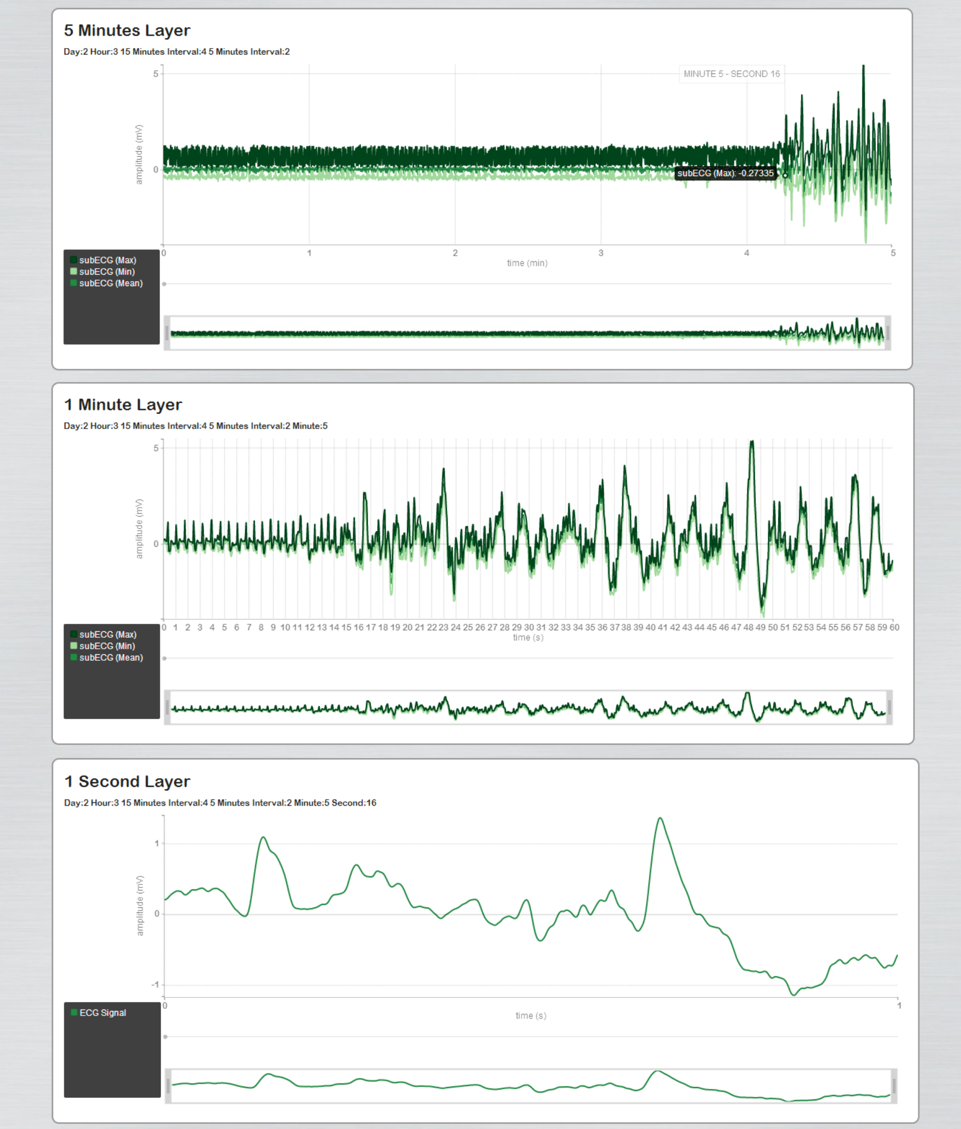


Figure 5.3: Previous Figure continuation.

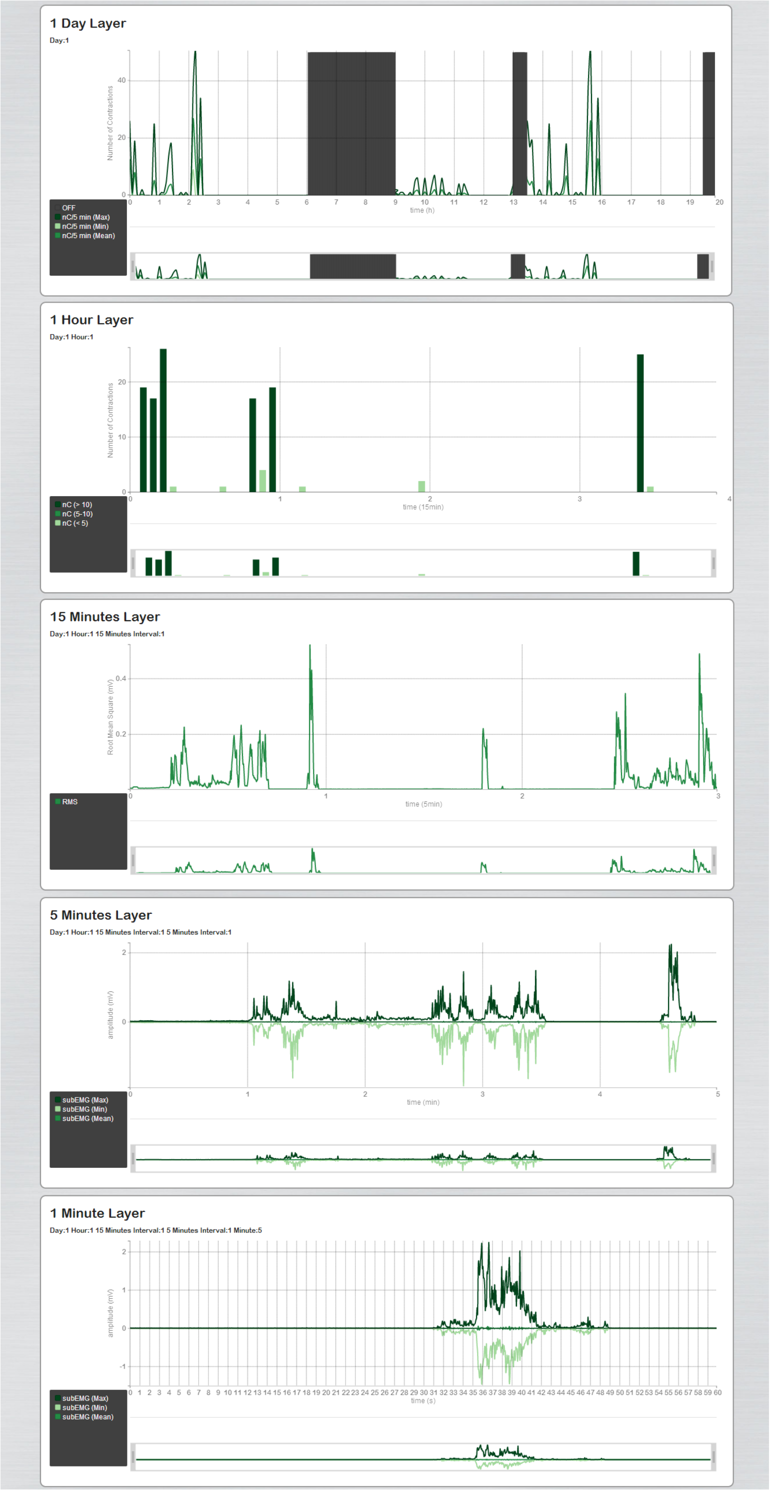


Figure 5.4: EMG signal analysis result. The analysis of the first layer displayed enables to get the general intervals where there was or not acquisition and the activity and inactivity intervals where the muscle in study did or did not perform any contraction. With the analysis of the 1 hour layer it is possible to observe that the number of contractions is higher in the first 15 minutes of the hour in study. The remaining layers show the EMG signal contractions with more detail.

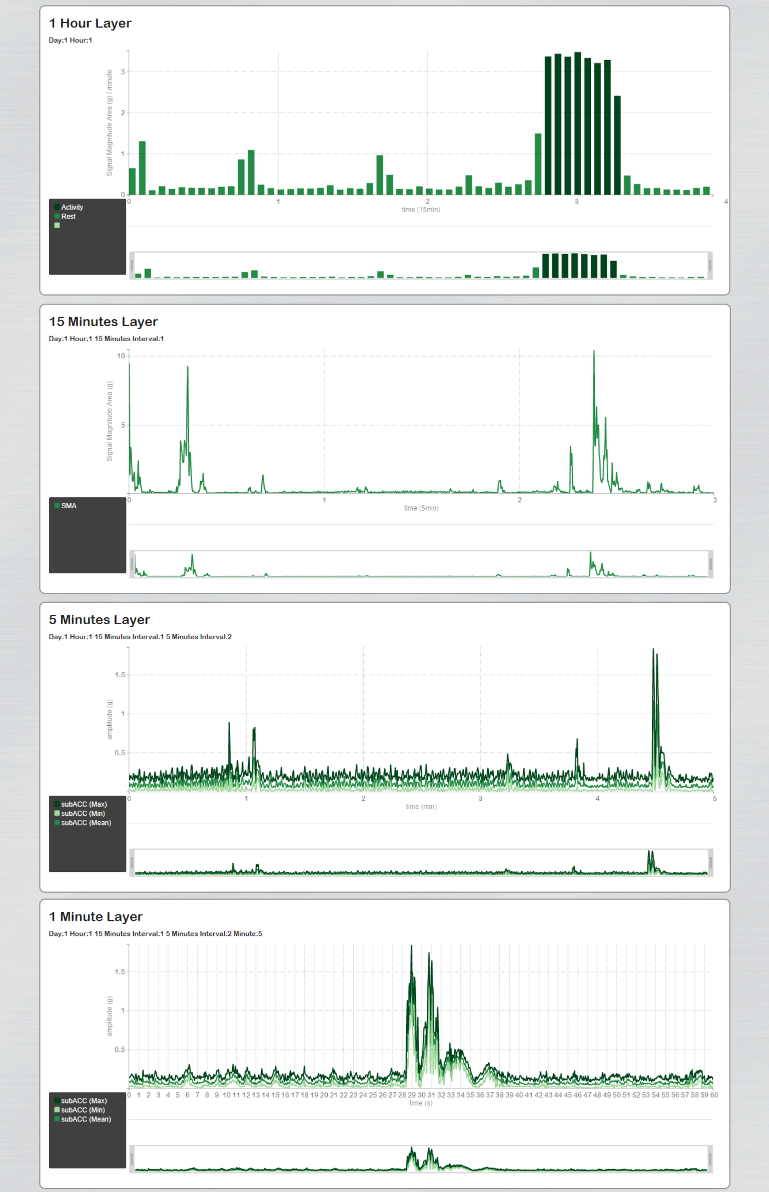


Figure 5.5: ACC signal analysis result. Here, it is possible to quantify the amount of time that the subject remained in each state. Through the navigation in the lowest layers it is possible to distinguish different activities.

6

Usability Evaluation

The development of the visualization application involved different stages already discussed. Finally, a first prototype was submitted to a usability evaluation with the aim of validating the developed work. In this chapter the selected test for assessing the usability of the system is revealed. The performed study is presented and the achieved results are discussed.

6.1 Usability Test Choice

In order to assess the usability of the developed work, a study was carried out with the SUS questionnaire. The SUS is a popular questionnaire developed by John Brooke in 1986, with the aim of providing a quick and global measure of the subjective perceptions of the usability of a system. This survey was developed according to the usability criteria defined by the ISO 9241-11 [9].

This scale has been the target of different studies that intended to validate it. Several studies have proved that SUS provides similar or more reliable results when compared with other surveys. In [7] it is achieved a coefficient alpha (measure of internal consistency) of 0.91 with a study of 2324 cases. This value indicates the high reliability of SUS. It was also proved that it can be applied regardless of the technology or system in study [7], [9]. Other studies conclude that not only usability but also learnability are measured by SUS [9].

6.2 The System Usability Scale (SUS) Questionnaire and its Score System

The SUS questionnaire consists of 10 items each having a five-point scale that ranges from *Strongly Disagree* (1) to *Strongly Agree* (5). There are five positive statements and five negative statements, which alternate in order to avoid response biases. The SUS questionnaire is presented in Table 6.1.

Table 6.1: The original SUS questionnaire.

	Statements
1	I think that I would like to use this system frequently.
2	I found the system unnecessarily complex.
3	I thought the system was easy to use.
4	I think that I would need the support of a technical person to be able to use this system.
5	I found the various functions in this system were well integrated.
6	I thought there was too much inconsistency in this system.
7	I would imagine that most people would learn to use this system very quickly.
8	I found the system very cumbersome to use.
9	I felt very confident using the system.
10	I needed to learn a lot of things before I could get going with this system.

The SUS scores are calculated considering first each item's score contribution that ranges from 0 to 4:

- for positively worded items (1, 3, 5, 7 and 9), the score contribution is the scale position minus 1.
- for negatively worded items (2, 4, 6, 8 and 10), the score contribution is 5 minus the scale position.

To get the overall SUS score, the sum of each item score contribution is multiplied by 2.5. Thus, the final scores range from 0 to 100. It is important to take into consideration that the individual statements are not supposed to have analytic value in themselves.

The typical minimum acceptable score is 70. While anything below this value presents usability issues that are cause of concern, systems scored in the 80s are considered good [9].

6.3 Performed Evaluation

The usability study performed included 10 participants with ages between 23 and 28 years. Although the number of participants could be considered a small sample, it was proved in previous studies that the SUS enables to get a confident measure of the perceived usability with a small group of participants [9].

The participants of the study belong to the final target of users of the developed tool and no one had any experience with the system. The application was developed for researchers, doctors and health technicians. However, this study only comprised biosignals researchers (in this case - biomedical engineers).

A short set of instructions were given at the beginning of the test. It was asked to the participants to:

- check the box that reflects their immediate response to each statement.
- do not dwell too long on any statement.
- simply answer 3 when not knowing how to respond.

The participants were also instructed to use the system freely, opening a file and exploring it with the available interaction tools.

In the evaluation carried out the word "cumbersome" in question 8 was replaced by the word "awkward". This replacement is supported by different studies that proved the word "cumbersome" is not clear enough for non-English speakers, and also its replacement does not change the final results [9].

6.4 Results and Discussion

The average score of the performed tests was 80.0. The minimum score was 70.0 and the maximum score was 90.0. These results indicate a good system with a high perceived usability.

In the end of the test the majority of the participants showed concerned about the questions 4 and 10. They agreed that some basic concepts were needed in order to perform a suitable analysis, however, these concepts should already be known by the final user. They also agreed that a necessary guidance should be performed at the first use of the system, still, it would not be necessary after this first use.

Nevertheless, the participants stated that the system is intuitive, useful and suitable. They also indicated that they were likely to recommend it.

7

Conclusions

This final chapter summarizes the developed work and presents an overview of its general results and accomplishments. It also outlines the future goals of the developed research.

7.1 General Contributions and Results

The main contribution of this work is the introduction of a novel visualization interface which allows the professionals, who work with biosignals, to get insight into the large data sets acquired by the enhanced monitoring systems that have been introduced in the last years. The developed application enables the draw of conclusions by its users and their direct interaction with the data, overcoming the demand of a new biosignals analysis platform.

To achieve the proposed objectives, an investigation of visualization techniques that support large data sets was first performed. Considering the advantages of web-based visualizations, a research for toolkits and plugins for different rendering techniques was also done.

Thereafter, the application framework was designed. The requirements of the visualization tool were stablished and the use of a web server was justified. The server was implemented and the layout of the client-server model was developed. Lastly, the better data format to manage large scale data sets was selected and the created architecture for these files was presented.

A visualization model was followed and its central process, the visualization process, was first developed. Here, the domain problem was settled and the possible abstractions were studied, then the suitable approach was followed. The application layout was also defined, which comprises a multi-level visualization and the usage of the best visualization rendering elements and techniques. The following process studied was the perception and cognition one. A user friendly and intuitive design was demanding. The colour-blind effect was taken into account and a responsive design was also integrated. The interactive exploration process was finally considered. Different interaction techniques were implemented and special attention was given to annotation handle due to its importance in the data mining process.

Three case studies were considered. An acquisition system was designed and ECG, EMG and ACC signals were acquired. Each one of these biosignals was introduced and analysed with the approach followed by the visualization model implemented.

The required processing algorithms were developed and the biosignals were visualized in the designed visualization application. Several performance tests were carried out in order to ensure the correct application response.

Finally, a usability evaluation was done. The usability of the system was assessed with the SUS questionnaire and with 10 participants, getting an average score of 80.0 (a minimum of 70.0 and a maximum of 90.0). This score indicates a good system with a high usability performance.

Therefore, the described achievements contribute for the biosignals analysis improvement (specially in big data cases). The developed application interconnects its users with the data, without users having to deal with signal processing algorithms directly what substantially improves their experience.

7.2 Future Work

Although the developed application has proved to be a suitable approach for the intended objective, further investigation should be performed.

- **More case studies:** The case studies presented in this work represent the two possible types of biosignals events, cyclic and sporadic. Three of the most used biosignals in biomedical research were chosen. However, another set of biosignals such as Electroencephalography (EEG), Electrodermal Activity (EDA) or Respiratory (RESP) signals should be considered and analysed in the developed application enabling its use in more research studies.

- **WebGL rendering:** The use of this type of rendering elements will improve the visualization application, however, for reasons already stated it could not be implemented at the time of this work. Therefore, the continuous search for WebGL rendering tools should be carried on and when possible these elements should replace the implemented SVG elements.

- **Additional usability validation:** Despite the usability test enabling the use of only a small sample of the final users, the performed tests were carried out only by biomedical engineers researchers. Since the final users target is also composed of health science professionals, such as doctors and technicians, the test should be also performed by users belonging to this group.

- **Adapt the application for patients:** Informing patients about their health state, in some clinical cases, could help their recovery. These applications motivate the patients and improve their treatments.
- **Integration of the processing in the acquisition system (real time processing):** The use of the developed application system presupposes a previous processing of the acquired biosignals. However, if the acquisition system performed the vast majority of signal processing in the acquisition moment, a considerably amount of time spent in processing could be reduced [38].

Bibliography

[1] ActiGraph Software Department, *ActiLife 6 user's manual*, Apr. 2012.

[2] W. Aigner, S. Miksch, and H. Schumann, *Visualization of Time-Oriented Data*, 1st ed. Springer.

[3] M. Alemu, S. P. Arjunan, and D. K. Kumar, "Observing exercise induced heart rate variability response", in *Biosignals and Biorobotics Conference (BRC), 2011 ISSNIP*, IEEE, 2011, pp. 1–6.

[4] E. Alpaydin, *Introduction to Machine Learning*, 2nd ed. Cambridge, Massachusetts: The MIT Press, 2010.

[5] Andrew Collette, *Python and HDF5*. O'Reilly, 2013.

[6] S. P. Arjunan, D. Kumar, K. Wheeler, H. Shimada, and G. Naik, "Spectral properties of surface EMG and muscle conduction velocity: a study based on sEMG model", in *Biosignals and Biorobotics Conference (BRC), 2011 ISSNIP*, IEEE, 2011, 1–4.

[7] A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the system usability scale", *International Journal of Human-Computer Interaction*, vol. 24, no. 6, pp. 574–594, Jul. 2008.

[8] D. M. Beazley, *Python Essential Reference*, 4th ed., ser. Developer's Library. Pearson Education, 2009.

[9] J. Brooke, "SUS: a retrospective", *Journal of Usability Studies*, vol. 8, no. 2, 29–40, 2013.

[10] P. Buono, A. Aris, C. Plaisant, A. Khella, and B. Shneiderman, "Interactive pattern search in time series", in *Electronic Imaging 2005*, 2005, 175–186.

[11] T. V. Camata, J. L. Dantas, T. Abrão, M. A. Brunetto, A. C. Moraes, and L. R. Altimari, "Fourier and wavelet spectral analysis of EMG signals in supramaximal constant load dynamic exercise", in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, IEEE, 2010, 1364–1367.

[12] J. L. Carus, V. Pelaez, S. Garcia, M. A. Fernandez, G. Diaz, and E. Alvarez, "A non-invasive and autonomous physical activity measurement system for the elderly", IEEE, Dec. 2013, pp. 619–624.

[13] S. Cheng, Z. Jiang, Q. Qi, S. Li, and X. Meng, "The polar parallel coordinates method for time-series data visualization", IEEE, Aug. 2012, pp. 179–182.

[14] *Datavisualization.ch selected tools*. [Online]. Available: http://selection.datavisualization.ch/ (visited on 01/20/2014).

[15] J. Dubois, L. Cottret, A. Ghozlane, D. Auber, F. Bringaud, P. Thebault, F. Jourdan, and R. Bourqui, "Systrip: a visual environment for the investigation of time-series data in the context of metabolic networks", IEEE, Jul. 2012, pp. 204–213.

[16] J. Duckett, *HTML and CSS: design and build websites*, 1st ed. John Wiley & Sons, 2011.

[17] ——, *JavaScript and JQuery: Interactive Front-End Web Development*, 1st ed. John Wiley & Sons, 2014.

[18] D. Figo, P. C. Diniz, D. R. Ferreira, and J. M. P. Cardoso, "Preprocessing techniques for context recognition from accelerometer data", en, *Personal and Ubiquitous Computing*, vol. 14, no. 7, pp. 645–662, Oct. 2010.

[19] T.-c. Fu, "A review on time series data mining", *Engineering Applications of Artificial Intelligence*, vol. 24, no. 1, pp. 164–181, Feb. 2011.

[20] A. Godfrey, R. Conway, D. Meagher, and G. ÓLaighin, "Direct measurement of human movement by accelerometry", *Medical Engineering & Physics*, vol. 30, no. 10, pp. 1364–1386, Dec. 2008.

[21] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet : components of a new research resource for complex physiologic signals", *Circulation*, vol. 101, no. 23, e215–e220, Jun. 2000.

[22] A. K. Golnska, "Poincaré plots in analysis of selected biomedical signals", *Studies in Logic, Grammar and Rhetoric*, vol. 35, no. 1, Jan. 2013.

[23] R. Gomes, C. Cavaco, H. Gamboa, and R. Matias, "Cloud parallel computing for large scale biosignal analysis and event driven visualization", *Submitted to Biomedical Signal Processing and Control*, 2014.

[24] R. Gomes, N. Nunes, J. Sousa, and H. Gamboa, "Long term biosignals visualization and processing", in *BIOSIGNALS*, 2012, 402–405.

[25] R. Goya-Esteban, I. Mora-Jimenez, J. L. Rojo-Alvarez, O. Barquero-Pérez, S. Manzano-Martinez, F. Pastor-Pérez, D. Pascual-Figal, and A. Garcia-Alberola, "Rhythmometric analysis of heart rate variability indices during long term monitoring", in *Computers in Cardiology, 2009*, IEEE, 2009, 57–60.

[26] M. Harrower and C. A. Brewer, "ColorBrewer.org: an online tool for selecting colour schemes for maps", en, *The Cartographic Journal*, vol. 40, no. 1, pp. 27–37, Jun. 2003.

[27] J. Heer, N. Kong, and M. Agrawala, "Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations", in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2009, 1303–1312.

[28] T. Hilbel, R. L. Lux, J. Dietzsch, M. Schliephake, and H. A. Katus, "Performance and productivity benefits using multi-core processors for the analysis of digital long-term ECG recordings", in *Computers in Cardiology, 2008*, IEEE, 2008, 1069–1072.

[29] R. A. Hoshi, C. M. Pastre, L. C. M. Vanderlei, and M. F. Godoy, "Poincare plot indexes of heart rate variability: relationships with other nonlinear variables", en, *Autonomic Neuroscience*, vol. 177, no. 2, pp. 271–274, Oct. 2013.

[30] D. L. Hudson and M. E. Cohen, "Intelligent analysis of biosignals", in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, 2006, 323–326.

[31] J. Hunter, D. Dale, E. Firing, and M. Droettboom, *Matplotlib, release 1.3.1*, Oct. 2013.

[32] N. P. N. Iliinsky and J. Steele, *Designing data visualizations*, English. Farnham: O'Reilly, 2011.

[33] W. Javed, B. McDonnel, and N. Elmqvist, "Graphical perception of multiple time series", *Visualization and Computer Graphics, IEEE Transactions on*, vol. 16, no. 6, 927–934, 2010.

[34] D.-U. Jeong, S.-J. Kim, and W.-Y. Chung, "Classification of posture and movement using a 3-axis accelerometer", IEEE, Nov. 2007, pp. 837–844.

[35] Z. Jiang, S. Cheng, X. Meng, and Z. Zhang, "Reseach on time-series data visualization method based on parameterized parallel coordinates and color mapping fuction", English, in *ICSAI 2012*, New York [u.a.], 2012.

[36] H.-C. Jung, J.-H. Moon, D.-H. Baek, J.-H. Lee, Y.-Y. Choi, J.-S. Hong, and S.-H. Lee, "CNT/PDMS composite flexible dry electrodesfor long-term ECG monitoring", *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1472–1479, May 2012.

[37] E. Kaniusas, *Biomedical Signals and Sensors I*, English, ser. Biological and Medical Physics, Biomedical Engineering. Springer Berlin Heidelberg, 2012, pp. 1–26.

[38] D. Karantonis, M. Narayanan, M. Mathie, N. Lovell, and B. Celler, "Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring", en, *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 1, pp. 156–167, Jan. 2006.

[39] D. A. Keim, "Information visualization and visual data mining", *IEEE Transactions on visualization and computer graphics*, vol. 7, no. 1, pp. 100–107, 2002.

[40] C.-C. Lan, Y.-H. Hsueh, and R.-Y. Hu, "Real-time fall detecting system using a triaxial accelerometer for home care", IEEE, May 2012, pp. 1077–1080.

[41] Y. Lee and others, "Spatiotemporal analysis of EMG signals for muscle rehabilitation monitoring system", in *Consumer Electronics (GCCE), 2013 IEEE 2nd Global Conference on*, IEEE, 2013, 1–2.

[42] J. Lin, E. Keogh, S. Lonardi, J. P. Lankford, and D. M. Nystrom, "VizTree: a tool for visually mining and monitoring massive time series databases", in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, 2004, 1269–1272.

[43] J. Lin, E. Keogh, S. Lonardi, and L. Wei, "Experiencing SAX: a novel symbolic representation of time series", *Data Mining and Knowledge Discovery*, vol. 15, pp. 107–144, Oct. 2007.

[44] I. Machado, R. Gomes, H. Gamboa, and V. Paixão, "Human activity recognition from triaxial accelerometer data - feature extraction and selection methods for clustering of physical activities", in *BIOSIGNALS*.

[45] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, and P. J. Schwartz, "Heart rate variability standards of measurement, physiological interpretation, and clinical use", *European heart journal*, vol. 17, no. 3, pp. 354–381, 1996.

[46] T. F. L. de Medeiros, A. B. Cavalvanti, E. V. C. de Lima Borges, I. L. P. Andrezza, B. E. S. Cavalcante, and L. V. Batista, "Heart arrhythmia classification using the PPM algorithm", in *Biosignals and Biorobotics Conference (BRC), 2011 ISSNIP*, IEEE, 2011, 1–5.

[47] T. Munzner, "Visualization", in *Fundamentals of Computer Graphics*, 2009, pp. 675–720.

[48] S. Murray, *Interactive data visualization for the web: [an introduction to designing with D3]*, English. Sebastopol, CA: O'Reilly, 2013.

[49] N. Neophytou, A. Kyriakides, and C. Pitris, "ECG analysis in the time-frequency domain", in *Bioinformatics & Bioengineering (BIBE), 2012 IEEE 12th International Conference on*, IEEE, 2012, 80–84.

[50] A. V. Nguyen, R. Wynden, and Y. Sun, "HBase, MapReduce, and integrated data visualization for processing clinical signal data.", in *AAAI Spring Symposium: Computational Physiology*, 2011.

[51] M. Pakhira, "Requirements of a graphical system", in *Computer Graphics Multimedia and Animation*, 2nd, PHI Learning Pvt. Ltd., 2010.

[52] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm.", *IEEE Trans Biomed Eng*, vol. 32, no. 3, pp. 230–236, Mar. 1985.

[53] Y.-T. Peng, C.-Y. Lin, M.-T. Sun, and C. A. Landis, "Multimodality sensor system for long-term sleep quality monitoring", *IEEE Transactions on Biomedical Circuits and Systems*, vol. 1, no. 3, pp. 217–227, Sep. 2007.

[54] R. P. Radecki and M. A. Medow, "Cognitive debiasing through sparklines in clinical data displays", in *AMIA Annu Symp Proc*, vol. 1085, 2007.

[55] C. A. Ratanamahatana, J. Lin, D. Gunopulos, E. Keogh, M. Vlachos, and G. Das, "Mining time series data", in *Data Mining and Knowledge Discovery Handbook*, Springer, 2010, 1049–1077.

[56] H. Reijner, "The development of the horizon graph", in *Electronic Proceedings of the VisWeek Workshop From Theory to Practice: Design, Vision and Visualization*, 2008.

[57] S. M. Rissanen, M. Kankaanpaa, M. P. Tarvainen, V. Novak, P. Novak, K. Hu, B. Manor, O. Airaksinen, and P. A. Karjalainen, "Analysis of EMG and acceleration signals for quantifying the effects of deep brain stimulation in parkinson disease", *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 9, pp. 2545–2553, Sep. 2011.

[58] C. Rossant and K. D. Harris, "Hardware-accelerated interactive data visualization for neuroscience in python", *Frontiers in Neuroinformatics*, vol. 7, 2013.

[59] J. Sedlak, D. Spulak, R. Cmejla, R. Bacakova, M. Chrastkova, and B. Kracmar, "Segmentation of surface EMG signals", in *Applied Electronics (AE), 2013 International Conference on*, IEEE, 2013, 1–4.

[60] S. B. Soares, R. R. Coelho, and J. Nadal, "The use of cross correlation function in onset detection of electromyographic signals", in *Biosignals and Biorobotics Conference (BRC), 2013 ISSNIP*, IEEE, 2013, 1–5.

[61] C. Tominski, J. Abello, and H. Schumann, "Axes-based visualizations with radial layouts", in *Proceedings of the 2004 ACM symposium on Applied computing*, 2004, 1242–1247.

[62] E. Tufte, *Beautiful Evidence*. Graphics Pr, 2006.

[63] J. J. Van Wijk, "The value of visualization", in *Visualization, 2005. VIS 05. IEEE*, 2005, 79–86.

[64] J. J. Van Wijk and E. R. Van Selow, "Cluster and calendar based visualization of time series data", in *Information Visualization, 1999.(Info Vis' 99) Proceedings. 1999 IEEE Symposium on*, 1999, 4–9.

[65] T. M. Vieira, R. Merletti, and L. Mesin, "Automatic segmentation of surface EMG images: improving the estimation of neuromuscular activity", en, *Journal of Biomechanics*, vol. 43, no. 11, pp. 2149–2158, Aug. 2010.

[66] S. de Waele, G.-J. de Vries, and M. Jager, "Experiences with adaptive statistical models for biosignals in daily life", in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, IEEE, 2009, 1–6. (visited on 08/17/2014).

[67] V. Wang, F. Salim, and P. Moskovits, "Introduction to HTML5 WebSocket", in *The Definitive Guide to HTML5 WebSocket*, Springer, 2013, 1–12.

[68] P. Warden, *Big Data Glossary*, English. Sebastopol: O'Reilly Media, 2011.

[69] M. Weber, M. Alexa, and W. Müller, "Visualizing time-series on spirals.", in *Infovis*, vol. 1, 2001, 7–14.

[70] Welch Allyn, *Expert holter software system - directions for use*.

[71] L. Wilkinson, *The Grammar of Graphics*, English. New York: Springer, 2005.

[72] J. L. Williams, *Learning HTML5 Game Programming: A Hands-on Guide to Building Online Games Using Canvas, SVG, and WebGL*, English, ser. Addison-Wesley Learning Series. Pearson, 2011.

[73] L. Xiaoyan, Z. Ping, and A. Alexander, "Teager–kaiser energy operation of surface EMG improves muscle activity onset detection", *Annals of Biomedical Engineering*, vol. 35, no. 9,

[74] C.-C. Yang and Y.-L. Hsu, "A review of accelerometry-based wearable motion detectors for physical activity monitoring", *Sensors*, vol. 10, no. 8, pp. 7772–7788, Aug. 2010.

A

Publications

During the development of this work some publications were submitted.

The publication entitled “New Visualization Model for Large Scale Biosignals Analysis” was accepted to the *BIOSIGNALS 2015* conference of the “8th International Joint Conference on Biomedical Engineering Systems and Technologies” (BIOSTEC 2015).

The publication entitled “Cloud parallel computing for large scale biosignal analysis and event driven visualization” was submitted to the journal “Biomedical Signal Processing and Control” from Elsevier.

Finally, a publication entitled “A reliable classification system for neuromusculoskeletal gait disorders" was accepted in the conference “Clinical Movement Analysis World Conference 2014”.

New Visualization Model for Large Scale Biosignals Analysis

Catarina Cavaco[1], Ricardo Gomes[1], Hugo Gamboa[1,2] and Ricardo Matias[3,4]
[1] *CEFITEC, Physics Department, Faculdade de Ciências e Tecnologias - Universidade Nova de Lisboa, Lisbon, Portugal*
[2] *PLUX Wireless Biosignals, Lisbon, Portugal*
[3] *School of Health Care, Polytechnic Institute of Setúbal, Setúbal, Portugal*
[4] *Neuromechanics Research Group, Interdisciplinary Centre for the Study of Human Performance (CIPER), Faculdade de Motricidade Humana - Universidade de Lisboa, Lisbon, Portugal*
cat.cavaco@gmail.com, ricardo.baptista.gomes@gmail.com, hgamboa@plux.info, ricardo.matias@ess.ips.pt

Keywords: Long-term Biosignals, Big Data, Biosignals Visualization, Biosignals Annotation, Medical Monitoring.

Abstract: The development of new resources in the medical field, such as wearable sensors, allowed the improvement of biosignals monitoring. Acquired data is then an important source of information to clinicians and researchers. Thus, extracting useful information from data is a task of the greatest importance that involves a variety of concepts and methods, such as data visualization and machine learning. However, these methods present several limitations mainly when dealing with big data. In this paper we present an innovative web-based application for biosignals visualization and annotation in a fast and user friendly way overcoming the detected limitations. Three case studies are presented and a usability study supports the reliability of the implemented work.

1 INTRODUCTION

The technological innovation in medical systems has been of the utmost importance in the monitoring improvement of human body signals, so-called biosignals. There are several types of biosignals resulting of the electrical, magnetic, chemical or mechanical activity during biological events such as heart beat or muscle activity. They can also be classified considering their nature, application or their characteristics (Kaniusas, 2012).

Biosignals monitoring can be done through the use of non-invasive wearable sensors which combined with systems allow the storage of the data acquired. Relevant information can then be extracted from this data to support clinicians and researchers decision-making, as well as to inform patients. Therefore, to achieve the goal of extracting relevant information from the data, a variety of concepts and methods, such as data visualization or machine learning are involved (Aigner et al., 2011).

Since humans main input sense is visual, data visualization is considered essential in signal analysis (Aigner et al., 2011). The integration of the human visual perception with the current massive computational capacities results in this concept which supports the examining, understanding and transmitting of the vital information carried by a signal (Iliinsky and Steele, 2011). However, data visualization presents some limitations that have to be considered. Besides the computation capacity limited by the memory and time to run an algorithm, the display is restricted by the number of pixels available to show the data. On the human side, the limitations comprise the human perceptual and cognitive capacities which can result in incorrect data interpretations both in time and space (too fast or too dense for the correct perception) (Munzner, 2009).

Machine learning algorithms, applied for optimization of performance criteria, can follow supervised or semi-supervised approaches that depend on the use of labelled data (Alpaydin, 2010). For this reason, this task has to be as most accurately as possible. Nevertheless, the annotation of the biosignals is a demanding task that depends on human subjective intervention and requires specific knowledge.

The demand for studies that result in large amounts of data, such as sleep analysis or neuromuscular diseases monitoring, is increasing. In this case, difficulties become evident. Dealing with massive sources of data, which can be considered big data, increase the complexity of the problems described before and its processing with traditional applications presents several limitations (Hudson and

Cohen, 2006).

Considering the extreme importance of biosignals analysis and the outlined hurdles, the aim of this work was to present an innovating solution for large dataset visualization and annotation in the context of biosignals. This was achieved through the better conjugation between processing and storage capabilities of computers and the visual, creative and knowledge capabilities of humans. In this study we present a novel web-based application for biosignals visualization and annotation in a fast and user friendly way.

The remainder of the paper is organised as follows. Section 2 reviews related work on visualization and annotation of large datasets of biosignals. In section 3, it is presented the application framework. Section 4 refers to the developed visualization model and Section 5 presents the case studies considered. Section 6 presents the usability study carried out. Finally, in section 7 the main conclusions are drawn and directions for further research are given.

2 RELATED WORK

The emergence of a variety of visualization tools has aimed to counter the gap verified between our ability to collect and store data, and to analyse it. Cluster and Calendar based Visualization tool (Van Wijk and Van Selow, 1999), TimeSearcher2 (Buono et al., 2005) and VizTree (Lin et al., 2004) are three examples of available applications, considering different analytical methods, that excel in research of applications for commercial, academic or research use.

The high concern in clinical systems for improving the physical and psychological wellness has resulted in the advent of crucial systems specifics for biosignals. The PhysioNet (Goldberger et al., 2000) offers free web access to a large biosignals databases that can include ground-truth information, and it also comprises a wide collection of software for viewing and analysing biosignals. The OpenSignals software enables the visualization and analysis of the biosignals acquired by a wearable hub that along with this software constitute the biosignalsplux system (PLUX, 2014), an advanced biosignal monitoring platform for sports and biomedical research. The ActiLife6 is the visualization software that integrates the ActiGraph system (ActiGraph, 2012), the most used actigraphy monitoring system in research and clinical trials involving physical activity and sleep assessment. In cardiology, one of the most common examinations is the Holter, an ambulatory Electrocardiography (ECG) for a minimum 24-hour period, conducted with the purpose of screening for ECG disturbances. The Welch Allyn Holter System (Welch Allyn, 2007) is one of the available systems to perform Holter examinations.

Notwithstanding the high development of visualization tools, they face yet some issues, particularly when dealing with big data. The possible integration of applications in real life platforms to monitoring diseases increases the demand for novel solutions.

3 APPLICATION FRAMEWORK

3.1 System Requirements

The development of the visualization application took into account some base requirements. Therefore, the developed application had to:

- be applied to all kind of biosignals
- enable the possibility to explore up to 10 days of continuous acquisitions
- show the time lapses where the signal acquisition was interrupted
- allow the handling of annotations
- present a fast and user friendly interface

Lastly, the proposed model had to represent a commitment between usability and performance, allowing the user to analyse a biosignal without having to deal with signal processing algorithms directly.

3.2 System Architecture

The implemented visualization tool is a web-based platform. This decision is justified by the fact that the web standards currently available provide some of the best tools for the creation of rich graphical user interfaces and it also eliminates complex installation and configuration procedures.

A client-server model was developed. The local server was implemented in Python language (Beazley, 2009) and the communication between the visualization platform and the server is done with WebSockets (Wang et al., 2013). The client sends request messages to the server which responds by serving a message containing the requested information. These messages are on JavaScript Object Notation (JSON) structure. The flow of information is schematically represented in Figure 1. The requests sent by the browser intend to get the signal parameters or specific data intervals, to handle annotations and to access a detailed analysis of a specific interval. Considering these requests, it was implemented different python Application Programming Interfaces (APIs).

The generated responses are then interpreted by a series of JavaScript libraries and the results dynamically displayed on the page.

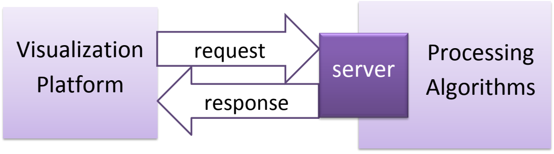


Figure 1: Information Flow.

Through the use of Python language to access and manipulate the data and JavaScript language (Duckett, 2014) to present the data and deal with the user-interface tasks, a highest performance is reached. While Hypertext Markup Language 5 (HTML5) (Murray, 2013) and Cascading Style Sheets 3 (CSS3) (Murray, 2013) allow the definition of the overall layout of the platform, the jQuery library (Duckett, 2014) enable to handle user interface events.

3.3 Data Architecture

The manipulation of very large amounts of data must involve suitable data format for storage and access this data. Although presenting some limitations, the Hierarchical Data Format 5 (HDF5) has proven to be the best option for the intended task (Gomes et al., 2012).

Before submitting a record file to the visualization tool it should be previously processed. The processing API involves filtering, subsampling, and events detection. As a result, different groups of data are created in the HDF5 file. Each group is organised by *n* different datasets that correspond to the *n* days of the record. The previous computation of some variables will speed up the visualization since the APIs previously referred only access the data instead of computing it (an example is later discussed).

4 VISUALIZATION MODEL

A generic model for visualization is proposed in (Van Wijk, 2005). In this model the central process is the visualization where data is transformed into a time varying image according to a specification. Then, the resulting image is perceived by a user, leading to an increase of users knowledge. Finally, the interactive exploration of the image enables to adapt the previous specification based on the current knowledge in order to further explore the data. Figure 2 schematically represents this model.

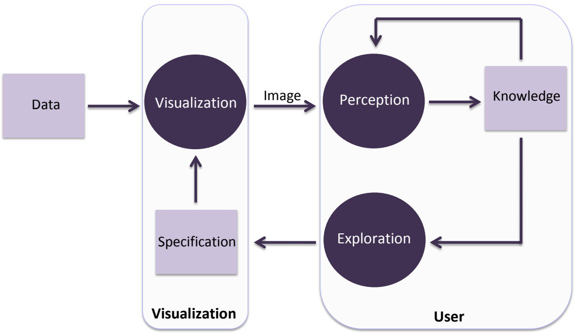


Figure 2: Simple model of visualization. While circles denote processes, boxes denote containers. Adapted from (Van Wijk, 2005).

4.1 Visualization Process

What, why and how are three questions that have to be answered when developing or choosing a visualization method (Aigner et al., 2011). Therefore, the development of a visualization process begins with the characterization of the domain problem.

4.1.1 Domain Problem Characterization

As already stated, the benefits of long-term monitoring (more than 24 hours of acquisition) have drawn considerable attention in medicine and healthcare. A 7-day ambulatory ECG monitoring have shown the efficiency to identify patients with atrial fibrillation, after a stroke or transient ischemic attack, which are not detected with the habitual 24-hour recordings (Goya-Esteban et al., 2009). Recently, in (Jung et al., 2012), the long-term health monitoring importance was reinforced with the development of an electrode which allows up to 7 days of continuous measurements of electrical biosignals such as ECG and electromyography (EMG).

However, dealing with the analysis of massive sources of information remains a challenge that has to be overcome.

For example, displaying the 24 hours of an ECG signal recorded with a modern Holter recorder, which uses sampling rates of 1000Hz or higher (Hilbel et al., 2008), results in a falling attempt of drawing more than 80 million data points. Since a standard computer's screen has only some thousands of available pixels (Pakhira, 2010), displaying the previously mentioned signals not only exceeds the capabilities of the visualization device, but also results in a massive time and memory consuming rate. Even if the visualization is possible, the representation of such a large amount of data points will surpass the human perceptual capacity resulting in incorrect interpretations of the data both in time and space. Figure 3 represents this last case through the plot of an ECG signal acquired during 1 hour with a sampling rate of 1000Hz, what corresponds to a display of more than 3 million data points. Here, it is possible to observe the overcrowded and cluttered display that does not give the proper information to the user.

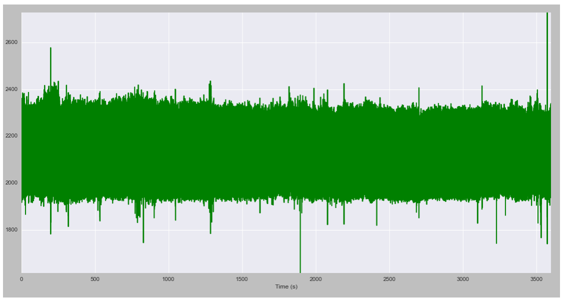


Figure 3: Overcrowded and cluttered display of an ECG signal acquired during 1 hour with a sampling rate of 1000Hz.

4.1.2 Abstraction Design

After a specific domain problem has been identified, it has to be abstracted into a more generic representation. Despite not always being performed, it is an essential process when dealing with massive data sets.

The created abstraction took into account that biosignals only contain crucial information in specific time intervals - in cyclic or sporadic events. Therefore, considering only these events, it is possible to highlight the key information in the biosignal and to suppress the irrelevant details. Furthermore, this will support the variability analysis in a physiological network. (West, 2013) states its importance.

As a result, considering only the portions of the signal which contain important information, instead of the whole signal, considerably reduces the number of data points to plot.

Consequently, categorical abstractions of the numerical data, which incorporate knowledge about the analysed biosignal, were performed in order to achieve the intended results.

4.1.3 Visualization in Layers

The proposed model comprises a multi-level visualization architecture. This architecture enables the easy search and focus in the interest regions through a simple navigation in the biosignal. While the first layer displayed gives the user a global overview of the whole biosignal, the others provide a more detailed visualization of the selected interval of the above layer (by default the selected interval is the first one).

In the application there are seven standard information layers divided into a defined number of intervals. The choice of the standard information layers had in consideration the system requirements and the fact that each layer should be divided into an integer number of intervals with the size of the layer immediately below.

The chosen layers, their number of intervals and type of data representation that each layer represents are described in Table 1.

Table 1: Model Layers.

Layer	**Number of Intervals**	**Data Representation**
10 days	10	Events
1 day	24	Events
1 hour	4	Events
15 minutes	3	Events
5 minutes	5	Subsampled
1 minute	60	Subsampled
1 second	1	Raw

The total number of information layers displayed in the platform depends on the total duration of the biosignal that is being analysed. The first layer to be displayed corresponds to the lowest layer that can represent the whole signal. All the layers below this one will also be displayed. Figure 9 ilustrates this case; the signal considered only has 20 hours of record, then the first layer to be displayed was the 1 day layer.

4.1.4 Visualization Techniques

After designing the abstractions and the layout for the visualization process model it was studied the visualization techniques that most efficiently communicate the data information. Three different visualization techniques were selected among a variety of available ones. Besides line and bar plots, two standard techniques in data visualization, horizon plots, were also considered. An horizon plot is a stacked graph that enables the performance comparison of a large number of time-dependent variables. It is built by adding color bands to a line graph and mirroring the negative values respecting the x-axis. Thereafter, using a technique called two-tone pseudo coloring the color bands are overlaid in the graph. This technique is then integrated with the small multiples technique which enables the display of a series of small graphs stacked one above the other (Heer et al., 2009).

The main interface combines only the line and bar plots to show the data selected. While the bar plots are used for the 1 hour layers, the line plots are used for the remaining ones (Figure 9). The horizon plots are used in the analysis page, later discussed, enabling the comparison of a variety of features in a restricted space (Figure 6).

The visualization elements are rendered in Scalable Vector Graphics (SVG) which are more reliable and flexible than the HTML canvas elements (Murray, 2013). These elements were created with the Data-Driven Documents (D3) JavaScript library and the visualization techniques described were rendered with two D3 plugins - the Cubism and the Rickshaw plugins (Murray, 2013).

The Web Graphics Library (WebGL) elements were also considered. Despite providing a similar rendering functionality in a faster way due to its interaction with the graphics processing unit, the toolboxes currently available to perform this type of rendering are still too complex or not stable enough to provide a suitable visualization.

4.2 The Perception and Cognition Process

The graphical perception refers to the ability to decode the information encoded on graphs (Heer et al., 2009). Therefore, in order to enhance the graphical perception some variables, such as positions, shapes and colours, were carefully assessed. The platform also includes a fast and responsive design and thoughtful default colours.

4.3 The Interactive Exploration Process

The direct interaction with the data enables the user to focus on some details according to his objectives.

Basic interaction techniques are often used to improve the visualization exploration and for this reason they are familiar to the majority of the users. The choice of these techniques comprised the zoom and pan techniques (represented in Figure 4), the use of tooltips (also represented in Figure 4) and the presence of modal dialogs to alert or guide the user. The platform also enables the saving of the exploration carried out in a pdf format.

4.3.1 Navigation

The default data to be displayed in the different layers corresponds to the data of the first interval in the layer above, with exception of the first layer where the complete interval is displayed. Hence, an appropriate navigation technique was required. A double click inside an interval of one layer will switch the layers below that will present data corresponding to the selected interval. This innovative type of browsing through the time axis is exemplified in Figure 5.

4.3.2 Annotations

Annotations are visually represented in the platform by a timestamp placed at the beginning of the annotation time frame in a division below the x axis. When the timestamp is clicked, it shows its message and a rectangular grey region highlighting the time frame of the annotation arises in the graph.

The platform enables both the creation and the removal of annotations. When a layer is loaded, either for the first time or if it is updated, every annotation that belongs to the selected time interval is displayed. Figure 5 illustrates the presence of annotations in the platform.

4.3.3 Further Analysis

The choice of key signal features can be a fundamental method to discover important information hidden on the signal. Therefore, in the developed application, it is possible to request for a more detailed analysis of some layers. Figure 6 illustrates an example.

The features selected to analyse a specific layer depend on the biosignal that is being analysed, however they are grouped in the same domains: temporal, statistical and spectral.

5 CASE STUDIES

5.1 Data Acquisition and Processing

The acquisition system consisted in a set of biomedical sensors, a wireless acquisition unit and a smartphone. A surface EMG sensor at the gastrocnemius muscles of the right leg, a triaxial accelerometry (ACC) sensor at the right hip, and a 3 lead ECG sensor at the left side of the chest were used along with the bioPLUX research system. The biosignals were acquired with a sampling frequency of 1000 Hz and with a 12 bit resolution and the data was sent via Bluetooth to the smartphone, saved in a text file and then converted to an HDF5 file.

Three acquisitions were carried out in a home atmosphere performing daily living activities. While the first one recorded an ECG and an ACC signal during 6 hours, the others recorded an EMG signal during 6 hours and 4 hours respectively. The signals were then processed.

In order to provide large files to use in the developed tool, the last signals were replicated or combined. The ECG and the ACC signals were replicated until achieve a duration of 10 days, in the ECG case, and 2 days 9 hours 32 minutes and 55 seconds, in the

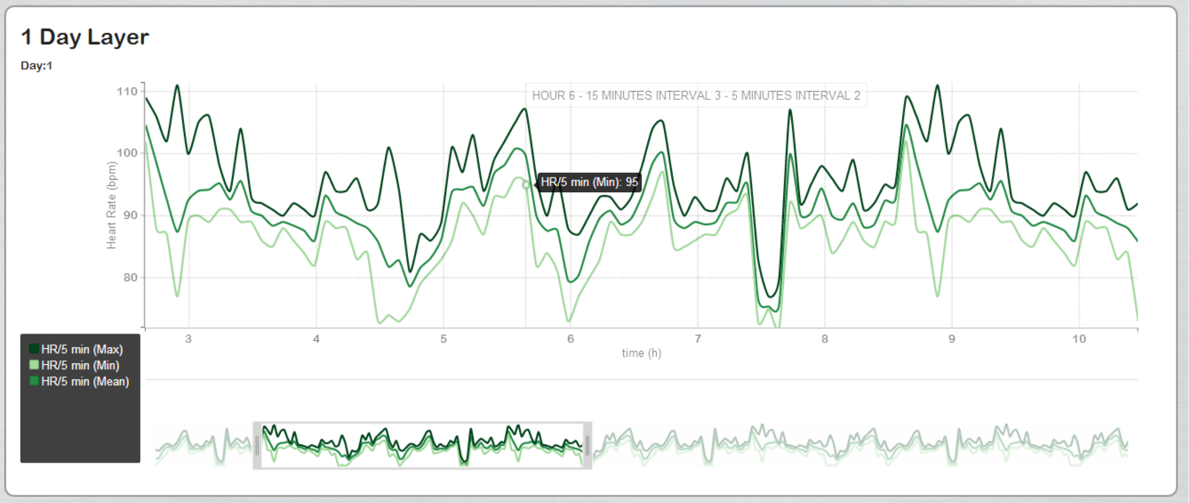


Figure 4: Zoom and pan of the previous graph. A tooltip is also presented, displaying the information related with this specific graph.

ACC case. The EMG signal consisted in the combination of the two EMG acquisitions with intervals where the acquisition was interrupted. Thus, the EMG signal was composed by the following sequence: the second performed acquisition, 3 hours without acquisition, the third performed acquisition, 30 minutes without acquisition and the second performed acquisition.

5.2 Electrocardiography

The ECG signal is a cyclic signal that consists in the recording of the electrical activity of the heart. Therefore, it provides a fundamental way of cardiac monitoring allowing the detection of cardiac abnormalities (Kaniusas, 2012).

5.2.1 Events

The abstraction for the ECG signal was performed through the consideration of the R peak in the QRS complex, one of the most important to analyse in an ECG waveform (Neophytou et al., 2012). Although the main event to consider is the positions where these peaks occur, it was also computed another event, the number of peaks that occur per minute (the heart rate). This secondary event could be easily computed through the access to the first event, however, its implementation during the processing allows a fast access to the data during the use of the platform, since it has to access a lower number of data samples and does not need to perform any other calculation. The automated R peak detection was performed with the Pan and Tompkins detection algorithm (Pan and Tompkins, 1985).

5.2.2 Features Selection

The selection of features to provide a further analysis of the ECG signal was done considering a derivation of the main event above considered. The Heart Rate Variability (HRV) represents the variation of the intervals between heartbeats (RR intervals) and its standard features provide a vital source of signal information.

The selected features are the triangular index, the short and long diameters of a Poincaré plot, the average and standard deviation of all RR intervals, the percentage of pairs of adjacent RR intervals that differ by more than 50 ms, the square root of the mean of the sum of the squares of differences between adjacent RR intervals, the very low, low and high frequency powers and the ratio of low to high frequency powers (Malik et al., 1996; Alemu et al., 2011; Hoshi et al., 2013).

5.2.3 Results

Figures 7 and 8 represent the result of one possible analysis of the ECG signal considered. Through the zoom, pan, the creation of annotations and the use of tooltips was possible to find a probably artefact caused by the chest muscles contraction. The double clicking in the graphs allow to switch the data in the layers below until the desired event be displayed.

5.3 Electromyography

An EMG signal is an electrical signal generated during a muscle contraction. Therefore, it enables the quantification of the neuromuscular function and for

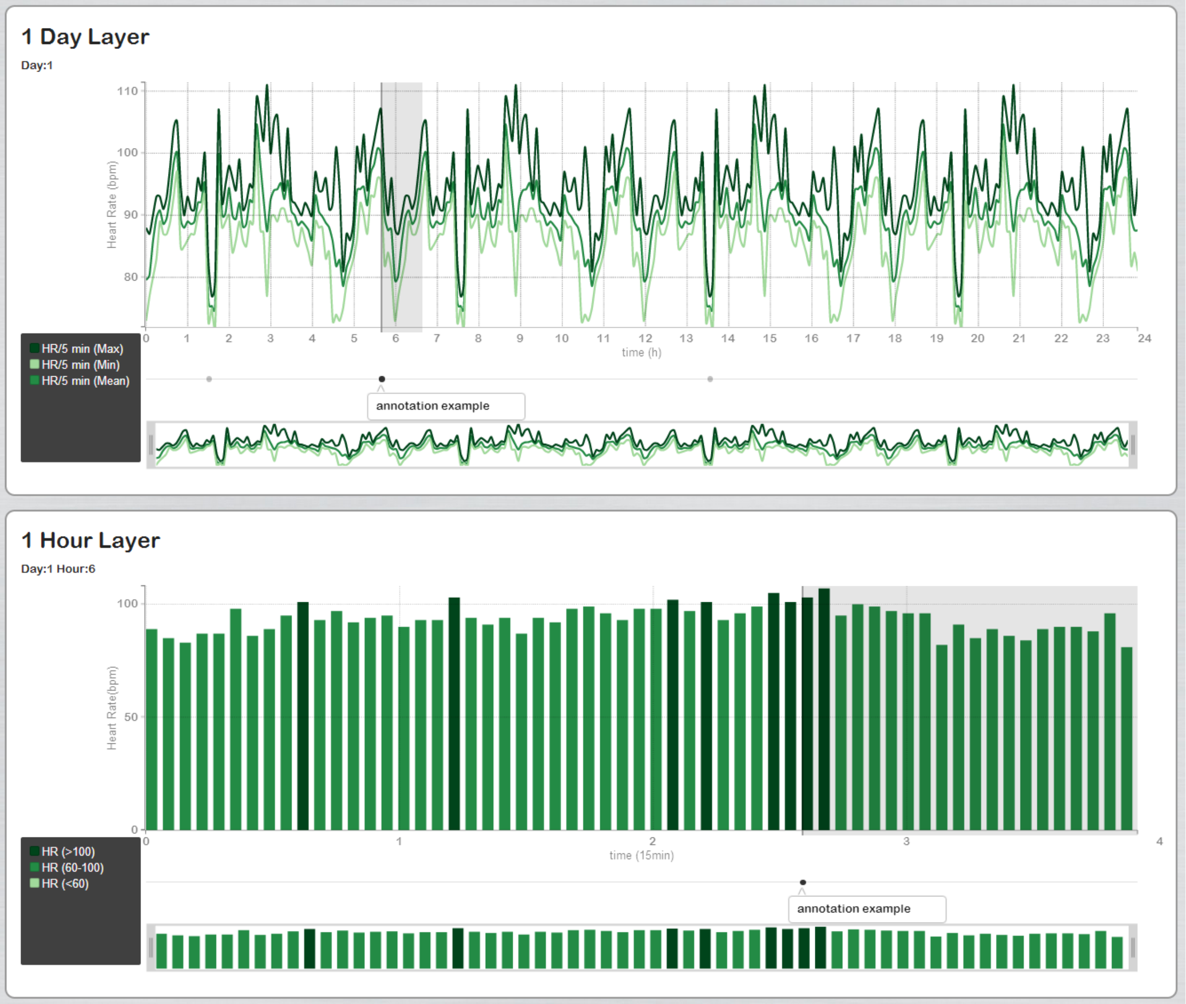


Figure 5: Navigation and Annotations. The selection of the sixth interval in the 1 day layer, i.e. the sixth hour of that day, will switch the data in the 1 hour layer in order to correspond to the data of the sixth hour. The 1 day layer presents three annotations, however only the second one belongs to the interval displayed at the 1 hour layer.

this reason these signals are mostly used to measure the degree of muscle activation and to access the neurophysiologic mechanisms of fatigue (Camata et al., 2010; Arjunan et al., 2011).

5.3.1 Events

The event with most importance in the EMG analysis is therefore the signal activation (onset and offset times). However, the EMG signal is also usually analysed by the root mean square value (Soares et al., 2013; Vieira et al., 2010). The detection of the onset time of muscle activity is performed with the Teager-Kaiser energy operation method (Xiaoyan et al., 2007) and the root mean square values are computed with 900 samples window.

5.3.2 Features Selection

The further analysis of the EMG signal is performed by the zero crossing rate, the short and long diameters of a Poincaré plot, the standard deviation and the median frequency features (Lee and others, 2013; Golnska, 2013; Camata et al., 2010; Rissanen et al., 2011; Arjunan et al., 2011).

5.3.3 Results

Figure 9 presents the result of one possible analysis of the EMG signal considered. The analysis of the first layer displayed enables to get the general intervals where there was or not acquisition and the inactivity hours where the muscle in study did not perform any contraction. The analysis of the 1 hour layer indi-

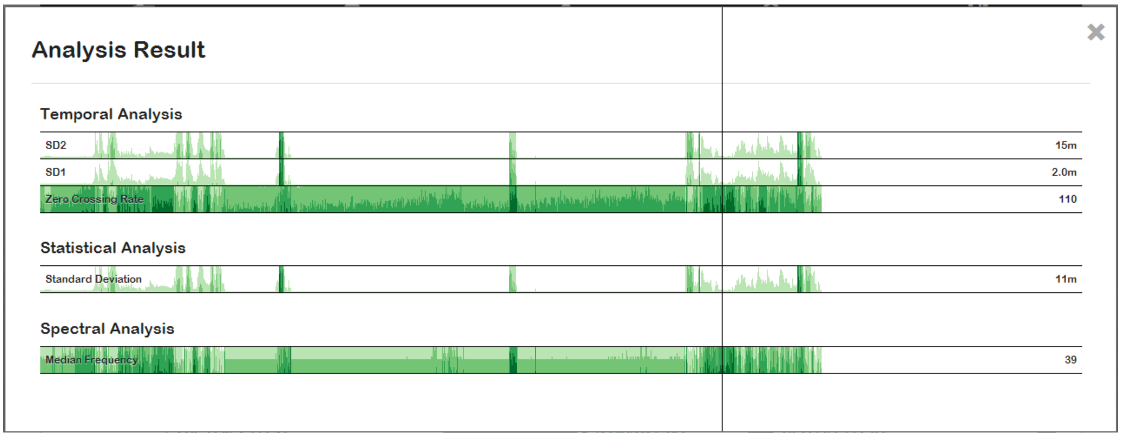


Figure 6: Further Analysis Result of the 15 minutes layer of an EMG signal.

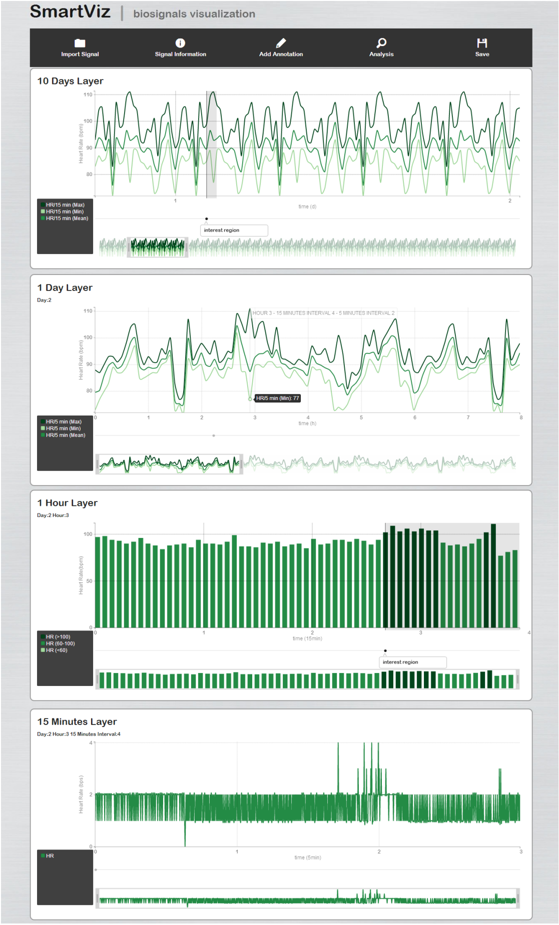


Figure 7: ECG signal analysis result.

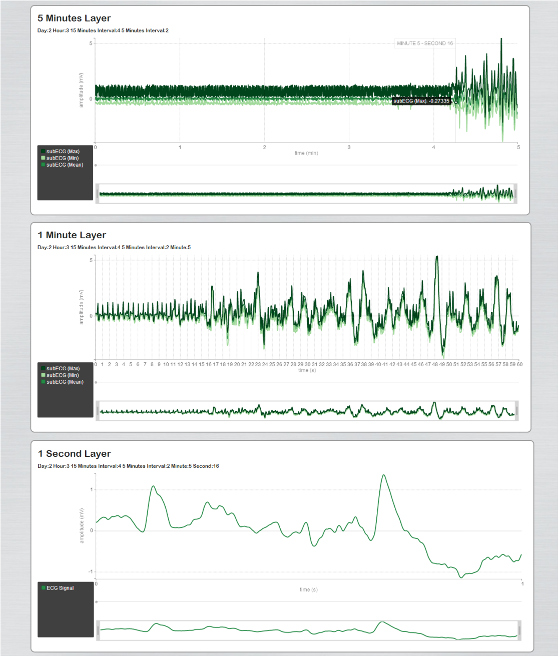


Figure 8: Continuation of the previous Figure.

cates that the number of contractions is higher in the first 15 minutes of the hour in study. The remaining layers show the EMG signal in which the contractions can be analysed with more detail. The last layer is not presented for not presenting relevant information for this case.

5.4 Accelerometry

The ACC signal provides the measurement of the applied acceleration acting along a reference axis. Therefore, its analysis provides crucial information that can be used in functional status and monitor falling studies (Jeong et al., 2007; Lan et al., 2012; Karantonis et al., 2006), sleep analysis (ActiGraph, 2012) and in neuromuscular diseases diagnosis (Machado et al., 2014).

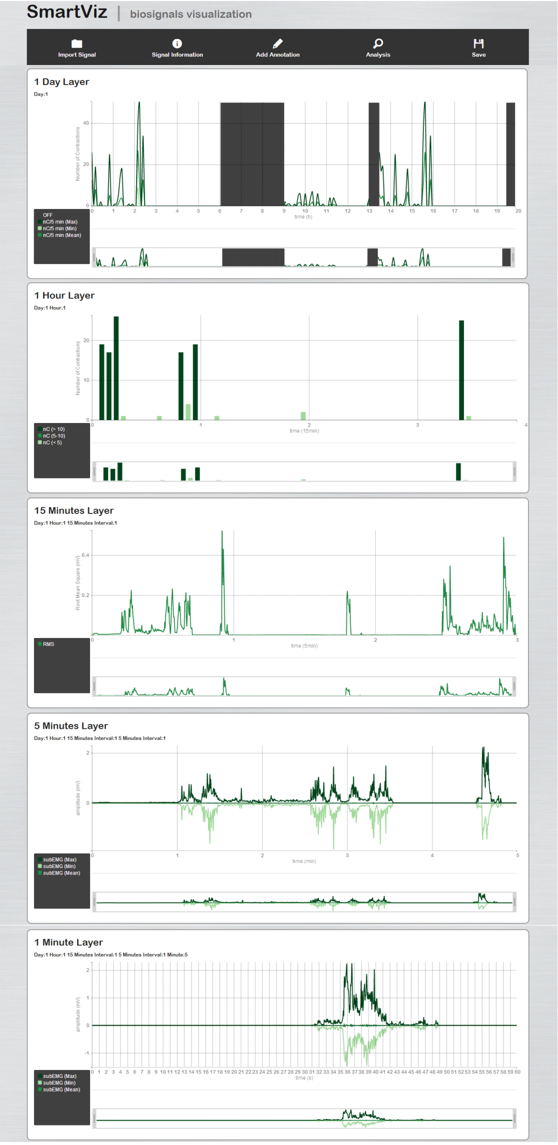


Figure 9: EMG signal analysis result.

5.4.1 Events

The abstraction for the ACC signals was settled considering that the signal represent different activities performed by the subject. These activities can be classified as static states (lying, sitting or standing) and dynamic states (walking or running) (ActiGraph, 2012; Figo et al., 2010). These two states are usually distinguish by the Signal Magnitude Area (SMA) feature (Equation 1). When the SMA value exceeded a preset threshold thus the subject is classified as being in a dynamic state (Jeong et al., 2007; Lan et al., 2012; Karantonis et al., 2006; Carus et al., 2013; Figo et al., 2010). The SMA was computed for each second of data.

$$SMA = \frac{1}{t}\left(\int_0^t |x(t)|dt + \int_0^t |y(t)|dt + \int_0^t |z(t)|dt\right) \quad (1)$$

5.4.2 Features Selection

The selected features to perform the further analysis of the ACC signal were the autocorrelation, the root mean square, the mean, the standard deviation and the median and fundamental frequencies (Machado et al., 2014; Figo et al., 2010). These features were computed with the total acceleration signal.

5.4.3 Results

The result of the ACC signal analysis is similar to the EMG signal, however instead of activations, here we present the idea of static or dynamic states.

6 USABILITY EVALUATION

In order to assess the usability of the developed work, a study with the System Usability Scale (SUS) (Brooke, 2013) was carried out. The high reliability of SUS is presented in different studies (Bangor et al., 2008). The SUS is composed by 10 items each having a five-point scale that ranges from *Strongly Disagree* to *Strongly Agree*. There are five positive statements and five negative statements, which alternate in order to avoid response biases. The SUS questionnaire and its score system is described in (Bangor et al., 2008).

To the date the usability study performed included 10 participants with ages between 23 and 28 years who had no experience with the application. All the participants belonged to the final target of users of the developed tool - health science professionals. The participants were instructed to use freely the system, opening a file and exploring it with the available interaction tools.

The average score of the performed tests was 80.0. The minimum score was 70.0 and the maximum score was 90.0. These results indicate a good system with a high perceived usability (Brooke, 2013). In the end of the test the majority of the participants agreed that some basic concepts were needed in order to perform a suitable analysis, however, these concepts should already be known by the final user. They also agreed that a necessary guidance should be provided at the first use of the system, still, it would not be necessary after this first use. Nevertheless, the participants stated that the system is intuitive, useful and suitable, they also indicated that they were likely to recommend it.

7 CONCLUSIONS AND FUTURE WORK

Considering the demand for long-term studies that result in large amounts of data and the meaning of visualization in signal analysis, the presented application provides a powerful contribution to the current need.

This work introduces a novel visualization interface which allows the professionals, who work with biosignals, to get insight into the large data sets acquired by the enhanced monitoring systems that have been introduced in the last years. The developed application enables the draw of conclusions by its users and their direct interaction with the data, overcoming the demand of a new biosignals analysis platform.

The visualization model followed includes a webserver, a multi-level layer layout, an abstraction approach and the implementation of different visualization techniques. Different explorative interaction techniques were also developed in which stands out the handling of annotations. The application layout was also studied and a responsive design as well as the choice of safe colours were considered.

Three case studies were presented and several performance tests were done. Finally, an usability evaluation was carried out in which an average score of 80.0 was achieved, which indicates a good usability performance.

Therefore, the developed application interconnects its users with the data, without users having to deal with signal processing algorithms directly what substantially improves their experience.

Despite the developed application has proven to be a suitable approach, further research can be performed.

We propose the use of more case studies. Despite of the considered signals represent the two possible types of biosignals events, cyclic and sporadic, the consideration of another set of biosignals such as electroencephalogram, electrodermal activity or respiration signals will further validate the developed work enabling its use in more research studies.

The search for WebGL rendering tools should be carrying on and when possible these elements should replace the implemented SVG elements.

Despite the SUS provides a confident measure of the perceived usability with a small group of participants (Brooke, 2013), it was carried out only with biomedical engineers. Hence, this study should also be performed by doctors and technicians such as physiotherapists.

The integration of the processing in the acquisition system would provide a real time processing which would reduce the time spent in processing.

REFERENCES

ActiGraph, S. D. (2012). ActiLife 6 users manual.

Aigner, W., Miksch, S., Schumann, H., and Tominski, C. (2011). *Visualization of Time-Oriented Data*. Springer, 1 edition.

Alemu, M., Arjunan, S. P., and Kumar, D. K. (2011). Observing exercise induced heart rate variability response. In *Biosignals and Biorobotics Conference (BRC), 2011 ISSNIP*, pages 1–6. IEEE.

Alpaydin, E. (2010). *Introduction to Machine Learning*. The MIT Press, Cambridge, Massachusetts, 2 edition.

Arjunan, S. P., Kumar, D., Wheeler, K., Shimada, H., and Naik, G. (2011). Spectral properties of surface EMG and muscle conduction velocity: A study based on sEMG model. In *Biosignals and Biorobotics Conference (BRC), 2011 ISSNIP*, page 14. IEEE.

Bangor, A., Kortum, P. T., and Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, 24(6):574–594.

Beazley, D. M. (2009). *Python Essential Reference*. Developer's Library. Pearson Education, 4 edition.

Brooke, J. (2013). SUS: a retrospective. *Journal of Usability Studies*, 8(2):29–40.

Buono, P., Aris, A., Plaisant, C., Khella, A., and Shneiderman, B. (2005). Interactive pattern search in time series. In *Electronic Imaging 2005*, pages 175–186.

Camata, T. V., Dantas, J. L., Abro, T., Brunetto, M. A., Moraes, A. C., and Altimari, L. R. (2010). Fourier and wavelet spectral analysis of EMG signals in supramaximal constant load dynamic exercise. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, page 13641367. IEEE.

Carus, J. L., Pelaez, V., Garcia, S., Fernandez, M. A., Diaz, G., and Alvarez, E. (2013). A non-invasive and autonomous physical activity measurement system for the elderly. pages 619–624. IEEE.

Duckett, J. (2014). *JavaScript and JQuery: Interactive Front-End Web Development*. John Wiley & Sons, 1 edition.

Figo, D., Diniz, P. C., Ferreira, D. R., and Cardoso, J. M. P. (2010). Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing*, 14(7):645–662.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet : Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

Golnska, A. K. (2013). Poincar plots in analysis of selected biomedical signals. *Studies in Logic, Grammar and Rhetoric*, 35(1).

Gomes, R., Nunes, N., Sousa, J., and Gamboa, H. (2012). Long term biosignals visualization and processing. In *BIOSIGNALS*, pages 402–405.

Goya-Esteban, R., Mora-Jimenez, I., Rojo-Alvarez, J. L., Barquero-Prez, O., Manzano-Martinez, S., Pastor-Prez, F., Pascual-Figal, D., and Garcia-Alberola, A. (2009). Rhythmometric analysis of heart rate variability indices during long term monitoring. In *Computers in Cardiology, 2009*, page 5760. IEEE.

Heer, J., Kong, N., and Agrawala, M. (2009). Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1303–1312.

Hilbel, T., Lux, R. L., Dietzsch, J., Schliephake, M., and Katus, H. A. (2008). Performance and productivity benefits using multi-core processors for the analysis of digital long-term ECG recordings. In *Computers in Cardiology, 2008*, pages 1069–1072. IEEE.

Hoshi, R. A., Pastre, C. M., Vanderlei, L. C. M., and Godoy, M. F. (2013). Poincare plot indexes of heart rate variability: Relationships with other nonlinear variables. *Autonomic Neuroscience*, 177(2):271–274.

Hudson, D. L. and Cohen, M. E. (2006). Intelligent analysis of biosignals. In *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, pages 323–326.

Iliinsky, N. P. N. and Steele, J. (2011). *Designing data visualizations*. O'Reilly, Farnham.

Jeong, D.-U., Kim, S.-J., and Chung, W.-Y. (2007). Classification of posture and movement using a 3-axis accelerometer. pages 837–844. IEEE.

Jung, H.-C., Moon, J.-H., Baek, D.-H., Lee, J.-H., Choi, Y.-Y., Hong, J.-S., and Lee, S.-H. (2012). CNT/PDMS composite flexible dry electrodesfor long-term ECG monitoring. *IEEE Transactions on Biomedical Engineering*, 59(5):1472–1479.

Kaniusas, E. (2012). Fundamentals of biosignals. In *Biomedical Signals and Sensors I*, Biological and Medical Physics, Biomedical Engineering, pages 1–26. Springer Berlin Heidelberg.

Karantonis, D., Narayanan, M., Mathie, M., Lovell, N., and Celler, B. (2006). Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. *IEEE Transactions on Information Technology in Biomedicine*, 10(1):156–167.

Lan, C.-C., Hsueh, Y.-H., and Hu, R.-Y. (2012). Real-time fall detecting system using a tri-axial accelerometer for home care. pages 1077–1080. IEEE.

Lee, Y. and others (2013). Spatiotemporal analysis of EMG signals for muscle rehabilitation monitoring system. In *Consumer Electronics (GCCE), 2013 IEEE 2nd Global Conference on*, pages 1–2. IEEE.

Lin, J., Keogh, E., Lonardi, S., Lankford, J. P., and Nystrom, D. M. (2004). VizTree: a tool for visually mining and monitoring massive time series databases. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 1269–1272.

Machado, I., Gomes, R., Gamboa, H., and Paixão, V. (2014). Human activity recognition from triaxial accelerometer data. In *BIOSIGNALS*.

Malik, M., Bigger, J. T., Camm, A. J., Kleiger, R. E., Malliani, A., Moss, A. J., and Schwartz, P. J. (1996). Heart rate variability standards of measurement, physiological interpretation, and clinical use. *European heart journal*, 17(3):354–381.

Munzner, T. (2009). Visualization. In *Fundamentals of Computer Graphics*, pages 675–720.

Murray, S. (2013). *Interactive data visualization for the web: [an introduction to designing with D3]*. O'Reilly, Sebastopol, CA.

Neophytou, N., Kyriakides, A., and Pitris, C. (2012). ECG analysis in the time-frequency domain. In *Bioinformatics & Bioengineering (BIBE), 2012 IEEE 12th International Conference on*, page 8084. IEEE.

Pakhira, M. (2010). Requirements of a graphical system. In *Computer Graphics Multimedia and Animation*. PHI Learning Pvt. Ltd., 2nd edition.

Pan, J. and Tompkins, W. J. (1985). A real-time QRS detection algorithm. *IEEE Trans Biomed Eng*, 32(3):230–236.

PLUX, w. b. S. (2014). biosignalsplux.

Rissanen, S. M., Kankaanpaa, M., Tarvainen, M. P., Novak, V., Novak, P., Hu, K., Manor, B., Airaksinen, O., and Karjalainen, P. A. (2011). Analysis of EMG and acceleration signals for quantifying the effects of deep brain stimulation in parkinson disease. *IEEE Transactions on Biomedical Engineering*, 58(9):2545–2553.

Soares, S. B., Coelho, R. R., and Nadal, J. (2013). The use of cross correlation function in onset detection of electromyographic signals. In *Biosignals and Biorobotics Conference (BRC), 2013 ISSNIP*, pages 1–5. IEEE.

Van Wijk, J. J. (2005). The value of visualization. In *Visualization, 2005. VIS 05. IEEE*, pages 79–86.

Van Wijk, J. J. and Van Selow, E. R. (1999). Cluster and calendar based visualization of time series data. In *Information Visualization, 1999.(Info Vis' 99) Proceedings. 1999 IEEE Symposium on*, page 49.

Vieira, T. M., Merletti, R., and Mesin, L. (2010). Automatic segmentation of surface EMG images: Improving the estimation of neuromuscular activity. *Journal of Biomechanics*, 43(11):2149–2158.

Wang, V., Salim, F., and Moskovits, P. (2013). Introduction to HTML5 WebSocket. In *The Definitive Guide to HTML5 WebSocket*. Springer.

Welch Allyn (2007). Expert holter software system - directions for use.

West, B. (2013). *Fractal Physiology and Chaos in Medicine*. World Scientific, 2 edition.

Xiaoyan, L., Ping, Z., and Alexander, A. (2007). Teager-kaiser energy operation of surface emg improves muscle activity onset detection. *Annals of Biomedical Engineering*, 35(9):1532–1538.

Cloud parallel computing for large scale biosignal analysis and event driven visualization

Ricardo Gomes*, Catarina Cavaco, Ricardo Matias, and Hugo Gamboa,

***Abstract*—The gradual aging of the worlds population and the rising costs of health increased the need for the creation of innovative solutions that promote an easy and efficient healthcare service. The appearance of various types of wearable sensors enables the continuous monitoring of biosignals, which is an important source of information for physicians and researchers in the evaluation of individuals' health status and functioning. As a result, the daily-generated amount of data coming from these devices has increased dramatically. To capitalize this new ability to measure a large number of biological signals, the extraction of useful information from the data is of the utmost importance. To this end, a variety of methods may be adopted, such as the visual inspection of data, processing for extraction of important parameters or machine learning techniques. However, these methods present a variety of limitations, especially when dealing with Big Data. In this study, we present a new solution based on parallel processing, cloud computing and web based technologies. The experimental results show that the proposed approach can significantly enhance biosignal processing in terms of speed and allows long-term biosignals' data analysis through a novel, events based, visualization technique.**

***Index Terms*—Biosignals, Big Data visualization, medical monitoring, parallel processing, cloud computing.**

I. Introduction

The development of innovative sensing technologies in physiology, aiming at the assessment of the human health status has attracted a lot of interest both on the personal level due to the emergence of the quantified self devices, and on the clinical perspective from the pervasive health domain.

Biosignals, which are descriptions of physiological phenomena and can assume a large variety of types, are important not only in the field of classical applications concerning medical diagnosis and subsequent intervention, but also for more recent applications such as elder individuals monitoring, driver monitoring, or simply for those who want to know what changes are ocurring inside their bodies in a quantified self approach, since they reflect human health and wellbeing [1].

In spite of carrying very useful information regarding the comprehension of several physiologic mechanisms, the base recorded signals, typically called raw data, has information that does not depend on the specific biological event that is the object of the study.

As an example, to assess the heart rate of a subject, a normal procedure is to acquire the Electrocardiographic signal. This signal may have information from external sources, such as noise from the power supply, or from muscular electrical activity. However, in such a complex signal, the only information that must be extracted in order to determine the heart rate is the instants when the heartbeats occurred. This information extraction example typically generates a data reduction by a factor equal to the sampling frequency (assuming a regular heart rate of 60BPM) [2].

The recent and ever increasing popularity of mobile devices is a factor that brings new opportunities in the field of pervasive healthcare. These devices allow users to interact with several wireless acquisition units, and acquire information from the devices own sensors. However, they still have several limitations when it comes to address the processing of large datasets, such as computational power, storage space, and battery life. As an answer to this problem, mobile cloud computing (MCC) has appeared as a new computing paradigm, which provides users an online access to virtually unlimited computing power and storage capacity [3]. As a consequence of the new technologies available to record this type of data, the increasing growth in the volume of acquired multimodal biosignal data is becoming a key tool to help in healthcare of several diseases, such as epilepsy and sleep medicine, as well as in other areas of investigation where a close follow up of the patients' signals is necessary. The usage of such large amounts of data in an efficient way needs new algorithms which deal with long-term datasets through the use of cloud computing and programming paradigms that potentiate this concept and enable research collaborations for faster and better scientific results [4].

The advent of multicore general-purpose personal computers, multicore graphical processing units (GPU) and cloud computing services, such as Amazon Web Services Elastic Compute Cloud (EC2) are becoming popular tools for programmers who want to develop applications that need large computational power [5].

The power of cloud computing for users who need resources to achieve compute intensive tasks is immense, as we can see in Figure 1. With few steps, everyone can use computational resources from hardware that the cloud computing company manages, in exchange for a relatively low price, without having to set up and maintain large clusters, which comprises large time and cost expenditures.

Several approaches in the area of big data processing have been developed, taking into account the aforementioned evolution. One of these approaches is parallel processing, which allows programming tasks to run in parallel. This paradigm enables the execution of a task to be distributed by several

R. Gomes and C. Cavaco are with the Department of Physics, FCT-UNL, Lisbon, Portugal e-mail: , cat.cavaco@gmail.com.

H. Gamboa is with PLUX Wireless Biosignals, Lisbon, Portugal e-mail:hgamboa@plux.info

R. Matias is with School of Health Care, Polytechnic Institute of Setubal , Portugal e-mail: ricardo.matias@ess.ips.pt

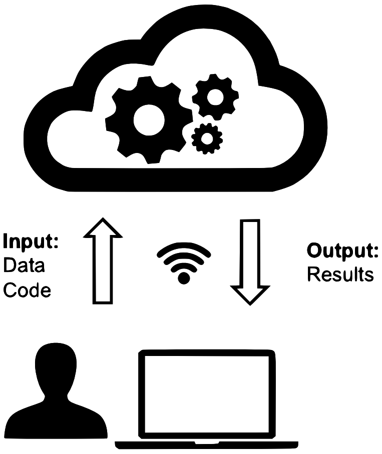


Fig. 1. Executing a processing task using the cloud.

computer cores or nodes, using message passing techniques, speeding up the execution and allowing software developers to take advantage not only of the existing computing resources in single computers, but also to leverage the power of current networks for task distribution over a cluster of computers. Nonetheless, to fully exploit the power of multicore platforms, programmers have to parallelize their applications by partitioning the workload into tasks and running them on different cores, which communicate with each other, a process that is complex, even for experienced programmers. This fact has been a barrier, slowing down the development of such applications [6].

In order to achieve the goal of extracting useful information from data, a variety of concepts and methods can be applied. Since humans main input sense is visual, data visualization is considered essential in this process [7]. This highly promising method consists in the integration of the human visual perception with the processing power of computers, in order to examine, understand and transmit the core information of the signal [8], [9]. However, data visualization presents some limitations either on the computer side or the human side. Besides the computation capacity, already discussed, the display capacity is also a limited resource that restricts the number of pixels that can be used to show the data. On the human side, the limitations comprise the human perceptual and cognitive capacity, which is also a limited resource that can result in incorrect interpretations of the data both in time and space (too fast or too dense for the correct human perception) [10]. Therefore, the better conjugation between the capabilities of computers and humans has to be achieved in the visualization process.

In the area of biosignals monitoring, one of the most widely acquired biosignals used for patient monitoring and diagnosis is the Electrocardiogram (ECG) [11]. The ECG is a record of the direction and magnitude of the electrical activity generated by the depolarization and repolarization of the hearts atria and ventricles. Being a periodic signal, each cycle is characterized by the existence of three different waves: P, QRS and T wave. The analysis of this signal is of the utmost importance in the detection of the patients heart condition, allowing the detection of cardiac anomalies [12]. Since a long time, the feature extraction from this signal has been object of research. As a result, many advanced techniques have been proposed and developed. The detection of the QRS complex, which is the result of depolarization in ventricles, is normally used to analyze a persons heart rate. In order to assess the heart rate variability, the detection of the variation of intervals between heartbeats (RR intervals) is performed. This variability is related with the function of the individuals autonomic nervous system [4].

In this study, we propose a new solution to promote the analysis of big data in the form of biosignals leveraging the power of parallel processing, cloud computing and web technologies for data visualization. Using the proposed tools, the users will be allowed to continuously acquire and visualize biosignals from wearable sensors and still have the possibility to deal with the large data collections. For this purpose, we have implemented as an example case, an ECG processing framework that runs in the cloud, using the parallel processing concept for task distribution over multiple nodes.

The rest of the paper is organized as follows. Section II provides a survey of related work on the biosignals processing using the cloud and parallel processing and also on visualization applications. Section III describes the architecture of the proposed cloud based biosignal processing and visualization tools. Section IV presents our approach for the processing of very large biological signals. Section V gives an insight on the proposed solution for biosignals visualization. Section VI illustrates a set of experiments to evaluate the efficacy and efficiency of the proposed solution, and finally, section VII presents the conclusions and future work.

II. Related Work

A. Biosignals Processing

Interest in the biosignals area comes from more than twenty centuries ago, and with it, the development of signal processing tools has also evolved through time [1]. Advances in personal mobile devices or desktop computers, which have now multiple processing cores, along with the rise of cloud computing services are nowadays tools that must be leveraged in biosignals processing. If programmers take advantage of the hardware architectures available and engineer new ways of creating parallelizable processing routines, we will be able to take full advantage of the available resources with faster programs. Konstantinidis *et al.* implemented the parallel processing paradigm to biomedical signal processing, accelerating the computation of the Correlation Dimension of multivariate neurophysiological recordings and the Skin Conductance Level using GPUs. The obtained results showed that the parallel implementation using graphical processors led to a speed-up factor of 29 when compared to a simple implementation using only C programming language [13]. In the research area of cloud computing for biosignals monitoring, Xia *et al.* developed a system for cloud based ECG quality evaluation, parameters extraction and data enhancement. The system demonstrated the possibility to upload data to the cloud from a mobile application client and get the results faster that they would be obtained using only the device [14]. Pandey *et al.* designed a cloud based realtime health monitoring and analysis system, which collects health data into an information repository and enables analysis on the data using software services hosted in the Cloud [15]. Other example of how cloud computing can help solving problems related to big data in the health area is presented by Sahoo *et al.*. In their work, a tool for real-time user interaction with electrophysiological signals, in particular with the ECG was developed. To achieve this goal, they developed novel parallelization approaches using Apache Hadoop architecture for signal processing [4]. One key element for successful parallel applications is the efficiency in message exchanging between workers of a parallel program. Hung *et al.* researched the creation a standard message passing programming to embedded multicore platforms, based on the popular Message Passing Interface (MPI) system [6].

B. Biosignals Visualization

The emergence of a variety of visualization tools, during the last decade, has aimed to counter the gap verified between our ability to collect and store data and to analyze it. Today, several applications, considering different analytical methods, are available for commercial, academic or research use. A common tool for time series visual exploration that excels in research is TimeSearcher2 [16]. This tool allows users to visualize long time series of multiple heterogeneous variables by a combination of filter and pattern search capabilities. Its main feature is the use of two types of rectangular query regions - the timeboxes and the searchboxes. While the first one filters the data and reduces the scope of the search, the second one enables to perform a specific pattern search anywhere in the remaining data. Besides flexible, intuitive and interactive, this application becomes cumbersome when dealing with extensive data sets and presents a limited scalability. By contrast to the previously described concept of query-by-example arises the concept of pattern discover implemented in VizTree [17]. This is a visualization system for massive time series data sets based on symbolic aggregate approximation (SAX). The data is discretized into a fixed length of subsequences, converted to symbols and then the symbols are concatenated to form symbol strings. Finally, a modified suffix tree is built, where frequency and other properties of patterns can be differentiated by colors and other visual properties. VizTree benefits from the ability to scale very large databases and to discover non-trivial patterns. Since a biosignal is a time series, as well as the share prices or the air temperature on successive days, the applications described for time series in general are also suitable for biosignals. However, the high concern in clinical systems for improving the physical and psychological wellness has resulted in the emergence of a variety of new and crucial systems specific for biosignals. OpenSignals, developed by PLUX, is a software application which enables biosignals visualization in real-time or from a previously recorded data set, even large. In a user-friendly interface, the entire signals are displayed. Then, an efficient navigation through the signals can be performed by zooming and panning. This tool also provides an Electromyography or HRV analysis and has the possibility of open different biosignals from a recording [18]. In cardiology, one of the most common examinations is the Holter, an ambulatory electrocardiography for a minimum 24-hour period, conducted with the purpose of screening for ECG disturbances. One of the applications used to analyze ECG signals in Holter examinations is the Welch Allyn Holter System Application [19]. It allows the selection of specific tasks enabling, for example, the review of the entire record with color-coded abnormal events or the conducting of specific assessments such as ST or HRV measurements. In the end, this software creates a report that includes selected ECG strips, some measurements and technicians comments. In spite of the high development tools to enable signals visualization, the visualization and processing of time series data sets faces yet some issues, particularly when dealing with big data. The integration of applications in real world scenarios aiming to monitor patients increases the demand for novel solutions.

III. General architecture

Biosignal monitoring has been used through the years to assess the patients condition in ambulatory, as a diagnosis tool for medical specialists and also as a follow up technique to control the health status of the elderly or chronic patients [20]. In some scenarios, the continuous acquisition of biosignals from such patients is a need that can currently be achieved through the use of wearable sensors, which enable biosignals to be recorded in an efficient and comfortable environment [21]. However, these pervasive health monitoring applications generate huge amounts of data that needs to be processed and visualized efficiently to give the medical specialists quality feedback about patients condition. Using regular personal mobile or desktop computers is an option for signal processing. Nonetheless, this approach does not enable a timely delivery of results to the medical specialists and researchers. In spite of the creation of multiple algorithms for signal processing, which enables researchers and clinicians to gather important information extracted from the patients vital signals, one of the most reliable methods of signal analysis is still the visual inspection. In the context of big data, however, the visualization of such large datasets is limited, since the existing visualization softwares normally display only small portions of the biosignal, and a richer visualization tool is not available to address data with huge sizes.

In order to create a solution to the identified problem of long-term biosignals processing and visualization, we propose a new architecture, based on cloud parallel processing and in the popular web technologies to help solving the problem of analyzing very large biosignal datasets, and efficiently visualize the data. In this work, we chose the ECG signal as a case study. A parallel processing approach, where the Pan and Tompkins ECG QRS detection algorithm [22] was adapted to run in the cloud and generate processing results for very large datasets with good performance and in satisfactory time was developed. Users can then have feedback about the obtained results of this algorithm, in a simple and effective visualization tool based on the web. The biosignal visualization is based on events and in the processed data, making it possible to have a detailed overview of very long term biosignals (week level continuous acquisition).

IV. Implementing and parallelizing processing algorithms for events detection in biosignals

ECG signals are one of the most important sources of information widely used in multiple medical health assessments. Only in Europe, cardiovascular disease is causing almost 4.1million deaths per year, or 46% of all deaths among men and women, being the leading cause of mortality [23]. When correctly acquired and processed, ECG signals can be a key input for medical decisions regarding patients with several types of heart problems [24]. Not only in ECG signal processing, but also for other types of biosignals, due to their vital importance in situations where a persons health is at risk, speed and accuracy are very important factors that must be achieved in the execution of signal processing algorithms [13].

During the workflow of processing ECG signals, there are some normally common stages to all types of applications, such as signal preprocessing (noise and artifacts removal), feature extraction, heartbeat classification and diagnosis. In this study we intended to process and visualize ECG records of up to 10 days of continuous data. For that, we adopted and implemented in python programming language the Pan and Tompkins QRS detection algorithm, which is a popular QRS complex-based heartbeat detection approach with an indicated predictive accuracy of 99.3% [22]. However, this algorithm was developed for sequential execution, and challenges were addresed to integrate them into a parallel computational workflow. Since we are dealing with very large datasets, and we want to use a parallel approach, processing the entire signal is not possible. Thus, in order to parallelize the peak detection algorithm it is necessary to slice the signal in smaller portions and process them separately, merging the results after that. Our parallel approach was inspired by the work in [25], where the very large input signal (10 days of duration, equivalent to 8.64 x 10 108 samples) is sliced in 60-second intervals. In order to avoid loosing QRS complexes during the detection, we have introduced an overlapping factor of 200 miliseconds, assuring that all the peaks are detected. The 10 days dataset used for this work was not acquired in a real scenario. Since such a large file was unavailable at the time we started this work, a smaller dataset with approximately 1 hour of duration was replicated, in order to generate the necessary file. For data storage, the HDF5 file format [26] was chosen, due to its ability to deal with very large datasets and store data in an hierarchical format. Moreover, there is a python library that enables simple conversion of the data in this file format to python arrays, facilitating data indexing and manipulation. The development of the parallel processing algorithm uses a python library developed by Vitalii Vanovschy for task distribution in SMP (symmetric multiprocessor) or cluster architectures called parallel python [27]. All the code was implemented and deployed to the cloud by using Amazon Web Services c3.8xlarge instances. This type of instance is compute optimized, and has 32 virtual central processing units (vCPU).

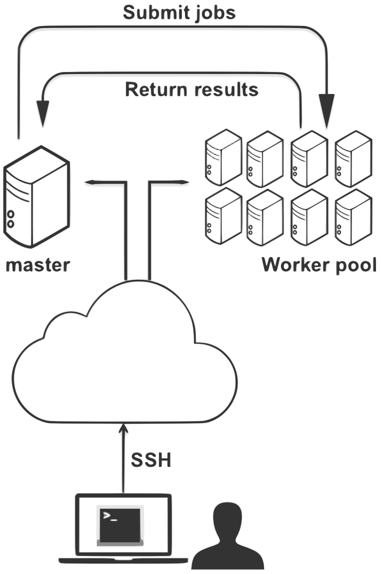


Fig. 2. Setup of the parallel processing framwork in the cloud.

The development steps for the parallel processing of long-term ECG signals are described in Figure 2.

As we can see in Figure 2, to run the ECG peak detection code in parallel, we need to follow these steps:

- connect to each EC2 worker instance, in order to launch parallel python worker process. This connection is done through Secure Shell (SSH), a network protocol, used to for remote command-line login. Each worker process is based on parallel python's ppserver.py command, with which it is possible to order the computer to listen on a specific port for the server's commands.
- connect (SSH) to the master EC2 instance, upload data to be processed and code to be executed. In this code, the parallel python master process is launched, distributing the data to be processed by the workers listening on the specified port.

The tasks executed by the workers are called jobs. When the job server is created in the master instance, a pool of job specifications is created. Each job contains as arguments the input data to be processed, and the code that the worker needs to apply to this data. After defining the jobs, the server starts sending them to the workers. When the worker terminates the job, asks the server for a new one. This enables load balancing, since the workers that complete their jobs faster do not wait for other workers to finish theirs. It must be noted that each worker instance has access to 32 processing cores, which enables task parallelization through a large farm of computing resources.

As it was mentioned, for the algorithm to work in the cloud, it was also necessary to transfer the raw data to the parallel computing master instance. The detected peaks obtained when the processing is finished are checked to avoid repetitions (which might be caused due to the overlapping factor) and saved on the HDF5 file. This file can then be downloaded from the cloud instance for later use with the developed visualization tool. Further specifications on the data structure used to save data are described next.

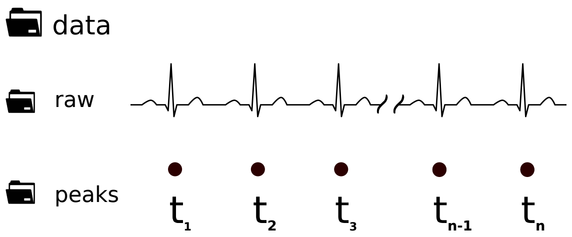


Fig. 3. Data structure for raw and processed data storage.

A. Data structure for processing results storage

Processing big datasets is a complex task. The computational resources that need to be used to achieve such task, and the time consumed suggests that saving all the processing results is mandatory. Thus, in this work, the main processing results (detected QRS peaks) are saved in the file that already contains raw data. The chosen file format is HDF5, which enables the hierarchical organization of data, with the creation of groups and datasets, making it possible to organize data in an easy and user friendly way. The chosen structure for data is described in Figure 3.

As represented in Figure 3, data is organized in two main groups: one with the raw data signal, and other containing the ECG peaks detected.

Due to the data sizes that these files might have (up to 3.5Gb), data manipulation in the cloud must be specially studied, since data storage and data transfer are part of the cloud computing paid services. In order to minimize the costs and the time that is needed to transfer such large files to the cloud, compressing the files is a good option that can result in a factor of 4 reduction in the file size.

V. New visualization model based on biosignal events

A. Core Idea

Visualizing a biosignal, as already stated, is a key point in signal analysis. However, displaying an entire biosignal, especially when considering a full day or full week of data, does not give much information to the user and exceeds the capabilities of the visualization device. For example, displaying a biosignal acquired with a sampling frequency of 1000Hz during only 1 hour would result in drawing more than 3 million data points in a screen with only some thousands of available pixels. When considering the same signal acquired during 1 day it would result in drawing more than 80 million data points and during 10 days in more than 800 million. In the visualization process, when developing or choosing a visualization method, it is essential to properly identify the domain problem. Then, this specific domain has to be abstracted into a more generic representation. In fact, abstraction is a crucial part in big data visualization and often involves the transformation of the original raw data into a derived dimension different from the original data type. During most of the time, the biosignals collected have no information of interest. The crucial information is contained in cyclic or sporadic events, which we call events. Consequently, the visualization model developed was based on events instead of considering the whole signal what reduces considerably the number of data points to display. This was achieved through categorical abstractions of the numerical data, which incorporate knowledge about the analyzed biosignal.

TABLE I
Information Layers

Information Layer	Duration of each Interval	Number of Intervals
10 Days	1 Day	10
1 Day	1 Hour	24
1 Hour	10 Minutes	6
10 Minutes	1 Minute	10
1 Minute	10 Seconds	6
10 Seconds	1 Second	10
1 Second	1 Second	1

B. System Architecture

The model proposed can be applied to all type of biosignals and gives the user the possibility of explore biosignals from acquisitions with up to 10 days of duration with detail in a user-friendly interface. We believe that the model proposed represents a commitment between usability and performance.

1) Web-based Visualization: The proposed visualization model comprises a web-based platform, which takes advantage of the advanced connectivity capabilities introduced by HTML5. A remote server was implemented in Python language and it communicates with the platform using WebSockets. This communication is based on a Message Passing Protocol (MPP) where the messages are on JSON (JavaScript Object Notation) a lightweight data-interchange format. Therefore, the browser program sends request messages to the server program that responds by serving a message with a JSON structure containing the requested information. Through the use of Python language to access and manipulate the data and JavaScript language to present the data and deal with the user interface tasks, a highest performance is reached. In order to easily access data, we created a python API that contains functions for data indexing and for the calculation of important parameters in signal analysis. HTML5 and CSS3 allowed the definition of the overall layout of the platform. To handle user interface events Bootstrap and jQuery libraries were also applied.

2) Visualization in Layers: The proposed model also considers different information layers, enabling an easier navigation by the biosignal with more or less detail in a stepwise manner. In the platform there are seven standard information layers divided into a defined number of intervals. The choice of the standard information layers took into account the maximum biosignal duration that should be shown (10 days) and the more adequate intervals in which each layer could be divided. The layers, its maximum size and number of intervals are described in Table I.

While the first layer shown gives the user a global overview of the whole biosignal, the remaining layers present the data corresponding to the selected interval of the above layer (by default the selected interval is the first one). The total number of information layers displayed in the platform depends on the total duration of the biosignal that is being analysed. For example, if the biosignal has 5 hours of total duration, the model will present only 6 layers where the 1 Day layer will be the first to be shown.

3) Visualization Techniques: Three different visualization techniques were considered in this work. Besides line plots, a standard technique, dots plots and heat maps were also considered. Dots plots is a common technique to visualize time series, however in this work, this technique was slightly modified to enable the use of dots with different colors and sizes what will allow a better perception of the data. Whereas different colors allow the distinction of dots with different information, different sizes allow the comparison of the information represented by dots with the same color. The organization of the dots in the layer is done considering the interval to which the dot belongs to (columns organization) and the type of information carried by the dot (rows and colors organization). Heat maps technique was also considered. In our approach, each cell represents an interval. A similar organization to the dots plot is followed, however, the same type of information can present a different tone of the row color. The different tones are applied according to the deviation of the value represented by the cell to the mean of the values with the same information in the whole layer. Figure 4a and 4b represent a layer scheme that use a dots plot and a heat map respectively. To custom the web-based visualization, we used the Data-Driven-Documents (D3) JavaScript library, a fast library that supports large data sets and dynamic behaviors.

4) Interaction Methods: The interaction of the user with the application enables the exploration of the biosignal and the focus on some details according to the user exploration objectives. In the platform it is possible the use of tooltips, the navigation in the layers by selecting a defined interval and the request for a more detailed analysis, in a different page, for each layer. When the mouse pointer moves over a dot, a cell of the heat map or a line a tooltip with information is shown. For a dot or a heat map cell this information consists in the interval, the event and the mean and standard deviation values of the feature selected for that event. In case of a line plot, the information only comprises the instant of time and the value of that point in the line. A double clicking one a dot or a line will select an interval. This interval will be highlighted and the layers bellow will change their information according to the selected interval. In Figure 4 we can see a tooltip and a highlight in three different 10 days layers, each one representing a different technique used in this work. The request for a more detailed analysis can be done with the choice of the layer that the user wants to analyze after a click in a button identified for this purpose. This will open a popup window with more information about that part of the biosignal.

C. Adjustment to an ECG Signal

Although the possibility of adapting the visualization model presented to any type of biosignals, this was first applied to an ECG signal, one of the most important biosignals that can be simply accessed.

1) Event Type: The main event to consider in the developed model for ECG signals visualization was chosen considering the importance of each wave in an ECG basic waveform. The R peak, in the QRS complex, is one of the most important to analyze [28]. For this reason, the R peaks were the main events to consider in the developed tool.

As already exposed, the HRV is usually derived from the RR intervals of the ECG, however, it only considers normal sinus to normal sinus (NN) interbeat intervals.

Therefore, considering the R peaks the main event in this application and the HRV its derivation, the best method to represent these events, resides on the presentation of different HRV standard features which provide a vital source of signal information.

2) Feature Selection: The HRV standard measures are usually divided into two broad categories: time domain measures and frequency domain measures. The first category is also divided into statistical measures and geometric measures [29]. These features have been widely studied and are well documented in the literature. Actually, the literature indicates two major trends in them. While statistical features are used to characterize the distribution of heart periods, frequency features are usually used to relate physiological mechanisms [30]. Table II presents the selected features for the analysis of ECG signals in the proposed visualization model[29], [30], [31], [32]. In long-term analysis (24 hours), all features presented in Table II may be used, however, for short-term analysis the features SDANN, SDNNIDX and ULF are not used and the VLF range change to the interval from 0 to 0.04 Hz. Despite not being computed from the HRV, the features SD1 and SD2, the short and long diameter of a Poincare plot respectively, are also used to describe the fast and slower components of HRV [32] .

VI. Experimental Results

A. Processing Results

In order to evaluate the performance of the proposed tools for cloud-based parallel biosignal processing, and to efficiently demonstrate the value and potential of the suggested approach for very long term datasets, we conducted a set of tests to assess the performance of the developed framework. Since in the processing side, our purpose was to create a fast and efficient tool for long-term biosignals processing, when developing the prototype for ECG beat detection, we did not intent to study the algorithms accuracy (since it has been reported to have 99.3% of accuracy [22]), but to accelerate the processing step of our framework. With this in mind, we generated a semi-artificial dataset, containing data that would correspond to a continuous acquisition of the ECG signal with a sampling rate of 1000Hz.

All the tests we have done were implemented using Amazon Web Services Elastic Compute Cloud (AWS EC2). More specifically, we have chosen the compute optimized c3.8xlarge instances. These instances provide large computing capacity to the users, since they have an equivalent (vCPU) to 32 processing cores, with 60 Gigabytes of RAM.

Table III shows the results of the different processing tests, varying the number of cloud machines used to run the tasks.

TABLE II
SELECTED FEATURES OF HRV

Domain	Measure	Description
Statistical	AVNN	Average of all NN intervals
	SDNN	Standard deviation of all NN intervals
	SDANN	Standard deviation of the averages of NN intervals in all 5-minute segments of the entire recording
	SDNNIDX	Mean of the standard deviations of NN intervals in all 5-minute segments of the entire recording
	RMSSD	Square root of the mean of the sum of the squares of differences between adjacent NN intervals
	pNN50	Percentage of pairs of adjacent NN intervals that differ by more than 50 ms
Geometrical	Triangular Index	Ratio of the number of NN intervals by the height of the discrete scale histogram with bins of 1/128 s
Frequency	TOTPWR	Total spectral power of all NN intervals up to 0.4 Hz
	ULF	Total spectral power of all NN intervals up to 0.003 Hz
	VLF	Total spectral power of all NN intervals between 0.003 and 0.04 Hz
	LF	Total spectral power of all NN intervals between 0.04 and 0.15 Hz
	HF	Total spectral power of all NN intervals between 0.15 and 0.4 Hz
	LF/HF	Ratio of low to high frequency power

TABLE III
PARALLEL PROCESSING RESULTS

Computing instance type	Number of machines	Total cores	Execution time (min)
2.3GHz Intel Core i5, 8Gb	1	4	220
2.5GHz Intel Xeon E5-2670 v2, 60Gb	1	32	16.5
	5	160	3.2
	10	320	1.7
	15	480	1.5

The most significant impact of the proposed solution is to improve the execution speed of such complex task. Using 15 cloud computing machines with the aforementioned characteristics, the peak detection algorithm ran 145 times faster than using a standard quadcore laptop. Taking into account that the 15 machines have a total of 480 processing cores, the speed up factor was even bigger than the ratio between the number of cores of cloud approach and of the laptop approach. This boost might have been caused by the greater processing capabilities of the compute optimized cloud instances used in the study.

However, using the cloud for big data processing in parallel also has its disadvantages. The main disadvantage is the time that is needed to upload the data to be processed. A regular Internet connection takes hours to accomplish the upload of an hdf5 file containing data from 10 days of acquisition, which size is approximately 3.5Gb. An alternative to the upload of the entire file is to do a real-time upload of the data during the biosignals acquisition.

B. Visualization Results

In order to represent the highest layers (10 and 1 days) three different approaches were done using the visualization techniques described in the visualization techniques section.

1) Dots approach: A first approach was done through the use of dots plots. In this case the main interface will present 3 dots with different colors by interval in each layer. These 3 dots represent the heart rate, the LF/HF and the Triangular Index of the considered interval. The size of the dot will be related to the value that it represents when compared with the others dots of the same row. The extra analysis that can be provided presents the values of the remaining features for the whole layer. Figure 4a represent this approach in a 10 days layer.

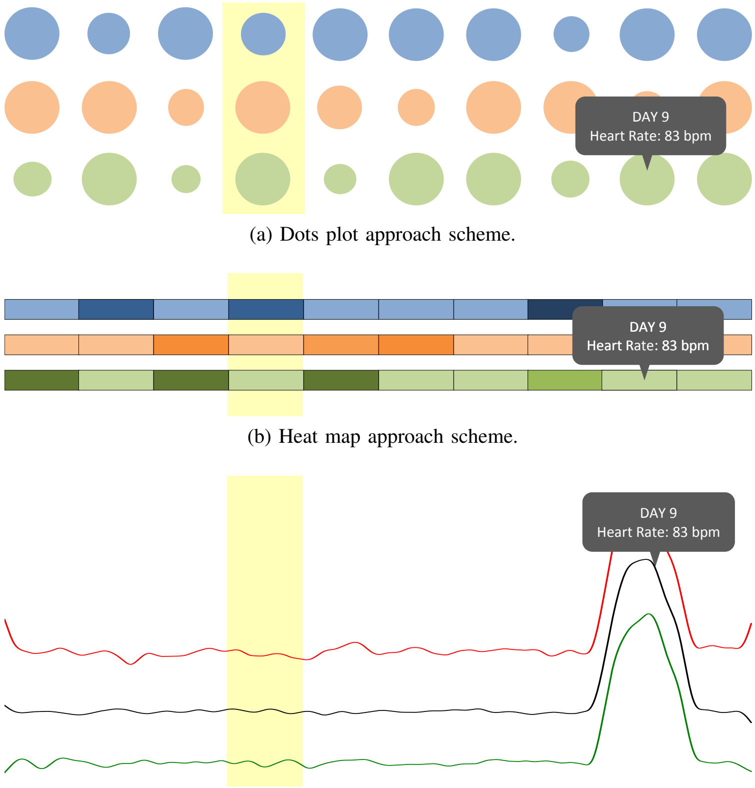
2) Heat maps approach: This approach makes use of the heat maps described in visualization techniques section. As well as the previous approach the main interface will present only the results of three selected features by interval in each layer. Here, each heart beat cell can also present two arrows pointing to the zones where outliers occur (one maximum and one minimum). The size of the arrows will enable the comparison of outliers values between intervals. Despite the fact that each feature is represented by a color, a tone gradient is used in each row. On the left is presented the name and value of the feature considering the whole signal. The remaining features are presented by the same method in the analysis page. Figure4b represents this approach in a 10 days layer.

3) Lines approach: This last approach shows the heart rate and its outliers per interval. Three lines are displayed. While the middle line represents the heart rate of the interval, the upper represent the upper outliers and the lower the lower outliers. The remaining features are presented by the previous method in the analysis page. Figure4c represents this approach in a 10 days layer.

The remaining layers, for all the described approaches, use line plots to represent the respective information. While the 1 hour layer represents the heart rate and the 10 minutes layer the RR tachogram, the remaining layers represent the subsampling or raw ECG signal.

Figure 5 represents a scheme of the main interface using the dots approach. The first two layers are represented through the chosen approach in which the different colors allow the distinction of 3 different features. The remaining layers are represented by line plots, as already explained. By order, they represent the heart rate, the tachogram and the last ones the signal.

All the approaches present strong and weak points when



(a) Dots plot approach scheme.

(b) Heat map approach scheme.

(c) Line plot approach scheme.

Fig. 4. Different approaches of a10 days layer with the fourth day selected and a tooltip at the ninth day.

compared with each others. Despite of allowing a fast perception of the whole signal, the dots plot approach is not ideal for the representation of variables that do not suffer considerable variations. Since in the higher layer each dot represents information from a very large time interval, a visualization approach that shows outliers or variations in the data is probably more indicated. The heat map approach enables to show the same information in a condensed way. It also allows the perception the instants where outliers occurred, since this approach uses a continuous time axis.

Therefore, this novel visualization model enables the visualization of a long-term (up to 10 days) biosignal. Since the visualization is based on events the signal is represented in a way that is possible to get useful information of it. The proposed visualization model is also specialized for biosignals and can be applied for all types of this signal.

VII. Conclusion

Several health problems need long-term biosignals monitoring in order to assess the patients condition continuously. Specialists must analyze the large amounts of data generated in clinical studies and in other applications. Thus, it is very important to provide the medical community with tools for large data processing. Despite of the recent technical evolution, standard computers do not guarantee the timely processing of such data, due to their limited processing power. Therefore, the cloud computing concept, which has been growing in importance lately, promises to be a part of the solution to the problem of processing very large datasets. Another problem that clinicians and researchers face when analyzing huge amounts of data is related to data visualization, since the duration of these type of biosignal acquisitions is too long to display the entire dataset at once, due not only to the lack of capacity to draw this large amount of data, but also to the humans capacity to visualize and interpret data.

In this work we propose an innovative solution that addresses both the identified problems: processing and analysis (visualization) of extremely large biosignal records, which can't be analyzed with th standard toos that currently exist. For this, we projected a new framework that leverages the power of parallel processing applied on cloud computing resources and containing tools for easily exploring this datasets in a user-friendly way.

Based on a case study on ECG signal monitoring, we demonstrate the application of the proposed analysis framework in a real scenario simulation. The experimental results are aligned with the objectives of this work: the proposed approach allowed to readily process and efficiently visualize a record containing 10 days of high sampling frequency data.

The proposed approach shows several advantages, providing new tools for biosignal processing and analysis, however, there are still work to do in order to improve the proposed solution. There are currently concerns with the exposure of users to security and privacy threats, which need to be specifically addressed in our cloud computing framework. Besides, the long time that is needed to upload a large file to the cloud

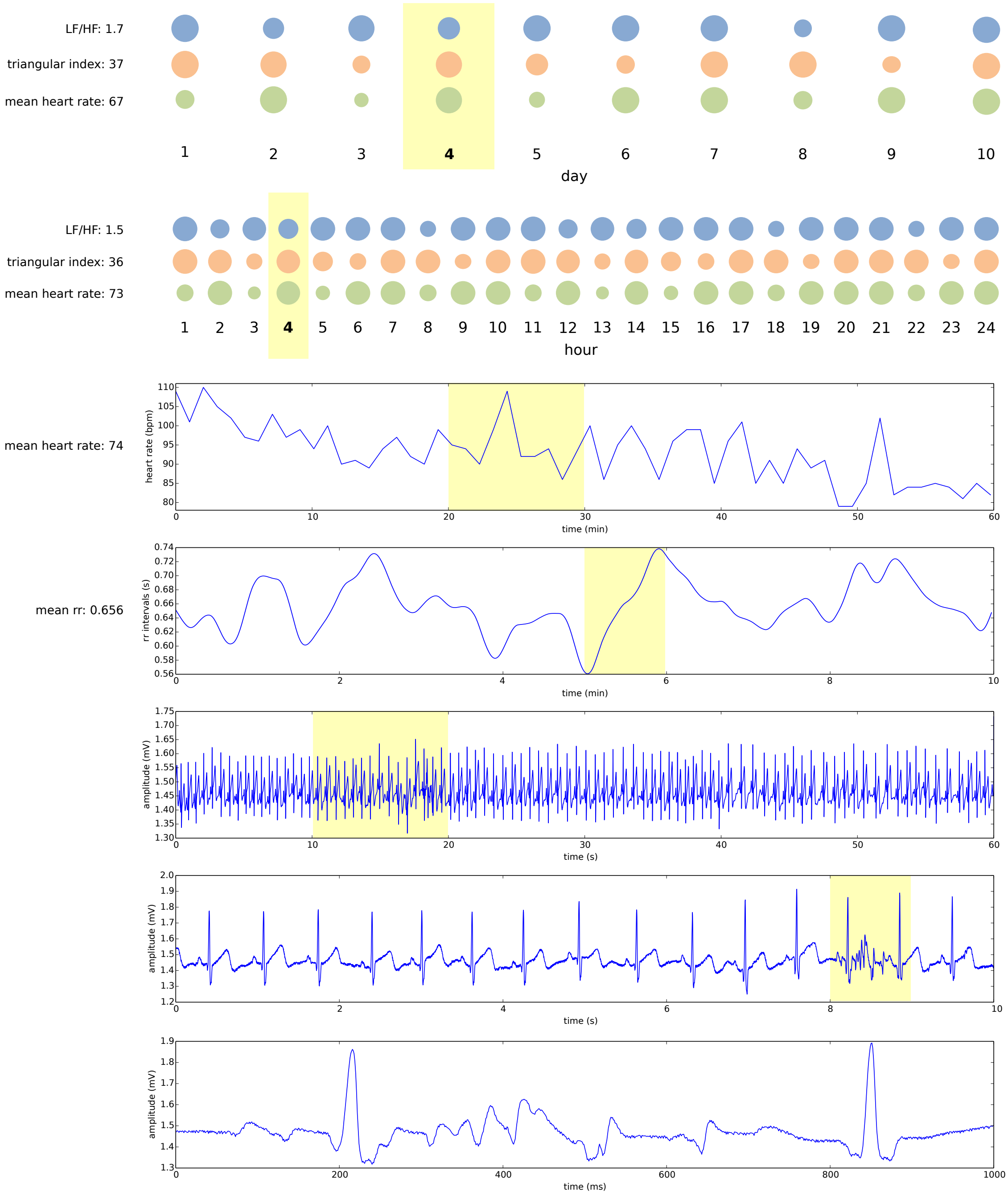


Fig. 5. Case study of the visualization tool with all 7 layers and dots approach.

is another item to address in the future development of our work.

In our future work, we will also improve the cloud computing interaction, turning it more efficient. For this, we must implement a high level API that enables the framework to control the cloud instances that are running, launch new instances or stop running instances, run the master or the worker processes on each nodes, between others. Another task that must be taken into account is the adaptation of the designed architecture to new types of signals.

References

[1] E. Kaniusas, "Fundamentals of biosignals," in *Biomedical Signals and Sensors I*, ser. Biological and Medical Physics, Biomedical Engineering. Springer Berlin Heidelberg, 2012, pp. 1–26.

[2] H. Liang, J. Bronzino, and D. Peterson, *Biosignal Processing: Principles and Practices*, ser. Biomedical Engineering Series. Taylor & Francis, 2012. [Online]. Available: http://books.google.pt/books?id=MzeUQ5Se6tsC

[3] X. Wang, Q. Gui, B. Liu, Z. Jin, and Y. Chen, "Enabling smart personalized healthcare: A hybrid mobile-cloud approach for ECG telemonitoring," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 3, pp. 739–745, May 2014.

[4] S. S. Sahoo, C. Jayapandian, G. Garg, F. Kaffashi, S. Chung, A. Bozorgi, C.-H. Chen, K. Loparo, S. D. Lhatoo, and G.-Q. Zhang, "Heart beats in

the cloud: distributed analysis of electrophysiological 'big data' using cloud computing for epilepsy clinical research," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 263–271, Mar. 2014.

[5] J. Ekanayake and G. Fox, "High performance parallel computing with clouds and cloud technologies," in *Cloud Computing*. Springer, 2010, p. 2038.

[6] S.-H. Hung, P.-H. Chiu, C.-H. Tu, W.-T. Chou, and W.-L. Yang, "Message-passing programming for embedded multicore signal-processing platforms," *Journal of Signal Processing Systems*, vol. 75, no. 2, pp. 123–139, May 2014.

[7] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski, *Visualization of Time-Oriented Data*, 1st ed. Springer, 2011.

[8] N. P. N. Iliinsky and J. Steele, *Designing data visualizations*. O'Reilly, 2011.

[9] D. A. Keim, "Information visualization and visual data mining," *IEEE Transactions on visualization and computer graphics*, vol. 7, no. 1, pp. 100–107, 2002.

[10] T. Munzner, "Visualization," in *Fundamentals of Computer Graphics*, 2009, pp. 675–720.

[11] J. D. Bronzino, *The Biomedical Engineering Handbook*, 2nd ed. CRC press, 1999, vol. 1.

[12] S. S. Mehta and N. S. Lingayat, "Development of entropy based algorithm for cardiac beat detection in 12-lead electrocardiogram," *Signal Processing*, vol. 87, no. 12, pp. 3190–3201, Dec. 2007.

[13] E. Konstantinidis, C. Frantzidis, L. Tzimkas, C. Pappas, and P. Bamidis, "Accelerating biomedical signal processing algorithms with parallel programming on graphic processor units," in *Information Technology and Applications in Biomedicine, 2009. ITAB 2009. 9th International Conference on*, Nov 2009, pp. 1–4.

[14] H. Xia, I. Asif, and X. Zhao, "Cloud-ECG for real time ECG monitoring and analysis," *Computer Methods and Programs in Biomedicine*, vol. 110, no. 3, pp. 253–259, Jun. 2013.

[15] S. Pandey, W. Voorsluys, S. Niu, A. Khandoker, and R. Buyya, "An autonomic cloud environment for hosting ecg data analysis services." *Future Generation Comp. Syst.*, vol. 28, no. 1, pp. 147–154, 2012.

[16] P. Buono, A. Aris, C. Plaisant, A. Khella, and B. Shneiderman, "Interactive pattern search in time series," in *Electronic Imaging 2005*, 2005, pp. 175–186.

[17] J. Lin, E. Keogh, S. Lonardi, J. P. Lankford, and D. M. Nystrom, "VizTree: a tool for visually mining and monitoring massive time series databases," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, 2004, pp. 1269–1272.

[18] R. Gomes, N. Nunes, J. Sousa, and H. Gamboa, "Long term biosignals visualization and processing," in *BIOSIGNALS*, 2012, pp. 402–405.

[19] Welch Allyn, "Expert holter software system - directions for use."

[20] N. Suryadevara and S. Mukhopadhyay, "Wireless sensor network based home monitoring system for wellness determination of elderly," *Sensors Journal, IEEE*, vol. 12, no. 6, pp. 1965–1972, June 2012.

[21] A. Pantelopoulos and N. Bourbakis, "A survey on wearable sensor-based systems for health monitoring and prognosis," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 40, no. 1, pp. 1–12, Jan 2010.

[22] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm." *IEEE Trans Biomed Eng*, vol. 32, no. 3, pp. 230–236, Mar. 1985. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/3997178

[23] M. Nichols, N. Townsend, P. Scarborough, and M. Rayner, "Cardiovascular disease in europe: epidemiological update," *European Heart Journal*, vol. 34, no. 39, pp. 3028–3034, Oct. 2013.

[24] A. Gacek and W. Pedrycz, *ECG Signal Processing, Classification and Interpretation: A Comprehensive Framework of Computational Intelligence*, ser. SpringerLink : Bücher. Springer, 2011. [Online]. Available: http://books.google.pt/books?id=lPTiGqPKY94C

[25] R. Gomes, N. Nunes, J. Sousa, and H. Gamboa, "Long term biosignals visualization and processing." in *BIOSIGNALS*, S. V. Huffel, C. M. B. A. Correia, A. L. N. Fred, and H. Gamboa, Eds. SciTePress, 2012, pp. 402–405.

[26] HDF group, "HDF5 - HDF group [online] Available at:http://www.hdfgroup.org/HDF5/ [Accessed 30 June 2014]," 2014.

[27] Vitalli Vanovschi, "Parallel Python [online] Available at: http://www.parallelpython.com/ [Accessed 30 June 2014]," 2014.

[28] O. Adeluyi and J.-A. Lee, "R-READER: a lightweight algorithm for rapid detection of ECG signal r-peaks," in *Engineering and Industries (ICEI), 2011 International Conference on*. IEEE, 2011, pp. 1–5.

[29] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, and P. J. Schwartz, "Heart rate variability standards of measurement, physiological interpretation, and clinical use," *European heart journal*, vol. 17, no. 3, pp. 354–381, 1996.

[30] M. Alemu, S. P. Arjunan, and D. K. Kumar, "Observing exercise induced heart rate variability response," in *Biosignals and Biorobotics Conference (BRC), 2011 ISSNIP*. IEEE, 2011, pp. 1–6.

[31] "Geometry of the poincar plot of *RR* intervals and its asymmetry in healthy adults," vol. 28, no. 3.

[32] R. A. Hoshi, C. M. Pastre, L. C. M. Vanderlei, and M. F. Godoy, "Poincar plot indexes of heart rate variability: Relationships with other nonlinear variables," *Autonomic Neuroscience*, vol. 177, no. 2, pp. 271–274, Oct. 2013.

Ricardo Gomes Master in Biomedical Engineering at Faculdade de Ciências e Tecnologia of Universidade Nova de Lisboa (FCT-UNL). He is a currently researcher at the Physics department of the FCT-UNL in collaboration with PLUX Wireless Biosignals and a DPhil candidate in Biomedical engineering. His research goals are: biosignals manipulation, analysis and processing, and development of embedded systems.

Catarina Cavaco has a degree in Biomedical Engineering Sciences from the Faculdade de Ciências e Tecnologia (FCT) of the Universidade Nova de Lisboa (UNL). At the moment she is working on her Master's Thesis in Biomedical Engineering at the same institution. Her research includes the development of tools that enable the visualization of long-term biosignals in clinical studies.

Ricardo Matias B.S.(Honors) in Physiotherapy by the School of Health Care - Polytechnic Institute of Setbal and Ph.D in Human Kinetics by the Faculty of Human Kinetics - University of Lisbon. Ricardo is currently a researcher of the latter institution's Neuromechanics of Human Movement Research Group and member of the scientific advisory board in a wireless biosignals company. His research merges three-dimensional musculoskeletal modeling and advanced statistical pattern recognition techniques to classify neuromusculoskeletal disorders and support clinical decision-making. He is also involved in the development of new clinical real-time solutions to optimize human function and health.

Hugo Gamboa (Member of IEEE) is an Assistant Professor at the Physics Department of the Sciences and Technology Faculty of the Universidade Nova de Lisboa and member of CEFITEC. PhD in Electrical and Computer Engineering from Instituto Superior Tcnico, Technical University of Lisbon, He is a founder and President of PLUX, a technology-based innovative startup in the field of systems and wireless medical sensors.