Digital processing of medical images: application in synthetic cardiac datasets using the CRISP_DM methodology

Procesamiento digital de imágenes médicas: aplicación a bases de datos sintéticas cardiacas usando la metodología CRISP-DM

Yudith Contreras MSc[1], https://orcid.org/0000-0003-4358-730X, Miguel Vera MSc, PhD[1,2], https://orcid.org/0000-0001-7167-6356, Yoleidy Huérfano MSc[2], https://orcid.org/0000-0003-0415-6654, Oscar Valbuena MSc[3], https://orcid.org/0000-0003-3080-8839, Williams Salazar MD[4], https://orcid.org/0000-0001-5669-6105, María Isabel Vera BSc[4], https://orcid.org/0000-0003-1135-6283, Maryury Borrero MSc[1], https://orcid.org/0000-0003-3025-1321, Doris Barrera MSc[1], https://orcid.org/0000-0002-6443-6757, Carlos Hernández MSc[1], https://orcid.org/0000-0001-8906-1982, Ángel Valentín Molina MSc[5], https://orcid.org/0000-0001-9604-7222, Luis Javier Martínez PhD[5], https://orcid.org/0000-0003-0917-9847, Frank Sáenz MSc[6], https://orcid.org/0000-0001-9604-7220, Marisela Vivas MSc, PhD[7], https://orcid.org/0000-0002-8941-4562, Juan Salazar MSc[1], https://orcid.org/0000-0001-6826-203X, Elkin Gelvez MSc[1], https://orcid.org/0000-0001-5157-3341.

*[1]Universidad Simón Bolívar, Facultad de Ciencias Básicas y Biomédicas, Cúcuta, Colombia. *E-mail de correspondencia: m.avera@unisimonbolivar.edu.co*

[2]Grupo de Investigación en Procesamiento Computacional de Datos (GIPCD-ULA), Universidad de Los Andes-Táchira, Venezuela.

[3]Grupo de Investigación en Educación Matemática, Matemática y Estadística (EDUMATEST), Facultad de Ciencias Básicas, Universidad de Pamplona.

[4]Servicio de Neurología, Hospital Central de San Cristóbal- Táchira, Venezuela.

[5]Grupo de Investigación en Ingeniería Clínica - HUS (GINIC-HUS), Vicerrectoría de Investigación, Universidad ECCI.

[6]Universidad Simón Bolívar, Facultad de Ingeniería, Cúcuta, Colombia.

[7]Universidad Simón Bolívar, Departamento de Ciencias Sociales y Humanas, Cúcuta, Colombia.

Resumen

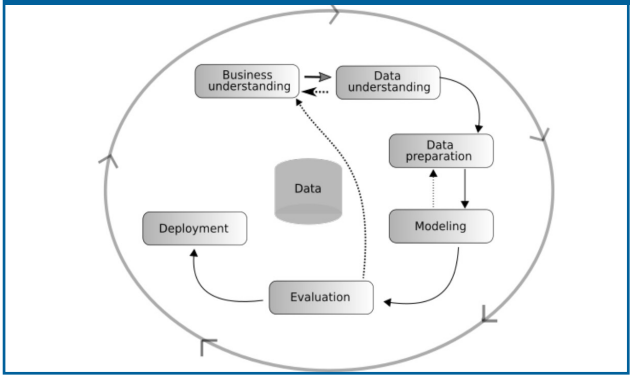
In this work an adaptation of the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, in the context of digital medical image processing is proposed. Specifically, synthetic images reported in the literature are used as numerical phantoms. Construction of the synthetic images was inspired by a detailed analysis of some of the imperfections found in the real multilayer cardiac computed tomography images. Of all the imperfections considered, only Poisson noise was selected and incorporated into a synthetic database. An example is presented in which images contaminated with Poisson noise are processed and then subject to two classical digital smoothing techniques, identified as Gaussian filter and anisotropic diffusion filter. Additionally, the peak of the signal-to-noise ratio (PSNR) is considered as a metric to analyze the performance of these filters.

Keywords—CRISP-DM Methodology, Synthetic cardiac images, Computerized tomography, Noise, Artifacts

Introduction

According to Moine[1]the CRISP-DM methodology was developed in the year 2000 by a group of companies, including SPSS, NCR and Daimler Chrysler,. It is based on the phases shown in figure 1.

Figure 1. Scheme that links the phases of the CRISP_DM methodology. (Taken from Moine[1]).



Moyne asserts that according to a study published in 2007[2], this CRISP-DM is the most used reference methodology in the development of data mining projects,

It is important to note that the adaptation and development of the phases shown in figure 1 in the context of digital processing of medical images is embodied in the various sections of this paper.

Phase 1. Problem understanding: Technical aspects and medical context of the problem.

The advent of diverse imaging modalities and their incorporation into the clinical routine, opens a world of possibilities in the areas of diagnosis, treatment and monitoring of diseases that affect the normal activities of human beings.

In the medical context, useful information is routinely extracted from images acquired by various imaging modalities such as Ultrasound Scanner (US), Magnetic Resonance Imaging (MRI), conventional Computed Tomography (CT), Multi-Slice Computed Tomography (MSCT), Positron Emission Tomography (PET), Single Photon Emission Computed Tomography (SPECT), Electron microscopy, and Endoscopy, among others. Usually, medical specialists focus their attention on a particular scene, organ, object or region in an image and resort to their experience to establish a diagnosis[3].

Often, these specialists must develop a process of delineation of the organ under examination (manual segmentation) in order to obtain parameters of clinical interest that will support their diagnosis. During this development of manual segmentation, they must process a significant number of layers or two-dimensional images, which are organized and systematically grouped together to make up the structure of the organ subject to analysis.

Formally, image segmentation is a technique of digital image processing that allows users to obtain an accurate description of the shape of objects present in a scene. But segmentation is also the partition of an image in many non-overlapping regions[4]. After applying the segmentation process, it is possible to generate as many independent regions as objects in the considered image.

In addition, it is important to highlight that medical image segmentation constitutes an open problem since each modality of imaging generates enormous amounts of information. Typically, a segmentation image contains a series of imperfections that can be caused by multiple factors, such as noise, artifacts, non-homogeneity of the tissues that constitute the various organs of the human body.

As a consequence, all the imaging modalities generate images with multiple imperfections. This constitutes the main obstacle to overcome when trying to identify or characterize the pathologies linked to any organ of the human body.

Phase 2. Data understanding: Identification of imperfections of the databases considered.

Once the access to the databases is granted, it is necessary to develop a detailed analysis of the images that embody the anatomical structure of interest.

The aforementioned analysis is crucial because it allows the identification of specific problems or imperfections in the image not associated with the anatomical structure of interest to the researcher. Thus, for example, it is possible establish without ambiguity: a) The type of noise (white, Gaussian, Rician, and so forth), b) Specific artifacts that affect the image quality (staircase, partial volume, star, dark band, ring, among others), c) Homogeneity or not of the information, and d) Problems related to the contrast of intensity between anatomical structures.

This phase of the methodology, called data understanding, will be illustrated by an example in which the process is used to generate the numerical phantoms. This type of phantoms can be considered as an artificial model in which the structure and/or characteristics of a real model are incorporated[3]. Particularly, Poisson noise present in cardiac MSCT images will be incorporated into a synthetic database (DB), called numerical phantoms.

Construction of numerical phantoms or synthetic databases

With the aim of constructing the synthetic databases, all the images from a cardiac MSCT real database are selected, in which both the endocardium and the left ventricular wall (LV) are present and an analysis of their outstanding characteristics is made. The information obtained from the aforementioned analysis is used to construct two numerical phantoms called: DB original and DB Poisson. The construction process of each mentioned DB is described below.

A. Original DB

To generate this database, the endocardium and LV wall are simulated by an internal cone and an external cone, respectively. In this sense, both cones are merged into a single volumetric image (3D), unsigned integer type, of 12 bits and with a spatial resolution of 256×256×50 voxels. This means that the original DB is made up of 50 images, each of which has a spatial resolution of 256×256 pixels. The value 256 is equivalent to 50% of the size of a typical cardiac tomography layer; while the value 50 corresponds, approximately, to 20% of the number of layers of the actual data analyzed.

As 50 layers are required, 50 circular layers for both the internal and external cones must be constructed, thus, 100 circular layers of variable radius are generated. The

common center of all these layers is located at the coordinates (128,128) pixels. For the 50 circular layers of the internal cone, the radius considers values between 11 and 60 pixels with a unit step size; while and for the 50 layers of the external cone, the same step size is used, but the radius range is between 51 and 100 pixels. The ordered grouping of these layers allows the construction of the mentioned cones.

The intensity value of the circular layers of the internal cone was 1500. This value corresponds to the average of the voxel intensities of the myocardium. On the other hand, the intensity value of the external cone is fixed at 1000. This value is generated by averaging the gray levels of the voxels of the LV wall of the real cardiac DB.

The gray level of the internal cone mantle is matched to the average of the values 1500 and 1000; whereas, the intensity or gray level of the outer mantle is obtained by averaging the values 1000 and 0 (the value 0 corresponds to the intensity of the background of each image), that is, the outer mantle has a gray level of 500.

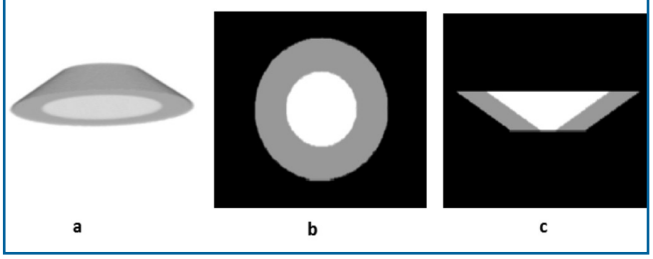
To enter into the cones the information about the intensities or levels of gray mentioned, each of the generated circular layers is processed with a filling algorithm. Thus, the original DB is obtained by means of the arithmetic sum of the external cone and an auxiliary cone that is constructed with the purpose of preserving the design parameters of the internal cone. Table 1 presents the parameters corresponding to the constructed cones.

Table 1. Parameters of the cones used to simulate the features of the LV.

	Circular layer intensity	Mantle intensity
Auxiliary Cone	500	750
External Cone	1000	500
Internal Cone	1500	1250

Additionally, figure 2 shows the orthogonal views of images that belong to the original DB created.

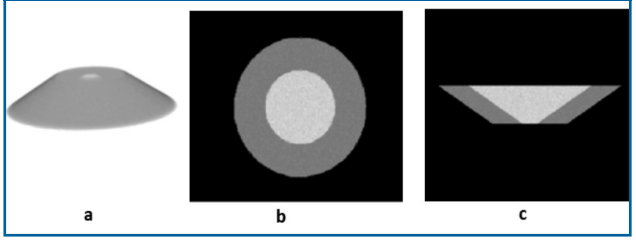
Figure 2. Original DB. (a) 3D view. (b) Axial view. (c) Coronal view.



B. Poisson DB

To construct the Poisson DB, the original DB is contaminated with Poisson noise using the algorithm proposed by Devroye[5]. Figure 3 presents images that reveal the orthogonal views of this database.

Figure 3. Poisson DB. (a) 3D view. (b) Axial view. (c) Coronal view.



Phase 3. Data preparing: Fundamentals of some filtering techniques for noise reduction.

Based on the identification of the problems and imperfections associated with the medical images which make up the databases, a set of procedures is established to improve the quality of the information of the data considered.

In this sense, a careful and detailed analysis is carried out of the digital image processing techniques that according to the specialized literature, can exhibit the best performance addressing each of the identified problems. Thus, for the sole purpose of illustrating the Poisson noise minimization process present in the synthetic databases, the filtering techniques considered in this work are described below.

The process of image filtering consists of the application of algorithms which are usually called filters, which are characterized by modifying to a certain degree the characteristics or attributes of an input image in order to minimize the possible imperfections present in it[6]. In addition, filters operate on images, in the frequency and spatial domains, in order to: a) enhance some type of desired information which may be linked, for example, with a structure or object of interest; b) minimize or suppress unwanted information which may correspond to artifacts, noise, background or other different objects than the object of interest[7].

After the filter is applied, an output image is generated in which some attributes present in the input image may appear smoothed or enhanced. Accordingly, in a preliminary way, filtering techniques can be classified as filtering techniques for the removal of unwanted information and filtering techniques for the enhancement of information of interest. Additionally, the most common filters for removing unwanted information (called low-pass filters) apply smoothing operations, oriented mainly to the noise elimination present in the considered image. Among them are Gaussian and anisotropic diffusion filters. Such filters are presented below.

Filter based on anisotropic diffusion

The anisotropic diffusion filters and its discrete implementation based on the approximation of partial derivatives, through finite differences, were introduced in image processing by Perona and Malik[8]. The purpose of applying such filters is to soften the information contained within the regions delimited by the edges of the objects present in an image. Anisotropic diffusion filters[9] can be modeled mathematically by equation 1.

$$\frac{\partial I(x,t)}{\partial t} = \nabla[c(x,t)\nabla I(x,t)] \quad (1)$$

Where ∇I(x, t) is the gradient of the image in the voxel x during the iteration (time) t, ∂I(x, t)/∂t is the partial derivative of I(x, t), and c(x, t) is the conductivity function given by equation 2.

$$c(x,t) = -\frac{||\nabla I(x,t)||}{k^2} \quad (2)$$

Where k is the conductivity parameter.

As shown in equations 1 and 2, the anisotropic filters use an edge detector that guides the diffusion process. Normally, such equations are solved numerically using finite differences, by means of an explicit scheme that allows for the softening of the image, in an iterative way, in each increment of time. In that sense, time controls the number of iterations.

Gaussian filter

The Gaussian filter is a linear spatial technique that has been used classically to minimize noise present in images. There is a connection between the amount of noise that is attenuated by the application of this filter and the blurring of the image. This type of filter uses a discrete Gaussian distribution which can be expressed by means of a mask or Gaussian kernel, of arbitrary size[10]. For softening, for example, a 3D image, the scalars in the aforementioned kernel can be obtained according to equation 3.

$$G(i,j,k) = \frac{1}{(\sqrt{2\pi})^3 \sigma_i \ \sigma_j \ \sigma_k} e^{-\left(\frac{i^2}{2\sigma_i^2}+\frac{j^2}{2\sigma_j^2}+\frac{k^2}{2\sigma_k^2}\right)} \quad (3)$$

Where $0 \leq i, j, k \leq (n-1)$, n is the size of the Gaussian kernel, and σ_i, σ_j and σ_k are the standard deviations for each spatial dimension.

In practice, Gaussian filtering is implemented by convolving the original image with the referred Gaussian kernel. The parameters of this filter are: the standard deviation of each of the spatial dimensions and the radius (r) that defines the mask size (n), given by equation 4 where r is an arbitrary scalar.

$$n = 2r + 1 n = 2r + 1 \quad (4)$$

Phase 4. Modeling: Presentation of the models to filter the Poisson DBS.

After reviewing the filtering techniques for noise removal, we proceed to design and implement a flexible model that allows the systematic minimization of the effect of this imperfection in the DBS. One of the desirable characteristics of the model is that it should be possible to represent it by means of a flow or block diagram that allows identifying the elements that compose it.

Usually, the model allows the implementation of a digital image processing strategy based on:

a) Filtering techniques based on filters: low-pass, band-pass, reject-band and/or high-pass.

b) Algorithms linked to the definition of regions of interest based on many classic operators (morphological, algebraic and/or geometric operators) as emerging (operators based on machine learning).

c) Segmentation techniques that consider methods, variational, grouping and/or hybrid.

From these sections mentioned above, as an example, only the part of literal a) corresponding to the consideration of low-pass filters or smoother filters will be implemented, In this sense, when carrying out filtering processes, usually, attention is focused on the type of noise that the images to be processed have. In response to this, synthetic bases contaminated with Poisson noise are considered.

In view of the above, by means of figure 4, the block diagram corresponding to the anisotropic smoothing technique is presented; while figure 5 illustrates the scheme relating to the application of the Gaussian filter.

In view of the above, by means of figure 4, the block diagram corresponding to the anisotropic smoothing technique is presented; while figure 5 illustrates the scheme relating to the application of the Gaussian filter.

Figure 4. Pipeline for filtering the Poisson DB by anisotropic diffusion.

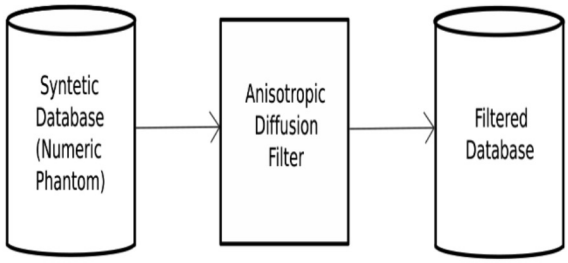
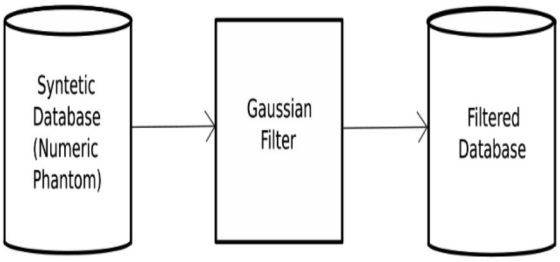


Figure 5. Pipeline for filtering the Poisson DB by Gaussian filtering.



Phase 5. Evaluation: Parameters tuning

Once the model has been designed, it becomes necessary to execute the tests of boxes (white, gray and/or black) to make judgments about the performance, simulated or real, of the strategy that will finally allow the implementation of data mining. The mentioned tests must include the process of generation of the optimal parameters (tuning process), associated with each of the filtering techniques considered. In this sense, the parameters for the Gaussian and anisotropic filters are tuned to minimize this type of noise.

Parameters linked to the Gaussian filter

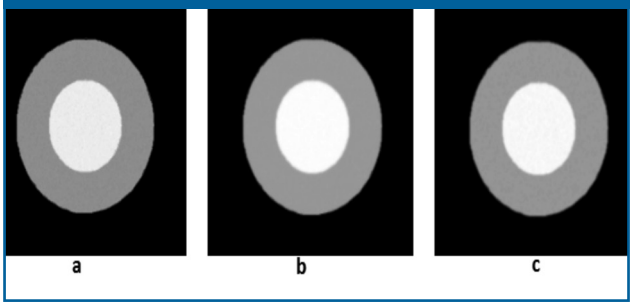
There has to be a compromise between the noise that is removed by this filter and image deformation. This deformation strongly depends on the standard deviation used in the convolution mask of Gaussian smoothing. In response to this, the standard deviation of the synthetic base subjected to filtering was considered. Additionally, in order not to produce an excessive deformation of the volume to be filtered, 3D neighborhoods of sizes (3,3,3), (5,5,5), (7,7,7) and (9,9,9) voxels were selected. The consideration of these neighborhoods allowed for the generation of 3 Gaussian filtered versions of the Poisson DB.

Parameters linked to the anisotropic diffusion filter

In the presence of noisy contours, diffusion filters have a tendency to degrade the edges of the images they process in proportion to the number of iterations[9]. For this reason, the number of iterations must be chosen carefully, in such a way that the aforementioned degradation is not excessive[3]. The number of iterations (Iter) considered was 1, 3, 5 and 7. This range of iterations includes relatively low values which determine that the mentioned degradation is within bounds. On the other hand, the parameter k varied from 0.1 to 1.1 with a step of 0.2. For the selection of these values, the work developed by Coupé et al.[9] was considered. Taking into account the combination of values selected for Iter and k, 24 anisotropically smoothed versions were obtained for each synthetic base processed.

Additionally, it is important to note that Poisson noise can be seen in any of the Poisson DB views: axial, coronal and sagittal. For this reason, only the axial view was chosen to visualize the filtering process. In this sense, through figure 6, the axial views of the Poisson DB and its filtered versions are presented after independent application of the anisotropic diffusion filter (figure 6b) and Gaussian smoothing (figure 6c) techniques.

Figure 6. Filtering applied to the Poisson DB. (a) Original Image. (b) Image filtered with anisotropic smoothing. (c) Smoothed image with Gaussian Filtering.



Analyzing figure 6, it is possible to observe that the anisotropic filter and Gaussian smoothing have reduced, to some degree, the Poisson noise impact on the considered DB. Moreover, if only qualitative information is available, it is very difficult to determine which of the techniques does the best job in minimizing that type of noise. For this reason, in order to establish which of the filters exhibits a better behavior with Poisson noise, quantitative information provided by the calculation of the peak of the signal-to-noise ratio (PSNR) is used.

One of the characteristics of the PSNR is that it allows us to establish the quality of an image after being submitted to a filtering process. The PSNR can be calculated using equation 5.

$$\mathrm{PSNR} = 20 * \log\left(\frac{Higher\ level\ of\ gray}{RMSE}\right) \quad (5)$$

Where RMSE is the square root of the estimated squared half error comparing the reference image (original DB) and the unfiltered image corresponding to the Poisson DB.

The results are shown in table 2, based on the PSNR.

Table 2. Performance of the filters with Poisson noise considering the PSNR.

Filtering technique	PSNR (dB)
Anisotropic diffusion	41.50
Gaussian smoothing	42.23

The values presented in table 2 show that, for the processed database, the Gaussian filter generates the highest PSNR, followed by the anisotropic smoothing. This indicates that Gaussian smoothing can be more effective for restoring images contaminated with Poisson noise, when compared to the other considered technique.

According to Lei and Sewchand[11] and Lu et al.[12], Poissonian noise, recreated in synthetic images, can be approximated by a Gaussian distribution. This could be one of the reasons why the Gaussian filter exhibits better performance than the other two filtering techniques considered in this section.

Phase 6. Implementation: Computational infrastructure.

After executing the evaluation phase, the last phase for the model implementation may be executed. In this last

phase, considerations must be made relative to the type of software and computational infrastructure that will be used for the implementing the model.

In this sense, for the computational implementation of the tuned model, an object-oriented programming paradigm was used, that included the use of free software tools, such as: VTK[13] and FLTK[14] libraries. In addition to the integrated development environment called Visual Studio version 10.0, to access C++[15], which requires a commercial license.

Conclusions

An example of adaptation of a data mining methodology (CRISP_DM) to the context of medical images is presented. Specifically, in synthetic images constructed considering Poisson noise normally present in real databases of cardiac computed tomography.

The work shows the effectiveness of the anisotropic diffusion filter and Gaussian smoothing to reduce the impact of Poisson noise on the quality of the synthetic images considered.

The quantitative results generated for the Gaussian filters and anisotropic diffusion, correspond to their optimal parameters as generated by intonation on the Poisson synthetic base, that is, those that gave the best values for the PSNR.

These results add force to the fact that anisotropic diffusion filter and Gaussian smoothing have been used, classically, as general noise removal techniques. In the example developed, the Gaussian filter exhibited better performance, reducing more Poisson noise that the anisotropic diffusion filter.

Acknowledgment
The authors would like to thank for the financial support given by the Simón Bolívar University –Colombia, through the 2016-16 code project.

References

1. Moine J. Methodologies for the discovery of knowledge in databases: a comparative study. [Master´s thesis]. Mar de Plata-Argentina: University of la Plata, 2013.

2. Dnuggets K (2007). Poll: ¿What main methodology are you using for data mining? Recovered in 7 de noviembre de 2010, de http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm.

3. Vera M. Segmentation of cardiac structures in multi-slice computed tomography images. [Doctoral thesis]. Merida-Venezuela: Los Andes University, 2014.

4. Shapiro L, Stockman G. Computer Vision. 1 edition. Upper Saddle River, NJ: Pearson; 2001.

5. Devroye L. Non–Uniform Random Variate Generation. New York, USA: Springer–Verlag, 1986.

6. Pratt W. Digital Image Processing. USA: John Wiley & Sons Inc, 2007.

7. González R., Woods R. Digital Image Processing. USA: Prentice Hall, 2001.

8. Perona P., Malik J. Scalespace and edge detection using anisotropic diffusion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990:12 (7), pp. 629–639.

9. Coupé P., Yger P., Prima S., Hellier P., Kervrann C., Barillot C. An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. IEEE Transactions on Medical Imaging, 2008: 27 (4), pp. 425–441.

10. Meijering H. Image enhancement in digital X–ray angiography. [Tesis Doctoral], Utrecht University, Netherlands, 2000.

11. Lei T., Sewchand W. Statistical approach to x–ray CT imaging and its applications in image analysis. statistical analysis of x–ray CT imaging. IEEE Transactions on Medical Imaging, 1992: 11 (1), pp. 53–61.

12. Lu H., Li X., Hsiao I., Liang Z. Analytical noise treatment for low-dose ct projection data by penalized weighted least-square smoothing in the kl domain. Proceedings of SPIE Medical Imaging, 2002: 4682, pp. 146–152.

13. Schroeder W., Martin K., Lorensen B. The Visualization Toolkit, An Object-Oriented Approach to 3D Graphics. New York: Prentice Hall, 2001.

14. Fast Light Toolkit (FLTK). Web page available on line: http://fltk.org/ last access: Oct, 2017.

15. B. Stroustrup, The C++ Programming Language. MA, USA: Addison–Wesley, 2000.