33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Inter-observer Variability Assessment of a Left Ventricle Segmentation Tool Applied to 4D MDCT Images of the Heart

Samuel Silva, Joaquim Madeira, Beatriz Sousa Santos and Carlos Ferreira

Abstract— **Multiple detector row computed tomography (MDCT) cardiac angiography provides a large amount of data concerning multiple cardiac phases which are not often considered. Segmentation is a first step towards exploring how this additional data can be used to perform left ventricle functional analysis or myocardial perfusion assessment. We present preliminary results regarding the assessment of inter-observer variability for a semi-automatic (multi-phase) segmentation tool for the left-ventricle.**

I. INTRODUCTION

Left ventricle (LV) functional analysis is of paramount importance for cardiac assessment. With MDCT scanners, 4D cardiac exams, typically including 10+ cardiac volumes distributed along one cardiac cycle, can be performed. Several studies (e.g., [1]) have shown that MDCT exams also allow computing several LV functional parameters (e.g., ejection fraction) which compare to those obtained using well established image modalities for cardiac analysis, such as magnetic resonance imaging (MRI) or echocardiography.

With the availability of multiple cardiac phases (besides end-systole and end-diastole), it is possible to explore how different parameters vary along the cardiac cycle, thus allowing a more complete analysis and providing a chance to explore new parameters and analysis techniques. For this, it is necessary to segment the relevant structures for all the cardiac phases, while dealing with a large amount of data (approx. 1.5 GB per exam).

Segmenting a large number of cardiac phases can be a tiresome task. Even if the segmentation is performed automatically, there is always a need to revise/edit it to ensure its correctness. Tools such as CardioViz3D [2], which allow analysing cardiac data, are not suited for 4D analysis since it cannot be performed as an integrated process, using knowledge from previously segmented cardiac phases to improve current phase segmentation and minimize required user interaction. One must not forget that, as important as the segmentation methods are the tools which allow user interaction to guide the method or allow correcting the results [3].

One of the most important aspects regarding such auxiliary tools is to perform an evaluation to assess output quality, i.e., how good are the results which can be obtained with the help of such a tool concerning aspects such as intra and inter-observer variability.

For that purpose of inter-observer assessment a quantitative evaluation has been carried out involving three radiographers which segmented the left ventricle for 24 MDCT cardiac angiography volumes (cardiac phases), using Cardio-Analyser [4]. The segmentations were then compared using the Jaccard similarity metric and agreement assessed using the Williams' index, yielding good results.

After a brief presentation of the context involving this evaluation study we describe its main features. A set of results is then presented and discussed. Finally, some conclusions and ideas for future work are presented.

II. CONTEXT

In this section we provide a brief description of the used image data, the steps involved in LV segmentation using CardioAnalyser and the main goals for the presented evaluation.

A. Exams

The MDCT cardiac angiography exams considered include 11 cardiac phases uniformly distributed along the cardiac cycle, the first obtained at 5% of the cardiac cycle and then at 10% intervals up to 95% plus an additional cardiac phase at 60%. The image volume corresponding to each cardiac phase has a $512 \times 512 \times \approx 256$ resolution.

The different cardiac phases have varying image quality. The 60% phase has the best quality since it corresponds to the diastolic phase, when the heart has less movement, and a higher dosage of radiation is applied in order to improve image quality for coronary assessment. Image quality is worse in the remaining phases, particularly around the end-systole, due to heart movement and radiation dosage reduction.

B. CardioAnalyzer

The segmentation protocol featured in the software application CardioAnalyzer [4] allows users to segment the left ventricle (LV) in cardiac angiography MDCT exams for each of the cardiac phases available. This is accomplished by using a semi-automatic left ventricle segmentation method and a set of steps which guide users along the process of revising (and editing, if needed) the proposed segmentations.

CardioAnalyzer starts by proposing a first segmentation for the 60% phase (which has better image quality and hence is easier to segment). Based on this first segmentation, suitable view planes for cardiac LV analysis are proposed.

S. Silva is supported by grant SFRH/BD/38073/2007 awarded by the Portuguese Science and Technology Foundation (FCT).

S. Silva (sss@ua.pt), J. Madeira and B. Sousa Santos are with the Dep. of Electronics, Telecommunications and Informatics, University of Aveiro, Aveiro, PT and with the Institute of Electronics And Telematics Engineering of Aveiro, Aveiro, PT

C. Ferreira is with the Dep. of Economics, Management and Industrial Engineering, Univ. of Aveiro, Aveiro, PT

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

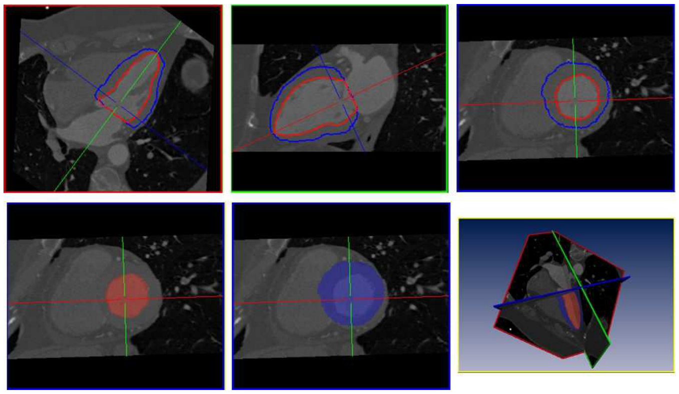

Fig. 1. Different views used by radiographers to analyze the cardiac volumes. Top row, left to right: four-chambers, two-chambers and short-axis view.

These are adjusted in order to provide the typical analysis planes, four-chambers, two-chambers and short-axis (see figure 1), instead of the standard orthogonal planes (axial, sagital and coronal). The radiographer is then allowed to correct the view planes. These will be used throughout the segmentation process given that spatial coherence is kept along the different cardiac phases, i.e., the LV keeps the same orientation and position.

The radiographer can start adjusting the proposed segmentation by changing the stopping plane (mitral valve plane) using a simple slider. This step is intended to allow solving serious problems when the mitral valve plane has not been detected by the segmentation algorithm, or it has been erroneously detected too early.

Finally, the radiographer can use a 3D editing tool [5] to perform the final corrections to the segmentation and approves the results. The segmentation of the endocardium (more accurately, the blood pool) is followed by the segmentation of the epicardium, and they have to be individually approved.

The segmentation approved for the 60% phase is then used to propose segmentations for the remaining cardiac phases basically by allowing the definition of a tighter volume of interest which is particularly important when segmenting the epicardium due to a poor differentiation between the LV wall and, for example, the right ventricle (RV). These initial segmentations can then be edited and approved by the radiographer.

C. Goal

The main goal of our study is to evaluate CardioAnalyzer as a tool for left ventricle segmentation from multiphasic MDCT cardiac angiography. For this purpose we are interested in assessing the degree of reproducibility (precision) it provides along with the time taken to perform the segmentation (computation time + user intervention) and the time taken by the radiographers to correct and approve the proposed segmentations.

Regarding reproducibility, two main aspects should be evaluated: intra-observer and inter-observer variability. The work presented here mainly concerns inter-observer variability.

It should also be noted that we do not aim to have a sophisticated segmentation method (i.e., fully automatic) but a simple, reliable and reasonably fast way to obtain left-ventricle segmentations, validated by expert radiographers, to allow left-ventricle analysis.

III. EVALUATION STUDY

In what follows a brief description of the main aspects of the performed evaluation study are provided.

Subjects — Three experienced radiographers (from now on generically referred to as radiographer A, radiographer B and radiographer C) who have everyday experience acquiring, segmenting and analyzing MDCT cardiac angiography images have participated in this evaluation.

Test Data Sets — Four exams have been chosen from the set of exams performed during one week time at the cardiology service. Care has been taken not to include exams which presented any serious acquisition artifact, but this was the only rejection criterion used.

Since most of the cardiac phases available are similar (diastolic stage of the cardiac cycle) and in order not to extend the evaluation time, overloading users with a large number of segmentations per exam, 6 cardiac phases were selected: the reference phase (60%), the end-systole (typically at 25%) and its neighbor phases (15% and 35%) which are usually significantly different and, finally, the end-diastole (typically at 95%) and a last phase, midway between the reference and end-diastole (75%).

Protocol — All radiographers were asked to use CardioAnalyzer to segment the endocardium and epicardium (one at a time) for all the exams and phases. CardioAnalyzer was modified to automatically present the proper exam to the user, when started, and to exit after the segmentation was performed.

The order in which the exams were presented to each user was randomly selected to avoid any effect related to the sequence in which exams were segmented.

All radiographers received an introductory explanation about the complete evaluation process and started the evaluation with a training exam (not considered for analysis). This allowed the radiographers to get acquainted with the different steps of the segmentation process, the tasks they had to perform and the editing tool they could use. Given the possible influence of ambient conditions, such as lighting, computer screen resolution and brightness, all radiographers performed the segmentation in the same computer and location (where they perform their everyday segmentation and analysis tasks).

At the end of each exam segmentation the radiographer had to rate image quality, the difficulty felt during segmentation and the overall satisfaction level concerning the obtained segmentations.

The radiographer was allowed to interrupt the evaluation at the end of each exam, after answering the corresponding questions and resume it later. This was aimed at reducing fatigue effects and providing easier inclusion of the evaluation tasks in the radiographer's daily schedule.

A. Comparison Metrics and Method

Several measures are described in the literature which allow comparing volumes [6]. To assess the similarity between two segmentations in this preliminary study the Jaccard similarity measure defined by $Jac = \frac{|X \cap Y|}{|X \cup Y|}$ has been used. On the other hand there is no ground truth which we might use to perform validation. All segmentations are performed by qualified radiographers and any could be considered a proper ground truth. In these situations it is common to obtain a consensus [7] from the segmentations provided by all raters comparing each segmentation to that consensus. To deal with such issue we used the Williams' index,

$$WI_i = \frac{(n-2)\displaystyle\sum_{j\neq i}^{n}\delta_{ij}}{2\displaystyle\sum_{j\neq i}^{n}\sum_{k\neq i}^{j-1}\delta_{jk}} \tag{1}$$

where δ is the similarity/dissimilarity for a pair of segmentations using a similarity measure (we used Jaccard [8]) and n is the number of raters (segmentations). Instead of determining the consensus it compares the mean agreement between one rater and each of the remaining raters with the mean agreement between all possible pairs in the group. If this index is close or above 1 the remaining segmentations are, at least, as similar to segmentation i as they are to each other.

To perform the comparisons the volume contained inside the endocardium and inside the epicardium borders was considered (see bottom row in figure 1). The contours shown to the radiographer, during segmentation, are the outline of such volumes.

IV. RESULTS

The average times to perform different tasks during the evaluation study are presented in table I. Notice that the average time needed to perform the segmentation of six cardiac phases (1 exam) was less than 18 minutes and included segmenting 6 endocardia and 6 epicardia. This is a very good time when compared to semi-automatic segmentation times reported in the literature. For instance, Coche et al. [9] report 15-20 min. to segment two phases (end-systolic and end-diastolic) using a semi-automatic method.

The total interaction time refers to the average time taken by the radiographer interacting with the segmentation method (revising and editing the segmentation) while the evaluation time concerns the average time for the whole evaluation study (including loading images from disk, processing and answering the questions at the end of each exam segmentation). Considering the total evaluation time, and considering that 24 segmentations were performed, one can estimate and average time around 4 min. to perform each segmentation (already including processing and image loading times).

TABLE I

AVERAGE TIME (OVER THE THREE SUBJECTS) TO PERFORM DIFFERENT TASKS DURING THE EVALUATION STUDY.

	time (s)	std. dev.	median (s)
endocarium	72	70	48
epicardium	90	110	66
phase	177	171	129
exam	1064	493	883
interaction time	4257	1698	4545
evaluation time	6124	2104	7307

The results obtained using the Jaccard similarity measure (figure 2) show evidence (values ≥ 0.85; with 1 meaning complete match) that for each radiographer the segmentations have a high degree of similarity when compared individually with the corresponding segmentations performed by the remaining radiographers.

The Williams' agreement index (figure 3) provides additional data concerning the segmentations as it presents how each segmentation can be compared with the set composed of the remaining two. Overall, the results show good agreement (with the Williams' index close or above 1). Nevertheless, the Williams index showed some poorer results (as low as 0.72) for some of the segmentations (see figure 3).

The left ventricle poses some segmentation difficulties regarding the mitral valve and outgoing tract regions, since there are different (valid) segmentation criteria which might include or not the outgoing tract and consider different ways of defining the segmentation close to the valve. This might result in considerable segmentation differences. Since the lowest Williams' index values happened for radiographer C, different criteria used by this radiographer might explain the poorer agreement found towards the remaining radiographers. In fact, visual comparison of some of the segmentations presenting the worst agreement values confirmed that the main difference was located in the mitral valve/outgoing tract region.

TABLE II

RATINGS GIVEN BY RADIOGRAPHERS REGARDING IMAGE QUALITY AND USER SATISFACTION.

Question	1 ...	... 5	median
Exam quality	Very Poor	Very Good	4
Endocardium segmentation	Very Hard	Very Easy	4
Epicardium segmentation	Very Hard	Very Easy	3
Overall satisf. with seg.	Dissatisfied	Satisfied	3.5

Regarding the ratings given by the radiographers, concerning image quality and user satisfaction, table II presents the median values for each item. Since we avoided images with any acquisition artifacts a good rating concerning image quality was expected. The rating concerning epicardium segmentation shows evidence that the epicardium was considered harder to segment than the endocardium (considered easy to segment). This can be explained given that it is not often easy to identify, in the septal region, for example, where the epicardium ends and the inside of the right ventricle begins. Nevertheless, we also consider that we can improve

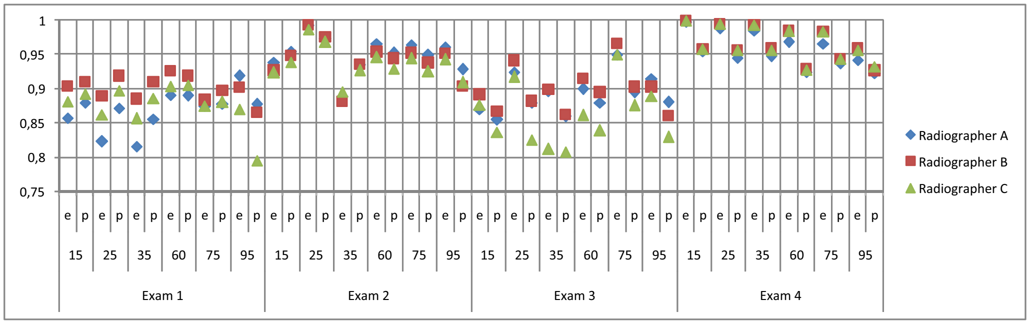

Fig. 2. Mean Jaccard similarity metric value obtained by the radiographers for each cardiac phase (e – endocarium; p – epicardium) and exam.

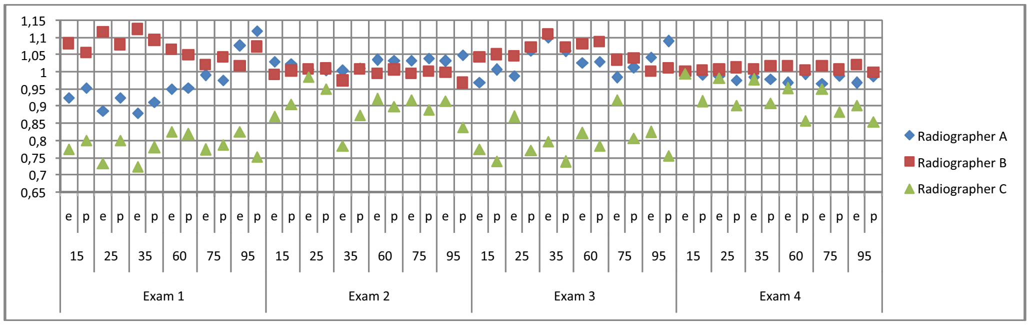

Fig. 3. Williams' agreement index (using Jaccard as similarity measure) computed for all radiographers for each cardiac phase (e – endocarium; p – epicardium) and exam.

our segmentation method to further help the radiographers in that task. Finally, concerning the overall satisfaction the radiographers rated it positively (3.5).

V. CONCLUSIONS AND FUTURE WORK

The results obtained using the Jaccard similarity measure and the William's index show that there is a small variability between observers (mean Jaccard over all subjects, for each phase ≥ 0.85) and a good agreement between each subject and the remaining two (mean Williams' index ≥ 0.91). Considering the lower agreement values found for radiographer C, inspecting some of the worst cases confirmed different segmentation criteria used by this radiographer concerning the mitral valve level and the outgoing tract. To depict such differences in the analysis data one possible approach it to compare the segmentations considering the different myocardial segments [10]. This would allow a more detailed (regional) analysis and therefore properly characterize the region responsible for the variability.

Concerning user satisfaction there is clearly room for improvement. Talking with the radiographers, one possible reason for this moderate level of satisfaction might be their unfamiliarity with the 3D editing tool [5] (no similar tool exists in the workstations they use daily). Satisfaction might be improved by additional training, resulting in greater user proficiency and confidence or, if required, tool improvement.

The presented results concern inter-observer comparison but intra-observer variability must also be assessed. For that purpose, the evaluation study is being repeated (time had to be allowed between both studies to discard memory effects) to obtain the required data.

REFERENCES

[1] A. Mahnken, P. Bruners, S. Stanzel, R. Koos, G. Mühlenbruch, R. Günther, and P. Reinartz, "Functional imaging in the assessment of myocardial infarction: MR imaging vs. MDCT vs. SPECT," *European Journal of Radiology*, vol. 71, no. 3, pp. 480–485, 2009.

[2] N. Toussaint, T. Mansi, H. Delingette, N. Ayache, and M. Sermesant, "An integrated platform for dynamic cardiac simulation and image processing: Application to personalised tetralogy of fallot simulation," in *Proc. Eurographics Workshop on Vis. Comp. for Biomed.*, 2008.

[3] S. D. Olabarriaga and A. W. M. Smeulders, "Interaction in the segmentation of medical images: A survey," *Medical Image Analysis*, vol. 5, no. 2, pp. 127–142, 2001.

[4] S. Silva, J. Madeira, B. Sousa Santos, and A. Silva, "Cardioanalyser: A software tool for segmentation and analysis of the left ventricle from 4D MDCT images of the heart," in *Proc. MediVis'10*, pp. 629–634, 2010.

[5] S. Silva, B. Sousa Santos, J. Madeira, and A. Silva, "A 3D tool for leflt ventricle segmentation editing," in *Proc. ICIAR 2010, LNCS 6112*, (Póvoa do Varzim, Portugal), pp. 79–88, 2010.

[6] S. Silva, B. Santos, C. Ferreira, J. Madeira, and A. Silva, "A preparatory study to choose similarity metrics for left-ventricle segmentations comparison," in *SPIE Medical Imaging, vol. 7963*, p. 796326, 2011.

[7] S. Vanbelle and A. Albert, "Agreement between an isolated rater and a group of raters," *Stat. Neerlandica*, vol. 63, no. 1, pp. 82–100, 2009.

[8] S. Bouix, M. Martin-Fernandez, L. Ungar, M. Nakamura, M.-S. Koo, R. McCarley, and M. Shenton, "On evaluating brain tissue classifiers without a ground truth," *NeuroImage*, vol. 36, pp. 1207–1224, 2007.

[9] E. Coche, M. Walker, F. Zech, and R. Crombrugghe, "Quantitative right and left ventricular functional analysis during gated whole-chest MDCT: A feasibility study comparing automatic segmentation to semi-manual contouring (in press)," *European J. of Radiology*, vol. 74, no. 3, pp. 138–143, 2010.

[10] M. D. Cerqueira, N. J. Weissmn, V. Dilsizian, and e. al., "Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: A statement for healthcare professional from the cardiac imaging comitee of the council on clinical cardiology of the american heart association," *Circulation*, vol. 105, pp. 539–542, 2002.