Clinical Expert Delineation of 3D Left Ventricular Echocardiograms for the CETUS Segmentation Challenge

Alexandros Papachristidis[1], Marcel L. Geleijnse[2], Elena Galli[3], Brecht Heyde[4], Martino Alessandrini[4], Daniel Barbosa[5], Erwan Donal[3], Mark J. Monaghan[1], Olivier Bernard[6], Jan D'hooge[4] and Johan G. Bosch[2]

[1] Cardiology, King's College Hospital, London, United Kingdom
[2] Cardiology/Biomedical Engineering, Erasmus MC, Rotterdam, Netherlands
[3] Cardiology, CHU Rennes, University of Rennes, France
[4] Cardiovascular Sciences, Catholic University of Leuven, Belgium
[5] Life and Health Sciences Research Institute (ICVS), University of Minho, Portugal
[6] CREATIS, University of Lyon, France

Abstract. Within the framework of the CETUS challenge, forty-five 3D echocardiographic datasets have been acquired and segmented independently by three clinical experts from different hospitals. The goal was to generate a well-established ground truth of validated expert contours on this broad range of images from different ultrasound vendors, for a number of common pathologies. Image data were acquired and segmented according to a specifically designed protocol. Since there is no clear standard or guideline for segmentation for 3DUS, we defined a tracing consensus which results in clinically acceptable and reproducible contours. Tracing was performed in four longitudinal and five transversal 3D-derived 2D planes in ED and ES. 3D contours were constructed from these tracings. If the contours or their clinical parameters differed by more than a predefined level, the tracings were compared and the experts would reach a consensus interpretation on the best segmentation. One or more experts would then adapt their tracings. Final distance differences in contours were 0.77±0.18mm for the training set and 0.82±0.27mm for the testing set. For the training set, 69% of contours were retraced. For the testing set, 76% of contours were retraced. The used protocol resulted in well-established ground truth contours.

1 Introduction

Three-dimensional echocardiography provides significant advantages over 2D and is currently applied in several aspects of cardiology [1]. The most common indication for performing echocardiography in adults is the evaluation of left ventricular (LV) size and function [2]. The use of 3D imaging for this purpose eliminates geometrical assumptions and misinterpretation errors caused by foreshortened views in 2D mode [3]. Automated segmentation of the left ventricle of the heart in 3D cardiac ultrasound images has been a subject of scientific research for the last 20 years [4]. Although many academic methods have been published and several commercial solutions exist, there has hardly been any comparison of different methods on the same datasets [5]. There is currently no standard dataset for testing these segmentation methods. Therefore, reports of algorithms are mostly incomparable since they have been evaluated on very different datasets.

The Challenge on Endocardial Three-dimensional Ultrasound Segmentation (CETUS), a grand challenge workshop associated with the MICCAI 2014 symposium, aims to address this issue. The CETUS challenge provides a series of clinically realistic 3D datasets as well as well-established reference contours based on manual tracings from three different expert echocardiography centers. A segmentation competition is organized based on this set, where all competing methods can be evaluated on the same data. Establishing a well-defined ground truth segmentation was an essential part of the challenge preparations. However, there are no clear clinical guidelines for endocardial tracing in 3D echocardiography [1]. Therefore, considerable effort was spent to define a consistent tracing method for segmentation of 3D echocardiographic data. For the ground truth in the CETUS study, we desired a contour definition that would be in line with clinical standards for 2D tracing [6]. A detailed tracing guideline was set up at the beginning of the study. This guideline was refined during the training phase (tracing of the first 15 patients) and used to solve conflicts during consensus discussions.

2 Methods

2.1 Acquisition Protocol

The setup of the challenge included the involvement of cardiologists via the European Association of Cardiovascular Imaging (EACVI) from cardiology centers with significant experience in 3D echocardiography. The acquisition and segmentation was planned to include 45 patients divided into 3 subgroups as follows: 15 healthy individuals, 15 patients with previous myocardial infarction at least 3 months before the time of scanning and 15 patients with dilated cardiomyopathy. Exclusion criteria were poor image quality defined as a) significant stitching or other type of artefact affecting the tracking of endocardium throughout the cardiac cycle; b) poor visualization of LV wall or wall out of image sector to an extend that the image can no longer be manually analysed with good confidence; c) patients with left bundle branch block (LBBB) or visually dyssynchronous LV.

The images were acquired by trained personnel in three different hospitals, using echocardiography machines from three different vendors (GE Vivid E9 (version 12) with a 4V probe, Philips iE33 with an X3-1 or X5-1 probe and Siemens SC2000 with a 4Z1c probe). Machine settings were optimized to achieve the maximum quality of images while keeping volume rate above 16Hz. All three hospitals acquired with two different ultrasound systems and were asked to acquire five patients from each patient group, so that patient group, hospital and ultrasound systems were equally distributed. The 45 patients were equally divided over three batches: *Training*, *Testing 1* and *Testing 2*, for the different parts of the challenge. Each batch had a similar distribution of pathologies, hospitals and ultrasound machines. Acquired data were fully anonymized and handled within the regulations set by the local ethical committees of each hospital.

An identification code number was given to each echocardiographic dataset. The data were transferred to the central site at Leuven University and were pre-processed for analysis. All data was converted to a general 4D image representation format (RAW) without loss of resolution. End-diastolic (ED) and end-systolic (ES) frames were identified. In order to uniformize the contouring process and ease the comparison, all volumes were pre-oriented prior to distribution by defining LV long axis, LV apex, LV base and the right ventricle (RV) insertion point. From this, nine standard anatomical planes were defined: four longitudinal planes through the long axis under 45 degrees angles and five transversal (short-axis) planes divided equally along the long axis.

2.2 Tracing Procedure

For the tracings, a custom non-commercial tracing package for 3D echocardiograms was used, developed by the University of Leuven and tested in an earlier study [7]. This Speqle3D software was customized (BH, MA) to accommodate special requirements of the challenge. Several additional features were implemented to facilitate the manual analyses and improve standardization, tracing quality and consistency. Speqle3D was supplied to all cardiologists (AP, MG, EG) and initial training was given including specific guidance provided in a written protocol. Troubleshooting was provided via remote on-line sessions. Each operator independently traced the endocardial border in the nine predefined planes, in both ED and ES instances. To guarantee direct comparisons between operators, they were not allowed to select their own views or change ED/ES frames. They were encouraged to play the full loop before and during tracing to ensure tracing consistency between ED and ES. For each longitudinal plane 9 to 15 points were set at the endocardial border, starting from the MV plane. The software then created a continuous contour by joining the individual points using b-spline interpolation. The operator could consecutively move, delete or add points to adjust the contour and delineate the endocardium as accurately as possible. In the transverse planes 6 to 10 points were set using a similar process. Contours would be closed in short-axis planes and open at the base for long-axis planes. Special consideration was taken for consistency of endocardial markers between longitudinal and transverse planes (Fig.1).

2.3 Tracing Protocol

A set of guidelines for performing the tracing were defined, by comparing the tracing conventions of the different centers and identifying possible conflicts. A convention was defined for LV wall, mitral valve (MV) plane, trabeculations, papillary muscles and apex. Basic points were to a) include trabeculae and papillary muscles in the LV cavity; b) keep tissue consistency between ED and ES planes; c) draw up to the mitral valve hinge points on the inside of the bright

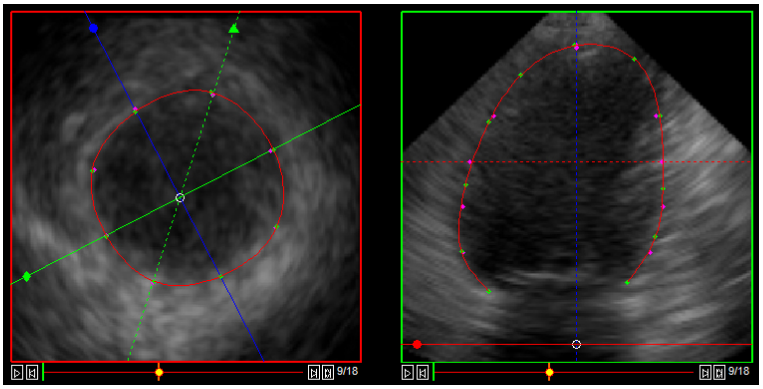

Fig. 1. Example of manual drawing in Speqle3D software. Left: Transverse plane (short axis). Right: Longitudinal plane. The green dots represent the points set on endocardium border by the operator on the actual plane. The red line represents the endocardial contour created by b-spline interpolation of the green dots. The pink dots represent the cross section points of the contours in the orthogonal planes.

ridge at the point where the valve leaflet is hinging; d) partly exclude left ventricular outflow tract (LVOT) from the cavity by drawing from septal MV hinge point to septal wall to create a smooth shape (Fig. 2); e) draw apex high up near epicardium both in ED and ES taking into consideration that there should be little displacement of the true apex point.

2.4 Evaluation of correspondence

After all three experts had submitted their segmentations, 3D shapes were generated from the nine 2D contours in ED by a spherical harmonics interpolation [7] and a standard set of vertices was generated from them. The process was repeated for the nine 2D contours in ES. The 3D contours of the three experts were compared pairwise and mean absolute distances, Hausdorff distances, Dice coefficients and LV volume and ejection fraction differences were calculated. Values were rounded to nearest integer. To qualify for consensus, between all operators the following criteria had to be met: relative difference in LV volumes $\leq$10%, Hausdorff distances $\leq$5mm and absolute difference in LVEF $\leq$5%-point.

2.5 Consensus and revisions

All three operators were asked to review the tracings of datasets which did not meet the consensus criteria. The contours of the three experts were superimposed for each one of the 18 pre-defined planes (Fig. 2). After careful evaluation of the 3 different approaches each operator would come with suggestions regarding the best endocardial tracings. Following discussion, one or more of the operators would revise their contours towards the consensus. Then the evaluation process would be repeated and slightly milder consensus criteria were applied: the average of the three pairwise observer differences was evaluated, and Hausdorff distances ≤7mm were accepted. If the tracings were still not accepted, a further revision cycle was initiated. For the *Testing 1* subgroup, there was only one round of discussion. In only two cases, the three operators did not agree on a final contour fully within the (relaxed) consensus criteria. These tracings were then accepted, in the context of persistent observer interpretation difference. From the final contours, an average 3D contour was generated for each dataset, which served as the expert ground truth in the CETUS challenge.

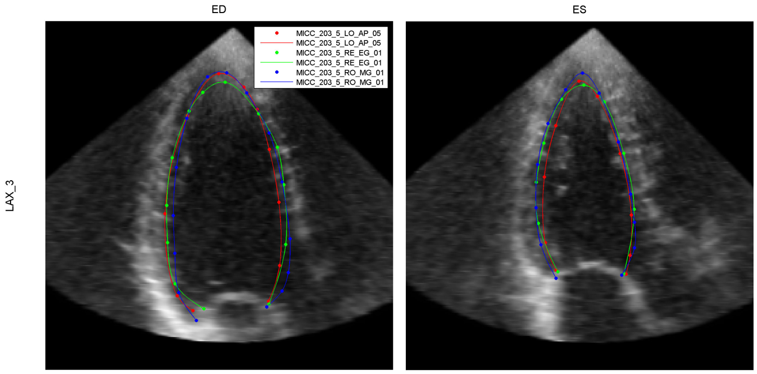

Fig. 2. Superimposition of the three expert manual contours (colour coded) on a single image in end-diastole (left panel) and end-systole (right panel). Note the drawing of the LVOT (right end of contours).

3 Results

3.1 Characteristics of datasets

Characteristics of the CETUS *Training* and *Testing 1* sets are given in table 1. Image quality was assessed by one expert as good, fair or poor and was slightly better in the *Training* set. It can be seen that ED volume as well as ES volume are significantly different ($p < 0.01$) between both datasets: the *Training* set generally has higher volumes. These are unwanted effects of the low number of patients in the set.

3.2 Tracing and retracing

For the first 30 patients, consensus was reached after the first tracings in only 2 patients (1 in the *Training* and 1 in the *Testing 1* set) and the remaining 28 patients were reconsidered. In the *Training* set, in 3 patients one expert retraced, in 5 patients two experts, and in 6 patients all experts retraced some contours. In total, 69% of expert contours were retraced. In the *Testing 1* set, these numbers were 1, 6 and 7 respectively, and 76% was retraced. In contradiction to what

Table 1. Characteristics of *Training* and *Testing 1* set. Results as mean ± standard deviation. *: average significantly different between sets (unpaired t-test, $p < 0.01$)

	ED Volume	**ES Volume**	**Ej. Frac.**	**Im. Qual.**
	(ml)	(ml)	(%)	(good/fair/poor)
All CETUS (N=30)	174.1 ± 84.6	115.5 ± 78.2	38.25 ± 13.74	10 / 11 / 9
Training (N=15)	213.2 ± 96.6	151.0 ± 91.1	33.13 ± 15.26	6 / 6 / 3
Testing 1 (N=15)	$135.0 \pm 47.2^*$	$80.0 \pm 40.6^*$	43.37 ± 10.09	4 / 5 / 6

was anticipated, the number of contour corrections did not change between the *Training* and *Testing 1* sets; apparently there was no learning effect and a similar number of corrections was needed to reach consensus.

3.3 Interobserver variability

The mean values of mean absolute distances (MAD), Hausdorff distances (HD), Modified Dice coefficients (ModDice), and the correlation coefficients (R), bias and limits of agreement (LOA) of ejection fraction (EF) and LV volumes were calculated with respect to the ground truth (average observer) and are presented in Table 2 for the *Training* set (left) and *Testing 1* set (right). Results are presented in the format of the MIDAS output for comparison with the contestant scores. Values are presented over independent observers before the consensus (i) and after consensus (c). For all error measures the differences were significantly reduced after consensus (paired t-test, $p \ll 0.01$). There was in general no significant difference in the final variability between the *Testing 1* and *Training* datasets after consensus. Only the Dice ED difference and the absolute ES volume difference were significant (unpaired t-test, $p < 0.01$), but this may be linked to the significant differences between the sets themselves, as found in Table 1.

4 Discussion

The obtained final consensus expert contours showed variability that was within the predefined limits for all but two cases (where the final variability was only marginally above the limits). The chosen approach was useful to obtain a well-established consensus ground truth that was successfully used in the CETUS challenge. Nevertheless, the observation that most contours needed some corrections, and that there was no clear improvement between the *Training* and *Testing 1* sets, probably means that the limits that we set for contouring consensus were narrower than what is achievable by trained experts on such 3D ultrasound data. One should realize that the contouring process in this study was fully manual and therefore more challenging in terms of variability with respect to previous studies [3, 5] where semi-automatic methods were used. Another reason might be that the quality of images in the *Training* datasets was better compared to that in the *Testing 1* datasets. The variability in the uncorrected images (as given by rows ”i” in table 2) probably provides a more realistic estimate of interobserver variability in our data. One should also realize that even a marginal correction in one plane was considered a retracing, and counted as strong as a total redraw in 18 planes. Therefore, the actual amount of corrections may be overrated.

Table 2. Interobserver variability on CETUS *Training* and *Testing 1* set, shown as differences from the CETUS ground truth (average±standard deviation). i: independent observers (before consensus); c: after consensus. *:average of c significantly smaller than of i (paired t-test, $p < 0.01$). $^+$:average of *Testing 1* c significantly different from *Training* c (unpaired t-test, $p < 0.01$).

		Training			***Testing 1***	
	MAD	**HD**	**ModDice**	**MAD**	**HD**	**ModDice**
ED i	1.15±0.60	3.70±1.34	.049±.022	1.01±0.30	3.37±0.87	.051±.015
c	0.77±0.18*	2.89±0.89*	.034±.008*	0.82±0.27*	2.72±0.69*	.043±.014^{*+}
ES i	1.18±0.52	3.85±1.21	.060±.026	1.01±0.38	3.30±0.94	.062±.021
c	0.78±0.19*	2.92±0.85*	.042±.013*	0.74±0.21*	2.65±0.63*	.047±.014*
	EDVol	**ESVol**	**EF**	**EDVol**	**ESVol**	**EF**
R i	0.992	0.993	0.977	0.981	0.987	0.959
c	0.999	0.999	0.996	0.993	0.997	0.978
Bias i	−4.491	−1.740	0.536	−0.636	−0.500	0.133
c	−0.811	−0.709	0.082	−0.387	−0.318	0.111
LOA i	[−44.6; 35.6]	[−44.1; 40.6]	[−8.2; 7.2]	[−18.8; 17.5]	[−14.9; 13.9]	[−5.9; 6.2]
c	[−20.9; 19.3]	[−16.3; 14.9]	[−3.0; 3.2]	[−12.8; 12.0]	[−8.7; 8.1]	[−4.1; 4.4]

4.1 Limitations

The current results are obtained on image data of reasonably good image quality. However in every-day clinical practice cardiologists and echocardiographers face the challenge of sub-optimal image quality. As has been demonstrated previously, the image quality is related to the biases in 3D LV volumes [8]. This is expected to be reflected in automated LV tracing as well. Finally, we tested the agreement and training-related improvement in inter-observer variability between experienced operators from centers with significant volume of 3D echocardiography studies. Whether the results may apply to the general cardiology and echocardiography community remains under question.

5 Conclusions

The described protocol produces expert contours with small variability. Tracing using the Speqle3D platform was quite effective and operators found it easy to use. Consensus evaluation was done in all patients, and resulted in retracing by one or more experts when differences in pre-defined parameters were above the agreed cut-off values. The level of agreement between operators as measured by differences in tracing distances and clinical calculations (LV volumes and EF) did not improve after a training period and establishment of specific guidance. However, following discussion and retracing of datasets as necessary, the level of agreement improved significantly. The used protocol resulted in well-established ground truth contours. The next step will be to use these ground truth contours to compare algorithms for automatic quantification of LV.

6 Acknowledgments

We would like to thank the European Association of Cardiovascular Imaging (EACVI) for recruiting the cardiologists involved in the CETUS study and the sonographers and cardiologists of the cooperating centers (especially Oana Mirea and Wim Vletter) for their contribution.

References

[1] Lang RM, Badano LP, Tsang W, et al. EAE/ASE recommendations for image acquisition and display using three-dimensional echocardiography. Eur Heart J Cardiovasc Imaging. 2012 Jan;13(1):1-46.

[2] Monaghan MJ. Role of real time 3D echocardiography in evaluating the left ventricle. Heart 2006; 92(1):131-136.

[3] Lang RM, Mor-Avi V, Dent JM, Kramer CM. Three-dimensional echocardiography: is it ready for everyday clinical use? JACC Cardiovasc Imaging 2009; 2(1):114-117.

[4] Leung KYE, Bosch JG. Automated border detection in three-dimensional echocardiography: principles and promises. Eur J Echocardiography 2010; 11(2):97-108.

[5] Soliman OII, Krenning BJ, Geleijnse ML, Nemes A, Bosch JG, et al. Quantification of left ventricular volumes and function in patients with cardiomyopathies by real-time three-dimensional echocardiography: a head-to-head comparison between two different semiautomated endocardial border detection algorithms. J Am Soc Echocardiography 2007; 20(9):1042-1049.

[6] Lang RM, Bierig M, Dereveux RB, et al. Recommendations for Chamber Quantification: A Report from the American Society of Echocardiographys Guidelines and Standards Committee and the Chamber Quantification Writing Group, Developed in Conjunction with the European Association of Echocardiography, a Branch of the European Society of Cardiology. Eur J Echocardiogr. 2006 Mar; 7(2):79-108.

[7] Heyde B, Barbosa D, Claus P, Maes F, Dhooge J. Three-dimensional cardiac motion estimation based on non-rigid image registration using a novel transformation model adapted to the heart. Proc STACOM 2012; LNCS 7746:142150.

[8] Pouleur A-C, le Polain de Waroux J-B, Pasquet A, et al. Assessment of left ventricular mass and volumes by three-dimensional echocardiography in patients with or without wall motion abnormalities: comparison against cine magnetic Resonance Imaging. Heart 2008; 94(8):1050-1057.