Comparison of Three Methods to Estimate Regional Wall Motion on the Evalechocard Database of Echocardiographic Image Sequences

N Kachenoura[1,2], F Frouin[1,2], L Sarry[3], C Tilmant[3,4], T Corpetti[5], H Guillemet[6], O Nardi[7], A Delouche[1,8], B Diebold[1,8]

[1]Inserm UMR 678, Paris, France
[2]Université Pierre et Marie Curie, Paris, France
[3]Inserm ERI 14, Clermont-Ferrand, France
[4]ISIMA, Aubière, France
[5]CNRS UMR 6554, Rennes, France
[6]Aptéryx, Issy-les-Moulineaux, France
[7]Hôpital Raymond Poincaré, Garches, France
[8]Hôpital Européen George sPompidou, Paris, France

Abstract

The objective of this paper is to propose a framework in order to make the comparison of image processing algorithms effective. This framework was applied to three methods developed for automated regional wall motion scoring and they were compared to a reference scoring on a database of echocardiographic images (Evalechocard). Firstly, 200 annotated echocardiograms on hundred patients were used for the training stage; secondly the algorithms were blinded tested on 100 additional echocardiograms. Results obtained by the three methods are presented, using different metrics to compare them. This evaluation procedure enables a real progress in the assessment of each method and helps to understand its limits and its potentialities. Test results have shown the difficult cases and could be further used to improve the methods. Moreover the annotated database is now open to any research group who wants to test its own methods.

1. Introduction

Evaluation is a critical point to validate a new image processing method in medical imaging. First of all, it should be linked to a medical task. For many years, the comparison of different image processing algorithms was seldom reported. Moreover, each new method was applied frequently to a few cases, which are not sufficient to demonstrate the interest on a clinical point of view. The French TechnoVision program has proposed a framework to perform one evaluation round of different image processing algorithms on databases having a large number of cases. The Evalechocard project was focused on the regional wall motion of the left ventricle, as it can be assessed from echocardiograms. Indeed, regional wall motion score (RWMS) is an important index of prognostic in case of ischemic diseases. Moreover, ultrasound is the first imaging modality to evaluate it, thanks to its high temporal resolution, its low cost and its harmlessness.

The segmentation of the left ventricle into 17 segments proposed by the AHA [1] is now the standard mode to assess RWMS; its application to two-chamber and four-chamber apical views yields seven segments per view. The wall motion is usually scored using four classes: normal (N), hypokinetic (H), akinetic (A), and dyskinetic (D). Conventional way of scoring is based on visual examination.

Thus, performance in RWMS depends on the experience of the reader, which requires an intensive training. Intra and inter operator variability has already been widely discussed, and it is more pronounced when choosing operators from various centers [2]. To overcome these limits, some quantitative methods have been proposed. But their actual impact on scoring has not really been studied. Moreover the performances of various methods can not be estimated satisfactorily when each method uses its own data for evaluation. In that sense, the approach that we proposed is innovative, since it aims at comparing three various methods using the same database.

The paper is organized as follows: the section 2 presents the definition of the annotated database, a quick overview of the three algorithms which have been tested, the description of the learning and the test stages, and the metrics which were defined to evaluate the performance of each method. Results obtained during the test phase are presented in the section 3. The section 4 indicates the main teachings of this evaluation procedure and provides some prospects.

2. Methods

2.1. The Evalechocard Database

For each patient, two-chamber and four-chamber apical views were selected. Images were acquired in B mode, using harmonic mode and were recorded digitally. Each record is an image sequence corresponding to one full cardiac cycle, the starting point being defined from the QRS complex. Preprocessing included the selection of three landmarks (extremities of the mitral valve and apex) to provide a segmentation of the left ventricle into seven regions, according to the method described in [3]. This simple segmentation was provided to experts and to image processing algorithms. Moreover, two ultrasound devices from two different manufacturers were used, in order to introduce some diversity. Indeed, it appears as a first step toward multi-center study. However, for this trial, the algorithms were not blinded to the ultrasound device. Image sequences and their relating information were stored using the Interfile format. Each segment of each view was annotated by expert consensus, as described in [4].

The database was split into two parts: 200 studies were dedicated to the learning phase and the 100 remaining studies were kept for the test phase. Studies acquired with each ultrasound device were present in the two bases, with similar proportion. Moreover, RWM abnormalities were equivalent distributed for both bases. The complementary information including landmark coordinates for the segmentation and reference RWMS for the learning examples was stored in a companion text file. The resulting database, called Evalechocard, was distributed to the participants of the project via a Web secured access.

2.2. Automated methods to estimate RWMS

Three methods which were proposed by different research laboratories were considered for the evaluation procedure: the FALVE method [5], the PAMM method [6] and a method based on the estimation of a Dense displacement Field and a further Principal Component Analysis of regional time displacement curves [7], which is further called DFPCA. All these methods were designed to resume the information contained in the image sequences (30 to 80 images).

The FALVE method estimates two static images relating to the direction and amplitude of the local wall motion. The PAMM method estimates four static images; two correspond to the amplitude of the local wall motion, as FALVE images do; the other two images correspond to additional chronological information of the local wall motion. Both methods have shown interesting features to distinguish between normal and pathological motion through the visual inspection of the static images that they produce [5, 6]. Due to the potential interest of these static images, a quantified approach was developed to extract relevant indices directly from these computed static images. Thus, a regional amplitude index is estimated from FALVE images, as it was detailed in [8], while a regional mixed index (mean contraction time to amplitude ratio) is computed from PAMM images, as it was presented in [9].

The DFPCA approach computes for each pixel, its displacement from one image to the next with the algorithm proposed in [10]. Moreover, the estimation is restricted to the myocardial area. This condition requires the temporal tracking of the endocardial contour, which is achieved using level sets with a priori knowledge relating to the endocardial shape (semi elliptical model). From the displacement field in each pixel of the myocardium, and the segmentation into seven regions provided by the experts, mean regional time-radial displacement curves and time-tangential displacement curves are computed. Finally, a classification of these curves leads to RWM scoring.

2.3. Training stage and test stage

The training stage was organized as follows: the 200 echocardiographic images and the corresponding RWMS as they were defined by experts were provided to the participants. This first step lasted about one year. Each participant freely organized the training.

For the FALVE and the PAMM methods, different indices were tested (amplitude with various weightings, various combinations of time and amplitude indices). But, one single index was retained, and the variations in this index introduced by the localization of each segment were minimized by adequate weightings. For the classification task, three thresholds were defined which respectively separated dyskinetic, akinetic, hypokinetic and normal classes. The choice of the three thresholds was refined using leave-one-out procedures. The index which provided the best classification was retained. Moreover, several developments were undertaken to minimize the number of large errors, i.e. the segments which were heavily misclassified, with an error greater than one.

For DFPCA, the training was based on a multi-parametric approach. A functional Principal Component Analysis was firstly applied to regional time-displacements curves, and the first components were selected. Thereafter, a supervised classification was defined, using support vector machines in the subspace obtained after PCA. The classification was performed separately for each segment location (seven per view, two views per patient).

The test stage was organized in a much more restricted time (within three weeks), and all users were blinded to reference scores. Only one run per participant was submitted. Only results obtained in the test stage are reported in the paper, because of the variability of the learning strategies.

2.4. Estimation of the performance

The comparison of the scores provided by each method being tested with the reference scores was reported in contingency tables. From these tables, absolute agreement (AA), relative agreement (segments within more or less one class) (RA), and weighted kappa coefficient (κ) were extracted. Moreover, the percentage of classified segments (CS) was indicated.

In addition, some more global indices (per patient) were computed, by grouping all the segments belonging to the same patient and estimating a mean wall motion score (varying between one for normal to four for dyskinetic). The evaluation of this index was done using 1) the correlation coefficient (r) of the linear regression and 2) the mean difference - or bias (b) and 3) the standard deviation of the difference (sd) between two methods, the latter indices deriving from Bland-Altman representation.

For a simple graphical interpretation of all these metrics evaluating performance, a scaling was introduced, where 100 corresponded to the best result. Thus, (1-b).100 was considered instead of b and d=(1-sd/2).100 instead of sd.

3. Results

3.1. Contingency tables

Tables 1, 2, and 3 give the contingency tables for the scoring established by the three tested methods (columns) versus the reference scoring (rows) on the 700 segments included in the test database. The N.A. stands for Not Available and corresponds to segments without score.

Table 1. Contingency table obtained for the amplitude index computed from FALVE images.

	N	H	A	D	N.A.
N	288	60	17	7	0
H	78	49	38	6	0
A	7	21	29	16	0
D	1	16	21	18	0
N.A.	3	4	17	4	0

Table 2. Contingency table obtained for the mean contraction time to amplitude ratio computed from PAMM images.

	N	H	A	D	N.A.
N	273	72	7	3	17
H	77	55	23	8	8
A	6	29	19	15	4
D	1	15	17	20	3
N.A.	2	7	10	4	5

Table 3. Contingency table obtained for the DFPCA approach.

	N	H	A	D	N.A.
N	275	75	12	10	0
H	65	85	14	7	0
A	15	30	14	14	0
D	18	16	11	11	0
N.A.	4	16	6	2	0

3.2. Comparison of the three methods

The figure 1 gives an overview of the performance of the three methods, according to the previously defined performance indices.

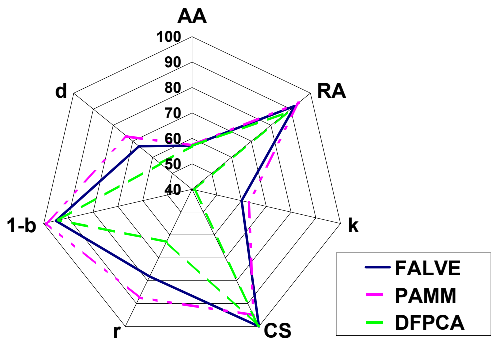

Figure 1. Radar chart of global performances of the FALVE, PAMM, and DFPCA methods.

The results are equivalent for the three methods when considering the absolute agreement (57%). However, a clear superiority appears for the method based on PAMM, when considering the other indices: relative agreement, kappa values and correlation coefficient have their highest values, while bias and standard deviation are reduced.

4. Discussion and conclusions

When developing new image processing methods, the importance of evaluation is often underestimated. Thus, it is difficult to draw conclusions relating the actual impact of the proposed method. To overcome this problem, our proposal consists in using a common database with a significant number of records. Thus, conclusions that can be drawn from a comparison are stronger.

In this paper, three methods: FALVE, PAMM and DFPCA were tested for the automated evaluation of RWMA. The results that were obtained on the test database show the superiority of the approach based on the PAMM method.

Moreover, some useful teachings can be derived from this evaluation campaign. Since the learning strategy was very different for each group, it has no sense to compare the results that were obtained on the learning database. But, it is important to stress that the three methods improved their robustness during this learning phase. Another point to mention is that the test stage evaluates the global performance of the image processing and the learning strategy. Thus, the observed differences in the performance are not only due to the image processing quantification, but also to the learning. For instance, the learning was performed for all the segments, irrespective of their localization for the FALVE and the PAMM methods while it was done separately for each type of localization for the DFPCA method. A problem with the later approach was obviously due to the reduced number of dyskinetic and akinetic cases for some specific localization. Another difference in the learning was the introduction of rules to define unclassifiable segments that was proposed for PAMM images [9]. This possibility could avoid some aberrant classifications, and was not used by the two other approaches.

Future work will include the correction of large errors of the algorithms and the development of new learning strategies in order to improve the global performance. Then the Evalechocard database should be upgraded with new studies, from various manufacturers, to enable a second test. Moreover, this database could be used for testing different tasks (detection of mitral valve opening, segmentation of the left ventricle), provided that it could be annotated by experts for the corresponding task. Moreover the annotated database can be distributed to any person who sends a requirement at:

evalecho@imed.jussieu.fr.

Acknowledgements

This work was supported in part by the Ministère de la Recherche (MNERT) and Délégation Générale de l’Armement (DGA) under Grant 04 R 428 and 04 R 429.

References

[1] Cerqueira MD, Weissman NJ, Dilsizian V, Jacobs AK, Kaul S, Laskey WK, Pennell DJ, Rumberger JA, Ryan T, Verani MS. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: a statement for healthcare professionals from the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association. Circulation 2002;105(4):539-42.

[2] Hoffmann R, von Bardeleben S, Kasprzak J, Borges A, ten Cate F, Firschke C, Lafitte S, Al-Saadi N, Kuntz-Hehner S, Horstick G, et al. Analysis of regional left ventricular function by cineventriculography, cardiac magnetic resonance imaging, and unenhanced and contrast-enhanced echocardiography: a multicenter comparison of methods. J Am Coll Cardiol. 2006;47(1):121-8.

[3] Ruiz-Dominguez C, Kachenoura N, Mulé S, Tenenhaus A, Delouche A, Nardi O, Gérard O, Diebold B, Herment A, Frouin F. Classification of segmental wall motion in echocardiography using quantified parametric images. In: Frangi A et al., editor. FIMH'05. Berlin: Springer Verlag; 2005. p. 477-486.

[4] Frouin F, Kachenoura N, Delouche A, Dumee P, Kalikian T, Guillemet H, Sarry L, Nardi O, Diebold B. EVALECHOCARD: a database in echocardiography for the comparison of methods dedicated to the estimation of regional wall motion abnormalities. In: Computers in Cardiology; 2006: 517-20.

[5] Diebold B, Delouche A, Abergel E, Raffoul H, Diebold H, Frouin F. Optimization of factor analysis of the left ventricle in echocardiography for detecting wall motion abnormalities. Ultrasound in Med & Biol 2005;31(12):1597-1606.

[6] Ruiz-Dominguez C, Kachenoura N, De Cesare A, Delouche A, Lim P, Gérard O, Herment A, Diebold B, Frouin F. Assessment of left ventricular contraction by Parametric Analysis of Main Motion (PAMM): theory and application for echocardiography. Phys Med Biol 2005;50:3277-3296.

[7] Tilmant C, Sarry L, Motreff P, Geoffroy E, Lusson JR, Boire JY. Detection of myocardium contractility defect by parietal and regional tracking in echocardiography. ITBM-RBM 2005;26(4):282-284.

[8] Frouin F, Ruiz-Dominguez C, Kalikian T, Kachenoura N, Delouche A, Herment A, Nardi O, Diebold B. Quantification of parametric images to assess segmental wall motion of the left ventricle in echocardiography. In: IEEE Computers in Cardiology. Lyon; 2005. p. 137-140.

[9] Kachenoura N, Delouche A, Ruiz-Dominguez C, Mulé S, Balvay D, Kalikian T, Herment A, Nardi O, Frouin F, Diebold B. Automatic scoring of segmental wall motion in echocardiography using quantified parametric images. In: Computers in Cardiology; 2006; p. 721-24.

[10] Corpetti T, Menin E, Perez P. Dense estimation of fluid flows, IEEE Trans Pattern Anal. Mach. Intell., 2002;24(3):365-80.

Address for correspondence

Frédérique Frouin
Inserm UMR 678 Laboratoire d’Imagerie Fonctionnelle CHU Pitié-Salpêtrière 91 boulevard de l’Hôpital 75634 Paris cedex France
frouin@imed.jussieu.fr