3D active shape modeling for cardiac MR and CT image segmentation
Assen, Hans Christiaan van

Citation

Version: Corrected Publisher’s Version
License: Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from: https://hdl.handle.net/1887/4460

Note: To cite this publication please use the final published version (if applicable).
Segmentation Performance Assessment of a Statistical Shape Model Built From Autolandmarked Training Data

This chapter was extended from:

*Segmentation Performance Assessment of a Statistical Shape Model Built From Autolandmarked Training Data*

S. Ordás, H.C. van Assen, E. Zacur, M.G. Danilouchkine, B.P.F. Lelieveldt and A.F. Frangi

To be submitted
Abstract

This paper presents an evaluation of a statistical shape model built from autolabeled training data in terms of the ultimate goal of cardiac MRI segmentation. We compared the segmentation accuracy achieved by a state-of-the-art model-based segmentation algorithm (3D-ASM driven by fuzzy inference) using three shape models built with different PDM parameterizations: radial sampling of each shape in the training set, radial sampling of the surface of an atlas representation and subsequent propagation of the landmarks to every training shape, and our autolandmarking technique, which differs from the latter in the use of uniform triangulation. Both 1- and 2-chamber cardiac models were built and tested. Only in the much simpler representation of a single ventricle the segmentation algorithm allowed for differentiating the subtle differences implied by imposing specific shape correspondence criteria in the model building procedure.

6.1 Introduction

In the last few years, model-based approaches have been used for several applications in cardiovascular image analysis [76]. Since they can capture a priori knowledge concerning motion and deformation, they are well suited for the segmentation of the cardiac chambers in image sequences. Statistical models of shape (ASM) [19] and appearance (AAM) [27] variability are two model-driven segmentation schemes initially put forward by Cootes et al. and subsequently further extended and applied by several groups to a variety of medical applications. In ASMs, the model primarily holds information about the shape in a Point Distribution Model (PDM), in which the allowed variations are depicted in terms of either Principal Component Analysis (PCA) or Independent Component Analysis (ICA) [14]. An AAM also embeds a PDM, but describes the image appearance of the object in a combined shape-appearance statistical analysis. 3D extensions of these algorithms, like 3D-ASMs [65, 77, 78] and 3D-AAMs [31, 33] have evidenced an encouraging performance in segmenting the left ventricle (LV) on MR/CT and MR/US. In a cardiac 3D-PDM, a set of landmarks has to be placed in a consistent way over a large database of training shapes, in order to guarantee that the final model will gather representative statistics of the underlying population. Even if a dense set of anatomical landmarks would exist in the object class and could be easily identified, manually specifying uniquely corresponding boundary points along the training set turns into an impractical, error-prone and subjective task. This is particularly evident in 3D, where the number of landmarks necessary to build a shape model increases dramatically. In early work, the required closely corresponding landmarks were solely defined manually; but in the last few years, there have been proposed several techniques to obtain such matches across sets of shapes or images in an automatic manner. For instance, correspondences across images were found based on geodesics and local geometry features [79], defined in the parameter domain of a spherical harmonic parametrization [35], optimized with the minimum description length of the resulting shape model [56], or by means of an elastically deformable model [16, 80]. An evaluation of some of these 3D correspondence establishing methods is provided in [17]. Some other approaches use non-rigid surface-based [81, 82] or image-based [38] registration schemes. In the first 3D applications of ASMs and AAMs [65, 77, 78], the shape models were built by ad hoc landmarking procedures without imposing specific shape correspondence criteria. In
6.2 Construction of the Statistical Shape Models

6.2.1 Training Data Set

The statistical shape models used in this research were built from a data set comprising 90 functional MRI studies of healthy and diseased hearts. Details concerning type and proportion of pathologies, as well as acquisition protocol, can be found in [75]. Manual delineations corresponding to five phases evenly distributed over the complete cardiac cycle were considered. Thus, it was possible to construct both single-phase (90 samples) and multi-phase (450 samples) PDMs for 1- and 2-chamber representations. The single ventricle model was built from the endo- and epicardial boundaries of the LV, namely LV-endo and LV-epi, respectively. The bi-ventricular model was built from the same LV-endo contour definition, plus the RV-endo and HEART-epi contours which stand for the boundaries of the right ventricle endocardium and the epicardium of the whole heart (from the base to the apex), respectively.
6.2.2 Model Building

The general layout of the model building methodology is to align and deform all the images of the training set to an atlas that can be interpreted as a mean shape. Once all necessary transformations are obtained, they are inverted and used to propagate any number of arbitrarily sampled landmarks on the atlas surface to the coordinate system of each subject. The transformations are made up of a concatenation of a global (obtained by rigid registration with nine degrees of freedom: translation, rotation, and anisotropic scaling) and local (using non-rigid registration) contributions. Any (automatically generated) set of landmarks in the atlas can be propagated to the training shapes by inverting these transformations. In this way, while it is still necessary to manually draw the contours in each training image, the technique relieves from manual landmark definition and for establishing the point correspondence across the training set. In the final step of the method, the autolandmarked shapes are normalized with respect to a reference coordinate frame, eliminating differences across objects due to rotation, translation and size. As a result, the remaining differences are solely shape-related, and PCA can be performed. The exploited algorithm can easily be set to build multi-part (open or closed) models of different configurations (e.g. 1-, 2- or 4-chamber). Moreover, its generality would allow for using it with other modalities (e.g. SPECT, CT) and organs with shape variability close to that of the heart (e.g. liver, kidneys). For a detailed description of the method the reader is referred to [18].

6.3 PDM Parameterizations

In the following paragraphs we describe the three shape models built for the segmentation performance tests.

1. Ray-Shooting on Samples (rss)
   This model does not follow any correspondence criterion other than sampling the short axis contours in the training set, slice-by-slice at constant angular increments. An approach like this was followed e.g. in [31, 65]. From base to apex, a fixed number of slices are evenly placed in all training samples. In the 2-chamber representation, the RV-endo subpart does not have a convex shape. Therefore, in order to obtain an even distribution of landmarks, equi-spaced
2. Ray-Shooting on Atlas (rsa)

The second model is built by sampling the atlas surfaces in the same way as in rss, and propagating the resulting landmarks to each shape in the training set by inverting the global and local transformations mentioned in Section 6.2.2.

3. Autolandmarked Atlas (almk)

In the third model, the landmarks are defined on the atlas surface by means of a triangulation. Adaptive mesh decimation (simplification) is performed to reduce the number of elements. The vertices of this mesh are propagated to each training shape in the same way as done for rsa. The landmark definition in the atlas surface includes the apices of the subparts. Consequently, this model has subparts with closed apices.

In summary, rss differs from the other two on the definition of landmark correspondence per subject instead of on an atlas, and rsa differs from alm in the configuration of the landmarks on the atlas. The landmarked shapes of the three PDM parameterizations were subsequently aligned with the Procrustes algorithm [57, 58, 83] using either similarity or rigid transformations, and finally PCA was performed. The mean shapes of the resulting 2-chamber PDMs are illustrated in Figure 6.1, and a comparison of their principal modes of variation in Figure 6.2 (2-chamber). Only similarity aligned cases are illustrated.
6.4 Shape Model Characterization

To investigate the statistical behavior of the constructed PDMs, we explored the differences in compactness, generalization and specificity between them, as is usually assessed in shape analysis studies (e.g. [17]). We found that these shape model properties were quite similar for all temporal phases and different number of nodes (as long as the global shape is reasonably preserved). Therefore, only the ED temporal phase is reported, using the following number of nodes: for the 2-chamber models, 2048 landmarks were defined for both rss and rsa. A decimation performed on a first dense triangulation on the atlas surface, yielded a similar number of landmarks (2110) for almk. All three 1-chamber models had the same number of landmarks (2848). Therefore, a total of 12 shape models (2 representations: 1- and 2-chamber, 3 parameterizations: rss, rsa, almk and 2 alignments: rigid (not shown) and similarity) were scaled in order to obtain, after performing PCA, mean shapes with the same size. This brought similarity and rigid body aligned shapes into the same coordinate system and allowed for their comparison.

6.4.1 Shape Analysis

From the visual inspection of Figure 6.2, the principal modes of variation of the rsa and almk 2-chamber models correspond to physiologically plausible cardiac shapes: the first mode describes differences in the aspect ratio of the hearts, the second mode indicates the large variation of the right ventricle, and the third and fourth modes describe the bending and twisting of the left ventricle. Higher modes combine vertical blends and less global deformations with decreasing impact to the global shape deformation. The 1-chamber principal modes of variation (not shown) exhibit less overall shape variation. The compactness capacity (normalized by the number of modes) of the different shape models is summarized in Figure 6.3. From Figure 6.3(a), the 2-chamber rss model has the largest cumulative variance; but for the 1-chamber representation, Figure 6.3(b) indicates the contrary. With regard to the generalization ability (leave-one-out tests), from Figure 6.4(a) all 2-chamber models were able to reconstruct the excluded shape of their training sets in a similar way, but rss ended up with 16% less error than the others. This same behavior was evidenced in the 1-chamber model (Fig. 6.4(b)) but with a notable difference (47%). The specificity of a shape model indicates how close to the training set are its generated instances. This property can assess if the statistical model generates non plausible shapes. From Figure 6.5(a), the almk model is the most specific. Nevertheless, in the 1-chamber curves of Figure 6.5(b), the rss model is the most specific. All shown curves correspond to similarity aligned shape models. They always have less cumulative variance than corresponding models aligned as rigid body. This is because in the former, the global scaling is eliminated and a considerable amount of variation is explained with aspect ratio differences. In the generalization and specificity curves, models aligned with rigid and similarity transformations behaved similarly, which is logical since these properties are measured in a normalized space. The performed analysis revealed that the better shape model properties of the almk model observed in the 2-chamber case, were not longer valid in the simpler single ventricle representation, for which the rss model performed better. The rss sampling has been defined as a stack of evenly placed sets of landmarks which do not have freedom to twist out of the image planes. Therefore, the only allowed variations are for (in-plane) displacements of the contour points.
Figure 6.3: Compactness capacity. Comparison of the cumulative variance of shape models built with different parametrization methods (only similarity alignment case shown), versus the number of modes used for their construction. Curves correspond to the ED phase of 2-chamber (a) and 1-chamber (b) representations.

In this way, a larger proportion of their cumulative variance can be explained with only the first few modes. The rsa and almk models, on the other hand, owe their variance to non-rigid deformations (in any 3D direction) of the set of landmarks defined on the atlas surface. It was thus not expected to find large differences between these two models, as they constitute two different samplings of the same dense deformation field. Some differences turned-up though, and thus are solely related to the definition of the triangulations and the inclusion of the apex. The differences between 1- and 2-chamber shape model properties of the rss model came from the fact of having used an equi-spaced (or arc-length) sampling in the RV contour, with the consequent devotion of an important proportion of its total variability to tangential displacements of the landmarks, and not on real shape variation. In [73] an illustrative example of this case is described for the shape analysis of the corpus callosum.

6.5 Segmentation Performance Assessment

A series of fitting tests was carried out with the aim to explore the potential benefits of including automatically landmarked training data in a model-based segmentation algorithm, and to assess how the previously described shape analysis properties would translate into segmentation accuracy performance. Only similarity aligned shapes were considered, as they are granted of independence to global scaling, strictly leaving solely shape-related differences between them. The fuzzy inference driven 3D-ASM algorithm of van Assen et al. was used for the performance assessment evaluation. This method is a state-of-the-art approach. Recent applications in cardiac MRI segmentation can be found in [77, 78].
Figure 6.4: Generalization ability. Comparison of the generalization ability of shape models built with different parametrization methods (only similarity alignment case shown). Curves correspond to the ED phase of rss, rsa, and almk models for (a) 2-chamber and (b) 1-chamber representations.

Figure 6.5: Specificity. Comparison of the specificity ability of shape models built with different parametrization methods (only similarity alignment case shown). Curves correspond to the ED phase of rss, rsa, and almk models for (a) 2-chamber and (b) 1-chamber representations.

6.5.1 Evaluation Data Set

The data set used for the segmentation tests comprised 30 studies at the ED and ES temporal phases. Fifteen were short axis scans of healthy volunteers acquired at the Leiden University Medical Center (Leiden, The Netherlands) using the balanced FFE protocol on a Philips Gyroscan NT Intera, 1.5 T MR scanner (Philips Medical Systems, Best, Netherlands). The slice thickness was 8 mm, with a slice gap of 2 mm and in-plane pixel resolution of $1.36 \times 1.36$ mm$^2$. The other fifteen studies corresponded to patients from the same clinical center mentioned in Section 6.2.1 (not used for building the shape models) with myocardium infarction (10), LV hypertrophy (2), and pericarditis (3).
6.5.2 Segmentation Tests

In this section, the results of the segmentation tests are presented. The 3D-ASM algorithm was set to run for a fixed number of iterations (100) using 60 modes of variation (more than 95% of the corresponding cumulative variance of all shape models tested). In the 2-chamber runs we have seen that in some cases the automatic segmentation in the area around the RV apex failed. The confounding image clues provided by this region of the image hampered the overall performance of the algorithm, as the RV linkage prevented the model to sufficiently stretch towards the heart apex. We realized that this effect was over shadowing the differences in segmentation performance brought by the use of point correspondence. To have an idea of the magnitude of the problem, the LV-endo surface had a 16% less accuracy in the 2-chamber model than in the 1-chamber counterpart, using the same appearance model parameters, and being the only equally-defined constitutive subpart in both representations. Consequently, we decided to assess the performance evaluation only with the 1-chamber model. Two patient data sets were discarded from the assessment because their automatic segmentations were not comparable to the quality of the rest (for all models). The uncorrected field inhomogeneity in one case and a severe pericarditis in the other, confounded the appearance model of the algorithm. Table 6.1 presents the segmentation accuracy, distinguishing between the LV-endo and LV-epi surfaces. Errors are expressed as the mean unsigned point-to-surface (P2S) distances from regularly sampled points on the manually segmented contours, to the surfaces of the fitted shapes, for both epicardial and endocardial surfaces. Figure 6.6 shows a typical segmentation result.

For the 1-chamber model, the segmentation performance of the different models was also tested on a labeled version of the evaluation data set, by flood-filling the manual contours of these images with different scalar values. In this way, we aimed to assess the segmentation performance of the algorithm only in regard to the models properties, independent of the characteristics of the images and their impact on the segmentation process. The difference between this experiment and the generalization ability tests reported in Figure 6.4 is basically that here all the steps of the 3D-ASM
Table 6.1: Mean ± SD of the unsigned point-to-surface (P2S) errors in millimeters.

<table>
<thead>
<tr>
<th></th>
<th>LV-epi (ED)</th>
<th>LV-endo ED</th>
<th>LV-epi (ES)</th>
<th>LV-endo ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>rss</td>
<td>2.40±0.66</td>
<td>2.89±0.82</td>
<td>3.14±1.30</td>
<td>4.05±1.45</td>
</tr>
<tr>
<td>rsa</td>
<td>2.14±0.75</td>
<td>2.27±0.75</td>
<td>2.91±1.06</td>
<td>3.78±1.38</td>
</tr>
<tr>
<td>almk</td>
<td>1.92±0.54</td>
<td>1.98±0.54</td>
<td>2.77±0.89</td>
<td>3.60±1.09</td>
</tr>
</tbody>
</table>

Figure 6.7: Segmentation on labeled data. Mean unsigned point-to-surface (P2S) errors (in millimeters) for the epicardial (a) and endocardial (b) surfaces of the 1-chamber model, for different model construction methodologies and number of modes. These results correspond to the ED phase.

algorithm (i.e. model intersection with image slices, search for candidate points, force propagation, and model instance generation) are performed, while in the shape reconstruction tests, only the model instance generation stage is assessed. Therefore, the resulting segmentation errors on ideal data give an approximation of the minimum value expected for real data. These tests were carried out only at ED, using an increasing number of modes (from 0 to 89, in steps of five). In Figure 6.7 the results are shown for both subparts. The almk performed remarkably better.

6.6 Discussion

The segmentation of the cardiac chambers does not constitute an easy task. If we consider only the portion of the heart below the base, the unpredictable appearance of the right ventricle apex makes the automatic delineation in this part very difficult. The deformation of the 2-chamber model as a whole is therefore affected and consequently the segmentation accuracy deteriorated. We did not derive any conclusion from the results of the 2-chamber segmentation tests, other than noting that with the use of a more complex representation, the fitting strategy and the appearance model robustness become the principal issue to improve. We will continue in the line of a FIS-driven model-based approach as we believe it is feasible to improve the accuracy in this particular region with other settings and considerations. Nevertheless, the RV epicardium is very difficult to segment automatically. With the use of a much sim-
pler representation like the 1-chamber model, we quite improved the reliability of the segmentation and differences with regard to the point correspondence emerged (in particular, using ideal data). In the ED phase the segmentation was good. A significant improvement was evidenced in comparing the rss and almk models, in favor of the latter: 20% for LV-epi and 31.2% for LV-endo ($p < 0.001$ in a paired t-test). Nevertheless, the main reason of these results were principally related to having a closed apex in the almk and not to the point correspondence itself. A closed shape in the apex efficiently helps model stretching in the force propagation step. Therefore, we searched for differences between the rsa and rss models, that have the same global configuration of landmarks and only differ in the point correspondence criteria. We realized that differences indeed existed for ED: 10.8% for LV-epi and 21.5% for LV-endo ($p < 0.001$ in a paired t-Test). In the ES results, the improved performance using automatic point correspondence was not statistically significant. The reason for not also having quite good results in ES, is that in this phase the papillary muscles fuse together and there are not many candidate points positioned in the "valleys" that can make the LV-endo surface cut through them. This could be solved by modifying the algorithm settings (only for ES), by not positioning the candidates points exactly in the transition myocardium-blood, but more towards inside the myocardium. The generated model instance at ES therefore underestimates the true endocardial volume. Modifications in the segmentation algorithm are thus needed.

6.7 Conclusion

In the presented work, we evaluated the segmentation performance of a 3D model-based segmentation algorithm making use of a shape model built with automatically defined point correspondence in the training data. The performance of this model was compared to the performance of models with simpler representations like the commonly used radial sampling approach. The automatically defined point correspondence in the shape model contributed to enhancing the segmentation performance, but the differences were subtle and therefore only evidenced when a good appearance model and segmentation strategy were provided. This is an important finding telling that future efforts should concentrate on this fact. Shape models of the LV like the ones used in [31, 65, 84] were presumably "appropriate" for the "simplicity" of the LV geometry, or "task-related". From our results, these assumptions were valid. However, from the shape analysis assessment and by the visual inspection of the deformation patterns, we believe that in more complex representations (e.g. 2- or 4-chamber) it would be highly desirable to have a PDM parametrization like almk, but we were not able to demonstrate this fact in a segmentation task.