Automatic pericardium segmentation and quantification of epicardial fat from computed tomography angiography

Alexander Norlén
Jennifer Alvé
David Molnar
Olof Enqvist
Rauni Rossi Norrlund
John Brandberg
Göran Bergström
Fredrik Kahl

Automatic pericardium segmentation and quantification of epicardial fat from computed tomography angiography

Alexander Norlén,^a Jennifer Alvéⁿ,^{a,*} David Molnar,^b Olof Enqvist,^b Rauni Rossi Norrlund,^b John Brandberg,^b Göran Bergström,^b and Fredrik Kahl^{a,c}

^aChalmers University of Technology, Department of Signals and Systems, Hörsalsvägen 9-11, Gothenburg 412 96, Sweden

^bGothenburg University, Sahlgrenska Academy, Institute of Medicine, The Wallenberg Laboratory, Bruna stråket 16, Gothenburg 413 45, Sweden

^cLund University, Faculty of Engineering, Centre for Mathematical Sciences, Sölvegatan 18, Lund 221 00, Sweden

Abstract. Recent findings indicate a strong correlation between the risk of future heart disease and the volume of adipose tissue inside of the pericardium. So far, large-scale studies have been hindered by the fact that manual delineation of the pericardium is extremely time-consuming and that existing methods for automatic delineation lack accuracy. An efficient and fully automatic approach to pericardium segmentation and epicardial fat volume (EFV) estimation is presented, based on a variant of multi-atlas segmentation for spatial initialization and a random forest classifier for accurate pericardium detection. Experimental validation on a set of 30 manually delineated computer tomography angiography volumes shows a significant improvement on state-of-the-art in terms of EFV estimation [mean absolute EFV difference: 3.8 ml (4.7%), Pearson correlation: 0.99] with run times suitable for large-scale studies (52 s). Further, the results compare favorably with interobserver variability measured on 10 volumes. © 2016 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: [10.1117/1.JMI.3.3.034003](https://doi.org/10.1117/1.JMI.3.3.034003)]

Keywords: computed tomography angiography; segmentation; machine learning; epicardial fat quantification; pericardium.

Paper 16037RR received Mar. 3, 2016; accepted for publication Aug. 29, 2016; published online Sep. 15, 2016.

1 Introduction

Visceral adipose tissue, i.e., fat surrounding the internal organs, may be a marker for increased risk of different metabolic and cardiovascular diseases. Epicardial fat is the visceral fat depot enclosed by the pericardial sac. In other words, it is the fat located around the heart but inside of the pericardial sac that surrounds the heart (see Fig. 1). In recent years, several studies have shown a relationship between increased volume of epicardial fat and coronary artery disease, coronary plaque, adverse cardiovascular events, myocardial ischemia, and atrial fibrillation.¹ However, due to technical limitations in three-dimensional (3-D) segmentation of epicardial fat the studies are of limited size and information on the prognostic value of epicardial fat for development of ischemic heart disease is scarce.

The Swedish CardioPulmonary Bioimage Study² (SCAPIS) is a nationwide research project that started in 2012 in a collaboration between six universities in Sweden and their university hospitals. It is a large-scale study that aims at collecting CT, MR, and ultrasound images from 30,000 men and women between 50 and 64 years of age. This database gives an opportunity to investigate the importance of epicardial fat as a risk marker for heart disease. Hence, there is a need for a fully automated method for epicardial fat quantification that is suitable for a study of this magnitude.

In this paper, an efficient method for pericardium segmentation and epicardial fat volume (EFV) estimation from computed tomography angiography (CTA) is presented. The algorithm uses efficient feature-based multiatlas registrations for spatial initialization. Thereafter, the pericardium is detected by random

forest classification and then the target image is segmented as either inside or outside of the pericardium by global optimization through graph cuts. Finally, the amount of epicardial fat can be quantified by combining thresholding with the pericardium labeling. Our experimental results on pericardium segmentation and EFV estimation show that the algorithm yields very accurate segmentations and significantly outperforms previous results on pericardium segmentation with run-times suitable for large-scale studies. More importantly, the measurement errors compare favorably to the interobserver variability measured on a set of 10 patients delineated by two medical experts using the same time-consuming and accurate method for delineation.

1.1 Contributions

The main contribution of this work is an algorithm that efficiently produces accurate EFV estimations from CTA images, making large-scale studies of the relationship between epicardial fat and heart disease tractable.

The primary algorithmic contribution is how a generalized formulation of multi-atlas segmentation based on distance maps is incorporated into a random forest classification framework. More specifically, the voxel-wise distribution of the distance to the boundary of the region of interest is used to produce rotation invariant features for the random forest classifier, effectively reducing the dimensionality of the classification problem from three dimensions to one. This not only makes the process of classification easier but also normalizes the training data leading to more efficient use of the (often in medical image analysis) limited labeled data set.

*Address all correspondence to: Jennifer Alvéⁿ, E-mail: alven@chalmers.se

Fig. 1 (a) Slice of a CTA volume. (b) The manual delineation of the pericardium is visualized. The pericardium is a thin structure and is just barely visible in the scans. The epicardial fat is the fat tissue (dark gray) inside of the pericardium. To obtain an accurate estimate of the volume of the epicardial fat it is essential to reliably locate the pericardium border, particularly in dark gray regions.

Another contribution to the research community is that all the data (both CTA volumes and their manual delineations) will be released to facilitate comparisons of algorithms and further research.

1.2 Related Work

Recently, a few methods have been developed for automated pericardium segmentation. Shahzad et al.³ use multi-atlas segmentation with majority voting. Practically, the same method was applied for cardiac segmentation by Kirisli and Schaap.⁴ Both algorithms were based on intensity-based registration (ELASTIX). Although the approach can be parallelized over several clusters they report that one segmentation takes around 20 min on a high-end computer with eight cores. Dey et al.⁵ used another intensity-based registration algorithm (DEMONS) and proposed to speed up the segmentation time by coregistering the atlases beforehand and given an unlabeled image, they only perform one atlas registration. By measuring the difference between each atlas and the target image, a weight was calculated measuring the importance of the atlas for the decision fusion. The method is relatively fast but the results regarding the actual fat estimated by their algorithm is not presented. In Ref. 6, Spearman et al. present a semi-automated method for epicardial fat estimation that uses a prototype software from Siemens Medical Solutions for initialization. The method is reported to be model-based and trained on manually annotated CTA and native scans (i.e., taken before the contrast material is administered).

Of the methods aforementioned, two present results regarding the estimated EFV. In Ref. 3, Shahzad et al. report a Pearson correlation of 0.91 with the manually estimated EFV and a linear regression coefficient 95%CI between 0.75 and 0.90. In Ref. 6, Spearman et al. report a correlation of 0.89 and measures an EFV distribution of 98.9 ± 60.2 with their algorithm compared to 65.8 ± 37.0 measured manually. Although both report a fairly high correlation, one would expect that their algorithms should produce a regression coefficient closer to 1 and an estimated EFV distribution closer to the manually measured distribution.

The method recently presented by Ding et al.⁷ seems to be more accurate reporting a regression of 0.98 on their data set containing 50 CT volumes. Their work is an extension of the

work done by Dey et al.,⁵ where the initial multiatlas segmentation is deformed by active contours driven by white lines (representing the pericardium) detected by a difference-of-Gaussians approach. They report higher correlation to manual labeling than previous attempts ($R = 0.97$).

The algorithm proposed in this work and the method by Ding et al. are similar in that they both use a multi-atlas approach for spatial initialization followed by segmentation guided by a pericardium detector. However, there are three main differences: (i) the algorithm proposed in this work is trained and validated on CTA images instead of CT images and validated on a different patient cohort (30 compared to 50 volumes); (ii) the proposed algorithm utilizes a learned classifier (random forests) compared to a hand-crafted one (difference of Gaussians) for detecting the pericardium boundary. This makes the detector more versatile, both in capturing the less deterministic attenuation introduced by administering contrast material to the patient, and it also makes the algorithm more general (e.g., making the algorithm easy to adapt to images without contrast material). However, learned classifiers put greater demand on the amount of training data, which in the proposed algorithm is solved by producing rotation invariant features leading to more efficient use of (possible) limited data; (iii) a global optimization technique (graph cuts, in our case) has advantages compared to local optimization (active contour deformations) since it does not risk getting stuck in a local optimum. Our algorithm is slightly faster but the differences in reported run times are minor.

2 Data Set

Two sets of CTA volumes with corresponding delineations of the pericardium were produced. The first set consists of 20 volumes delineated by an expert. This set was used for development of the algorithm, both for training and cross-validation. We refer to this data set as the training set. The second set consists of 10 CTA volumes delineated by the same expert and by an additional expert. This set is used for measuring interobserver variability and for evaluation of the final algorithm. This set is referred to as the test set. The two sets of 30 volumes were selected from a total of 980 examinations, as detailed as follows.

2.1 Images

Computed tomography scanning is performed using a Somatom Definition Flash scanner with a Stellar detector (Siemens Healthcare, Forchheim, West Germany). Care Dose 4D, Care kV and SAFIRE are used for dose optimization. The information on epicardial fat was retrieved from images generated during a coronary CTA. Procedures have been described in detail in Ref. 2. Briefly, all cardiac imaging is electrocardiogram triggered. Heart rate is controlled at around 60 beats/min using a beta-blocker and maximal vasodilatation is induced using sublingual glyceryl nitrate. For coronary CTA, the contrast medium iohexol is administered (350 mg I/mL; Omnipaque; GE Healthcare, Stockholm, Sweden). The individual dose is 325 mg I/kg body weight and the injection time is 12 s. Five different acquisition protocols were used dependent on body weight, heart rate, and heart rate variability.

In total, 1111 subjects were recruited to the pilot study of which 980 performed a full coronary CTA examination.² A subset of 30 examinations were selected. The image set was chosen with equal representation of men and women and also to represent a range of different body mass indexes (BMIs). This was

Table 1 Demographics of the subjects present in the data sets used for training and evaluation of the algorithm.

Variable	Training set	Test set	Total
N	20	10	30
Sex, female (<i>n</i> , %)	10 (50)	5 (50)	15 (50)
Age (median, range)	57 (51 to 65)	58 (50 to 65)	57 (50 to 65)
BMI (median, range)	27.4 (17.4 to 40.2)	30.1 (17.9 to 40.1)	28.0 (17.4 to 40.2)

deemed suitable since EFV correlates with BMI. Demographics of these subjects are shown in Table 1. The images have resolutions ranging between $512 \times 512 \times 342$ and $512 \times 512 \times 458$ voxels with voxel dimensions between $0.32 \times 0.32 \times 0.30$ mm³ and $0.43 \times 0.43 \times 0.30$ mm³.

The study was approved by the ethics committee at Umeå University and adheres to the Declaration of Helsinki. Informed consent was collected from all subjects.²

2.2 Manual Delineations

The manual delineations were done by two medical experts, both specialists in thoracic radiology. The pericardium was delineated on every 10th slice in the three standard orthogonal planes (axial, coronal, and sagittal) independently. Delineation in two dimensions was preferred to a possible method of segmenting directly in three dimensions to (i) ensure maximal anatomical precision, as radiologists are more comfortable with viewing structures in two dimensions at a time, (ii) be able to precisely reproduce the circumstances for the two experts. The same slices were delineated by both experts.

During segmentation, if the pericardium was not clearly visible in parts of the actual slice, a decision was made where the pericardium was most probably located based on the neighboring slices and the experts' anatomical knowledge. This approach was particularly useful in the areas where many different anatomical structures are close to each other, e.g., the diaphragmatic surface of the pericardium.

Delineation in all three planes was mainly done because of the problem of delineating structures parallel to the plane of viewing, resulting in poor accuracy in these areas. The slice-wise segmentations, made in each orthogonal plane independently, were interpolated into three volumes. The final resulting volume was computed as the volume where two out of the three volumes overlapped, assuming that this would reduce the error, mainly stemming from the problem of tangential delineation mentioned above. The final volume was approved by the expert. We refer to the manual labeling as the gold standard.

3 Method

The developed algorithm consists of three main parts. The first part is the spatial initialization (Sec. 3.1) using efficient feature-based multi-atlas techniques. This first part serves as a global initialization for pericardium localization, reducing the need for an explicit shape model. A variant of multi-atlas representation (denoted as MADMAP) provides valuable information of the certainty of the voxels being inside or outside of the

pericardium. With this information, we can limit the pericardium search space to a small region around the pericardium surface.

The second part of the algorithm is the pericardium detection (Sec. 3.2). A random forest classifier is trained on the labeled atlas set to accurately detect the pericardium. The extracted image features used for training and classification are aligned along a direction estimated from the MADMAP to be perpendicular to the pericardium, practically reducing the pericardium detection problem to a line search. This approach also expands the effectively used amount of training data because it lets the forest learn what a pericardial neighborhood looks like irrespective of how it is oriented toward the image coordinate axes (an important consideration in medical image analysis where manually labeled data rarely is abundant). The classifier is trained to distinguish between four classes:

1. just inside of the pericardium,
2. just at the pericardium boundary,
3. just outside of the pericardium,
4. everything else.

This makes detailed information of what the boundary looks like available to the forest during training and produces a classifier with a high discriminating power.

The final part is segmentation (Sec. 3.3). The information from the global spatial initialization and from local and independent posteriors estimated by the random forest classifier is combined into a Markov random field (MRF). The globally optimal segmentation is computed through graph cuts. Figure 2 summarizes the main parts of the algorithm.

3.1 Spatial initialization

Multiatlas segmentation (see for example Ref. 8), which is used by almost all previous methods for pericardium segmentation (including this one), is a widely used method for organ segmentation in medical image analysis. An atlas is an image with a corresponding labeling $\mathcal{L}$. Standard multiatlas segmentation involves registering each atlas image to the target image, followed by transferring the atlas labeling to produce a vote map. The proposed algorithm includes a spatial initialization including feature-based registration and a generalized representation of the standard multiatlas vote map.

3.1.1 Feature-based registration

In medical applications, the registration methods are typically intensity-based and nonrigid, e.g., as in Ref. 3, which tend to be computationally very demanding. As our intention is to apply our framework to thousands of images, a more efficient method is required. In contrast to intensity-based methods, feature-based registration is less common in medical image analysis due to the conception that it is hard to detect salient features in medical images. However, as was shown by Svärm et al.,⁹ feature-based registration based on robust optimization techniques outperforms a variety of intensity-based methods in estimating affine transformations for whole-body CT scans as well as brain MR scans. Feature-based registration was both more efficient and less likely to produce large errors.

We use a 3-D version of the difference-of-Gaussians detector in SIFT¹⁰ together with the descriptor from SURF.¹¹ We use

Fig. 2 Visualization of the main parts of the algorithm. (a) Sagittal view of a target volume to be segmented. In this slice, the pericardium is barely visible as a thin white line in the fat tissue (dark gray). (b) The probability map constructed using the MADMAP, where white corresponds to a high probability of the voxel being inside of the pericardium and black corresponds to a low probability. The gray contour defines the region of uncertainty defined by the probability map. (c) The posterior probabilities of the voxels being just at the pericardium boundary estimated by the random forest classifier where white corresponds to a high probability of the voxel being at the pericardium boundary and black a low probability. (d) The gold standard (white contour) and the final segmentation (black contour). The gray contour defines the region of uncertainty.

rotation invariant features. The features are matched between the images using the ratio criterion used by Ref. 10, referred to as the Lowe criterion, i.e., we discard matches where the ratio of the distance from the closest neighbor to the distance of the second closest are larger than a threshold. Given the match hypotheses, RANSAC¹² is used to obtain the matches that are approximately consistent with an affine transformation. RANSAC is run with the l_1 -norm (truncated at a threshold) as a cost function and with 50,000 iterations. Only unique matches are allowed. If there is a matching conflict, then the match that is closest in descriptor space is used. Through this process a set of matches, mostly cleared from outliers, is obtained. We only use features in the atlas images that are within 10 mm or inside of the pericardium, thus completely ignoring other anatomical regions in the atlases.

Finally, the nonrigid deformations around the heart are estimated by registering the final feature matches (the ones considered as inliers by the RANSAC algorithm). We represent the deformations with B-spline and use an implementation based on the registration algorithm by Ref. 13 with a final B-spline grid size of 4 mm.

3.1.2 Multi-atlas distance map

The MADMAP is a generalized representation of the standard multi-atlas vote map. What is usually done is that the (binary)

manually labeled images produced by the experts are transformed into the space of the target image resulting in a vote map where the information at each voxel is the number of atlases that vote for this voxel being inside or outside of the region of interest. Exactly the same procedure is used here with the modification that instead of the atlas labels, the signed distances to the pericardium are transformed into the space of the target image. A similar approach was proposed in Ref. 14.

The proposed approach is a minor change to the standard one but it results in a major information gain regarding the multi-atlas registrations at no extra computational cost. For each voxel, the atlases vote for the signed distance to the boundary of the region of interest. Not only does this give us the possibility of estimating the actual distance to the real boundary, we also obtain a voxel-wise measure of uncertainty of the estimated signed distance (and by extension a measure of uncertainty of the atlas registrations around the voxel) by measuring the variance of the votes. This approach generalizes the standard multi-atlas voting procedure; the standard votes are obtained as a special case by only counting negative votes, e.g., majority voting fusion is obtained as all voxels where the median of the votes is less than zero.

The MADMAP (denoted M) is the object containing all distance votes. In this work, we use a compact representation of the MADMAP where we only save the voxel-wise median of the votes (denoted $\tilde{M}$) and the voxel-wise mean absolute deviation

from the median (denoted $D[M]$). We refer to this representation as the l_1 -norm representation of M . We also validate the accuracy of the algorithm when using the l_2 -norm, i.e., the voxel-wise mean and standard deviation of the distance votes. For simplicity, the l_1 -norm notation is used for the rest of the presentation of this algorithm.

The median $\bar{M}$ is an estimation of the distance transform of the pericardium in the target image. The median $\bar{M}$ and the deviation $D[M]$ are used to compute a probability of a location p being inside $\Pr(p \in \mathcal{L}|M)$ or outside $\Pr(p \in \mathcal{L}^c|M)$ of the pericardium by assuming a normally distributed measurement error. This probability map is used to define a region of uncertainty, i.e., locations that are not definitely inside and not definitely outside according to the MADMAP. For efficiency, the pericardium search is limited to this region. For a visualization, see Fig. 2(b).

3.2 Pericardium Detection

Multi-atlas registration is a robust method for spatial estimation of where the pericardium is approximately located. But since the pericardium does not constitute a clearly visible boundary for the region of interest, which would guide the registrations, the actual placement of the segmentation boundary will not be accurate. Therefore, we train a boundary detector that, given the spatial initialization from MADMAP, will respond to the image features that resemble the pericardium surface.

The boundary detector is based on random decision forests,^{15,16} a machine learning technique suitable for this classification tasks since it generalizes well to unseen data, naturally extends to multiclass classification problems and is computationally efficient.

3.2.1 Training

The forest is trained to distinguish between four classes. Just inside of the pericardium (between -1.5 and -0.5 mm from the pericardium), just at the pericardium boundary (-0.5 to

0.5 mm), just outside of the pericardium (0.5 to 1.5 mm) and background (any other location between -8 and 8 mm). These classes are denoted $c \in \{\text{in, on, out, bg}\}$. An equal amount of locations are randomly sampled from each class and the corresponding features are extracted. The axis aligned splitting function is used where the splitting function is defined as a hyperplane aligned along one of the axes. The hyperplane is defined by the axis and a threshold. The splitting functions are chosen as the function that maximizes the information gain (Shannon entropy). For training, a total of 40 million data points were extracted, evenly distributed over the classes and the atlases.

3.2.2 Features

The feature vector [the data point $\vec{v}(p)$ extracted at location p] consists of mean values and l_1 -variance of the image intensities and first- and second-order derivatives and gradient magnitudes of the image intensities, extracted from local regions around p . The regions are oriented along the normal direction of the pericardium surface as predicted by MADMAP, effectively reducing the dimensionality of the classification problem from three dimensions to one.

The feature elements at a location p are computed as follows. Let $G = \{G_{i,j,k}\}_{i,j,k=1}^3$ be an equidistant $3 \times 3 \times 3$ grid of points centered at the origin. The spacing between the points is 1 mm. Let R be the rotation that aligns the third dimension of G (indexed by k) with the gradient of $\bar{M}$ at p . Let $S_{i,j,k} = I(p + R \cdot sG_{i,j,k})$ be the intensities of the image I sampled at the location specified by the grid point $G_{i,j,k}$ (which has been centered at location p , scaled by a factor s , and rotated along the MADMAP gradient). A set of local image statistics $T(p, I, s)$ consists of

$$\text{Mean: } m = (1/27) \sum_{i,j,k=1}^3 S_{i,j,k} \text{ and } \sum_{i,j,k=1}^3 |S_{i,j,k} - m|,$$

$$\text{Means: } m_k = (1/9) \sum_{i,j=1}^3 S_{i,j,k} \text{ and } \sum_{i,j=1}^3 |m_k - S_{i,j,k}|, \quad k = \{1, 2, 3\},$$

$$\text{First gradient: } \sum_{k=1}^2 \sum_{i,j=1}^3 S_{i,j,k+1} - S_{i,j,k} \text{ and } \sum_{k=1}^2 \sum_{i,j=1}^3 |S_{i,j,k+1} - S_{i,j,k}|,$$

Fig. 3 Validation of the parameters of the multi-atlas initialization. The results are measured in dice index between the overlap of the epicardial volume (not only the fat volume). The results are presented as the mean of the 20 samples in the training set (solid line) and 95% confidence interval assuming a normal distribution (dashed line). (a) The effect of changing the Lowe threshold for the feature matching. (b) The effect of changing the inlier threshold when estimating the affine transformation with RANSAC. (c) The effect of using different numbers of atlases and the l_1 - and the l_2 -norm for the MADMAP construction. The atlases with the most inlier matches are chosen.

First gradients: $\sum_{i,j=1}^3 S_{i,j,k+1} - S_{i,j,k}$ and $\sum_{i,j=1}^3 |S_{i,j,k+1} - S_{i,j,k}|$, $k = \{1, 2\}$,
 Second-order derivative: $\sum_{i,j=1}^3 -S_{i,j,3} + 2S_{i,j,2} - S_{i,j,1}$ and $\sum_{i,j=1}^3 |-S_{i,j,3} + 2S_{i,j,2} - S_{i,j,1}|$.

The features are sampled from the CT volume I_0 , the same volume filtered with a Gaussian kernel with $\sigma = 1$ mm (I_1) and with $\sigma = 2$ mm (I_2). The complete list of features extracted from each location p is $I_0(p)$, $I_1(p)$, $I_2(p)$, $T(p, I_0, 1)$, $T(p, I_0, 1.5)$, $T(p, I_1, 1.5)$, $T(p, I_1, 2)$, $T(p, I_2, 2)$, and $T(p, I_2, 3)$. A total of 99 features.

3.3 Segmentation

By viewing the image I as an observation of a MRF^{17,18} and the realization of the field as the labeling $\mathcal{L}^*$ of the voxels, the labeling that maximizes the *a posteriori* probability can be inferred by minimizing an energy function of the form

$$E(\mathcal{L}^*|I) = \sum_{p \in \mathcal{P}} V_p(l_p; i_p) + \sum_{(p,q) \in \mathcal{N}} V_{p,q}(l_p, l_q; i_p, i_q), \quad (1)$$

where $\mathcal{P}$ is the set of all pixels (or voxels) in the image and $\mathcal{N}$ is the set of all neighbors. Here V_p is referred to as the unary cost and $V_{p,q}$ the pairwise cost. A function on this form can be formulated as a weighted graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$. If E in Eq. (1) is sub-modular, the globally optimal segmentation $\mathcal{L}^*$ can be computed in polynomial time.

The MADMAP has been used to compute the probability of a location being inside $\Pr(p \in \mathcal{L}|M)$ or outside $\Pr(p \notin \mathcal{L}|M)$ of the pericardium and a six-connected graph is constructed over the region of uncertainty. The set of locations $\mathcal{P}_v$ corresponding to the nodes $\mathcal{V}$ in the graph are classified by the random forest producing a distribution $\Pr[p \in c|\vec{v}(p)]$ over the set of classes $c \in C$, for each $p \in \mathcal{P}_v$. Figure 2(c) presents an example of what $\Pr[p \in \text{on}|\vec{v}(p)]$ can look like. To control the amount of influence, the MADMAP probabilities have on the final segmentation, we introduce the parameter μ and define the parameterized MADMAP probability M^{in} as

Table 2 Result comparison between the proposed method versus Expert 1 and Expert 2 versus Expert 1.

	Proposed versus Expert 1	Expert 2 versus Expert 1
Mean absolute EFV difference (ml)	2.68	5.10
Median absolute EFV difference (ml)	2.22	3.82
EFV (ml) (Expert 1: 108.44 $\pm$ 74.65)	109.22 $\pm$ 75.11	103.34 $\pm$ 74.82
Pearson correlation	0.9989	0.9986
Linear regression coefficient (95% CI)	1.01 (0.97, 1.04)	1.00 (0.96, 1.04)
Bland-Altman bias (ml) (95% CI)	0.78 (-6.31, 7.86)	-5.10 (-12.88, 2.67)
Dice (mean $\pm$ std)	0.91 $\pm$ 0.04	0.90 $\pm$ 0.04
Dice total volume (mean $\pm$ std)	0.97 $\pm$ 0.01	0.98 $\pm$ 0.00

Note: The comparisons are of the measured EFV in all cases except for dice total volume, where the overlap of the total epicardial volume is measured.

$$M^{\text{in}}(p) = \frac{1}{1 + \left[\frac{\Pr(p \notin \mathcal{L}|M)}{\Pr(p \in \mathcal{L}|M)} \right]^{\frac{1}{\mu}}}. \quad (2)$$

The unary costs of the MRF energy function in Eq. (1) are defined as

$$V_p(1) = -\log(M^{\text{in}}(p)\{1 - \Pr[p \in \text{out}|\vec{v}(p)]\}), \quad (3)$$

$$V_p(0) = -\log([1 - M^{\text{in}}(p)]\{1 - \Pr[p \in \text{in}|\vec{v}(p)]\}), \quad (4)$$

Fig. 4 Validation of the parameters of the random forest and the MRF. The results are presented as the mean absolute difference of EFV compared to the expert measurements of the 20 samples in the training set (solid line) and the standard deviation of the difference (dashed line). (a) The effect of training the forests with 5, 10, and 15 candidate features and using different number of decision levels for classification. (b) The effect of changing the number of trees used for classification. (c) The effect of changing the multi-atlas parameter μ . (d) The effect of changing the regularization parameter r .

Fig. 5 Comparison of the measurements of the EFV by Expert 1 and by the proposed algorithm on the test set (10 samples). (a) Correlation and regression analysis. (b) Bland–Altman plot.

where for shorthand notation, we have excluded the obvious dependence on I . In other words, the cost of assigning a node to inside the pericardium is small if the probability of p being inside is large according to the MADMAP and if the probability of p being just outside of the pericardium is small according to the random forest.

For each edge $\{p, q\} \in \mathcal{E}$, we define its location as the location between the nodes connected by the edge, i.e., $(p + q)/2$. It is classified by the random forest and we infer a probability of the edge being on the pericardium boundary, $\Pr\{(p + q)/2 \in \text{on}[\vec{v}[(p + q)/2]]\}$. The pairwise costs are then defined as

$$\begin{aligned} V_{p,q}(1, 0) &= V_{p,q}(0, 1) \\ &= \min[-r \log(\Pr\{(p + q)/2 \\ &\in \text{on}[\vec{v}[(p + q)/2]]\}), \tau]. \end{aligned} \quad (5)$$

Two parameters are introduced, the regularization r controlling the weighting between the unary and the pairwise costs and

the uncertainty threshold τ that specifies the maximum cost for an edge. Nodes with infinite unary costs are appended directly on the inside and outside of the region of uncertainty, forcing the boundary into this region.

The final segmentation $\mathcal{L}^*$ is chosen as the max-flow/min-cut over the graph [see Fig. 2(d)] and is inferred using the max-flow algorithm,¹⁹ which is a widely used method in computer vision.

3.4 Hyperparameter Optimization

The algorithm was validated using leave-one-out on the training set consisting of 20 CTA volumes of the heart and the corresponding gold standard.

The validation was first done on the hyperparameters of the MADMAP. A MADMAP was constructed for each of the images by registering the remaining atlases affinely to the target image. The atlases with the most inlier matches were registered nonrigidly and their distance transforms were subsequently propagated to the target image. The parameters of the registrations were optimized against the mean dice index of the total

Fig. 6 Comparison of the measurements of the EFV by Expert 1 and by Expert 2 on the test set (10 samples). (a) Correlation and regression analysis. (b) Bland–Altman plot.

Table 3 Result comparison between the proposed method versus Expert 1 and the compared method in Ref. 21 versus Expert 1.

	Proposed method	Compared method ²¹
Mean absolute EFV difference (ml)	2.68	21.86
Median absolute EFV difference (ml)	2.22	9.83
EFV (ml) (Expert 1: 108.44 ± 74.65)	109.22 ± 75.11	130.22 ± 98.59
Pearson correlation	0.9989	0.9911
Linear regression coefficient (95% CI)	1.01 (0.97, 1.04)	1.31 (1.17, 1.45)
Bland-Altman bias (ml) (95% CI)	0.78 (-6.31, 7.86)	21.77 (-30.26, 73.80)
Dice (mean ± std)	0.91 ± 0.04	0.82 ± 0.04
Dice total volume (mean ± std)	0.97 ± 0.01	0.95 ± 0.01

Note: The comparisons are of the measured EFV in all cases except for dice total volume, where the overlap of the total epicardial volume is measured.

epicardial volume computed between the region, where the median of the MADMAP was less than 0 (equivalent to majority voting in standard multiatlas segmentation) and the gold standard.

The best MADMAPs were then used as initialization for validating the hyperparameters of the pericardium detection and segmentation. The hyperparameters of the random forest and the MRF used for the final segmentation of the images were optimized against the mean absolute EFV difference. For each image, the algorithm was trained on the remaining images.

3.5 Epicardial Fat Volume Quantification

The intensity values of the voxels in a CTA image correspond directly to Hounsfield units (HU). Usually, the fat in the image is found by simple thresholding. In this work, fat is defined as all voxels with an attenuation between -192 and -30 HU,²⁰ which combined with the pericardium segmentation allows for quantification of the EFV.

4 Experiments and Results

4.1 Hyperparameter Optimization

Figure 3 presents some results from the MADMAP hyperparameter optimization. As can be seen, the results are not sensitive to the choice of Lowe threshold for the matching [see Fig. 3(a)]. A threshold of 0.975 was chosen for the Lowe criterion. The outlier matches were handled by RANSAC, where the inlier threshold was chosen to 15 mm [Fig. 3(b)]. Finally, we evaluated the effect of only using the atlases with the most inlier matches for the nonrigid registration and looked at the effects of using the mean of the MADMAP instead of the median (l_2 instead of l_1). The results are presented in Fig. 3(c). Interestingly, by only using a few of the atlases for the final construction of the MADMAP, the initialization gets slightly more robust and of course it makes it more efficient. Also, the l_1 -norm slightly outperforms the l_2 -norm, especially when using more atlases. The l_1 -norm and using the six atlases with most inliers were chosen for the final parameter set. The region of uncertainty is defined as the region where either $|\hat{M}(p)| < 8$ mm or $0.0001 < \Pr(p \in \mathcal{L}|\mathcal{M}) < 0.9999$.

Figure 4 shows some results obtained during the random forest and MRF hyperparameter optimization. The forest was trained using 5, 10, and 15 candidate features (the size of the random subsets of features used for optimization of the splitting functions). About 15 candidate features and 19 decision levels, the maximum allowed depth allowed by the current implementation, were chosen [see Fig. 4(a)]. Overtraining did not seem to be a concern. Interestingly, when the trees were trained in this manner, one obtains the same results with just a few trees [see Fig. 4(b)]. In fact, the results were stable using only one tree in

Fig. 7 Comparison of the measurements of the EFV by Expert 1 and by the compared method in Ref. 21 on the test set (10 samples). (a) Correlation and regression analysis. (b) Bland-Altman plot.

Table 4 Comparison between the measurements by the proposed method and Expert 1 on the complete set of 30 hearts (both the training and the test set).

Proposed versus Expert 1	
Mean absolute difference (ml)	3.84
Median absolute difference (ml)	1.88
EFV (ml) (Expert 1: 92.44 ± 51.86)	91.04 ± 51.26
Pearson correlation	0.9923
Linear regression coefficient (95% CI)	0.98 (0.93, 1.03)
Bland–Altman bias (ml) (95% CI)	-1.40 (-14.02, 11.21)
Dice (mean $\pm$ std)	0.91 ± 0.03
Dice total volume (mean $\pm$ std)	0.97 ± 0.01

Note: The comparisons are of the measured EFV in all cases except for dice total volume where the overlap of the total epicardial volume is measured.

the forests. To be sure, we chose 10 trees for the final parameter set.

The graph was not constructed directly on the voxels of the image but was first downsampled. An isometric node spacing of 1 mm made the algorithm more efficient and at the same time proved to provide enough detail to not affect the accuracy of the segmentation. The algorithm was not very sensitive to the multi-atlas parameter μ [see Fig. 4(c)] and it was set to 20. The pairwise cost parameters r and τ were set to 2.5 and 10, respectively. The effect of different r can be seen in Fig. 4(d).

4.2 Pericardium Segmentation and Epicardial Fat Volume Estimation

The proposed algorithm was trained on the 20 atlases in the training set and tested on the 10 samples in the test set, which were unseen during development of the algorithm. The pericardium in each of the 10 test samples was manually

delineated by the same expert (Expert 1) who delineated the samples in the training set and by another expert (Expert 2) for interobserver comparisons. The results are presented in Table 2. Regression analysis and Bland–Altman plots between the proposed method and Expert 1 and between Expert 2 and Expert 1 are visualized in Figs. 5 and 6, respectively. The average total segmentation time was 51.9 s (Intel Core i7-43930k @3.40GHz with 6 cores).

4.3 Comparison to State-of-the-Art Segmentation Method

In addition, our method was compared to the multiatlas-based segmentation described in Ref. 21 (using joint label fusion with corrective learning). Their method won the first place of the multiatlas labeling challenge at MICCAI 2012²² and was one of the top performers in the Segmentation: Algorithms, Theory, and Applications challenge at MICCAI 2013²³ including data from the Cardiac Atlas Project. For these challenges, their approach outperformed several other well-known label fusion approaches such as STAPLE.²⁴ The comparison was carried out using the same training (20 atlases) and the same test set (10 atlases) for both methods, and the same spatial initialization (as presented in Sec. 3.1). We used the authors' own implementation. The results of the comparison are presented in Table 3. Regression analysis and Bland–Altman plots between the compared method and Expert 1 are visualized in Fig. 7. As can be seen, the performance of the label fusion plus corrective learning does not reach the accuracy level of our approach.

4.4 Leave-One-Out Cross Validation

For completeness, we also present the comparison between the proposed method and Expert 1 after cross-validation on both the training set and the test set. For each image, the proposed method is trained on the 29 remaining images. This gives us a more comprehensive set consisting of 30 samples. The results are presented in Table 4 and regression and Bland–Altman analysis is presented in Fig. 8.

Fig. 8 Comparison of the measurements of the EFV by Expert 1 and by the proposed algorithm on the training and the test set (30 samples). (a) Correlation and regression analysis. (b) Bland–Altman plot.

5 Conclusions

In this work, we have presented a general segmentation framework that couples multitlas segmentation with a random forest boundary detector trained on labeled images in an atlas set. The algorithm is applied to the problem of pericardium segmentation (EFV estimation), which is a demanding problem because of the lack of salient image features around the segmentation boundary (the pericardium is a thin membrane, barely visible to the naked untrained eye).

The automated method performed extraordinary well on the test set producing a mean absolute difference of 2.7 ml and a correlation of 0.9989 compared to the manual measurements of Expert 1. There is no significant bias present between Expert 1 and the proposed method (Bland–Altman bias of 0.8 ml). The mean absolute difference between Expert 1 and Expert 2 was 5.10 ml with a correlation of 0.9986 indicating that the proposed algorithm actually could outperform the manual measurements of Expert 2 in terms of measuring the EFV as Expert 1. Further, the proposed method outperformed the popular label fusion scheme in Ref. 21, which has proved to produce state-of-the-art accuracy for diverse medical image segmentation tasks.

For a more comprehensive analysis, we also evaluated the algorithm on both the test and the training set (cross-validation with a total of 30 samples). The algorithm still produced state-of-the-art results with a mean absolute difference of 3.8 ml and a correlation of 0.9923 compared to the measurements of Expert 1.

The best previous method for EFV quantification, known to the authors, report a correlation of 0.97 and a 95% confidence interval between −18.43 and 14.91 ml measured on 50 CT images of the heart.⁷ By using our proposed method on CTA images, we report a correlation of 0.99 and a 95% confidence interval between −14.02 and 11.21 ml. Both algorithms have approximately the same run-times. Note should be taken to the fact that the methods are evaluated on different data sets and the results are therefore not directly comparable. Our algorithm is the first to produce accurate results on CTA images and it is general enough to easily be adapted to images without contrast material.

Since the proposed method produced state-of-the-art results for EFV quantification, outperformed the state-of-the-art segmentation method based on label fusion and compared favorably with the interobserver variability, we conclude that this algorithm can be used for large-scale studies of the prognostic importance of epicardial fat.

To further validate the algorithm, exposure to a larger population than 30 patients is necessary. Therefore, future work includes validating the algorithm on a set of (at least) 200 patients. To make the manual delineations tractable, the algorithm will be evaluated on randomly chosen slices of the CTA image, rather than the EFV of the complete volume.

Acknowledgments

This work was supported by the Swedish Research Council under Grant no. 2012-4215 and by the Swedish Heart-Lung Foundation. The authors declare there are no conflicts of interest pertaining to this manuscript.

References

1. D. Dey et al., “Epicardial and thoracic fat—noninvasive measurement and clinical implications,” *Cardiovasc. Diagn. Ther.* **2**, 85–93 (2012).

2. G. Bergström et al., “The Swedish cardiopulmonary bioimage study: objectives and design,” *J. Intern. Med.* **278**(6), 645–659 (2015).
3. R. Shahzad et al., “Automatic quantification of epicardial fat volume on non-enhanced cardiac CT scans using a multi-atlas segmentation approach,” *Med. Phys.* **40**, 091910 (2013).
4. H. Kirisli and M. Schaap, “Fully automatic cardiac segmentation from 3D CTA data: a multi-atlas based approach,” *Proc. SPIE* **7623**, 762305 (2010).
5. D. Dey et al., “Automated algorithm for atlas-based segmentation of the heart and pericardium from non-contrast CT,” *Proc. SPIE* **7623**, 762337 (2010).
6. J. V. Spearman et al., “Automated quantification of epicardial adipose tissue using CT angiography: evaluation of a prototype software,” *Eur. Radiol.* **24**, 519–526 (2014).
7. X. Ding et al., “Automated pericardium delineation and epicardial fat volume quantification from noncontrast CT,” *Med. Phys.* **42**(9), 5015–5026 (2015).
8. A. S. El-Baz et al., Eds., *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*, Vol. I, Springer Science + Business Media, New York (2011).
9. L. Svärm et al., “Improving robustness for inter-subject medical image registration using a feature-based approach,” in *Int. Symp. on Biomedical Imaging* (2015).
10. D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vision* **60**, 91–110 (2004).
11. H. Bay et al., “Speeded-up robust features (SURF),” *Comput. Vision Image Understanding* **110**, 346–359 (2008).
12. M. A. Fischler, R. C. Bolles, and J. D. Foley, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Commun. ACM* **24**, 381–395 (1981).
13. S. Lee, G. Wolberg, and S. Y. Shin, “Scattered data interpolation with multilevel b-splines,” *IEEE Trans. Visual Comput. Graphics* **3**, 228–244 (1997).
14. C. Sjöberg and A. Ahnesjö, “Multi-atlas based segmentation using probabilistic label fusion with adaptive weighting of image similarity measures,” *Comput. Meth. Programs Biomed.* **110**(3), 308–319 (2013).
15. L. Breiman, “Random forests,” *Mach. Learn.* **45**(1), 5–32 (2001).
16. A. Criminisi and J. Shotton, Eds., *Decision Forests for Computer Vision and Medical Image Analysis*, Springer, London (2013).
17. C. Sutton and A. McCallum, “An introduction to conditional random fields,” *Foundations and Trends® Mach. Learn.* **4**(4), 267–373 (2012).
18. C. Wang, N. Komodakis, and N. Paragios, “Markov random field modeling, inference & learning in computer vision & image understanding: a survey,” *Comput. Vision Image Understanding* **117**, 1610–1627 (2013).
19. Y. Boykov and V. Kolmogorov, “An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision,” *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 1124–1137 (2004).
20. B. Chowdhury et al., “A multicompartiment body composition technique based on computerized tomography,” *Intern. J. Obes.* **18**(4), 219–234 (1994).
21. H. Wang et al., “Multi-atlas segmentation with joint label fusion,” *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(3), 611–623 (2013).
22. B. A. Landman and S. K. Warfield, “MICCAI 2012 multi-atlas labeling challenge,” in *MICCAI 2012 Workshop on Multi-Atlas Labeling*, Nice, France (2012).
23. A. Asman et al., “MICCAI 2013 segmentation: algorithms, theory and applications (SATA) challenge results summary,” in *MICCAI 2013 Challenge Workshop on Segmentation: Algorithms, Theory and Applications* (2013).
24. S. K. Warfield, K. H. Zou, and W. M. Wells, “Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation,” *IEEE Trans. Med. Imaging* **23**(7), 903–921 (2004).

Alexander Norlén received his Master of Science degree in engineering physics from Lund University, Sweden, in 2014, and thereupon remained as a research project employee at the Computer Vision and Medical Image Analysis Group, Chalmers University of Technology, where he had done the research for his master’s thesis. Currently, he is a software developer at 3Shape AS in Copenhagen, Denmark.

Jennifer Alvén received her Master of Science degree in engineering mathematics from Lund University, Sweden, in 2015. She is a PhD student at the Computer Vision and Medical Image Analysis Group, Chalmers University of Technology. Her main research area is machine learning techniques in medical image analysis.

David Molnar received his Master of Science in medicine degree in 2001 and was granted his medical license in 2003. Currently, he is doing his PhD in radiology at the Department of Molecular and Clinical Medicine, Sahlgrenska University Hospital. He is a specialist in radiology in 2013 and a subspecialist in thoracic radiology in 2015. His main research interest is automated image interpretation in cardiac computed tomography.

Olof Enqvist received his Master of Science degree from Linköping University, Sweden, in 2006, and his PhD in mathematics from Lund University, Sweden, in 2011. He worked as a postdoctoral researcher at Lund University from 2011 to 2013. Since 2013, he has been an assistant professor at Chalmers University of Technology. Two common research themes are robust optimization techniques and medical image analysis.

Rauni Rossi Norrlund received her PhD degree in radiation physics and immunology: *Improving Experimental Tumor Radioimmunotargeting* from the Department of Diagnostic Radiology, University of Umeå, Sweden, in 1977. She received her medical doctor degree from the University of Tampere, Finland, in 1988. Her

specialist certifications are diagnostic radiology in 1994 and nuclear medicine in 2013. Her present position is a senior radiologist at the Thoracic Radiology Department, Sahlgrenska University Hospital, for the last two decades.

John Brandberg received his PhD from the Department of Radiology, Institute of Clinical Sciences at Sahlgrenska Academy in 2009. He is currently an adjunct lecturer at the same department.

Göran Bergström is the head of the Physiology Group, Wallenberg Laboratory, and senior consultant in clinical physiology at the Vascular Diagnostic Unit, Sahlgrenska University Hospital. He is the chair of the Swedish Cardiopulmonary Bioimage Study (SCAPIS), which aims to recruit and extensively phenotype 30,000 subjects aged 50 to 64 years at six Swedish university hospitals. The ultimate goal of SCAPIS is to reduce mortality and morbidity from cardiovascular disease, chronic obstructive pulmonary disease, and related metabolic disorders.

Fredrik Kahl received his PhD in mathematics from Lund University, Sweden, in 2001. He was a postdoctoral research fellow first at the Australian National University, then at UC San Diego in 2003 to 2005. Currently, he is a professor at Chalmers and Lund University. His research areas include geometric computer vision, medical image analysis, and optimization methods. In 2005, he was awarded the Marr Prize, and in 2008, he obtained an ERC Starting Grant from the European Research Council.