Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database

Revised version

Bram van Ginneken, Mikkel B. Stegmann, and Marco Loog *

11th August 2004

*B. van Ginneken and M. Loog are with the Image Sciences Institute, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands. E-mail: {bram,marco}@isi.uu.nl, URL: http://www.isi.uu.nl/. M.B. Stegmann is with Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark. E-mail: mbs@imm.dtu.dk, URL: http://www.imm.dtu.dk/.

Abstract

The task of segmenting the lung fields, the heart, and the clavicles in standard posterior-anterior chest radiographs is considered. Three supervised segmentation methods are compared: active shape models, active appearance models, both first proposed by Cootes *et al.* [1] and a multi-resolution pixel classification method that employs a multi-scale filter bank of Gaussian derivatives and a k-nearest-neighbors classifier. The methods have been tested on a publicly available database of 247 chest radiographs, in which all objects have been manually segmented by two human observers.

A parameter optimization for active shape models is presented, and it is shown that this optimization improves performance significantly. It is demonstrated that the standard active appearance model scheme performs poorly, but large improvements can be obtained by including areas outside the objects into the model.

For lung field segmentation, all methods perform well, with pixel classification giving the best results: a paired t-test showed no significant performance difference between pixel classification and an independent human observer. For heart segmentation, all methods perform comparably, but significantly worse than a human observer. Clavicle segmentation is a hard problem for all methods; best results are obtained with active shape models, but human performance is substantially better.

In addition, several hybrid systems are investigated. For heart segmentation, where the separate systems perform comparably, significantly better performance can be obtained by combining the results with majority voting.

As an application, the cardio-thoracic ratio is computed automatically from the segmentation results. Bland and Altman plots indicate that all methods perform well when compared to the gold standard, with confidence intervals from pixel classification and active appearance modelling very close to those of a human observer.

All results, including the manual segmentations, have been made publicly available to facilitate future comparative studies.

Index terms — Chest radiographs, Segmentation, Lung field segmentation, Heart segmentation, Clavicle segmentation, Active shape models, Active appearance models, Pixel classification.

Running title: Segmentation of anatomical structures in chest radiographs

1 Introduction

A large amount of literature in the medical image analysis research community is devoted to the topic of segmentation. Many methods have been developed and tested on a wide range of applications. Despite these efforts, or perhaps because of the large number of algorithms that have been proposed, it remains very difficult for a system designer to decide which approach is best suited for a particular segmentation task. Fortunately, there is a growing awareness in the medical image research community that evaluation and performance characterization of segmentation methods is a critical issue [2, 3]. Such evaluations are greatly facilitated by the availability of public image databases with manual annotations on which researchers can test and compare different algorithms. For this study, we have annotated a public database, and have made the manual segmentations available [4].

We compare three methods for segmenting five important anatomical structures in the single most acquired medical image: the standard posterior-anterior (PA) chest radiograph. To this end, these structures –the lung fields, the heart, and the clavicles– have been segmented manually by two observers independently in 247 radiographs from the publicly available JSRT (Japanese Society of Thoracic Radiology) database [5]. The fact that each object has been manually segmented twice allows one to use one manual segmentation as gold standard and compare the performance of automatic methods with that of an independent human observer. The web site of the annotated JSRT database [4] allows other researchers to upload the results of other segmentation algorithms applied to the database and we invite the medical image analysis research community to do so.

Accurate segmentation of anatomical structures in chest radiographs is essential for many analysis tasks considered in computer-aided diagnosis. These include various size measurements, the determination of the presence of pulmonary nodules or signs of interstitial lung disease. Knowledge about the location of the clavicles can be used to reduce false positive findings or to detect lesions hiding 'behind a clavicle' more reliably.

The methods considered here are active shape models (ASM) [6, 1], active appearance models (AAM) [7] and pixel classification (PC). ASM is a popular segmentation method, with many internal parameters. We consider how to tune

these parameters. AAM has recently found widespread application in medical image segmentation. In this work we use an implementation available in the public domain [8] and compare the standard AAM scheme with an extension in which the surroundings of objects are modelled as well. PC is a classical segmentation method, but the basic concept is so general that it can be implemented in many different ways. We propose an implementation in which both position and local image derivatives are used as input features and show how a multi-resolution implementation and an approximate k-nearest neighbor classifier lead to a relatively fast scheme that yields accurate segmentations. Finally, we also consider three hybrid approaches. The first one fuses the results of the best performing ASM, AAM and PC scheme by majority voting. The other hybrid schemes uses a "tissue map" produced from the probability output of the PC scheme as input for the ASM and AAM method, respectively.

Each of the methods examined here is *supervised*. This means that example images with the desired output need to be supplied for training. This makes the methods versatile; by supplying different training images and annotations, each method can be applied to many different segmentation tasks, including the ones investigated here. This is in contrast to rule-based schemes that are specifically designed to handle one segmentation task.

The article is organized as follows. Section 2 briefly reviews previous work on segmentation of lung fields, heart and clavicles in chest radiographs. Section 3 describes the data. The segmentation methods are presented in Section 4. Section 6 presents the results, followed by a discussion in Section 6. Section 7 concludes.

2 Previous work

Segmentation of lung fields in PA chest radiographs has received considerable attention in the literature. Rule-based schemes have been proposed by Li *et al.* [9], Armato *et al.* [10], Xu *et al.* [11, 12], Duryea and Boone [13], Pietka [14], and Brown *et al.* [15]. Lung segmentation by pixel classification using neural networks has been investigated by McNitt-Gray *et al.* [16], and Tsujii *et al.* [17]. Vittitoe *et al.* [18] developed a pixel classifier for the identification of lung regions using Markov random field modeling. An iterative pixel-based

classification method related to Markov random fields was presented in [19].Van Ginneken and Ter Haar Romeny proposed a hybrid method that combines a rule-based scheme with a pixel classifier [20]. ASM has been used for lung field segmentation in [21, 22].

Segmentation of the outline of the heart has been studied by several researchers, usually with the aim of detecting cardiomegaly (enlarged heart size). For this purpose, only parts of the heart border need to be known. Published methods typically use rule-based schemes, using edge detection and a geometrical model of the heart shape [23, 24, 25, 26, 27].

The segmentation of clavicles in chest radiographs has, to the best of our knowledge, not been studied before.

3 Materials

3.1 Image data

The chest radiographs are taken from the JSRT database [5]. This is a publicly available database with 247 PA chest radiographs collected from 13 institutions in Japan and one in the United States. The images were scanned from films to a size of 2048 $\times$ 2048 pixels, a spatial resolution of .175 mm/pixel and 12 bit gray levels. 154 images contain exactly one pulmonary lung nodule each; the other 93 images contain no lung nodules.

3.2 Object delineation

Each object has been delineated by clicking points along its boundary using a mouse pointer device. These points are connected by straight line segments. For the ASM and AAM segmentation methods, these contours need to be converted to a fixed number of corresponding points. To this end several additional, distinguishable points on the contour are clicked by the user, indicating anatomical or other characteristic landmarks. These characteristic points are assumed to correspond. After the complete boundary has been defined, all but the corresponding points are discarded and subsequent points are obtained by equidistantly sampling a certain fixed number of points along the contour between the aforementioned indicated points. This is illustrated in Fig. 1.

Two observers segmented 5 objects in each image. Observers were allowed to zoom and adjust brightness and image contrast, and could take unlimited time for segmentation. The first observer was a medical student, the second observer a computer science student specializing in medical image analysis. While neither of the observers were radiologists, both are familiar with medical images and medical image analysis and have a good background in human anatomy. Before the segmentations were made, both observers were instructed by an experienced radiologist until he was convinced that the segmentations produced by the observers were reliable. After segmenting all objects, each observer reviewed the results, and adjusted them to correct occasional errors and avoid bias due to learning effects. When in doubt, they reviewed cases with the radiologist and the radiologist provided the segmentation he believed to be correct. Review was necessary in about 10% of all cases. Both observers segmented the images and reviewed the results independently, but they did consult the same radiologist.

The segmentations of the first observer are taken as gold standard in this study, to which the segmentations of a computer algorithm and the second observer can be compared. The availability of a second observer allows for comparisons between 'human' and 'computer' results.

3.3 Anatomical structures

In this work we consider the right and left lung, the outline of the heart and the right and left clavicles[1]. It is important to carefully define what is meant by the outline of an anatomical structure in a projection image.

The intensity in each pixel is determined by the attenuation of the radiation by a column of body tissue. One could define the lung fields as the set of pixels for which the radiation has passed through the lung fields. However, this outline is impossible to determine from a frontal chest radiograph. Therefore we adopt the following definition for the lung fields: any pixel for which radiation passed through the lung, but not through the mediastinum, the heart, structures below the diaphragm, and the aorta. The vena cava superior, when visible, is not considered to be part of the mediastinum.

The heart is defined as those pixels for which radiation passes through the

[1]Note that, by convention, a chest radiograph is displayed as if one is facing the patient. This means that the right lung and clavicle are on the left in the image.

heart. From anatomical knowledge the heart border at the central top and bottom part can be drawn. The great hilar vessels can be assumed to lie on top of the heart.

For the clavicles, only those parts superimposed on the lungs and the rib cage have been indicated. The reason for this is that the peripheral parts of the clavicles are not always visible on a chest radiograph.

Fig. 1 shows one image and the annotated objects.

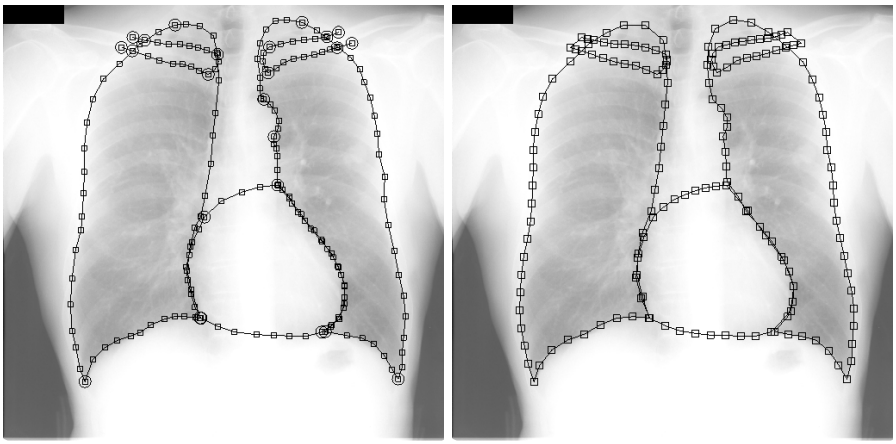


Figure 1: Left: the points indicated by the first observer on the first image of the JSRT database to delineate lung fields, the heart, and the clavicles. The anatomical or distinctive points are circled. The right lung contains 3 of these points, the left lung 5, the heart 4 and each clavicle 6. Right: the landmarks interpolated between the anatomical landmarks along the contours indicated on the left for use in the ASM and AAM segmentation method. The total number of landmarks is 166, with 44, 50, 26, 23 and 23 points in right lung, left lung, heart, right clavicle, and left clavicle, respectively.

4 Methods

4.1 Active Shape Model segmentation

The following is a brief description of the ASM segmentation algorithm. The purpose is mainly to point out the free parameters in the scheme; the specific

values for these parameters are listed in Table 1. Cootes *et al.* first introduced the term active shape model in [28] and [6]. However, [6] does not include the gray level appearance model and [28, 6] do not include the multi-resolution ASM scheme. Both of these components are essential to obtain good segmentation results with ASM in practice. Our implementation follows the description of the ASM method given in [1] to which the reader is referred for details.

The ASM scheme consists of three elements: a global shape model, a local, multi-resolution appearance model, and a multi-resolution search algorithm.

A set of objects in a training image is described by n corresponding points. These points are stored in a shape vector $\mathbf{x} = (x_1, y_1, \ldots, x_n, y_n)^T$. A set of these vectors can be aligned by translating, rotating and scaling them so as to minimize the sum of squared distances between the points (Procrustes alignment, [29, 1]). Alignment can also be omitted, which will include the variation in size and pose into the point distribution model which is subsequently constructed. Let $\overline{\mathbf{x}}$ denote the mean shape. The t principal components (modes of variation in the shape model) of the covariance matrix of the shape vectors are computed. The value of t is determined by f_v, the amount of variation in the training shapes one wants to explain. Shapes can now be written as

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{\Phi}_x \mathbf{b}_x, \tag{1}$$

where $\mathbf{\Phi}_x$ contains the modes of variation of the shape model and $\mathbf{b}_x$ holds the shape parameters. During the ASM search it is required to fit the shape model to a set of landmarks. This is done by projecting the shape on the eigenvectors in $\mathbf{\Phi}_x$ and truncating each projection in $\mathbf{b}_x$ to m times the standard deviation in that direction.

A local appearance model is constructed for each landmark. On either side of the contour at which the landmark is located, k pixels are sampled using a fixed step size of 1 pixel, which gives profiles of length $2k + 1$. Cootes *et al.* propose to use the normalized first derivatives of these profiles [1]. The derivatives are computed using finite differences; the normalization is such that the sum of absolute values equals 1. Note that this requires a notion of connectivity between the landmark points from which the direction perpendicular to the contour can be computed.

As a measure for the goodness of fit of a pixel profile encountered during

search, the Mahalanobis distance to the set of profiles sampled from the training set is computed. These profile models are constructed for L_{max} resolutions. A standard image pyramid [30] is used.

The search algorithm is a simple iterative scheme initialized by the mean shape. Each landmark is moved along the direction perpendicular to the contour to n_s positions on either side, evaluating a total of $2n_s + 1$ positions. The landmark is put at the position with the lowest Mahalanobis distance. After moving all landmarks, the shape model is fitted to the displaced points, yielding an updated segmentation. When a proportion p_{close} of points ends up within $n_s/2$ of its previous position, or when N_{max} iterations have been made, the search moves to the next resolution level, or ends. The highest resolution level in our experiments was 256×256 pixels. The use of higher resolutions did not improve performance.

In [1] suitable values are suggested for all parameters in the ASM scheme. They are listed in Table 1 and they have been used in the experiments, referred to as 'ASM default'. In order to investigate the effect of different settings, we performed pilot experiments on a small test set (to keep computation time within bounds) and varied all settings within a sensible range, also given in Table 1. The overall best setting was kept (last column in Table 1) and also used in the experiments, referred to as 'ASM tuned'.

4.2 Active Appearance Models

The active appearance model (AAM)segmentation and image interpretation method [7] has recently received a considerable amount of attention in the image analysis community [8]. AAM uses the same input as ASM, a set of training images in which a set of corresponding points has been indicated.

The major difference to ASM is that an AAM considers all objects pixels, compared to the border representation from ASM, in a combined model of shape and appearance. The search algorithm is also different. This section will summarize the traditional AAM framework, list the parameter settings and describe alterations we applied for this segmentation task. Our implementation was based on the freely available C++ AAM implementation described in [8].

An AAM is a generative model, which is capable of synthesising images of a given object class. By estimating a compact and specific basis from a training

set, model parameters can be adjusted to fit unseen images and hence perform both image interpretation and segmentation. The modelled object properties are shape — using the shape vectors $\mathbf{x}$ — and pixel intensities (called *texture*), denoted by $\mathbf{t}$. As in ASM, variability is modelled by means of principal component analyses (PCA). Prior to PCA modelling, shapes are Procrustes aligned and textures are warped into a shape-free reference frame and sampled. Usually only the convex hull of the shape is included into the texture model. It is also possible to model the inside of every closed contour. New instances for the shape can be generated by Eq. 1, and, similarly, we have for the texture

$$\mathbf{t} = \overline{\mathbf{t}} + \mathbf{\Phi}_t \mathbf{b}_t \tag{2}$$

where $\overline{\mathbf{t}}$ denotes the mean texture, $\mathbf{\Phi}_t$ are eigenvectors of the texture dispersions (both estimated from the training set) and $\mathbf{b}_t$ holds the texture model parameters. To recover any correlation between shape and texture and obtain a combined parameterization, $\mathbf{c}$, the values of $\mathbf{b}_x$ and $\mathbf{b}_t$ are combined in a third PCA,

$$\begin{pmatrix} \mathbf{W}_x \mathbf{\Phi}_x^{\mathrm{T}} (\mathbf{x} - \overline{\mathbf{x}}) \\ \mathbf{\Phi}_t^{\mathrm{T}} (\mathbf{t} - \overline{\mathbf{t}}) \end{pmatrix} = \begin{pmatrix} \mathbf{W}_x \mathbf{b}_x \\ \mathbf{b}_t \end{pmatrix} = \begin{pmatrix} \mathbf{\Phi}_{c,x} \\ \mathbf{\Phi}_{c,t} \end{pmatrix} \mathbf{c} = \mathbf{\Phi}_c \mathbf{c}. \tag{3}$$

Here, $\mathbf{W}_x$ is a diagonal matrix weighting pixel distances against intensities. Synthetic examples, parameterized by $\mathbf{c}$, are generated by

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{\Phi}_x \mathbf{W}_x^{-1} \mathbf{\Phi}_{c,x} \mathbf{c}$$

and

$$\mathbf{t} = \overline{\mathbf{t}} + \mathbf{\Phi}_t \mathbf{\Phi}_{c,t} \mathbf{c}$$

and rendered into an image by warping the pixel intensities of $\mathbf{t}$ into the geometry of the shape $\mathbf{x}$.

Using an iterative updating scheme the model parameters in $\mathbf{c}$ can be fitted rapidly to unseen images using the L_2-norm as a cost function. See [1, 7] for further details. As in ASM, a multi-resolution pyramid is used.

4.2.1 Parameter settings

This section lists the settings that were used in the AAM experiments. To determine these settings, pilot experiments were performed. Our experience suggests that the results are not sensitive to slight changes in these settings.

Segmentation experiments were carried out in a two-level image pyramid (128 × 128 and 256 × 256 pixels). The use of coarser start resolutions was investigated but did not improve performance.

The model was automatically initialized on the top level, by a sparse sampling in the observed distribution of training set pose. This sparseness is obtained by considering the convergence radius of each model parameter (inspired by [1]), thus avoiding any unnecessary sampling. Since rotation variation was completely covered by the convergence radius, no sampling was performed in this parameter. From the training data, it was estimated that the model should converge if initialized in a 2 by 2 grid around the mean position. Further, due to the variation in size over the training set, each of these four searches was started at 90%, 100%, and 110% of the mean size, respectively. Thus, 12 AAM searches were executed in each image and the search producing the best model-to-image fit was selected.

Both shape, texture and combined models were truncated at $f_v = .98$, thus including 98% of the variance. Bounds m on the combined eigenvalues were three standard deviations. Model searches had a limit of 30 iterations at each pyramid level. AAM parameter update matrices for pose and model parameter were calculated using Jacobian matrices. These were estimated using every 15th training case. Parameter displacements were as follows: *model parameters*: $\pm 0.5\sigma_i$, $\pm 0.25\sigma_i$ (σ_i denotes the standard deviation of i^{th} parameter), *x-y position*: ±2%, ±5% (of width and height, respectively), *scale*: ±2%, ±5%, *rotation*: ±2%, ±5% degrees. Displacements were carried on sequentially; i.e. one experiment for each displacement setting. The details of this process can be found in [7], and are further expanded in [8].

4.2.2 AAM with whiskers

In this particular application of segmenting chest radiographs, the objects in question are best characterized by their borders. They do not have much distinct and consistent interior features; the lungs show a pattern of ribs and vasculature

but the location of these structures relative to the points that make up the shape is not fixed, the heart is dense, but opaque, and no distinct structures can be observed. This behavior is common to many medical image analysis problems and poses a problem to the original AAM formulation where only the object's interior is included into the texture model. This means that the cost function can have minimum when the model is completely inside the actual object. To avoid this, information about the contour edges needs to be included into the texture model. We use the straightforward approach from ASMs; namely to add contour normals pointing outwards on each object. These normals are in this context denoted *whiskers* and are added implicitly during texture sampling with a scale relative to the current shape size. Texture samples obtained by sampling along whiskers are now concatenated to the texture vector, $\mathbf{t}$, with a uniform weight relating these to the conventional AAM texture samples obtained inside every closed contour. This provides a simple weighted method for modelling object proximity in an AAM. Unfortunately, this also introduces two additional free parameters, the length of the whiskers and the weighting of whisker samples.

The following parameters were chosen: whisker length was equal to distance between landmark 1 and 2 on the mean shape (sized to mean size) and texture samples from whiskers influenced the texture model with the same weight as the normal, interior texture samples. Pilot studies showed that moderate changes from the parameter set chosen above had no significant impact on the accuracy.

4.2.3 Refinement of AAM Search Results

AAMs provide a very fast search regime for matching the model to an unseen image using prior knowledge derived from the training set. However, due to the approximate nature this process will not always converge to the minimum of the cost function. A pragmatic solution to this problem is to refine the model fit by using a general-purpose optimization method. Assuming that the AAM search brings the model close to the actual minimum, this approach is feasible wrt. computation, despite the typical high-dimensional parameter space of AAMs. We have used a gradient-based method for this application. The cost function remained unchanged; the L_2-norm between model and image texture. It was optimized by a quasi-Newton method using the BFGS (Broyden, Fletcher, Goldfarb and Shanno) update of the Hessian, see e.g. [31]. Alternatively, to

avoid spurious minima, a random-sampling method such as Simulated Annealing can be employed. Refinement of AAMs has previously been employed to improve the model fit in cardiac and brain MRI by Stegmann *et al.* [32].

4.3 Pixel classification

Pixel classification (PC) is an established technique for image segmentation. It enjoys popularity in many areas of computer vision, .e.g. remote sensing [33]. Within medical imaging, it has been used extensively in multi spectral MR segmentation [34]. A recent example of an application to 3D MR brain segmentation can be found in [35]. In chest radiograph segmentation it has been used before in [16, 20] and, in the context of Markov random field segmentation in [18].

Sect. 4.3.1 and 4.3.2 describe a general multi-resolution implementation of PC that we developed. Components and parameters used for this particular segmentation problem are given in Sect. 4.3.3–4.3.5.

4.3.1 General algorithm

In PC, a training and a test stage can be distinguished. The train stage consists of

1. Choose a working resolution. Obtain a copy of each training image at this working resolution.

2. Choose a number of samples (positions) in each training image.

3. Compute a set of features (the input) for each sample. Possible features are the gray level value at that position or in the surroundings, filter outputs, and position values. Associate an output with each sample. This output lists to which classes this position belongs. Note that in our case pixels can belong to multiple classes simultaneously (e.g. left clavicle and left lung field).

4. (Optional) Compute a suitable transformation for the feature vectors. Examples of transformations are normalization, feature selection, feature extraction by PCA or whitening, or non-linear transformation to create new, extra features.

5. Train a classifier with the input feature vectors and the output; this classifier can map new input to output. In this work we require that the classifier can compute the posterior probability (the probability, given the input features) that a pixel belongs to each object class.

The test stage consists of

1. Obtain a copy of the test image at the working resolution.
2. Compute the features for each pixel in the image.
3. (Optional) Apply the transformation to each feature vector.
4. Obtain the posterior probabilities that the pixel belongs to each class, using the transformed feature vectors and the trained classifier.
5. A binary segmentation for each object is obtained by thresholding the output at .5. Optionally, postprocessing operations can be applied before and after binarization.

4.3.2 Multi-resolution PC

In the multi-resolution PC method, training is performed for a range of working resolutions. The test stage begins at the coarsest resolution and stores the posterior probabilities p_i for each class i and also stores $p_{min} = \min_i(p_i^*)$ where $p_i^* = \max(p_i, 1 - p_i)$, the chance that the pixel belongs to class i, or not, whichever is more likely. If p_{min} is close to 1, the classifier is confident about all the class labels of that pixel. If it is close to .5, the classifier is unsure about at least one of the labellings. The test stage continues at the next resolution level. In this level, the number of pixels is larger. The p_{min} values are linearly interpolated from the previous level, and only if $p_{min} < T$, a pixel is reclassified. Otherwise, the interpolated posterior labels of the coarser level are taken. This process continues until the finest resolution has been processed.

The rationale behind this strategy is that classification (step 4 in the test algorithm given above) is usually the computationally most expensive operation. As low resolution images contain less pixels, and in many applications a large area of the image are often easy to classify (p_{min} close to 1), using the multi-resolution scheme can speed up the PC process considerably. Estimating *all* p_i at a next level if *any* p_i^* is below T may seem superfluous. For the classifier of

our choice, however, this is as expensive as only estimating those p_i for which $p_i^* < T$.

Lower resolution images were created with the Gaussian pyramid [30], as is done in ASM. In the experiments just two resolution levels were considered, where the images were reduced to 128 by 128 and 256 by 256 pixels. The threshold T was conservatively set to .99. Higher resolutions increased computation time but did not improve performance; more lower resolution levels slightly decreased performance at hardly any computational gain.

4.3.3 Samples & Features

A rectangular grid of 64 by 64 pixels was placed over each training image to extract 4096 samples per image.

Spatial features, the (x, y) coordinates in the image, are used, because the structures we aim to segment have a characteristic location within a chest radiograph. Additionally, the output of Gaussian derivative filters of up to second order $(L, L_x, L_y, L_{xx}, L_{yy}, L_{xy})$ at five scales ($\sigma = 1, 2, 4, 8, 16$ pixels at the current resolution) are used to characterize local image structure. Finally, the gray value in the original images was taken as a feature.

This set of features was computed at each resolution level. As the pixel size of images is different at each level, the scale of the filters is different as well, to the effect that large apertures are used for a first coarse segmentation and finer apertures subsequently segment the finer details at higher resolution levels.

4.3.4 Classifier and feature transformations

The effect of feature transformations is closely related to the choice of classifier. A kNN classifier was used, with $k = 15$. The kNN classifier has the attractive property that, under certain statistical assumptions and in the case of infinite training data, the conditional error is $(1 + 1/k)R^*$, where R^* is the minimally achievable Bayes error [36]. Mount and Arya's tree-based kNN implementation [37], was employed which allows for a considerable speed up of the classification by calculating an approximate solution. The approximation is controlled by a variable ϵ, which means that the approximate nearest neighbors which the algorithm finds, are no more than $(1 + \epsilon)$ the distance away from the query point than the actual nearest neighbors are [37]). ϵ was set to 2. This did not

lead to a decrease in accuracy as compared to exact kNN with $\epsilon = 0$.

In pilot experiments, various feature selection and feature extraction methods were tested, but they did not yield a significant performance increase. Eventually, we only applied normalization, which means that a scaling factors per feature are determined so that each feature has unit variance in the training set.

An additional advantage of the kNN classifier, already hinted at above, is that the posterior probability for each object class can be determined using only one neighbor search. Note that the combination of $k = 15$ and $T = .99$ means that pixels are only not reclassified at a finer resolution level if all k neighbors have the same class label configuration (as $14/15 < .99$).

4.3.5 Post-processing

The obvious way to turn the soft classification p_i into binary masks is thresholding at a posterior probability of .5. However, this does not ensure connected objects; segmentations will often contain clouds of isolated pixels near the object's boundary. To ensure that the segmentation for each structure yields a single connected object, a simple post-processing procedure was developed. This procedure is the same for each object considered.

First the soft output is blurred with $\sigma = 0.7$ mm. This reduces the grainy appearance at object boundaries and can be interpreted as pooling of local evidence. Subsequently the largest connected object is selected, and holes in this object are filled.

4.4 Hybrid approaches

Different methods for segmentation are considered in this work. These methods may provide complementary information and if it is possible to combine this information effectively, a hybrid segmentation scheme with higher performance can be constructed. Three possible approaches to such a combination are envisaged.

First, one can consider the *output* of different methods only. All methods can output hard classification labels for each pixel, so it is a logical choice to work with this information. The *hybrid voting* scheme takes the classification labels of the best performing ASM, AAM and PC scheme and assigns pixels to objects according to majority voting. This is the most commonly used voting

rule for hard classifications. For more background and voting strategies that have been researched in the context of classifier fusion see [38].

A second approach is to take the output of one method as input for another scheme. An obvious approach is to use the posterior probabilities for each pixel as obtained from the PC method and convert these into an image where different objects have different gray value. To construct such an image, the posterior probabilities for a pixel to be right lung, left lung, heart, right or left clavicle were added, and in addition the probabilities for lung were multiplied by two (the latter operation is necessary to obtain contrast between heart/lung boundaries). This 'probability image' is used as input for the ASM segmentation method (this is referred to as the *hybrid ASM/PC* method) and the AAM segmentation method (the *hybrid AAM/PC* method). Clearly other output/input chains are conceivable. For example, output of ASM or AAM can be used as a feature for PC, or ASM and AAM may be combined.

A third option is to design a method which is comprised of a combination of elements from different schemes. Several systems proposed in the literature may be interpreted as such combinations [39] [40] [22]. In this work we don't consider these approaches.

5 Experiments and results

5.1 Point distribution model

The analysis of the shape vectors $\mathbf{x}$ gives insight in the typical variations in shape of lungs, heart and clavicles that occur in chest radiographs, and their correlation. This is an interesting result in its own right, and therefore the first few modes of variation are displayed in Fig. 2. In Fig. 3 the spread of each model point after Procrustes alignment is displayed. This is another way of visualizing which parts of the objects exhibit most shape variation.

5.2 Folds

The 247 cases in the JSRT database were split in two folds. One fold contained all 124 odd numbered images in the JSRT database. The other fold contained the 123 even numbered images. This division ensured that both folds contained an equal amount of normal cases and cases with a lung nodule. Images in one

fold were segmented with the images in the other fold as training set, and vice versa.

5.3 Performance measure

To measure the performance of a segmentation algorithm, a 'goodness' index is required. For a two class segmentation problem, one can distinguish true positive (TP) area (correctly classified as object), false positive (FP) area (classified as object, but in fact background), false negative (FN) area (classified as background, but in fact object), and true negative (TN) area (correctly classified as background). From these values, measures such as accuracy, sensitivity, specificity, kappa and overlap can be computed. In this work we use the intersection divided by union as an overlap measure, given by

$$\Omega = \frac{TP}{TP + FP + FN}. \tag{4}$$

This is a well accepted measure, but one should be aware that objects that are small or have a complex shape usually achieve a lower Ω than larger objects [41].

In addition, the mean absolute contour distance is computed. For each point on contour A, the closest point on contour B is computed; these values are averaged over all points; this is repeated with contours A and B interchanged to make the measure symmetric [41]. The distances are given in millimeters; one pixel on the 256 by 256 resolution images on which all experiments were performed corresponds to 1.4 mm.

For comparisons between methods, paired t-tests were used. Differences are considered significant if $p < 0.05$.

5.4 Evaluated methods

The five objects in each of the 247 images were segmented with 15 methods in total:

- First of all, the segmentations of the second human observer were used to compare computerized methods with human performance.
- As a reference method, we computed the performance when the mean shape of each object is taken as segmentation, independent of the actual

image contents. Clearly any method should outperform this 'a priori' segmentation.

- Two ASM systems were employed; ASM with the 'default' settings and the 'tuned' settings given in Table 1.

- For AAM, three systems were evaluated: the 'standard' system (Sect. 4.2.1); the version with whiskers added (Sect. 4.2.2) and finally, the system with whiskers refined by BFGS (Sect. 4.2.3).

- The results for pixel classification are given both with and without post-processing (Sect. 4.3.5).

- Three hybrid methods are employed: voting, and using the output of the post-processed PC system as input for the tuned ASM system and for the AAM method with whiskers refined by BFGS.

- To obtain upper bounds for the performance of ASM and AAM systems, the tuned ASM method was run, initialized from the gold standard; the ASM shape model was fitted directly to the gold standard and the AAM method with whiskers was method was run, initialized from the gold standard. Note that these results are supplied only for reference, obviously these system cannot be used in practice as they require the gold standard to be known.

To reduce the amount of figures and tables, the results are pooled (by averaging performance measures) for both lungs and both clavicles.

5.5 Segmentation results

Results of all 12 systems are listed in Table 2 and 3. In Fig. 4 and 5 the quantiles are shown graphically in box plots. In both these figures and tables, systems are sorted according to performance, and it is indicated when the difference between a system and the system next in rank is significant. For results of the three ASM/AAM systems that were started from the ground truth see Table 4 and 5.

In general, the best results *per system* were obtained by the tuned ASM system, the AAM system with whiskers and BFGS refinement added and the PC

system with postprocessing. However, for clavicle segmentation, post-processing did not significantly improve PC segmentation and use of the AAM BFGS refinement did not improve upon AAM with whiskers only. The voting system was clearly the best hybrid system considered.

For lung field segmentation, PC clearly outperforms ASM and AAM. There is no significant difference between PC and the human observer for both error measures investigated. Voting improves the mean boundary distance for lung field segmentation.

For heart segmentation, performance of the human observer is lowest among all objects, but significantly better than any computer method. AAM, PC and ASM are all close. Interestingly, the combination of these three through voting yields a system that is significantly better than any of its parts.

Clavicle segmentation proves to be a hard problem for any of the methods. The human observer greatly outperforms any computer method. Best results are obtained with ASM. The results of ASM are so much better than those of AAM and PC that the hybrid methods do not improve upon ASM.

Fig. 6 shows the results of the best performing ASM, AAM, PC and hybrid method for four cases. These images were selected in the following way. For each image, the overlap of each object when segmented with the ASM, AAM and PC system was averaged. All images were sorted on this 'overall overlap average'. The images ranking #1, #82, #165 and #247 are displayed, corresponding to an easy, relatively easy, relatively hard and a hard case, respectively.

5.6 Computation of the cardiothoracic ratio

A segmentation as such is hardly ever the final outcome of a computer analysis in medical imaging. The ultimate 'goodness' index for a segmentation is its usefulness for subsequent processing. One important diagnostic measure that can be directly calculated from a segmentation of lungs and heart in a chest radiograph is the cardiothoracic ratio (CTR), defined as the ratio of the transverse diameter of the heart to the transverse diameter of the thorax. A ratio above 0.5 is generally considered a sign of cardiomegaly, and this test is used frequently in clinical practice and clinical research (e.g. [42]). Automatic computation of the CTR has been investigated before [25, 26]. We computed the CTR from the gold standard, and compared the results with the second observer and the best

ASM, AAM and PC systems. Bland and Altman plots [43] are given in Fig. 7 together with the mean absolute difference and the 95% confidence intervals. Note that the confidence interval is tighter for the PC system than for the second observer. This is due to an outlier, though. If that outlier is removed, the confidence interval for the second observer shrinks to (-0.033,0.030). The confidence interval of PC is tighter than that of AAM, which is tighter than ASM. From the Bland and Altman plots it can be appreciated that there is more often substantial disagreement between the gold standard and computerized measures for cases with a large CTR.

5.7 Computation times

The ASM and PC segmentation were performed on a 2.8 GHz Intel PC with 2 GB RAM. The AAM experiments were carried out on a 1.1Ghz Athlon PC equipped with 768 MB RAM. All implementations were in C++, and in all cases there is room for optimizations. Computation time required for segmenting a single image was around 1 s for ASM, 30 s for PC, and 3 s for AAM.

6 Discussion

Some of the presented results obtained by computer algorithms are very close to human performance. Therefore we start this discussion by considering the limitations of manual segmentations which were used to determine the gold standard, and discuss the representativity of the data. Then the results for lung segmentation, heart segmentation, clavicle segmentation and the automatic determination of the CTR are discussed. After pointing out the fundamental differences between pixel classification, active shape and appearance models, we briefly consider some possibilities for improvements in each of the three methods.

Accuracy of the gold standard, representativeness of the data

Supervised segmentation methods require training data for which the 'truth' is available, and their performance will therefore depend on the quality of this 'truth'. In this work, manual segmentations from a single observer are taken as gold standard. It may be preferable to construct a gold standard from multiple

observers [44] [45]. For a large database as considered here however, obtaining multiple expert segmentations is impractical. There are two types of inaccuracies in the gold standard. Occasionally, the observer may misinterprete the superimposed shadows and follow the wrong edge or line in the image. Such *interpretation errors* occur mainly along the mediastinum and heart border and in some cases when edges of clavicles and ribs create a confusing pattern. Interpretation errors can lead to relatively large distances between boundaries drawn by human observers. The outlier for the second observer versus the gold standard in Fig. 7 is an example of an interpretation error (of the second observer, as was judged retrospectively). Interpretation errors are more likely to occur when the image contains pathology or unusual anatomy, and they can sometimes be attributed to the fact that the observers are not radiologists. The review process, however, eliminated most interpretation errors due to observer inexperience. Some errors of computer algorithms could be considered interpretation errors as well such as allotting areas of the stomach or bowels to the left lung, which happens when there is a lot of air in the stomach and the diaphragm below the left lung has a line-like instead of an edge-like appearance. Another example is following the wrong edge for the border between heart and left lung, of which some examples can be seen in Fig. 6.

The second type of inaccuracy could be described as *measurement error*. Clicking points along the boundary of hundreds of objects is a straining task for human operators; inevitably small errors are made. It is possible that supervised computerized methods such as the ones considered here can 'average away' such errors when building their statistical models. On close inspection, certain parts of the boundary of the lung fields found by PC are in fact judged to be more accurate than the gold standard. This may partly explain the fact that the best PC system (and the voting system) achieve better performance for right lung field segmentation than the second observer. Another reason for this fact may be that there are systematic differences between both observers - and the computer algorithms are trained and evaluated with segmentations from the same observer.

There are at least two reasons why the fact that there is no significant difference between a computer method and a human observer for lung field segmentation does not mean that this segmentation task can be considered 'solved'.

First, depending on the usage of the segmentation, the overlap measure Ω and the mean distance to contour may not be good measures of segmentation performance. Although the overall overlap is excellent, there are certain parts of the lung field which pose more problems for a computer than for a human observer. Second, the JSRT database contained only images of good technical quality, and very few images with gross abnormalities. Such images are much harder to segment for the considered computer methods than for humans, because grossly abnormal cases are usually individually unique and thus not represented in the training set.

Keeping these limitations in mind, let us consider the performance of the different methods for each of the segmentation tasks examined.

Lung segmentation

Of all objects, the overlap values obtained for the lungs are highest, for both the human observer and all automatic methods. PC and voting obtain better results than the human observer, although the difference is only significant for the voting system using the overlap as criterion. It is interesting to note how well the shapes produced by PC approximate lung shapes even though no shape information is encoded explicitly in the method. The left lung is more difficult to segment than the right lung because of the presence of the stomach below the diaphragm which may contain air, and the heart border which can be difficult to discern. There is no indication that any of the methods for lung segmentation proposed in the literature (Sect. 2) achieves segmentation accuracy comparable to human performance.

In most cases, ASM and AAM produce satisfactory results as well, but occasionally left lung segmentation proves problematic. Consider the difficult case on the right in Fig. 6, where the border of the enlarged heart is very close to the outer border of the left lung field. Moreover, the heart border is fuzzy, and therefore difficult lo locate precisely. ASM followed a different edge and included the heart in the lung field; this can be considered an interpretation error. AAM put the heart border somewhere halfway between the true border and the border followed by ASM, and pushed the border of the lung field outside the rib cage, probably as a result of shape modeling which does not allow the

heart border and the lower left lung border to be so close. PC, not hampered by a shape model that cannot deal with this uncommon shape, produces a very satisfying result. Note how the segmentation of the second observer deviates from the gold standard. The second observer probably made an interpretation error in this case. Note also that the Ω values for lungs and heart are similar for this case, although the CTR is very different.

The hybrid systems that use PC output for ASM and AAM outperform direct usage of ASM and AAM. This can be explained by the fact that PC works very well for lung segmentation and thus provides reliable input to ASM and AAM.

Heart segmentation

For the heart segmentation, the difference between the second observer and the automatic methods is much larger than for lung segmentation. The agreement between both human observers is much lower as well, the lowest for all objects. The reason for this is that the upper and lower heart border cannot be seen directly on the radiograph. The observers have to infer the location of the heart from the left and right border and anatomical knowledge. The upper heart border is known to be located just below the hilum where the pulmonary arteries enter the lungs. ASM, AAM and PC perform comparably. Their ranking depends on the evaluation criterion used. Apparently, there is complementary information in the three methods: the hybrid voting method is significantly better than any other method and comes quite close to human performance.

Clavicle segmentation

Contrary to the other objects, the clavicles are small. There is no clear difference in performance of any method between left and right clavicle. Clavicle segmentation is a difficult task for various reasons. The bone density can be low, so that the clavicles are hardly visible; there are other, similar edges from ribs in close proximity, and the orientation and position of the clavicles varies enormously. This can be seen from the shape model and the spread of the individual points in Fig. 2 and 3. As a result, segmentation of clavicles is a

challenging task. The difference between the computerized methods and the second observer is large.

ASM is the best method for clavicle segmentation, but occasionally there is hardly any overlap between the detected and actual clavicle, as can be seen in the box plots of Fig. 4 and the most difficult case in Fig. 6. Running a separate ASM to detect clavicles did not show a clear performance improvement. We believe that the problem of confusing edges and the large variation of clavicle position and orientation are the main reasons for failures with ASM.

AAM performs substantially poorer than ASM for clavicle segmentation, contrary to heart and lung segmentation where the results between the two methods were comparable. This behavior was anticipated for several reasons. We hypothesize that the dominating factor is the differences in weighting of the clavicles compared to the lung and heart regions. ASM uses independent, equally weighted texture models around each landmark. Thus, the relative 'importance' – wrt. the model-to-image cost function – of each subobject in an ASM is solely determined by its number of landmarks. In this application, each clavicle had approximately half the number of landmarks present in the corresponding lung contour (see Fig. 1). On the contrary, AAM optimizes a global model-to-image fit, accounting for all pixel samples on the object surfaces. This means that no equalization between areas or objects is performed. Consequently, small objects are easily sacrificed for a better fit of large objects. Further, objects with subtle intensity variations – and weakly defined borders – are sacrificed for objects with large intensity variation. Both issues pertain to segmentation of clavicles. In this application the non-global ASMs behavior has proved desirable for clavicle segmentation, but in other cases such strong priors may lead to problems, typically due to amplification of noise-contaminated signal parts. Using two landmark-based benchmarks, ASM was also shown to outperform AAM on face and brain data in work by Cootes *et al.* [46].

Secondly, since AAM requires mappings between examples to be homeomorphisms (continuous and invertible), layered objects moving independently will inherently cause problems. In the projection images analyzed here, the clavicle position with respect to the lung border is inconsistent. In two examples clavicles were actually above the lungs. Further, the medial endpoints of the clavicles can be inside or outside the lung fields. To obtain a perfect registra-

tion between such images an AAM would need to introduce folds, i.e. degenerate piecewise affine warps with inverted mesh normals. Replacing these with thin-plate splines [47] will inevitably also lead to folds. To solve this, objects need to be modeled as independent layers in the texture model. Rogers [48] has previously acknowledged this problem when modeling capillary images.

To assess the practical impact of this on the overlap measure, we have built an AAM for the lung and heart contours only. Warp degeneracy had no noticeable impact on the accuracy of lung and heart localization.

PC undersegments the clavicles, but hardly has any false positives. This can be explained by poor features, which make classification into the class with higher prior probability more likely. A lower threshold for the hard classification could prove helpful, but this has not been investigated in detail. The post-processing of the PC method is counterproductive when the clavicle segmentation produces two or more segments of similar size since only the largest segment is retained. This occurs in the two most difficult cases shown in Fig. 6.

As ASM is clearly the best performing method for this task, the hybrid approaches do not improve upon ASM.

Determination of the CTR

There is a variety of measures that can be computed directly from a segmentation of anatomical structures in a chest radiograph. Such measures can be of great clinical importance and their automatic computation may help in extracting more information from a routine chest examination. In this work, we considered the cardiothoracic ratio. Other possibilities are the area of heart and lungs, the total lung capacity (for which an additional lateral chest film is required) [49], the diaphragm length [50], and the vascular pedicle width (VPW) [51]. Measurement of the diaphragm length and the VPW requires knowledge about the location of certain landmarks in the image, which is known from the point positions obtained by ASM and AAM segmentation.

An automatic system to estimate the CTR was described by Nakamori *et al.* in [26], in which points along the heart boundary were detected by fitting a Fourier shape to image profiles. This system was used to compute the CTR in 400 radiographs in another study [27] where radiologists had to correct the

computer result in 20% of all cases. Automatic determination by the methods presented here is probably substantially more accurate.

Pixel classification versus active shape and appearance models

There are some fundamental differences between PC, ASM and AAM. PC does not have a shape model. Therefore it can produce unplausible shapes, and is likely to do so when the evidence obtained from image features is not conclusive. This can be observed from the heart borders and the clavicles in Fig. 6. PC does not require landmarks, only labels per pixel. In that sense it is more general, it can also be applied to tasks where it is difficult or impossible to set corresponding points. PC can also produce a soft classification, which cannot be obtained directly from ASM and AAM. On the other hand, ASM and AAM provide more information than just a binary segmentation; correspondences along contours are established between the search result and any training example. Thus, a registration is obtained and any anatomical landmarks defined by a position on a contour can be inferred on new examples using ASM or AAM.

Another important difference between PC and ASM/AAM is that the latter are based on linear appearance models whereas the PC system uses non-linear features and a non-linear classifier to map appearance and position characteristics to class labels. Had the PC system been restricted to features similar to those used in AAM and ASM (pixel values in the object and along profiles) and a linear classifier, the results would have been much worse.

PC does not employ an iterative optimization scheme. This avoids the problems typically associated with such schemes, such as ending up in local minima. Although PC is conceptually more simple, and easier to implement, it is computationally more demanding than ASM and AAM. The multi-resolution scheme proposed here, and the approximate kNN classifier make the method usable on these 2D data. With increasing computational power and dedicated optimizations for processing speed, segmentation of 3D data sets with complex PC systems will become routinely feasible as well.

Contrary to both ASM and PC, an AAM also establish a dense planar correspondence. This dense registration enables that every *interior* point on the model can be localized on an unseen image after AAM search. This combined

with a per-pixel statistical model, provides a starting point for e.g. detection of abnormalities in the lung field.

All in all, the choice for a particular segmentation algorithm can be motivated by more than the expected segmentation accuracy, such as computational demands, implementation complexities and the requirements for further analysis.

Improving ASM, AAM and PC

Changes or extensions to the ASM algorithm can address the shape model, the appearance model or the optimization algorithm. Tables 4 and 5 show that when the shape model is fitted to the gold standard, the mean overlap remains below the results of PC for the right lung and around the accuracy of the human observer for both lungs. For the clavicles, fitting the shape model leads to an overlap well below that of the human observer. Thus the shape model is not able to capture all shape variations in the test set. It is possible that more examples are needed, or that better results can be obtained with more flexible models. To test if the results of ASM and AAM could be improved by simply using more examples, we ran the tuned ASM algorithm using both folds for training (which will positively bias results because training and test data are not separated). It turned out that segmentation performance did not improve. For AAM, leave-one-out experiments were performed but again, this led to only very minor differences. For both systems the mean Ω improved by around 0.01. When ASM is started from the gold standard position, the results for all objects are surprisingly close to the actual results of the tuned ASM system. This indicates that the multi-resolution optimization procedure initialized with the mean shape performs adequately. Still, by comparing the shape model fit to the gold standard and the result of ASM initialized with the gold standard, it can be seen that the solution drifts away from the perfect initialization. The appearance model is thus not perfect. Non-linear appearance models might lead to better performance [52, 53, 22].

Contrary to ASM, changes to the internal parameters of AAM had little influence on the final result. Therefore there is no default and tuned setting presented for AAM. The extension with whiskers, however, was essential to

obtain good performance with AAM. Instead of BFGS, a refinement method based on random sampling should partially avoid the problem of falling into a local minimum. However, this is anticipated to be computationally more demanding due to the high dimensionality of the optimization space. Tables 4 and 5 also show that the lung accuracy obtained using AAM with BFGS refinement is very close to the upper bound. This suggests that the cost function hyper surface is indeed very flat and improvements in overlap for the heart and clavicles in Tables 4 and 5 is apparently obtained by starting the AAM search closer to the global minimum in this region.

The results of ASM and AAM segmentation also depend on the choice of landmarking scheme as this affects the quality of the shape model. Recently, methods for automatically obtaining corresponding landmarks from binary segmentations have been proposed, that optimize an information criterion [54]. Such landmarking strategies may be superior to the landmarking extraction procedure employed here.

The good performance for PC depends heavily on the features and the classifier. Clearly, there are many more feature sets and classifiers that could be evaluated, and feature extraction and selection techniques could be employed. This is an advantage of PC: it formulates segmentation in terms of a standard pattern recognition task. The full vocabulary of techniques from this field can be used.

The postprocessing stage of PC is simple and ad hoc. It ensures a single object, without holes and with a somewhat smooth border (due to blurring the posterior probabilities) but the settings of the procedure have not been trained and the optimal settings are unlikely to be the same for all objects. There are many more advanced possibilities to express the spatial correlation between nieghboring pixel labels than just Gaussian smoothing. Examples are iterative relabelling [19], relaxation labelling [55] or Markov random field models [56]. Such approaches will likely improve performance and ensure more satisfactory object shapes, at the expense of more computation time.

Spatial position is an important feature for PC. Without it, the left and right lung fields, for example, would virtually indistinguishable. The heart segmentation suffers from little image information, and position is a very important feature. The position feature are 'raw' positions, however. After the lung fields

have been segmented - which can be done very accurately, as has been demonstrated, the spatial position relative to the lung fields could be used instead of the raw (x, y) values. This may improve heart segmentation accuracy. For clavicle segmentation, it may not be too useful, as the position of the clavicles relative to the lung fields varies a lot (Fig. 7).

This shows that one must pay careful attention to parameter settings and variations on the basic algorithms that have been proposed in the literature when applying ASM or AAM in practice.

7 Conclusions

A large experimental study has been presented in which several versions of three fully automated supervised segmentation algorithms have been compared.

The methods were active shape models (ASM), active appearance models (AAM) and pixel classification (PC). The task was to segment lung fields, heart, and clavicles from standard chest radiographs. Results were evaluated quantitatively, and compared with the performance of an independent human observer. The images, manual annotations and results are available for the research community to facilitate further studies, and so is the AAM implementation [8].

The main conclusions are the following:

1. All methods produce results ranging from excellent to at least fairly accurate for all five segmentation tasks considered, using the same settings for all objects. This demonstrates the versatility and flexibility of general supervised image segmentation methods.

2. The cardiothoracic ratio can be determined automatically with high accuracy. Automatic CTR determination could be provided in a clinical workstation.

3. The best method for lung field segmentation is PC, or a combination of PC, ASM and AAM through voting - depending on the evaluation criterion used; for heart segmentation voting among PC, AAM and ASM outperforms other methods and the individual methods perform comparably; for clavicle segmentation ASM produces the best results.

4. A combination of ASM, AAM and PC through majority voting is in all cases superior to the use of PC output as input for ASM or AAM.

5. For segmentation of both right and left lung, PC and voting perform better than manual segmentation by an independent human observer although the differences are not always significant. In all other cases, the human observer performed significantly better than the best computer system. For clavicles, the differences between computerized methods and a human observer were largest. This indicates that accurate computerized segmentation of clavicles and heart in chest radiographs is still an open problem.

6. Significant performance improvements for all five segmentation tasks were obtained by tuning the parameters of ASM and by including information from outside the object in the AAM. Refining the AAM result with a general purpose optimization method improved performance for all objects except clavicles. This shows that one must pay careful attention to parameter settings and variations on the basic algorithms that have been proposed in the literature when applying ASM or AAM in practice.

7. The output of pixel classification can lead to ragged borders. A simple post-processing stage decreases the error significantly for lung fields and heart segmentation. The presented PC method obtains excellent results in these tasks, despite its lack of a shape model.

Acknowledgement

The authors gratefully acknowledge R. Nievelstein for supervising the manual segmentations and G. Mochel and A. Scheenstra for each clicking around 75,000 points to segment the images.

References

[1] T.F. Cootes and C.J. Taylor. Statistical models of appearance for computer vision. Technical report, Wolfson Image Analysis Unit, University of Manchester, 2001.

[2] P. Jannin, J.M. Fitzpatrick, D.J. Hawkes, X. Pennec, R. Shahidi, and M.W. Vannier. Validation of medical image processing in image-guided therapy. *IEEE Transactions on Medical Imaging*, 21(12):1455–1449, 2002.

[3] K.W. Bowyer, M.H. Loew, H.S. Stiehl, and M.A. Viergever. Methodology of evaluation in medical image computing. In *Rep. Dagstuhl Workshop*, 2001.

[4] Image Sciences Institute Research Databases. `http://www.isi.uu.nl/Research/Databases/`.

[5] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174:71–74, 2000.

[6] T. F. Cootes, C. J. Taylor, D. Cooper, and J. Graham. Active shape models – their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[7] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[8] M. B. Stegmann, B. K. Ersbøll, and R. Larsen. FAME – a flexible appearance modelling environment. *IEEE Transactions on Medical Imaging*, 22(10):1319–1331, 2003.

[9] L. Li, Y. Zheng, M. Kallergi, and R.A. Clark. Improved method for automatic identification of lung regions on chest radiographs. *Academic Radiology*, 8(7):629–638, 2001.

[10] S. G. Armato, M. L. Giger, and H. MacMahon. Automated lung segmentation in digitized postero-anterior chest radiographs. *Academic Radiology*, 4:245–255, 1998.

[11] X.W. Xu and K. Doi. Image feature analysis for computer-aided diagnosis: accurate determination of ribcage boundary in chest radiographs. *Medical Physics*, 22(5):617–626, 1995.

[12] X.W. Xu and K. Doi. Image feature analysis for computer-aided diagnosis: detection of right and left hemidiaphragm edges and delineation of lung field in chest radiographs. *Medical Physics*, 23(9):1613–1624, 1996.

[13] J. Duryea and J.M. Boone. A fully automatic algorithm for the segmentation of lung fields in digital chest radiographic images. *Medical Physics*, 22(2):183–191, 1995.

[14] E. Pietka. Lung segmentation in digital chest radiographs. *Journal of Digital Imaging*, 2:79–84, 1994.

[15] M.S. Brown, L.S. Wilson, B.D. Doust, R.W. Gill, and C. Sun. Knowledge-based method for segmentation and analysis of lung boundaries in chest X-ray images. *Computerized Medical Imaging and Graphics*, 22:463–477, 1998.

[16] M.F. McNitt-Gray, H.K. Huang, and J.W. Sayre. Feature selection in the pattern classification problem of digital chest radiograph segmentation. *IEEE Transactions on Medical Imaging*, 14(3):537–547, 1995.

[17] O. Tsujii, M.T. Freedman, and S.K Mun. Automated segmentation of anatomic regions in chest radiographs using an adaptive-sized hybrid neural network. *Medical Physics*, 25(6):998–1007, 1998.

[18] N.F. Vittitoe, R. Vargas-Voracek, and C.E. Floyd Jr. Identification of lung regions in chest radiographs using Markov Random Field modeling. *Medical Physics*, 25(6):976–985, 1998.

[19] M. Loog and B. van Ginneken. Supervised segmentation by iterated contextual pixel classification. In *Proceedings 16th International Conference on Pattern Recognition*, pages 925–928, 2002.

[20] B. van Ginneken and B. M. ter Haar Romeny. Automatic segmentation of lung fields in chest radiographs. *Medical Physics*, 27(10):2445–2455, 2000.

[21] B. van Ginneken, S. Katsuragawa, B. M. ter Haar Romeny, K. Doi, and M. A. Viergever. Automatic detection of abnormalities in chest radiographs using local texture analysis. *IEEE Transactions on Medical Imaging*, 21(2):139–149, 2002.

[22] B. van Ginneken, A. F. Frangi, J. J. Staal, B. M. ter Haar Romeny, and M. A. Viergever. Active shape model segmentation with optimal features. *IEEE Transactions on Medical Imaging*, 21(8):924–933, 2002.

[23] D.L. Hall, G.S. Lodwick, R.P. Kruger, and S.J. Dwyer III. Computer diagnosis of heart disease. *Radiological Clinics of North America*, 9(3):533–541, 1971.

[24] R.P. Kruger, J.R. Townes, D.L. Hall, S.J. Dwyer III, and G.S. Lodwick. Automated radiographic diagnosis via feature extraction and classification of cardiac size and shape descriptors. *IEEE Biomedical Transactions*, BME-19(3):174–186, 1972.

[25] N. Sezaki and K. Ukena. Automatic computation of the cardiothoracic ratio with application to mass screening. *IEEE Transactions on Biomedical Engineering*, BME-20(4):248–253, 1973.

[26] N. Nakamori, K. Doi, V. Sabeti, and H. MacMahon. Image feature analysis and computer-aided diagnosis in digital radiography: automated analysis of sizes of heart and lung in chest images. *Medical Physics*, 17(3):342–350, 1990.

[27] N. Nakamori, K. Doi, H. MacMahon, Y. Sasaki, and S.M. Montner. Effect of heart-size parameters computed from digital chest radiographs on detection of cardiomegaly: potential usefulness for computer-aided diagnosis. *Investigative Radiology*, 26(6):546–550, 1991.

[28] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam. The use of active shape models for locating structures in medical images. *Image and Vision Computing*, 12(6):355–366, 1994.

[29] C. Goodall. Procrustes methods in the statistical analysis of shapes. *Journal of the Royal Statistical Society B*, 53(2):285–339, 1991.

[30] P.J. Burt and E.H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, COM-31,4:532–540, 1983.

[31] R. Fletcher. *Practical Methods of Optimization*. John Wiley & Sons, 1987.

[32] M. B. Stegmann, R. Fisker, and B. K. Ersbøll. Extending and applying active appearance models for automated, high precision segmentation in different image modalities. In *Proc. 12th Scandinavian Conference on Image Analysis - SCIA 2001*, volume 1, pages 90–97, 2001.

[33] J.A. Richards and X. Jia. *Remote sensing digital image analysis: an introduction*. Springer Verlag, 3rd edition, 1999.

[34] J. C. Bezdek, L. O. Hall, and L. P. Clarke. Review of MR image segmentation techniques using pattern recognition. *Medical Physics*, 20:1033–1048, 1993.

[35] C. A. Cocosco, A. P. Zijdenbos, and A. C. Evans. A fully automatic and robust brain MRI tissue classification method. *Medical Image Analysis*, 7(4):513–527, 2003.

[36] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley and Sons, New York, 2nd edition, 2001.

[37] S. Arya, D.M. Mount, N.S. Netanyahu, R. Silverman, and A.Y. Wu. An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *Journal of the ACM*, 45(6):891–923, 1998.

[38] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[39] S. C. Mitchell, B. P. F. Lelieveldt, R. J. van der Geest, H. G. Bosch, J. H. C. Reiver, and M. Sonka. Multistage hybrid active appearance model matching: segmentation of left and right ventricles in cardiac MR images. *IEEE Transactions on Medical Imaging*, 20(5):415–423, 2001.

[40] S. Yan, C. Liu, S. Z. Li, H. Zhang, H. Shum, and Q. Cheng. Face alignment using texture-constrained active shape models. *Image and Vision Computing*, 21(1):69–75, 2003.

[41] M. Gerig, G. Jomier and M. Chakos. Valmet: a new validation tool for assessing and improving 3D object segmentation. In *MICCAI 2001*, number 2208 in Lecture Notes in Computer Science, pages 516–523. Springer, Berlin, 2001.

[42] M. T. Kearney, J. Nolan, A. J. Lee, P. W. Brooksby, R. Prescott, A. M. Shah, A. G. Zaman, D. L. Eckberg, H. S. Lindsay, and P. D. Batin. A prognostic index to predict long-term mortality in patients with mild to moderate chronic heart failure stabilised on angiotensin converting enzyme inhibitors. *European Journal of Heart Failure*, 5(4):489–497, 2003.

[43] J. M. Bland and D. G. Altman. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet*, 346(8982):1085–1087, 1995.

[44] S. K. Warfield, K. H. Zou, and Wells W. M. Validation of image segmentation and expert quality with an expectation-maximization algorithm. In *MICCAI*, volume 2488 of *Lecture notes in Computer Science*, pages 298–306, 2002.

[45] S. K. Warfield, K. H. Zou, and Wells W. M. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[46] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Comparing active shape models with active appearance models. In *Proc. British Machine Vision Conf.*, pages 173–182, 1999.

[47] F. L. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–85, 1989.

[48] M. Rogers. *Exploiting Weak Constraints on Object Structure and Appearance for Segmentation of 2-D Images*. PhD thesis, University of Manchester, 2001.

[49] H. J. Barnhard, J. A. Pierce, J. W. Joyce, and J. H. Bates. Roentgenographic determination of total lung capacity. A new method evaluated in health, emphysema and congestive heart failure. *American Journal of Medicine*, 28:51–60, 1960.

[50] F. Bellemare, J. Couture, M. Cordeau, P. Leblanc, and E. Lafontaine. Anatomic landmarks to estimate the length of the diaphragm from chest ra-

diographs: effects of emphysema and lung volume reduction surgery. *Chest*, 120(2):444–452, 2001.

[51] E. W. Ely and E. F. Haponik. Using the chest radiograph to determine intravascular volume status: the role of vascular pedicle width. *Chest*, 121(3):942–950, 2002.

[52] I. Scott, T. F. Cootes, and C. J. Taylor. Improving appearance model matching using local image structure. In *Information Processing in Medical Imaging*, volume 2732 of *Lecture Notes in Computer Science*, pages 258–269. Springer, 2003.

[53] M. de Bruijne, B. van Ginneken, W.J. Niessen, and M.A. Viergever. Adapting active shape models for 3D segmentation of tubular structures in medical images. In *Information Processing in Medical Imaging*, volume 2732 of *Lecture Notes in Computer Science*, pages 136–147. Springer, 2003.

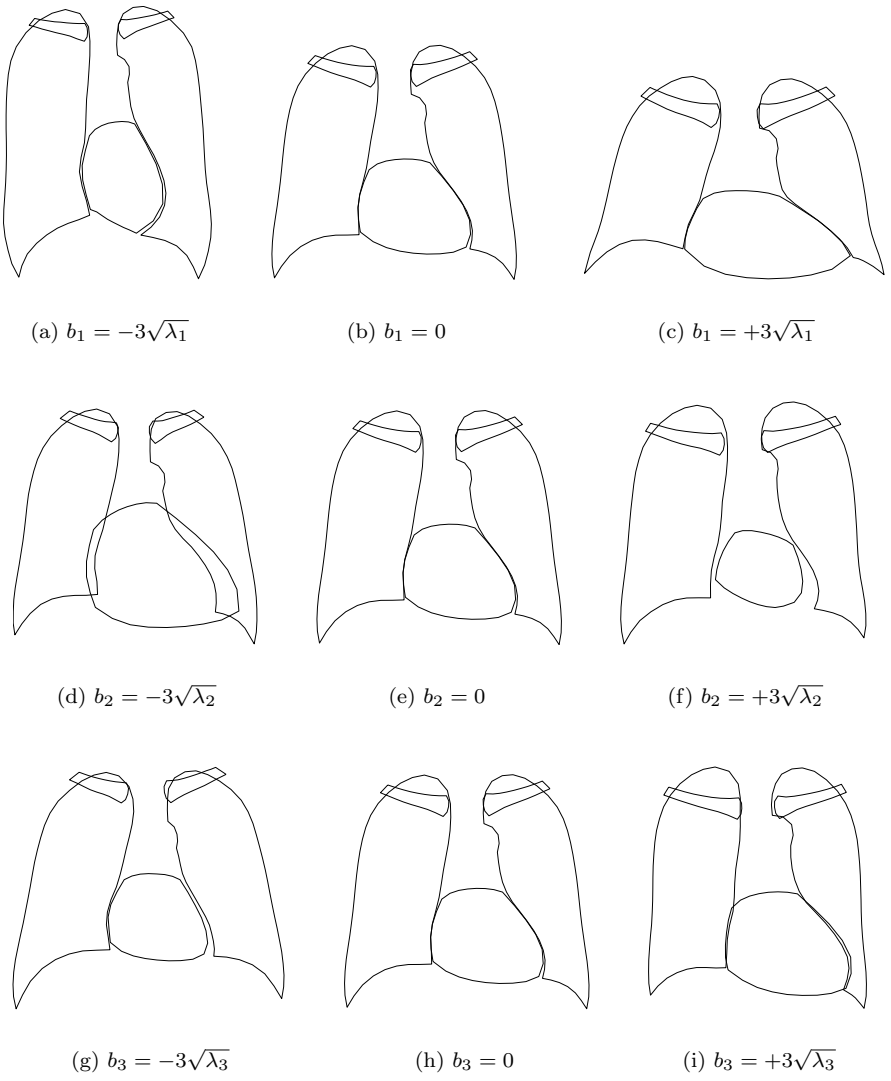
[54] R. H. Davies, C. J. Twining, T. F. Cootes, J. C. Waterton, and C. J. Taylor. A minimum description length approach to statistical shape modeling. *IEEE Transactions on Medical Imaging*, 21(5):525–537, 2002.

[55] J. Kittler and J. Illingworth. Relaxation labelling algorithms—a review. *Image and Vision Computing*, 3(4):206–216, 1985.

[56] G. Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Number 27 in Applications of mathematics. Springer-Verlag, Berlin, 2nd edition, 1995.

		standard	tested	tuned
Shape model				
align	use Procrustes shape alignment	`true`	`false` - `true`	`false`
f_v	variance to be explained by the shape model	.98	.95 - .999	.995
m	bounds on eigenvalues	3.0	2.0 - 3.0	2.5
Appearance model				
k	points in profile on either side of the point	3	1 - 9	5
L_{max}	resolution levels	4	1 - 6	5
Search algorithm				
n_s	positions to evaluate on either side of point	2	1 - 9	2
N_{max}	max. iterations per level	5	5 - 20	20
p_{close}	convergence criterion	.9	.9 - 1.1	1.1

Table 1: Parameters for ASM. The standard settings are the default values suggested in [1]. The test range was used in pilot experiments to determine optimally tuned settings. The tuned settings are in the last column. Note that $p_{close} > 1$ means that N_{max} iterations are always made.



(a) $b_1 = -3\sqrt{\lambda_1}$ (b) $b_1 = 0$ (c) $b_1 = +3\sqrt{\lambda_1}$

(d) $b_2 = -3\sqrt{\lambda_2}$ (e) $b_2 = 0$ (f) $b_2 = +3\sqrt{\lambda_2}$

(g) $b_3 = -3\sqrt{\lambda_3}$ (h) $b_3 = 0$ (i) $b_3 = +3\sqrt{\lambda_3}$

Figure 2: Mean shape deformation obtained by varying the first three modes of the shape model between -3 and +3 standard deviations.

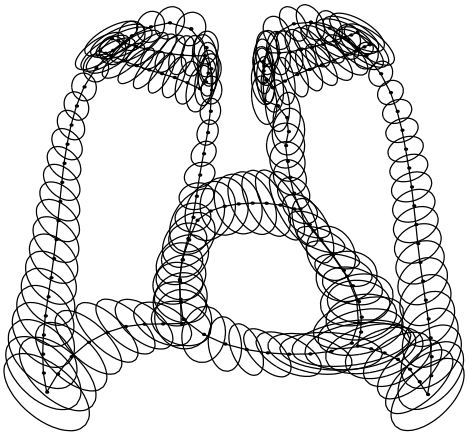


Figure 3: Independent principal component analysis for each model point after Procrustes alignment.

Lungs	$\mu \pm \sigma$	min	Q1	median	Q3	max
Hybrid voting*	0.949 ± 0.020	0.818	0.945	0.953	0.961	0.978
PC post-processed	0.945 ± 0.022	0.823	0.939	0.951	0.958	0.972
Human observer*	0.946 ± 0.018	0.822	0.939	0.949	0.958	0.972
PC*	0.938 ± 0.027	0.823	0.931	0.946	0.955	0.968
Hybrid ASM/PC	0.934 ± 0.037	0.706	0.931	0.945	0.952	0.968
Hybrid AAM/PC*	0.933 ± 0.026	0.762	0.926	0.939	0.950	0.966
ASM tuned*	0.927 ± 0.032	0.745	0.917	0.936	0.946	0.964
AAM whiskers BFGS*	0.922 ± 0.029	0.718	0.914	0.931	0.940	0.961
ASM default*	0.903 ± 0.057	0.601	0.887	0.924	0.937	0.960
AAM whiskers*	0.913 ± 0.032	0.754	0.902	0.921	0.935	0.958
AAM default*	0.847 ± 0.095	0.017	0.812	0.874	0.906	0.956
Mean shape	0.713 ± 0.075	0.460	0.664	0.713	0.768	0.891
Heart	$\mu \pm \sigma$	min	Q1	median	Q3	max
Human observer*	0.878 ± 0.054	0.571	0.843	0.888	0.916	0.965
Hybrid voting*	0.860 ± 0.056	0.651	0.833	0.870	0.900	0.959
Hybrid ASM/PC	0.836 ± 0.082	0.430	0.804	0.855	0.889	0.948
Hybrid AAM/PC	0.827 ± 0.084	0.499	0.791	0.846	0.888	0.957
AAM whiskers BFGS	0.834 ± 0.070	0.510	0.791	0.845	0.882	0.967
PC post-processed*	0.824 ± 0.077	0.500	0.783	0.844	0.877	0.932
PC	0.811 ± 0.077	0.497	0.769	0.832	0.862	0.914
ASM tuned*	0.814 ± 0.076	0.520	0.770	0.827	0.873	0.938
ASM default*	0.793 ± 0.119	0.220	0.755	0.824	0.872	0.954
AAM whiskers*	0.813 ± 0.080	0.489	0.770	0.823	0.874	0.938
AAM default*	0.775 ± 0.135	0.026	0.733	0.806	0.860	0.947
Mean shape	0.643 ± 0.147	0.221	0.550	0.665	0.754	0.921
Clavicles	$\mu \pm \sigma$	min	Q1	median	Q3	max
Human observer*	0.896 ± 0.037	0.707	0.880	0.905	0.922	0.952
ASM tuned	0.734 ± 0.137	0.093	0.705	0.776	0.822	0.912
Hybrid voting*	0.736 ± 0.106	0.091	0.701	0.762	0.801	0.904
ASM default*	0.690 ± 0.143	0.000	0.647	0.731	0.781	0.862
Hybrid ASM/PC	0.663 ± 0.157	0.033	0.595	0.712	0.773	0.891
AAM whiskers BFGS*	0.642 ± 0.171	0.003	0.588	0.689	0.761	0.861
Hybrid AAM/PC	0.613 ± 0.206	0.000	0.558	0.676	0.755	0.850
AAM whiskers	0.625 ± 0.171	0.000	0.578	0.674	0.728	0.870
PC post-processed	0.615 ± 0.123	0.223	0.554	0.639	0.706	0.837
PC*	0.618 ± 0.100	0.232	0.567	0.630	0.689	0.808
AAM default*	0.505 ± 0.234	0.000	0.393	0.575	0.679	0.834
Mean shape	0.303 ± 0.214	0.000	0.098	0.300	0.481	0.715

Table 2: Segmentation results for lungs, heart and clavicles, for each system considered. All results are in terms of the overlap Ω, as defined in Eq. (4). The systems are ranked according to the median Ω. A paired t-test has been applied to each system and the system below it in this ranking. If the difference is significant ($p < 0.05$), this is indicated with an asterix.

Lungs	$\mu \pm \sigma$	min	Q1	median	Q3	max
PC post-processed	1.61 ± 0.80	0.83	1.17	1.41	1.73	8.34
Hybrid voting	1.62 ± 0.66	0.85	1.27	1.45	1.78	7.72
Human observer*	1.64 ± 0.69	0.83	1.29	1.53	1.83	9.11
Hybrid ASM/PC	2.08 ± 1.40	0.91	1.41	1.68	2.09	11.57
Hybrid AAM/PC*	2.06 ± 0.84	0.99	1.56	1.85	2.25	7.31
ASM tuned	2.30 ± 1.03	1.07	1.62	1.95	2.62	7.67
AAM whiskers BFGS*	2.39 ± 1.07	1.15	1.78	2.13	2.61	12.09
PC*	3.25 ± 2.65	0.93	1.64	2.25	3.72	15.59
AAM whiskers*	2.70 ± 1.10	1.16	1.98	2.43	3.04	8.74
ASM default*	3.23 ± 2.21	1.17	1.92	2.52	3.78	16.57
AAM default*	5.10 ± 4.44	1.21	2.97	4.14	6.21	57.30
Mean shape	10.06 ± 3.18	3.50	7.68	10.00	12.05	23.77
Heart	$\mu \pm \sigma$	min	Q1	median	Q3	max
Human observer*	3.78 ± 1.82	0.96	2.50	3.40	4.87	16.26
Hybrid voting*	4.24 ± 1.87	1.18	2.88	3.90	5.07	12.83
PC post-processed	5.20 ± 2.59	1.88	3.37	4.52	6.33	18.72
Hybrid ASM/PC	5.24 ± 3.10	1.47	3.37	4.58	5.86	24.94
AAM whiskers BFGS	5.30 ± 2.58	0.94	3.57	4.82	6.76	19.79
Hybrid AAM/PC	5.57 ± 3.06	1.27	3.36	4.93	6.86	19.84
ASM tuned	5.96 ± 2.73	1.91	3.86	5.47	7.36	16.55
AAM whiskers	6.01 ± 2.88	1.73	3.93	5.58	7.60	19.38
PC	6.38 ± 2.94	2.41	4.31	5.67	7.58	21.64
ASM default*	6.81 ± 4.65	1.30	3.82	5.68	7.92	34.32
AAM default*	7.72 ± 6.79	1.54	4.24	6.15	8.73	71.70
Mean shape	13.00 ± 6.55	1.84	8.18	11.93	17.13	35.39
Clavicles	$\mu \pm \sigma$	min	Q1	median	Q3	max
Human observer*	0.68 ± 0.26	0.31	0.50	0.62	0.79	2.02
ASM tuned*	2.04 ± 1.36	0.55	1.26	1.58	2.24	9.13
Hybrid voting*	1.88 ± 0.93	0.66	1.32	1.61	2.19	7.92
ASM default	2.49 ± 2.02	0.92	1.51	1.95	2.74	24.52
Hybrid ASM/PC	2.78 ± 1.89	0.84	1.59	2.12	3.23	12.31
AAM whiskers BFGS*	3.02 ± 2.23	0.87	1.69	2.32	3.34	15.99
Hybrid AAM/PC*	3.49 ± 3.26	0.99	1.80	2.37	3.68	25.41
PC	2.83 ± 1.68	1.15	2.06	2.46	3.05	16.00
PC post-processed	2.90 ± 1.54	1.03	1.96	2.49	3.31	12.55
AAM whiskers*	3.38 ± 3.71	0.73	1.96	2.52	3.35	41.27
AAM default	6.30 ± 8.97	0.94	2.35	3.44	5.77	66.37
Mean shape	7.25 ± 4.24	1.87	4.60	6.39	8.82	33.67

Table 3: Segmentation results for lungs, heart and clavicles, for each system considered. All results are in terms of the mean absolute contour distance, given in millimeter. The systems are ranked according to the median mean absolute contour distance. A paired t-test has been applied to each system and the system below it in this ranking. If the difference is significant ($p < 0.05$), this is indicated with an asterix.

Lungs	$\mu \pm \sigma$	min	Q1	median	Q3	max
ASM GS	0.93 ± 0.03	0.72	0.92	0.94	0.94	0.96
shape model fit	0.95 ± 0.02	0.76	0.94	0.95	0.96	0.97
AAM GS	0.93 ± 0.02	0.84	0.93	0.94	0.94	0.96
Heart	$\mu \pm \sigma$	min	Q1	median	Q3	max
ASM GS	0.82 ± 0.08	0.50	0.77	0.82	0.88	0.95
shape model fit	0.94 ± 0.04	0.44	0.94	0.95	0.96	0.98
AAM GS	0.88 ± 0.05	0.66	0.86	0.89	0.92	0.96
Clavicles	$\mu \pm \sigma$	min	Q1	median	Q3	max
ASM GS	0.74 ± 0.13	0.19	0.71	0.78	0.81	0.90
shape model fit	0.82 ± 0.06	0.35	0.80	0.83	0.85	0.90
AAM GS	0.72 ± 0.08	0.26	0.68	0.74	0.78	0.88

Table 4: Segmentation results for lung, heart and clavicles, using the tuned ASM system initialized with the gold standard, fitting the shape model from the ASM system directly to the gold standard, and the AAM whiskers system initialized with the gold standard. These results provide upper bounds for ASM and AAM systems. All results are in terms of the overlap Ω, as defined in Eq. (4).

Lungs	$\mu \pm \sigma$	min	Q1	median	Q3	max
ASM GS	2.18 ± 0.89	1.10	1.66	1.93	2.40	7.70
shape model fit	1.56 ± 0.59	0.95	1.27	1.45	1.66	6.73
AAM GS	1.93 ± 0.47	1.08	1.61	1.85	2.16	4.48
Heart	$\mu \pm \sigma$	min	Q1	median	Q3	max
ASM GS	5.87 ± 2.93	1.40	3.77	5.61	7.47	17.00
shape model fit	1.67 ± 1.27	0.57	1.16	1.42	1.83	17.93
AAM GS	3.61 ± 1.53	1.11	2.50	3.38	4.51	10.98
Clavicles	$\mu \pm \sigma$	min	Q1	median	Q3	max
ASM GS	1.99 ± 1.22	0.69	1.28	1.64	2.12	7.41
shape model fit	1.28 ± 0.53	0.64	1.00	1.19	1.40	5.25
AAM GS	2.05 ± 0.73	0.84	1.57	1.92	2.35	6.66

Table 5: Segmentation results for lung, heart and clavicles, using the tuned ASM system initialized with the gold standard, fitting the shape model from the ASM system directly to the gold standard, and the AAM whiskers system initialized with the gold standard. These results provide upper bounds for ASM and AAM systems. All results are in terms of the mean absolute contour distance, given in millimeter.

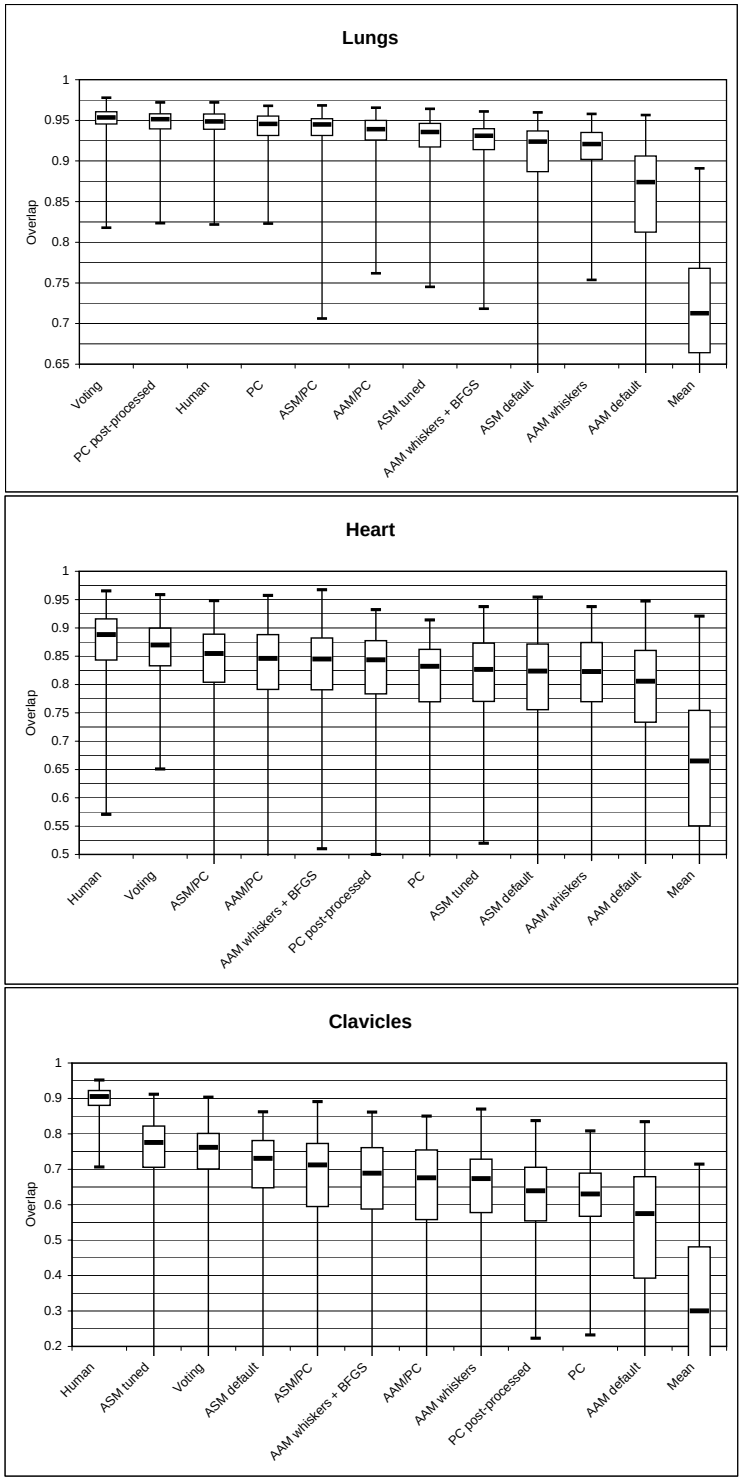


Figure 4: Box plots of the overlap Ω for lungs, heart and clavicles for all methods considered. The corresponding numerical values are listed in Table 2.

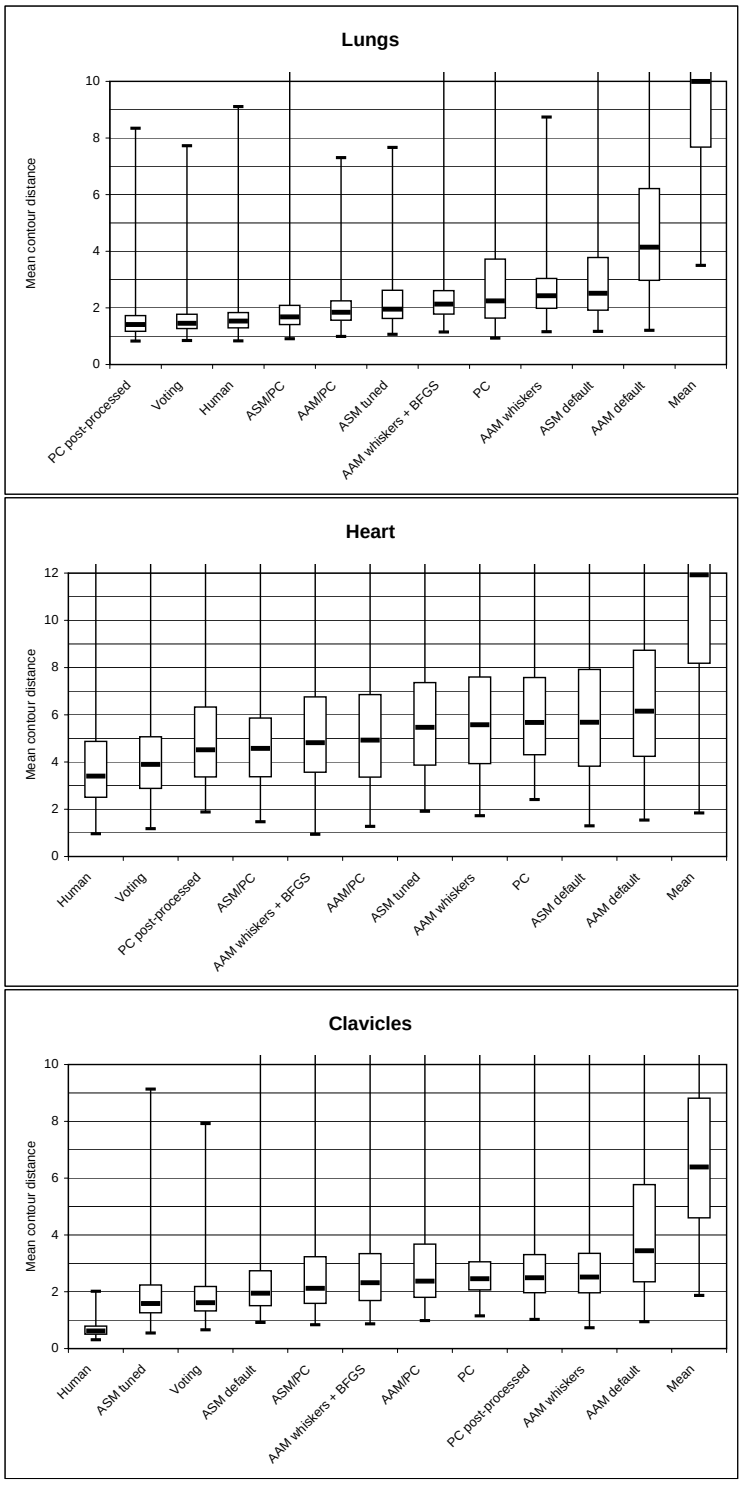


Figure 5: Box plots of the mean contour distance for lungs, heart and clavicles for all methods considered. The corresponding numerical values are listed in Table 3.

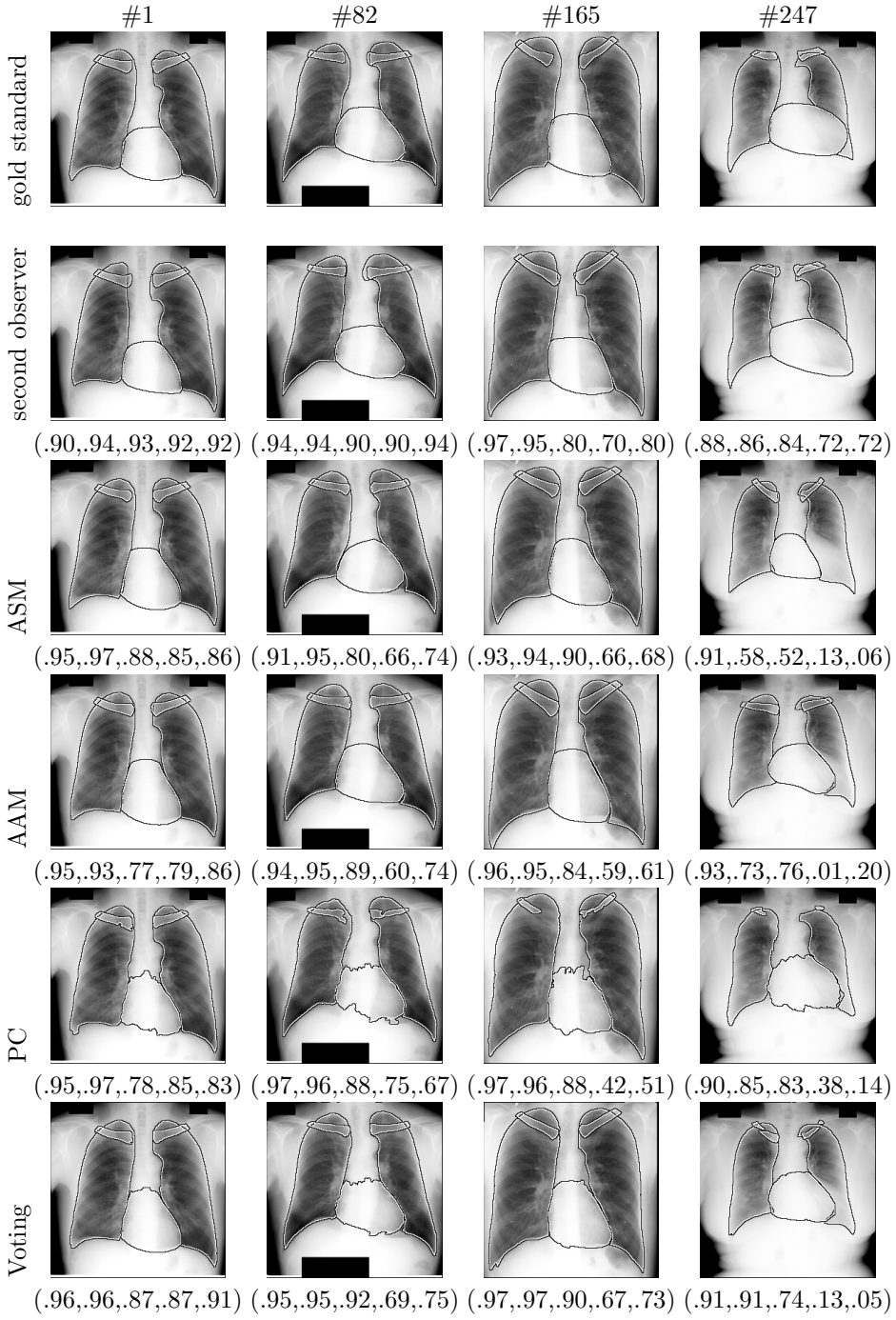
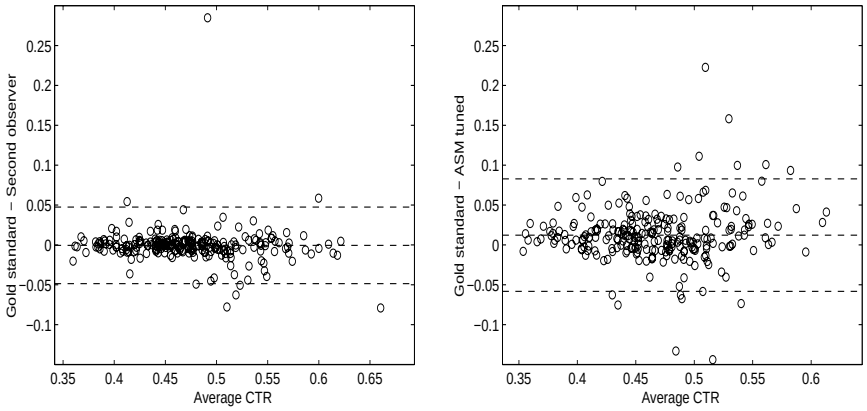
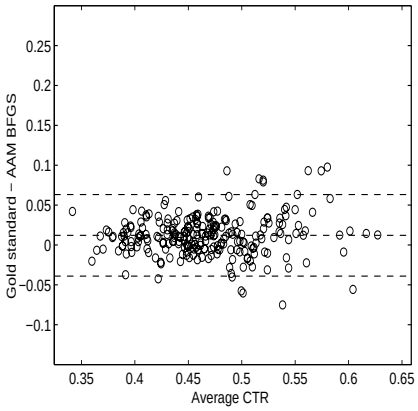


Figure 6: Segmentation results for the gold standard, the second observer, the best ASM, AAM, and PC systems, and the voting system which combines the three latter systems, respectively. Four cases are shown ranging from easy (left) to difficult (right). See the text for details on how this ranking has been computed. Below each image the overlap Ω is listed for the right lung, left lung, heart, right clavicle and left clavicle, respectively.

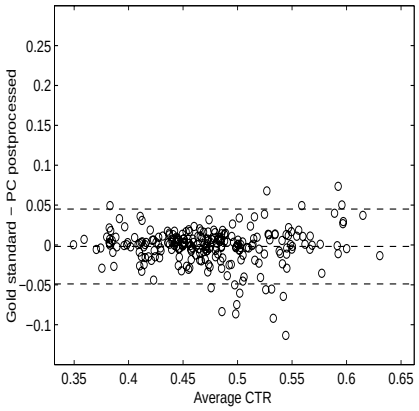


$\bar{d} = -0.00052$; (-0.049,0.047)

$\bar{d} = 0.012$; (-0.058,0.083)



$\bar{d} = 0.012$; (-0.039,0.063)



$\bar{d} = -0.0019$; (-0.048,0.045)

Figure 7: Bland and Altman plots of the cardio-thoracic ratio computed from the gold standard versus the second observer, ASM, AAM and PC. In the graphs and below them the mean difference and the 95% confidence intervals ($\bar{d}-2\sigma, \bar{d}+2\sigma$) are given.