A Probabilistic Model for Automatic Segmentation of the Esophagus in 3-D CT Scans

Johannes Feulner, S. Kevin Zhou, *Member, IEEE,* Matthias Hammon, Sascha Seifert, Martin Huber, Dorin Comaniciu, *Senior Member, IEEE,* Joachim Hornegger, *Member, IEEE,* Alexander Cavallaro

***Abstract*—Being able to segment the esophagus without user interaction from 3-D CT data is of high value to radiologists during oncological examinations of the mediastinum. The segmentation can serve as a guideline and prevent confusion with pathological tissue. However, limited contrast to surrounding structures and versatile shape and appearance make segmentation a challenging problem. This paper presents a multi-step method: First, a detector that is trained to learn a discriminative model of the appearance is combined with an explicit model of the distribution of respiratory and esophageal air. In the next step, prior shape knowledge is incorporated using a Markov chain model. We follow a "detect and connect" approach to obtain the maximum a posteriori estimate of the approximate esophagus shape from hypothesis about the esophagus contour in axial image slices. Finally, the surface of this approximation is non-rigidly deformed to better fit the boundary of the organ. The method is compared to an alternative approach that uses a particle filter instead of a Markov chain to infer the approximate esophagus shape, to the performance of a human observer and also to state of the art methods, which are all semiautomatic. Cross-validation on 144 CT scans showed that the Markov chain based approach clearly outperforms the particle filter. It segments the esophagus with a mean error of 1.80 mm in less than 16 s on a standard PC. This is only 1 mm above the inter observer variability and can compete with the results of previously published semiautomatic methods.**

***Index Terms*—esophagus, segmentation, tubular structure, Markov chain, tracking**

I. Introduction

DURING oncological examinations of the chest, radiologists are particularly interested in the region around the trachea and the esophagus [1]. These are natural gateways into the body and therefore often surrounded by lymph nodes, which need to be examined for all types of cancer. CT scans of the thorax are common practice for diagnosis and in order to assess whether treatment is effective. While the trachea is clearly visible in CT, the esophagus is much harder to see and can easily be confused with other structures, which is one reason that makes the interpretation of the CT images tedious. An automatic segmentation can serve as a guideline and provide valuable overview to a physician.

A segmentation is also useful for therapy planning. Atrial fibrillation, which is a major cause of stroke, can be treated with an ablation therapy in the heart. During this intervention, however, there is a small risk of an atrial-esophageal fistula. Then, air from the esophagus can enter the left atrium, which normally leads to the death of the patient [2]. A preoperative segmentation of the esophagus can be useful for intervention planning.

Automatic esophagus segmentation is a challenging problem. The wall of the esophagus consists of muscle tissue, which has a low contrast to vessels, other muscles and lymph nodes. Shape and appearance can vary a lot. It appears solid if it is empty, but it can also be filled with air, remains of orally given contrast agent, or both. Even for a human, it is often impossible to accurately delineate the boundaries given only a single slice. Fig. 1 shows two examples along with manual ground truth segmentation.

Up to now, the amount of publications on the topic is limited. In [3], a method is described which combines a spatial prior of the esophagus centerline with a histogram based appearance model. The centerline is extracted using a shortest path algorithm. Then, ellipses are fitted into axial slices by optimizing an energy function that is again histogram based and also has a regularization term for smooth transitions between neighboring slices. The method is semiautomatic and requires two manually placed points on the centerline and also a segmentation of the left atrium and the aorta as input. In [4], another semiautomatic segmentation method is proposed which also uses a spatial prior of the esophagus centerline. The prior is estimated relative to a set of axial 2-D contours of vertebrae, the trachea, the left main bronchi, the aorta and the heart that were segmented manually in seven reference slices. This is combined with a level set segmentation, which is initialized with the detected centerline. In [5], contour lines that were manually drawn in axial slices are interpolated in the frequency domain without using the image itself. In [6], the user draws one contour in an axial slice, and registration based on optical flow is used to propagate the contour to neighboring slices. The segmentation error was not evaluated quantitatively.

In the last years, discriminative learning has become popular for object detection [7], [8], [9]. In [10], we proposed a model based approach for esophagus segmentation which consists of multiple sub-steps. First, elliptical candidates of the esophagus contour are generated for each axial slice using a cascade of

J. Feulner and J. Hornegger are with the University of Erlangen-Nuremberg, Pattern Recognition Lab, Martensstr. 3, 91058 Erlangen, Germany (see http://www5.informatik.uni-erlangen.de)

S. K. Zhou and D. Comaniciu are with Siemens Corporate Research, 755 College Road East, Princeton, N.J. 08540, U.S.A.

J. Feulner, S. Seifert and M. Huber are with Siemens Corporate Technology, Günther-Scharowsky-Str. 1, 91058 Erlangen, Germany

M. Hammon and A. Cavallaro are with the University of Erlangen-Nuremberg, Radiology Institute, Maximiliansplatz 1, 91054 Erlangen, Germany

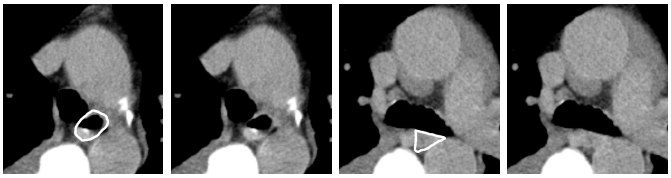


Fig. 1. Two axial slices with and without manual ground truth segmentation displayed as white contours. In the right example, it is hardly possible to accurately delineate the boundary given only one slice.

detectors based on discriminative learning techniques. This is combined with prior knowledge of the esophagus shape which is modeled with a Markov chain. This allows to efficiently infer the most likely path through the axial slices. Finally, a surface is generated and further refined using a detector that was trained to find the esophagus boundary. In [11], we extended this approach in multiple ways:

- A region of interest (ROI) detection step was added to make the segmentation work fully automatically on whole CT scans as they come from the scanner. In [10], the ROI was selected manually. In order to make sure that the esophagus is always inside, the ROI has to be made relatively large, which makes the detection problem harder because of the increased search space and more clutter that is visible in the larger ROI.
- An explicit model of respiratory and esophageal air was included because we noticed that esophageal air holes rather distract the detector, even though they are a clear hint to a human observer.
- An additional step was added after the first step of the detection cascade. It differs from the first step of the cascade only in the way the training examples are generated: The false positive detections of the first detector are used as negative training examples of the second.
- The quantitative results are considerably better, even though the detection problem on the larger ROI is harder.

This paper presents the method of [10] including the extensions of [11] in detail. It introduces a new variant of the Condensation algorithm for particle filtering and compares the Markov chain based "detect and connect" approach to the particle filtering based tracking approach. Particle filtering is commonly used for tracking over time, but can also track tubular structures. New experiments were added, including experiments on three additional databases. This enables a better comparability with previously published methods.

We evaluated our method on a large set of CT scans. Manual segmentations served as ground truth. Besides evaluating the influence of model parameters, we also investigated the inter observer variability of manual segmentations.

The paper is structured as follows: Section II describes the region of interest detection, ellipse candidate detection, path inference and surface generation steps of our method. Section III presents experiments and results, and section IV concludes the paper.

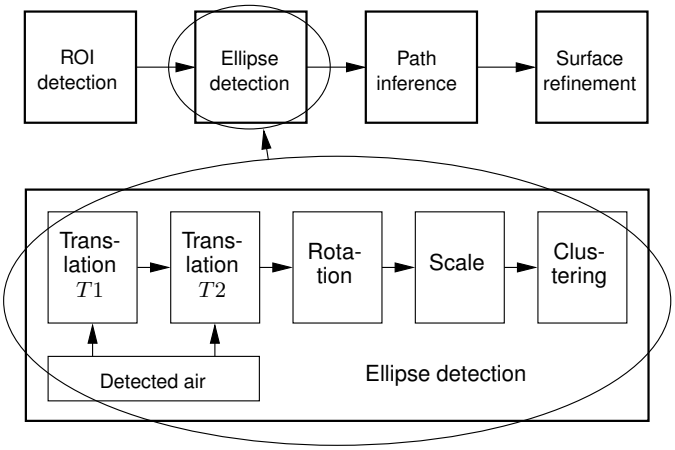


Fig. 2. Overview of the system and the steps involved in ellipse detection.

II. Esophagus Segmentation

A. Region of interest detection

CT scans containing the thoracic esophagus typically show at least the hole thorax and often also the abdominal region. To simplify and accelerate the actual segmentation, a region of interest (ROI) is automatically detected. The ROI is an axis aligned cuboidal region that is rigidly attached to an anatomic landmark. As landmark, the bifurcation of the trachea is used because it is close to the esophagus and can be detected robustly in CT as it is unique and rich in contrast. We follow the approach of [12] to find this landmark. Instead of directly constructing a single detector, a network of detectors for salient axial slices and landmarks is used to constrain the results to be anatomically reasonable. This helps to resolve ambiguities and improves robustness. The size of the ROI and the offset from the landmark was selected such that the esophagus was always inside in all datasets that were available for evaluation with a margin of at least 3 cm in x and y direction, where the x axis points to the left and the y axis to the back. The resulting cuboid has a cross section of $13.3 \times 15.6\text{cm}^2$. Along the z axis, which points upwards, the size is set to 26 cm. This ROI is relatively large which assures that the esophagus is not missed, but which also makes the detection problem harder as the region contains more clutter.

B. Ellipse detection

In the first steps of our method, the contour of the esophagus in an axial slice is approximated using an ellipse with parameters $\boldsymbol{e}$

$$\boldsymbol{e} = (\boldsymbol{t}, \theta, \boldsymbol{s}), \tag{1}$$

where $\boldsymbol{t} = (x, y)$ is the center, θ is the rotation angle within the slice, and $\boldsymbol{s} = (a, b)$ are the lengths of the semi major and semi minor axes ($a \geq b$). Using ellipses, we can get a good approximation of the contour with only five degrees of freedom.

For each axial slice, we want to find a set of ellipse candidates $\boldsymbol{e}^{(i)}$, $i = 1 \ldots N$, which are hypothesis of the true esophagus contour. Fig. 2 gives an overview of the ellipse detection process. Instead of searching the five dimensional

search space directly, we use a technique called marginal space learning which has been proposed in [9]. The idea is to prune large portions of the search space using classifiers that were trained on marginal spaces, which are translation $\boldsymbol{t}$ only and translation together with orientation $(\boldsymbol{t}, \theta)$ in this case. The classifiers form a cascade. At leach level, candidates with a poor detection score are rejected, and the remaining ones are propagated to the next level and augmented if the dimension increases. Translation is detected by two classifiers $T1$ and $T2$ that differ in the examples they are trained with. They also take into account the distribution of respiratory and esophageal air, which is further explained in section II-B1. Finally, the set of candidates is reduced in a clustering step.

As classifiers, we use probabilistic boosting trees (PBT) [8]. A PBT is a binary decision tree with a strong AdaBoost classifier at each node.

In step $T1$, we only consider translation and train a classifier to learn the probability

$$p(m = 1|\boldsymbol{H}(\boldsymbol{t})) \tag{2}$$

of whether there is an ellipse model instance with center $\boldsymbol{t}$ given a feature vector $\boldsymbol{H}(\boldsymbol{t})$ that was extracted at position $\boldsymbol{t}$. Here, m is the binary class label which is either one for "true" or zero for "false", and $\boldsymbol{H}$ are 3-D Haar-like features. These are combinations of simple cuboid filters similar to the rectangle filters described in [7]. Although being simple, these features are powerful because they can be computed very efficiently with the help of an integral image, which allows to search the whole volume exhaustively. A set of translation candidates $C_{T1} = \{\boldsymbol{t}_1 \dots \boldsymbol{t}_{N_{T1}}\}$ is generated from the N_{T1} positions with best detection score $p(m = 1|\boldsymbol{H}(\boldsymbol{t}))$.

In step $T2$, a second classifier is trained to also learn

$$p(m = 1|\boldsymbol{H}(\boldsymbol{t})). \tag{3}$$

Training requires two sets containing positive and negative examples. While the negative examples for the first classifier are generated by random sampling and rejecting samples too close to the ground truth annotations, the negatives of the second classifier are the false positives of the first one. The second classifier only considers the set C_{T1} and generates a new set C_{T2} containing the N_{T2} top position candidates, where $N_{T1} > N_{T2}$. This significantly improves the accuracy because the second one gets specialized on the difficult cases [11].

A third classifier is trained to learn the probability

$$p(m = 1|\boldsymbol{S}(\boldsymbol{t}, \theta)) \tag{4}$$

of a model instance given a vector $\boldsymbol{S}$ of steerable features [9] that depend on rotation and translation. These are point features like the image intensity, the gradient, and nonlinear combinations of them which are evaluated on a regular grid of size $7 \times 7 \times 3$ that is placed at position $\boldsymbol{t}$ and rotated according to θ. While these features are more expensive to compute compared to the Haar-like features, rotation detection is still efficient because only the set C_{T2} of position candidates needs to be considered. The result is a set containing N_{TR} rotation and translation candidates.

A fourth classifier is trained on the full search space to learn

$$p(m = 1|\boldsymbol{S}(\boldsymbol{t}, \theta, \boldsymbol{s})), \tag{5}$$

again using steerable features, but now the sampling pattern is also scaled according to $\boldsymbol{s}$. It generates the final set of ellipse candidates $C = \{\boldsymbol{e}_1 \dots \boldsymbol{e}_N\}$. Also for the last two classifiers, the negative training examples are generated from the false positives of the previous one.

1) Incorporating the distribution of air: In section II-B, the detection only relies on local features. As described in [11], this can lead to ambiguities in the presence of air bubbles in the esophagus. To a human, air bubbles, which are easy to see in CT, are a clear hint for the esophagus. But we observed that instead of learning a correlation between air bubbles and the esophagus, the classifier is rather distracted by esophageal air. The reason is that locally, esophageal air looks similar to respiratory air, which is a priori much more likely because the lung and the trachea cover a larger volume. A human, however, easily recognizes and excludes the respiratory organs.

We found that the detection performance can be improved by adding the knowledge that respiratory air cannot belong to the esophagus, while air elsewhere most likely does. This is modeled by a binary mask $B(\boldsymbol{t})$

$$B(\boldsymbol{t}) = \begin{cases} 0 & : \ \boldsymbol{t} \text{ belongs to a respiratory organ} \\ 1 & : \ \text{else} \end{cases} \tag{6}$$

and a probability map $A(\boldsymbol{t})$ of the esophagus that is generated from detected air holes. Respiratory air can be segmented easily in CT because it forms one connected region. In the first step, voxels with an attenuation coefficient below -625HU are labeled as air. To also include vessels and airways inside the lung, 2-D connected components in axial, sagittal and coronal slices with an area below 9 cm^2 are labeled as air as well. Now all 3-D connected components marked as air that touch the left, right, front or back border of the region of interest are removed. The removed regions are labeled as 0 in B. Elsewhere, B is set to 1. Regions filled with air that were not removed probably belong to the esophagus. These regions are marked in a second binary mask $E(\boldsymbol{t})$

$$E(\boldsymbol{t}) = \begin{cases} 1 & : \ \text{esophageal air at } \boldsymbol{t} \\ 0 & : \ \text{else.} \end{cases} \tag{7}$$

A similar method to detect air holes in the esophagus is described in [13]. If now an axial slice of E contains exactly one connected region marked as esophageal air, this is a very strong hint for the esophagus. Then, the corresponding axial slice of A is set to

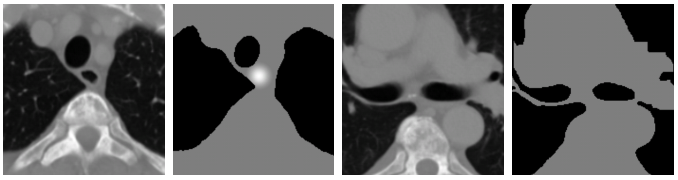
$$A(\boldsymbol{t}) = g(\|\boldsymbol{t} - \boldsymbol{p}\|_2) \tag{8}$$

$$g(r) = \max\left(\frac{e^{-\frac{r^2}{\sigma_a^2}} - e^{-\frac{w^2}{\sigma_a^2}}}{1 - e^{-\frac{w^2}{\sigma_a^2}}}, 0\right), \tag{9}$$

where $\boldsymbol{p}$ denotes the point of gravity of the region within the slice and g is a Gaussian with standard deviation σ_a that is deformed to have a maximum of 1 and limited support in $[-w, w]$. We selected a value of 7 mm as σ_a and 10 mm as w.

We now define a combined probability map $C(\boldsymbol{t})$ as

$$C(\boldsymbol{t}) = \frac{B(\boldsymbol{t}) + A(\boldsymbol{t})}{2} \tag{10}$$



(a) (b) (c) (d)

Fig. 3. Two examples of CT slices (a,c) along with their combined probability map $C(\boldsymbol{t})$ (b,d) generated from the distribution of air inside the volume. Left: The air hole is a clear hint for the esophagus. Right: No air hole is present, but respiratory air can be excluded.

and model the probability $p(m = 1|C(\boldsymbol{t}))$ of observing the esophagus at position $\boldsymbol{t}$ given the global distribution of air as being proportional to $C(\boldsymbol{t})$:

$$p(m = 1|C(\boldsymbol{t})) \propto C(\boldsymbol{t}). \tag{11}$$

During position detection, we are finally interested in the probability $p(m = 1|\boldsymbol{H}(\boldsymbol{t}), C(\boldsymbol{t}))$ of observing the esophagus at a certain location $\boldsymbol{t}$ given the Haar-like feature response $\boldsymbol{H}(\boldsymbol{t})$ and the information from the global distribution of air $C(\boldsymbol{t})$. In order to simplify the notation, we will omit the argument $\boldsymbol{t}$ in the remainder of this section. Using Bayes' rule, this can be rewritten as

$$p(m = 1|\boldsymbol{H}, C) = \frac{p(\boldsymbol{H}, C|m = 1)p(m = 1)}{p(\boldsymbol{H}, C)}. \tag{12}$$

Now we assume that the feature vector $\boldsymbol{H}$ is statistically independent from the distribution of air C. This is of course a simplifying assumption. The feature vector $\boldsymbol{H}$ is affected by the presence of air, and therefore $\boldsymbol{H}$ and C are to some extent statistically dependent. But this dependency is not very strong because $\boldsymbol{H}$ is extracted from a local neighborhood, while C captures the global distribution of air. Locally, respiratory and esophageal air look very similar, but globally, they can be well distinguished. The assumption is further justified by the fact that the map C does improve the performance as we will see, which means that $\boldsymbol{H}$ does not contain too much information about C. With this assumption, (12) can be transformed into

$$p(m = 1|\boldsymbol{H}, C) = \frac{p(\boldsymbol{H}|m = 1)p(C|m = 1)p(m = 1)}{p(\boldsymbol{H})p(C)} \tag{13}$$

$$= \frac{p(m = 1|\boldsymbol{H})p(m = 1|C)}{p(m = 1)}, \tag{14}$$

which is proportional to the product $p(m = 1|\boldsymbol{H})C$ of the classifier output and the probability map C. This means we can integrate C into a translation detector simply by multiplying it with the detection score. This is done for both detectors $T1$ and $T2$. In Fig. 3, the probability map C is visualized for two axial CT slices.

Regions filled with respiratory air are not considered by the detector. Therefore, we also do not generate negative training examples from these regions. This makes the learning problem easier because now air is a priori more likely to be part of the esophagus.

The final detection score of an ellipse candidate $\boldsymbol{e}$ is now

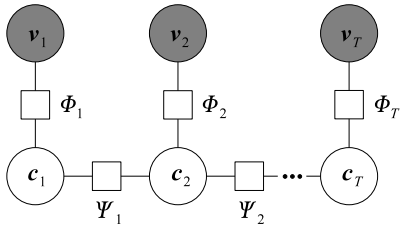


Fig. 4. Factor graph of the Markov chain model.

modeled as being proportional to the product

$$p(m = 1|\boldsymbol{e}) \propto p(m = 1|\boldsymbol{H}(\boldsymbol{t}), C(\boldsymbol{t})) \cdot p(m = 1|\boldsymbol{S}(\boldsymbol{t}, \theta))p(m = 1|\boldsymbol{S}(\boldsymbol{t}, \theta, \boldsymbol{s})) \tag{15}$$

of the translation, the rotation and the scale detection score. Here, we only take into account the score of the second translation detector $T2$ in order not to overemphasize translation. Detector $T1$ is only used to reject most samples at an early stage.

2) Clustering: In order to reduce subsequent search effort and to detect modes in the distribution of the candidates $\{\boldsymbol{e}^{(1)} \dots \boldsymbol{e}^{(N)}\}$, they are spatially clustered using an agglomerative hierarchical average-linkage clustering algorithm until a distance threshold d_{max} is reached. Two clusters are merged if the mean radius of inter-cluster pairs of points is below d_{max}. The result is a set of cluster centers $\{\boldsymbol{c}^{(1)} \dots \boldsymbol{c}^{(K)}\}$ with weights $\{\sigma^{(1)} \dots \sigma^{(K)}\}$, where the weight $\sigma^{(k)}$ of cluster center k is the sum of detections scores $p(m = 1|\boldsymbol{e})$ of its members.

C. Path inference

1) First order Markov model: So far, the axial slices of the volume image were treated separately. Shape knowledge is incorporated into a Markov chain model [14] of the esophagus and used to infer the most likely path through the axial slices. A factor graph [15] of the Markov model used here is depicted in Fig. 4. The variables $\boldsymbol{c}_1 \dots \boldsymbol{c}_T$ correspond to the axial slices of the image. Possible states of a variable $\boldsymbol{c}_t$ are the ellipses corresponding to the cluster centers $\boldsymbol{c}_t^{(k)}, k = 1 \dots K_t$ of slice t. Note that the slice index t and the 2-D ellipse center $\boldsymbol{t}$ are different variables. Each state variable $\boldsymbol{c}_t$ is conditioned on the observed image slice $\boldsymbol{v}_t$. The clique potentials (or factors) of the observation cliques are denoted with Φ_t. They are set to the scores of the cluster centers:

$$\Phi_t(\boldsymbol{c}_t^{(k)}, \boldsymbol{v}_t) = \sigma_t^{(k)}. \tag{16}$$

The clique potentials Ψ_t between adjacent state variables $\boldsymbol{c}_t, \boldsymbol{c}_{t+1}$ represent the prior shape knowledge. They are set to the transition distribution from one slice to another:

$$\Psi_t(\boldsymbol{c}_t, \boldsymbol{c}_{t+1}) = p(\boldsymbol{c}_{t+1}|\boldsymbol{c}_t). \tag{17}$$

To simplify the transition distribution, it was assumed that the transition of the translation parameters x, y is statistically independent from the other parameters. The same was assumed for the scale parameters. As the rotation parameter θ is not well defined for approximately circular ellipses, the transition

of rotation also depends on the scale parameters, but independence was assumed for translation and scale parameters. With these assumptions, the transition distribution can be factorized and becomes

$$p(\boldsymbol{c}_{t+1}|\boldsymbol{c}_t) = p(\boldsymbol{t}_{t+1}|\boldsymbol{t}_t) \tag{18}$$
$$\cdot p(\theta_{t+1}|\theta_t, \boldsymbol{s}_t) \tag{19}$$
$$\cdot p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t). \tag{20}$$

The translation transition (18) is modeled as a 2-D normal distribution

$$p(\boldsymbol{t}_{t+1}|\boldsymbol{t}_t) = \mathcal{N}(\Delta\boldsymbol{t}_t|\boldsymbol{\Sigma}_p, \boldsymbol{m}_p) \tag{21}$$

with $\Delta\boldsymbol{t}_t = \boldsymbol{t}_{t+1} - \boldsymbol{t}_t$, and the transition (20) of scale $\boldsymbol{s} = (a, b)$ as a 4-D normal distribution

$$p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t) = \mathcal{N}(\boldsymbol{s}_{t+1}, \boldsymbol{s}_t|\boldsymbol{\Sigma}_s, \boldsymbol{m}_s). \tag{22}$$

In (21) and (22), $\boldsymbol{\Sigma}_p$ and $\boldsymbol{\Sigma}_s$ are the covariance matrices, and $\boldsymbol{m}_p$ and $\boldsymbol{m}_s$ are the mean vectors of the two normal distributions.

The variance of the rotation transition is small for an elongated ellipse because the esophagus is usually smooth and not heavily twisted. However, the variance highly increases for more circular ellipses. The reason is that θ takes arbitrary values for a circle. This may results in big jumps of θ from slice to slice even though the shape of the esophagus contour hardly changes.

This is handled by modeling (19) with different normal distributions for elongated and more circular ellipses. In total, we use ten 1-D normal distributions, one for a certain interval of circularity, which is measured by the ratio $\frac{b}{a}$ of the length of the semi minor and the semi major axis:

$$p(\theta_{t+1}|\theta_t, a_t, b_t) \approx \mathcal{N}\left(\Delta\theta_t \middle| \sigma_r\left(\frac{b_t}{a_t}\right), m_r\left(\frac{b_t}{a_t}\right)\right). \tag{23}$$

Here, $\sigma_r(\frac{b}{a})$ is the standard deviation and $m_r(\frac{b}{a})$ the mean of the normal distribution that corresponds to the circularity value $\frac{b}{a}$. Fig. 5 shows samples of rotation transitions along with the circularity. It illustrates that (19) can be represented with a Gaussian for a fixed circularity value.

This is an approximation because a normal distribution has only one mode and unlimited support, but a rotation by 180° results into the same ellipse. It is however enough to only consider the range of $\Delta\theta$ between $-90°$ and $90°$, where the approximation works well.

The parameters of all normal distributions were estimated from manually annotated data. The annotations are contours of the esophagus drawn in each axial slice. For each slice, an ellipse is fitted into the contour points. We use the method described in [16] to non-iteratively compute the least squares solution. The transitions from one slice to the next are treated as samples and used to compute the mean vectors and covariance matrices.

The a posteriori joint distribution of all states $p(\boldsymbol{c}_{1:T}|\boldsymbol{v}_{1:T})$ is now given by the product of all factors of the factor graph.

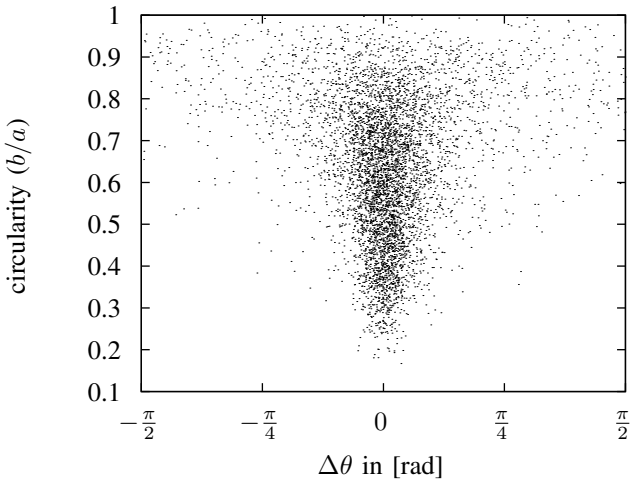


Fig. 5. Samples of rotation transitions from one axial slice to another of the ellipses fitted to the ground truth annotations. For a fixed circularity value $\frac{b}{a}$, the 1-D conditional probability density $p(\Delta\theta|\frac{b}{a})$ can be well approximated with a normal distribution. The standard deviation is higher for high circularity values. For $\frac{b}{a} = 1$, the distribution is uniform in $[-\frac{\pi}{2}, \frac{\pi}{2})$.

The maximum a posteriori (MAP) estimate

$$\begin{aligned}\hat{\boldsymbol{c}}_{1:T}^{(\mathrm{MAP})} &= \operatorname*{argmax}_{\boldsymbol{c}_{1:T}} p(\boldsymbol{c}_{1:T}|\boldsymbol{v}_{1:T}) \\ &= \operatorname*{argmax}_{\boldsymbol{c}_{1:T}} \Phi_1(\boldsymbol{c}_1, \boldsymbol{v}_1) \prod_{t=2}^{T} \Phi_t(\boldsymbol{c}_t, \boldsymbol{v}_t)\Psi_{t-1}(\boldsymbol{c}_{t-1}, \boldsymbol{c}_t)\end{aligned} \tag{24}$$

can be computed efficiently using dynamic programming.

2) Second order Markov model: Alternatively, the shape prior was modeled with a Markov chain that assumes a Markov order of two for the transition of translation. The transition of rotation and scale was handled as described in section II-C1 because we observed that the translation parameters are more continuous compared to rotation and scale, and therefore the shape prior should benefit most from a second order assumption here. Furthermore, generalizing (22) and (23) to a second order model leads to problems because of limited training data. Now, the factor Ψ_t corresponding to the state transition probability (17) becomes

$$\Psi_t(\boldsymbol{c}_{t-1}, \boldsymbol{c}_t, \boldsymbol{c}_{t+1}) = p(\boldsymbol{c}_{t+1}|\boldsymbol{c}_t, \boldsymbol{c}_{t-1}) \tag{25}$$
$$= p(\boldsymbol{t}_{t+1}|\boldsymbol{t}_t, \boldsymbol{t}_{t-1}) \tag{26}$$
$$\cdot p(\theta_{t+1}|\theta_t, \boldsymbol{s}_t) \tag{27}$$
$$\cdot p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t). \tag{28}$$

The translation transition is again modeled using a normal distribution, but now, also the second derivative

$$\Delta\Delta\boldsymbol{t}_t = \Delta\boldsymbol{t}_t - \Delta\boldsymbol{t}_{t-1} \tag{29}$$

of $\boldsymbol{t}$ with respect to z is considered, which corresponds to the curvature of the esophagus:

$$p(\boldsymbol{t}_{t+1}|\boldsymbol{t}_t, \boldsymbol{t}_{t-1}) = \mathcal{N}(\Delta\boldsymbol{t}_t, \Delta\Delta\boldsymbol{t}_t|\boldsymbol{\Sigma}_{p_2}, \boldsymbol{m}_{p_2}). \tag{30}$$

The MAP estimate

$$\hat{\boldsymbol{c}}_{1:T}^{(\mathrm{MAP})} = \underset{\boldsymbol{c}_{1:T}}{\operatorname{argmax}} \left(\Phi_1(\boldsymbol{c}_1, \boldsymbol{v}_1)\Phi_2(\boldsymbol{c}_2, \boldsymbol{v}_2)\Psi_1(\boldsymbol{c}_1, \boldsymbol{c}_2) \right. \quad (31)$$

$$\left. \cdot \prod_{t=3}^{T} \Phi_t(\boldsymbol{c}_t, \boldsymbol{v}_t)\Psi_{t-1}(\boldsymbol{c}_{t-2}, \boldsymbol{c}_{t-1}, \boldsymbol{c}_t) \right) \quad (32)$$

can be computed in the same way as in the single order case.

3) Particle filtering: We furthermore investigated to model and infer the esophagus path with a particle filter approach [17] instead of using a Markov chain. Particle filtering, which is also known as condensation, is popular for tracking applications. Probability distributions are represented non-parametrically with weighted samples which are called particles. In contrast to Kalman or extended Kalman filtering, the distributions may be multimodal, which allows to model different hypothesis of the true state. It is also becoming popular for tracking of tubular structures [18], [19].

We formulate the problem of inferring the esophagus shape in the framework of dynamic state estimation. Though the problem is not dynamic, we treat the vertical axis z of the volume image as time t. As before, an axial slice becomes the observation $\boldsymbol{v}_t$. The unknown state at time t is the ellipse parameter vector $\boldsymbol{e}_t$. Given the observation density $p(\boldsymbol{v}_t|\boldsymbol{e}_t)$ and the state transition density $p(\boldsymbol{e}_{t+1}|\boldsymbol{e}_t)$, the probability density $p(\boldsymbol{e}_{t+1}|\boldsymbol{v}_{1:t+1})$ of the state at time $t+1$ given all previous observations can be computed recursively as

$$p(\boldsymbol{e}_{t+1}|\boldsymbol{v}_{1:t+1}) = \frac{p(\boldsymbol{v}_{t+1}|\boldsymbol{e}_{t+1}) \int p(\boldsymbol{e}_{t+1}|\boldsymbol{e}_t)p(\boldsymbol{e}_t|\boldsymbol{v}_{1:t})d\boldsymbol{e}_t}{p(\boldsymbol{v}_{t+1}|\boldsymbol{v}_{1:t})} \quad (33)$$

for a Markov order of one. Here, we use the notation $\boldsymbol{v}_{1:t}$ for the sequence $\boldsymbol{v}_1 \dots \boldsymbol{v}_t$. This is the core equation of probabilistic dynamic state estimation. In the condensation algorithm [20], the integral of (33) is computed by drawing samples from the probability density $p(\boldsymbol{e}_t|\boldsymbol{v}_{1:t})$ that is represented by a set $\mathcal{S}_t$ of particles

$$\mathcal{S}_t = \left\{ \left(\boldsymbol{e}_t^{(i)}, P_t^{(i)} \right), i = 1 \dots I \right\}. \quad (34)$$

Each particle consists of a sample $\boldsymbol{e}_t^{(i)}$ and a weight $P_t^{(i)}$. The probability density $p(\boldsymbol{e}_t|\boldsymbol{v}_{1:t})$ is approximated as

$$p_{\mathcal{S}_t}(\boldsymbol{e}_t|\boldsymbol{v}_{1:t}) = \sum_{i=1}^{I} P_t^{(i)} \delta(\boldsymbol{e}_t - \boldsymbol{e}_t^{(i)}), \quad (35)$$

where δ is a window function. For particle filtering, it is common practice to simply use the Dirac window δ because the empiric probability density $p_{\mathcal{S}_t}$ is not evaluated but only used for drawing samples or computing moments.

In the Condensation algorithm [20], the samples are noisily propagated to the next time step using $p(\boldsymbol{e}_{t+1}|\boldsymbol{e}_t)$ and then weighted according to whether they fit to the new observation $\boldsymbol{v}_{t+1}$ using $p(\boldsymbol{v}_{t+1}|\boldsymbol{e}_{t+1})$. The algorithm is depicted in Fig. 6.

We found that the idea of marginal space learning can be nicely integrated into the particle filtering framework by factorizing both the observation density $p(\boldsymbol{v}_t|\boldsymbol{e}_t)$ and the state transition density $p(\boldsymbol{e}_{t+1}|\boldsymbol{e}_t)$, and expressing the factors of the observation density using classifiers that were trained on marginal spaces.

The observation density $p(\boldsymbol{v}|\boldsymbol{e})$ (to improve the readability, the subscript t is omitted in this section if it is the same for all variables) requires a generative model which is often not available in the context of object detection in images. Using Bayes' rule

$$p(\boldsymbol{v}|\boldsymbol{e}) = \frac{p(\boldsymbol{e}|\boldsymbol{v})p(\boldsymbol{v})}{p(\boldsymbol{e})}, \quad (36)$$

it can be transformed to a discriminative model $p(\boldsymbol{e}|\boldsymbol{v})$ and two priors for the image $\boldsymbol{v}$ and the parameter vector $\boldsymbol{e}$. By using Bayes' rule once again, $p(\boldsymbol{e}|\boldsymbol{v}) = p(\boldsymbol{t}, \theta, \boldsymbol{s}|\boldsymbol{v})$ can be factorized into

$$p(\boldsymbol{e}|\boldsymbol{v}) = p(\boldsymbol{t}|\boldsymbol{v})p(\theta|\boldsymbol{t}, \boldsymbol{v})p(\boldsymbol{s}|\theta, \boldsymbol{t}, \boldsymbol{v}). \quad (37)$$

For the state transition density $p(\boldsymbol{e}_{t+1}|\boldsymbol{e}_t)$, we use the factorization of (18). Together with (37), (33) can be rewritten as

$$p(\boldsymbol{e}_{t+1}|\boldsymbol{v}_{1:t+1}) \propto \frac{1}{p(\boldsymbol{e}_{t+1})} p(\boldsymbol{s}_{t+1}|\theta_{t+1}, \boldsymbol{t}_{t+1}, \boldsymbol{v}_{t+1}) \int_{\boldsymbol{s}_t} p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t)$$
$$p(\theta_{t+1}|\boldsymbol{t}_{t+1}, \boldsymbol{v}_{t+1}) \left[\int_{\theta_t} p(\theta_{t+1}|\theta_t, \boldsymbol{s}_t) \right.$$
$$p(\boldsymbol{t}_{t+1}|\boldsymbol{v}_{t+1}) \left(\int_{\boldsymbol{t}_t} p(\boldsymbol{t}_{t+1}|\boldsymbol{t}_t)p(\boldsymbol{e}_t|\boldsymbol{v}_{1:t}) \right.$$
$$\left. \left. d\boldsymbol{t}_t \right) d\theta_t \right] d\boldsymbol{s}_t. \quad (38)$$

Because $p(\boldsymbol{v})$ and $p(\boldsymbol{v}_{t+1}|\boldsymbol{v}_{1:t})$ do not depend on the parameters $\boldsymbol{e}$ that are to be estimated, they can be treated like constants. In (38), the integral of (33) over the state space was replaced by three nested integral over the scale, rotation and the translation subspace. Note that each weighted integral is very similar to the weighted integral in (33). Like in one iteration of the condensation algorithm, it can be carried out by drawing, propagating and weighting samples. But instead of propagating the samples across time, they are propagated across dimension of the state space within a single time step. This is especially useful for higher dimensional state spaces, like 5-D in this case: Filling a 5-D state space with particles would require a high number of particles. Here, we can reject particles already after the first 2-D integration if the translation parameters do not fit the new observation and concentrate on particles with promising translation parameters. The same can be done for the rotation parameters

Fig. 7 depicts the algorithm we used to compute (38). It is formulated to solve our specific problem, but the principle is always applicable if the state transition probability density (18) and the observation probability density (36) can be factorized and the factors are available.

The factors of (37) are now modeled as being proportional to the scores of the three detectors trained on the marginal spaces of translation, rotation, and scale

$$p(\boldsymbol{t}|\boldsymbol{v}) \propto p(m = 1|\boldsymbol{H}(\boldsymbol{t}), C(\boldsymbol{t})) \quad (39)$$

$$p(\theta|\boldsymbol{t}, \boldsymbol{v}) \propto p(m = 1|\boldsymbol{S}(\boldsymbol{t}, \theta)) \quad (40)$$

$$p(\boldsymbol{s}|\theta, \boldsymbol{t}, \boldsymbol{v}) \propto p(m = 1|\boldsymbol{S}(\boldsymbol{t}, \theta, \boldsymbol{s})), \quad (41)$$

Input: Old particle set

$$\mathcal{S}_t = \left\{ \left(\boldsymbol{e}_t^{(i)}, P_t^{(i)} \right), i = 1 \dots I \right\}$$

of time step t

Construct the new particle set

$$\mathcal{S}_{t+1} = \left\{ \left(\boldsymbol{e}_{t+1}^{(i)}, P_{t+1}^{(i)} \right), i = 1 \dots I \right\}$$

of the next time step from $\mathcal{S}_t$ as follows:

for i=1…I **do**

  Construct the particle $\left(\boldsymbol{e}_{t+1}^{(i)}, P_{t+1}^{(i)} \right)$:

  Select a sample $\boldsymbol{e}_t^{(j)}$ with probability $P_t^{(j)}$

  Predict by sampling from $p(\boldsymbol{e}_{t+1}|\boldsymbol{e}_t^{(j)})$ to choose the sample $\boldsymbol{e}_{t+1}^{(i)}$ of the next time step

  Weight the new sample by setting

$$P_{t+1}^{(i)} = p(\boldsymbol{v}_{t+1}|\boldsymbol{e}_{t+1}^{(i)})$$

end for

Normalize the weights s. th. $\sum_{i=1}^{I} P_{t+1}^{(i)} = 1$

Output: New particle set $\mathcal{S}_{t+1}$

Fig. 6. One iteration of the original Condensation algorithm [20].

and the prior $p(\boldsymbol{e})$ is assumed to be uniform. In contrast to the Markov chain approach, only a single translation detector is used here. A second one could be integrated by adding a fourth integral in equation (38).

Finally, we are interested in the MAP estimate

$$\hat{\boldsymbol{e}}_{1:T}^{(\text{MAP})} = \underset{\boldsymbol{e}_{1:T}}{\operatorname{argmax}}\, p(\boldsymbol{e}_{1:T}|\boldsymbol{v}_{1:T}). \tag{42}$$

This estimate can be easily obtained [17] from $p(\boldsymbol{e}_T|\boldsymbol{v}_{1:T})$ by finding the history $\boldsymbol{e}_{1:T}$ of the particle

$$\hat{\boldsymbol{e}}_T^{(\text{MAP})} = \underset{\boldsymbol{e}_T}{\operatorname{argmax}}\, p(\boldsymbol{e}_T|\boldsymbol{v}_{1:T}) \tag{43}$$

with the highest weight in the last time step T.

D. Surface generation

After the MAP estimate of the path has been detected, the sequence of ellipses is converted into a triangular mesh representation by sampling the ellipses and connecting neighboring point sets with a triangle strip.

The cross-section of the esophagus is generally not an ellipse, and the path obtained in section II-C often contains some inaccuracies. Therefore, the mesh model is further refined to better fit the surface of the organ.

A PBT classifier was trained to learn the boundary of the esophagus. The classifier uses steerable features as proposed in [9]. As for ellipse detection, the steerable features are sampled on a regular grid, but now with a size of $5 \times 5 \times 9$. For each mesh vertex, the sampling pattern is placed so that the vertex is in the center of the pattern and the longest axis points in direction of the mesh normal. Now the pattern is moved along the normal to find the maximal detector response and the new position of the vertex. Finally, the surface is

Input: Old particle set

$$\mathcal{S}_t = \left\{ \left(\boldsymbol{e}_t^{(i)} = (\boldsymbol{t}_t^{(i)}, \theta_t^{(i)}, \boldsymbol{s}_t^{(i)}), P_t^{(i)} \right), i = 1 \dots I \right\}$$

of time step t

STEP 1: Construct a first intermediate set of sub-particles

$$\mathcal{S}_{t+1,1} = \left\{ \left(\boldsymbol{t}_{t+1}^{(i)}, P_{t+1,1}^{(i)} \right), i = 1 \dots I \right\}$$

from the old particle set $\mathcal{S}_t$:

for i=1…I **do**

  Construct the sub-particle $\left(\boldsymbol{t}_{t+1}^{(i)}, P_{t+1,1}^{(i)} \right)$:

  Select a sample $\boldsymbol{e}_t^{(j)}$ with probability $P_t^{(j)}$

  Predict by sampling from $p(\boldsymbol{t}_{t+1}|\boldsymbol{e}_t^{(j)}) = p(\boldsymbol{t}_{t+1}|\boldsymbol{t}_t^{(j)})$ to choose $\boldsymbol{t}_{t+1}^{(i)}$

  Weight the new sample by setting $P_{t+1,1}^{(i)} = p(\boldsymbol{t}_{t+1}^{(i)}|\boldsymbol{v}_{t+1})$

end for

Normalize the weights s. th. $\sum_{i=1}^{I} P_{t+1,1}^{(i)} = 1$

STEP 2: Construct a second intermediate set of sub-particles

$$\mathcal{S}_{t+1,2} = \left\{ \left((\boldsymbol{t}_{t+1}^{(i)}, \theta_{t+1}^{(i)}), P_{t+1,2}^{(i)} \right), i = 1 \dots I \right\}$$

from the first intermediate set $\mathcal{S}_{t+1,1}$:

for i=1…I **do**

  Construct the sub-particle $\left((\boldsymbol{t}_{t+1}^{(i)}, \theta_{t+1}^{(i)}), P_{t+1,2}^{(i)} \right)$:

  Select a sample $\boldsymbol{t}_{t+1}^{(j)}$ with probability $P_{t+1,1}^{(j)}$

  Predict by first finding the full particle predecessor $\boldsymbol{e}_t^{(k)}$ of sub-particle $\boldsymbol{t}_{t+1}^{(j)}$ and then sampling from $p(\theta_{t+1}|\boldsymbol{e}_t^{(k)}) = p(\theta_{t+1}|\theta_t^{(k)}, \boldsymbol{s}_t^{(k)})$ to find $\theta_{t+1}^{(i)}$

  Weight with $P_{t+1,2}^{(i)} = p(\theta_{t+1}^{(i)}|\boldsymbol{t}_{t+1}^{(i)}, \boldsymbol{v}_{t+1})$

end for

Normalize the weights s. th. $\sum_{i=1}^{I} P_{t+1,2}^{(i)} = 1$

STEP 3: Construct the particle set of the next time step

$$\mathcal{S}_{t+1} = \left\{ \left(\boldsymbol{e}_{t+1}^{(i)}, P_{t+1}^{(i)} \right), i = 1 \dots I \right\}$$

from the second intermediate set $\mathcal{S}_{t+1,2}$:

for i=1…I **do**

  Construct the particle $\left(\boldsymbol{e}_{t+1}^{(i)}, P_{t+1}^{(i)} \right)$:

  Select a sample $(\boldsymbol{t}_{t+1}^{(j)}, \theta_{t+1}^{(j)})$ with probability $P_{t+1,2}^{(j)}$

  Predict by first finding the full particle predecessor $\boldsymbol{e}_t^{(k)}$ of sub-particle $(\boldsymbol{t}_{t+1}^{(j)}, \theta_{t+1}^{(j)})$ and then sampling from $p(\boldsymbol{s}_{t+1}|\boldsymbol{e}_t^{(k)}) = p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t^{(k)})$ to find $\boldsymbol{s}_{t+1}^{(i)}$

  Weight with $P_{t+1}^{(i)} = p(\boldsymbol{s}_{t+1}^{(i)}|\theta_{t+1}^{(i)}, \boldsymbol{t}_{t+1}^{(i)}, \boldsymbol{v}_{t+1})$

end for

Normalize the weights s. th. $\sum_{i=1}^{I} P_{t+1}^{(i)} = 1$

Output: New particle set $\mathcal{S}_{t+1}$.

Fig. 7. One iteration of our variant of the Condensation algorithm for the problem of ellipse tracking.

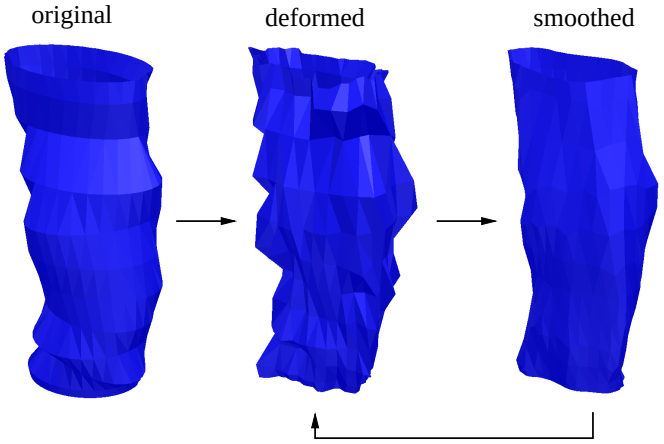


Fig. 8. Example of one boundary refinement iteration shown for a section of the esophagus. Left: The ellipses obtained in the path inference step, connected with triangle strips. Middle: Surface after displacing the vertices along their normals according to the classifier output. Right: Smoothed surface.

Scanner type	Number of datasets
Biograph 64	1
Definition AS+	1
Sensation 10	25
Sensation 16	6
Sensation 64	110
Volume Zoom	1

TABLE I
CT SCANNERS USED FOR DATA ACQUISITION.

passed through a Laplacian smoothing filter. Smoothing is necessary because the vertices are displaced independently from each other. This process of deformation and smoothing is repeated for a certain number of iterations that is varied in the experiments.

Fig. 8 shows the first iteration of boundary refinement for a section of an inferred path.

Fig. 9 summarizes the detection pipeline and shows example output for each step.

III. RESULTS

A. Results of cross-validation

The method has been evaluated on 144 CT scans of the thoracic or the thoracic and abdominal region. No patient was included twice. The voxel spacing in x and y direction was in the range of 0.7 mm to 0.8 mm. The spacing in (longitudinal) z direction was 5 mm. After ROI detection, the volumes were resampled to a voxel spacing of $0.7 \times 0.7 \times 5\text{mm}^3$. The data was acquired using six different CT scanner types listed in TABLE I. Out of the 144 datasets, 143 were reconstructed using soft tissue kernels, and one was reconstructed using a lung kernel (B70f). The filter kernels are listed in TABLE II. The accelerating voltage was 120kV in all cases, and the tube current ranged from 94mA to 575mA with a mean and standard deviation of 293.79 $\pm$ 87.25 mA.

Manual segmentations were available for all datasets. The segmentations typically ranged from the thyroid gland down to a level below the left atrium.

Filter kernel	Number of datasets
B30f	2
B31f	24
B31s	1
B40f	6
B41f	110
B70f	1

TABLE II
FILTER KERNELS USED FOR IMAGE RECONSTRUCTION.

Classifier	tree levels	weak classifiers	candidates
Translation 1	2	20	400
Translation 2	2	20	120
Rotation	2	20	50
Scale	2	20	200
Surface	5	20	n/a

TABLE III
PARAMETER SETTINGS FOR THE FIVE CLASSIFIERS OF THE DETECTION PIPELINE: THE NUMBER OF LEVELS IN THE PBT CLASSIFIER, THE NUMBER OF WEAK CLASSIFIERS PER ADABOOST NODE, AND THE NUMBER OF CANDIDATES GENERATED.

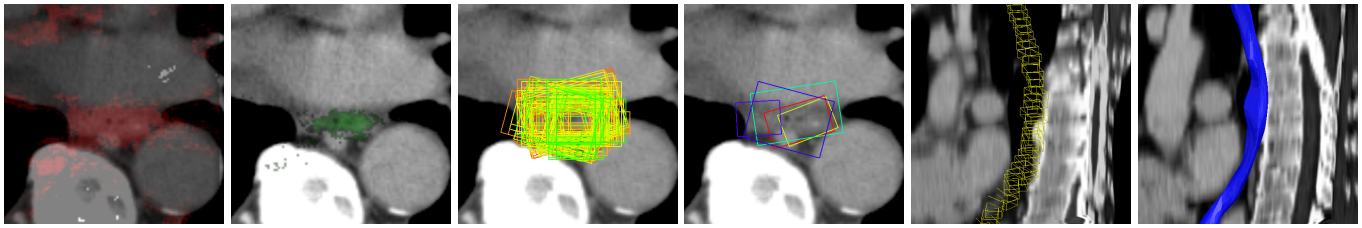
Among the scans, 34 were taken from patients suffering from lymphoma, which often causes enlarged lymph nodes in the mediastinal region. In some datasets, the esophagus contained remains of orally given contrast agent.

The accuracy was measured using threefold cross-validation. For each fold, all five classifiers for translation ($2\times$), orientation, scale and surface were trained on the training data, and the parameters of the Markov model were estimated from the same training data. The remaining data was used for testing. There was no overlap between training and testing data. For evaluation, the detector was run in z direction on the same interval covered by the manual annotation in order not to introduce artificial errors because of different lengths of the segmentations.

ROI detection succeeded in all of the 144 datasets, meaning that the bifurcation of the trachea was always detected with a reasonable accuracy. Due to the large ROI, the segmentation method can tolerate normal anatomical variations and detection errors.

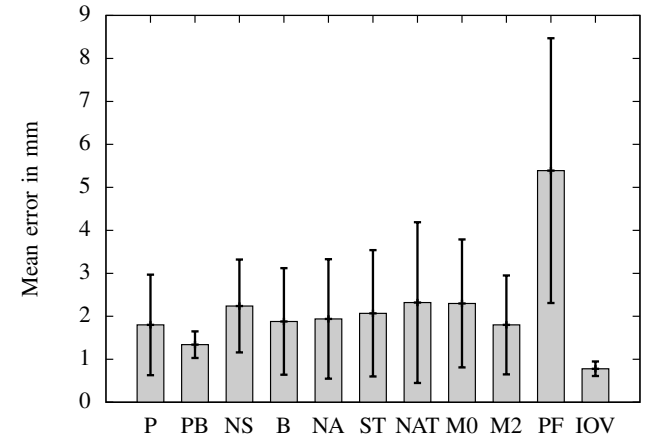
Method		mean err. in mm	Hausdorff dist. in mm
P	Proposed method	1.80 $\pm$ 1.17	12.62 $\pm$ 7.01
PB	Proposed method, best 80%	1.34 $\pm$ 0.31	9.65 $\pm$ 3.07
NS	No surface refinement	2.24 $\pm$ 1.08	12.93 $\pm$ 7.16
B	Only binary air model $B(\boldsymbol{t})$	1.88 $\pm$ 1.24	13.00 $\pm$ 7.88
NA	No air model	1.94 $\pm$ 1.39	13.06 $\pm$ 7.21
ST	Single translation class.	2.07 $\pm$ 1.47	14.50 $\pm$ 8.92
NAT	No air mdl., single trnsl. cls.	2.32 $\pm$ 1.87	15.02 $\pm$ 9.83
M0	Markov order 0	2.30 $\pm$ 1.49	17.29 $\pm$ 11.42
M2	Markov order 2	1.80 $\pm$ 1.15	12.65 $\pm$ 6.92
PF	Particle filtering	5.39 $\pm$ 3.08	22.32 $\pm$ 7.97
IOV	Inter observer variability	0.78 $\pm$ 0.17	7.29 $\pm$ 2.22

TABLE IV
RESULTS OF PERFORMANCE EVALUATION. SHOWN IS THE MEAN ERROR AND THE MEAN HAUSDORFF DISTANCE ALONG WITH THE CORRESPONDING STANDARD DEVIATIONS. FIRST ROW: THE PROPOSED METHOD USES THE FIRST ORDER MARKOV CHAIN APPROACH. ROWS 2-10: PROPOSED METHOD WITH AT LEAST ONE PARAMETER OR EXPERIMENTAL SETTING ALTERED.

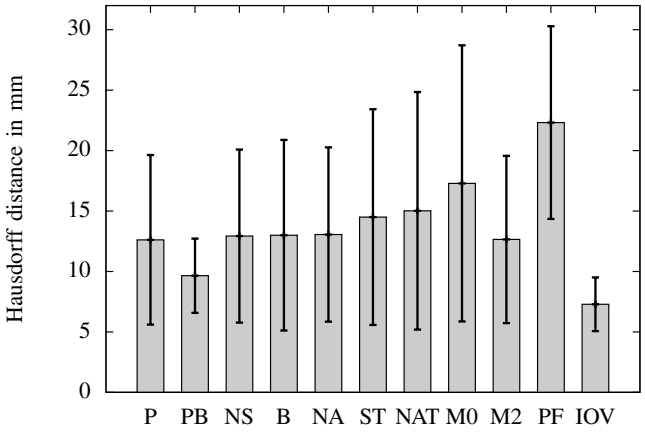


(a) (b) (c) (d) (e) (f)

Fig. 9. Example output for each step of the detection pipeline. (a): Score $p(m = 1|\boldsymbol{H})C$ of the first classifier for position multiplied with the probability map C. (b): Score $p(m = 1|\boldsymbol{H})C$ of the second classifier for position multiplied with the probability map C, evaluated for the best candidates from stage (a). (c): Candidate boxes after rotation and scale detection. The confidence of a box is color coded in HSV color space. Violet is lowest, red is highest score. (d): Cluster centers after clustering and merging. (e): Result of the path inference step. (f): Final surface after refinement.



(a)



(b)

Fig. 10. Mean segmentation error (a) and Hausdorff distance (b) of our segmentation method and different variants. The length of an error bar is two standard deviations. See text and TABLE IV for an explanation of the abbreviations.

Unless otherwise stated, parameters of the classifiers were set to the values displayed in TABLE III, the distance threshold d_{max} was set to 8 mm, and surface refinement was iterated two times.

TABLE IV and Fig. 10 show the results of performance evaluation. Two error measures were computed: The mean symmetric point-to-mesh distance, and the maximum symmetric point-to-mesh distance, which is also known as Hausdorff distance. Symmetric means that the distance between two meshes remains the same if the meshes are swapped.

Our proposed method (P), which uses the Markov chain model for path inference with a Markov order of one, segments the esophagus with a mean error of 1.80 mm and a standard deviation of 1.17 mm. The data used for evaluation also contains difficult and extreme cases. Here, our method occasionally failed to properly find the esophagus boundary. If the 20% most difficult cases are excluded (PB), the mean error was 1.34 mm.

The surface refinement step has a significant impact on the accuracy: If it is omitted (NS), the error raises to 2.24 mm.

To evaluate the effect of the soft probability map $A(\boldsymbol{t})$, we measured the accuracy when only the binary air model $B(\boldsymbol{t})$ is used (B). The resulting error is 1.88 mm, meaning that $A(\boldsymbol{t})$ improves the accuracy by 4.3%. If also $B(\boldsymbol{t})$ is omitted (NA), the error is 1.94 mm, which means that modelling the distribution of air explicitly leads to an improvement of 7.2%. Using a second classifier for translation improves performance by 13%: If it is omitted (ST), the error raises to 2.07 mm. Omitting both the air model and the second translation detector (NAT) leads to an error of 2.32 mm.

In addition to a Markov order of one, we measured the error for orders of zero (M0) and two (M2). An order of zero means the detected ellipses in neighboring slices do not influence each other. The second order Markov model as described in section II-C2 also takes the curvature of the esophagus into account. A Markov order of zero yields an mean error of 2.30 mm, which shows that the Markov model clearly improves detection performance by resolving ambiguities. A Markov order of two does not further improve the performance. Therefore, we propose to use an order of one as it does not introduce unnecessary complexity.

With a mean error of 5.39 mm, the particle filtering (PF) approach described in section II-C3 performs poorly. The resulting segmentation was usually completely off the true esophagus. We observed that particle filtering is much more prone to tracking loss compared to the Markov chain based "detect and connect" approach. The reason is that in the "detect and connect" approach, each slice is searched exhaustively in the position detection step, while the particle filter does not search exhaustively but only evaluates the particles. Occasionally, the esophagus is hard to see on a number of

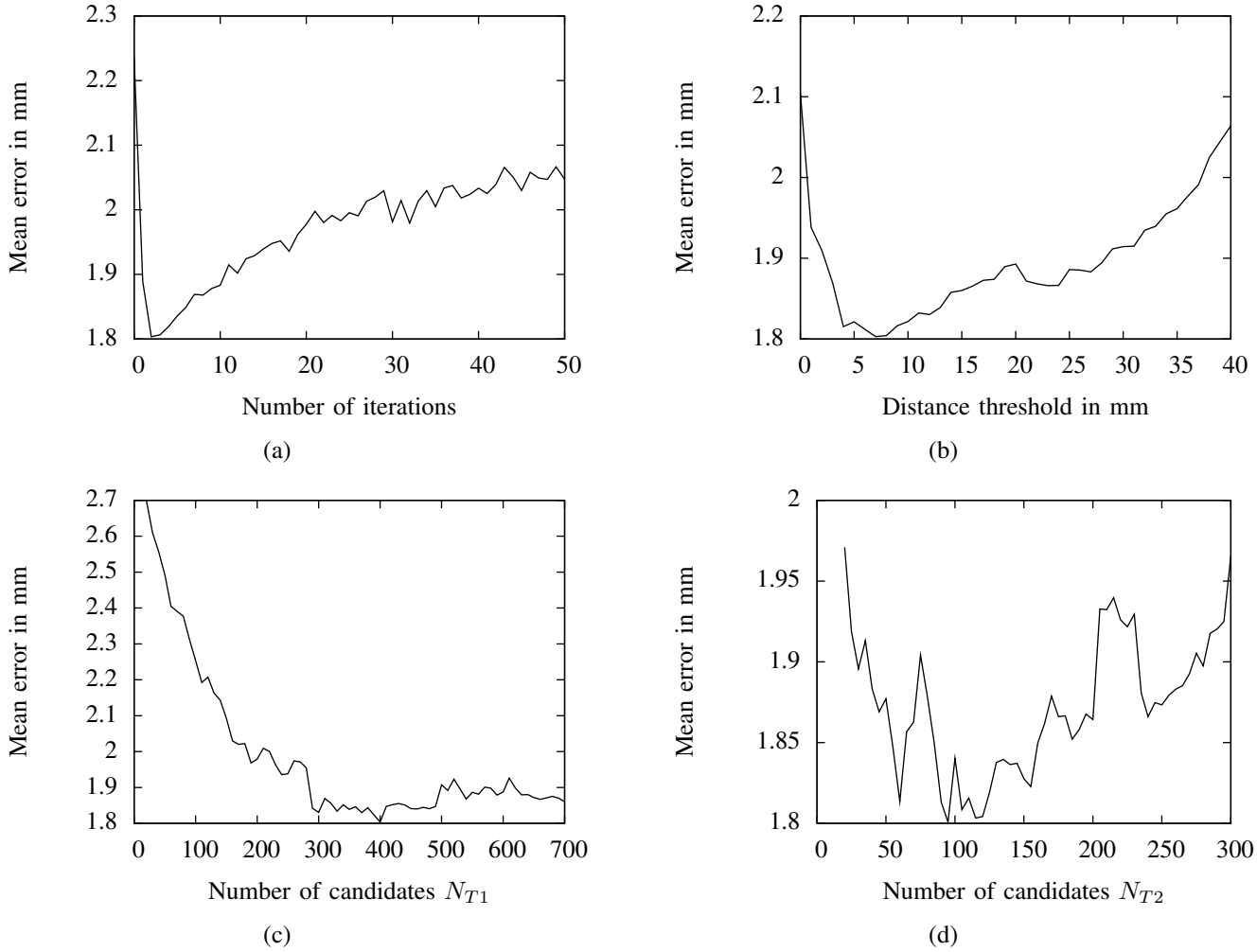


Fig. 11. Mean segmentation error for different values of the number of surface refinement iterations (a), the distance threshold used for clustering (b), the number N_{T1} of translation candidates of the first translation detector (c) and the number N_{T2} of translation candidates of the second translation detector.

slices. The PF often follows another structure, e.g. a vessel, and no particles remain on the true esophagus. Once the PF loses track, it usually does not recover. We observed that the "detect and connect" approach recovers much better after passing a difficult section.

In order to compare the performance of the detector to the performance of a human, we did an experiment on the inter observer variability (IOV): In ten datasets, the esophagus was manually segmented a second time by another person, and the second segmentations were treated like automatic ones. The mean error was 0.78 mm with a standard deviation of 0.17 mm.

Next, we evaluated how the presence of contrast agent in the esophagus affects the detector performance. For thorax-abdomen scans, patients usually drink 1.5 l of contrast agent over a 60 minutes period in preparation for the scan in order to contrast the digestive system, especially the intestine that is hard to see otherwise.

In 119 out of the 144 datasets, we did not see remains of orally given contrast agent (CA) inside the esophagus. In 19 datasets, the esophagus contains only small amounts of CA, and also only in some sections of the esophagus. Visually, the CA hardly makes a difference. In 6 datasets, the esophagus is filled with large amounts of contrast agent. The diameter of the esophagus is greatly increased. These 6 cases look very different from the remaining 138.

We don't have the medical findings of our images and therefore we neither do know from what a patient suffered and why s/he was scanned nor details about the examination. But in these 6 cases most likely 0.5 l of CA were administered when the patient was already lying on the table, directly before s/he was scanned. In combination with given Butylscopolamine, larger amounts of CA remain in the esophagus. This is typically done to contrast the upper digestive system and the esophagus itself. Some of the 6 patients may have suffered from achalasia, which means that the lower part of the esophagus does not properly open. It is also possible that parts of the esophagus were resected due to cancer and replaced with intestine tissue.

We refer to the 119 datasets as "not contrasted" (NC), to the 19 datasets as "hardly contrasted" (HC) and to the 6 datasets as "contrasted and dilated" (CD). The results can be found in TABLE V. The performance of our proposed method on the "not contrasted" and "hardly contrasted" datasets is similar

Subgroup		num. datasets	mean err. in mm	Hausdorff dist. in mm
NC	Not contrasted	119	1.62 ± 0.84	11.73 ± 6.44
HC	Hardly contrasted	19	1.69 ± 0.68	12.00 ± 5.33
CD	Contr. and dilated	6	5.75 ± 1.02	28.53 ± 1.88

TABLE V
DETECTOR PERFORMANCE DEPENDING ON WHETHER ORALLY GIVEN CONTRAST AGENT IS VISIBLE IN THE ESOPHAGUS.

and better than the 1.80 mm average segmentation error. On the six "contrasted and dilated" datasets, the performance is much worse. The esophagus is greatly increased in diameter in these cases and mostly bigger than the aorta. Furthermore, the contrast agent leads to a unusual appearance. Here, the automatic segmentation did not cover the whole cross section of the esophagus.

Generally, machine learning methods often have problems with rare extreme cases. We think that cases like these can be handled in principle by our method as long as there are enough examples in the training data, but we did not empirically verify this.

We also evaluated how the mean error depends on different parameter settings. The results can be seen in Fig. 11. When the number of surface refinement iterations is varied (a), at first the error steeply drops, reaches an optimum after 2-3 iterations, and rises again. We therefore kept the number of iterations fixed to two, which gives not only good results but is also computationally efficient. Fig. 11 (b) shows results for different values of the distance threshold d_{max} used for clustering. If this value is too low, the number of clusters K will be high, and for $d_{max} = 0$ equal to the number N of ellipse candidates. Then, clustering is unable to find modes in the distribution of the ellipse candidates. On the other hand, if d_{max} is too high, most or even all candidates fall into the same cluster. In this case, there are no different hypothesis about the esophagus contour in a slice any more, and the Markov model becomes ineffective. Values between 6 mm and 8 mm performed best. In (c), the number of candidates N_{T1} generated by the first translation detector is varied. Selecting a too low value introduces the risk of missing the true esophagus, while a very high value means that many false alarms are propagated to further levels of the detection cascade. A value of 400 is a reasonable choice. The number of candidates N_{T2} generated by the second translation detector (d) must be considerably lower than N_{T1}, otherwise the stage could be omitted. Low values are also computationally less expensive because less candidates have to be examined in the later stages of the cascade. But again a too low value bears the risk of losing the true esophagus. Here, a value of 120 was chosen.

B. Results on further datasets

We furthermore evaluated our method on other image databases that are listed in TABLE VI together with databases used for evaluation in [5], [4], [3]. The evaluation results are shown in TABLE VII.

The first row of TABLE VII shows the results of cross-validation of our method on the 144 datasets, referred to as "thick slice" data, as described above. Rows 2-4 show the performance of our method on three further test databases. In these three experiments, our model was trained on all 144 volume images from the "thick slice" database. We did not train on other databases because it is hard to obtain a sufficient number of ground truth segmentations.

Our 144 datasets all have a slice spacing of 5 mm. To obtain results on thin slice data, we evaluated our method on ten high-resolution datasets with a slice thickness in the range of 0.5-0.8 mm (second row of TABLE VII). An expert-reviewed ground truth segmentation was available for each of the ten datasets. With a mean segmentation error of 2.76 mm, the performance is considerably worse than the cross-validation error. One reason is an outlier case with a very high segmentation error caused by an air pocket of the lung that distracted the detector. The images are also noisier, and no thin slice data was included in the training set.

Next, our method was evaluated on publicly available data. Since we are not aware of a public chest CT database with images optimized for soft tissue, we selected a set of images from the database of the Lung Image Database Consortium (LIDC)[1]. We used all 28 volume images of the LIDC database that show the thorax and have a slice thickness in the range of 3-5 mm except one. This was excluded because of a rotated coordinate system, which is currently not handled by our method. In order to reduce the effort for the manual annotation, the data was not completely segmented. Instead, the contour of the esophagus was manually drawn in six cross-sectional slices and reviewed by an expert. Our method's mean error of 1.36 mm on this data is even better than the results of cross-validation. One reason is that there are no extreme cases among the datasets.

Finally, we were able to evaluate our method on a superset of the data that was used for evaluation in [5]. We use a superset because it is not known on which of the images the method of [5] was evaluated on. We also did not have the original ground truth data and manually annotated the esophagus contour in four axial slices in each volume image. Only four instead of six slices were annotated because the images cover a shorter segment of the esophagus. These annotations were reviewed by an expert as well. Even though the data is a superset and the ground truth may be slightly different, it still allows a relatively fair comparison. Our method achieved a mean segmentation error of 1.82 mm and a Dice coefficient of 0.72. In [5], a Dice coefficient in the range of 0.60-0.84 is reported. The result depends on whether the user draws two, three or five contours manually into axial slices. While the result of [5] is better if five contours are manually annotated, our automatic method outperforms [5] with two manually drawn contours.

For comparison, the results stated in [4] and [3] are shown in rows six and seven of TABLE VII. Our method performed considerably better than the mean error of 2.6 mm reported in [4] on three of four databases it was evaluated on. The Dice coefficient reported in [3] is better than ours. However, the comparability is limited because the methods of [5], [4], [3]

[1] https://imaging.nci.nih.gov

Database	num. datasets	slice thickness in mm	covered body region	manual segmentation type
Thick slice	144	5	thorax or thorax-abdomen	full
Thin slice	10	0.5-0.8	thorax or thorax-abdomen	full
LIDC	27	3-5	thorax	slice
Superset of Fieselmann et al. [5]	36	0.6-1.5	heart	slice
Fieselmann et al. [5]	8	0.6-1.5	heart	full
Kurugol et al. [4]	8	3.75	thorax	full
Rousson et al. [3]	20	unknown	heart	full

TABLE VI
ROWS 1-4: DATASETS USED FOR EVALUATION IN THIS PAPER. ROWS 5-8: DATASETS USED FOR EVALUATION IN PRIOR WORK.

Method	fully automatic	test data	training data	cross-validation	mean err. in mm	Hausdorff dist. in mm	Dice coeff.
Proposed method	yes	thick slice	thick slice	yes	1.80 ± 1.17	12.62 ± 7.01	0.74 ± 0.14
Proposed method	yes	thin slice	thick slice	no	2.76 ± 2.76	14.91 ± 15.47	0.67 ± 0.21
Proposed method	yes	LIDC	thick slice	no	1.36 ± 0.44	7.19 ± 2.78	0.73 ± 0.08
Proposed method	yes	superset of [5]	thick slice	no	1.82 ± 1.27	9.64 ± 6.24	0.72 ± 0.11
Fieselmann et al. [5]	no	Fieselmann et al. [5]	n/a	n/a	unknown	unknown	0.60-0.84*
Kurugol et al. [4]	no	Kurugol et al. [4]	Kurugol et al. [4]	yes	2.6± 2.1	unknown	unknown
Rousson et al. [3]	no	Rousson et al. [3]	Rousson et al. [3]	yes	unknown	unknown	0.80

TABLE VII
PERFORMANCE ON OTHER DATASETS AND COMPARISON WITH OTHER METHODS. *: DEPENDING ON THE AMOUNT OF USER INTERACTION.

ROI detec.	prob. map generation	ellipse detec.	path inference	refinement	**total**
6.96	1.13	7.40	$0.40 \cdot 10^{-3}$	0.34	**15.83**

TABLE VIII
COMPUTATION TIME IN SECONDS FOR DIFFERENT STEPS OF THE METHOD.

all require user interaction, while our method does not. In [5], [3], the focus was on the section of the esophagus close to the left atrium, which is relatively short. Given two points on the centerline as used in [3], the trivial centerline estimate, which is the linear interpolation of the points, can already be close to the true centerline.

C. Examples and computational requirements

Fig. 12 shows examples of segmentation results of the proposed method in blue along with the corresponding ground truth in green. Axial cross sections of two volumes are shown in (a) and (b). The yellow boxes show the result of the path inference step. They are the tight bounding boxes of the ellipses which approximate the esophagus contour. In (c), four example segmentations are displayed in 3-D. These 3-D images also visualize the size of the ROI. All shown datasets were not included in the training data.

TABLE VIII shows the computational requirements of the different steps of the proposed method. It was measured on a single CT scan of the entire torso on a standard PC with a 2.20 GHz dual core CPU. With 6.96 s, ROI detection is the second most time consuming step because the whole volume is searched exhaustively. The remaining steps only consider the ROI. Ellipse contour candidate detection including clustering takes 7.40 s and is the most time consuming step. Generating the probability map based on air and surface refinement is comparatively inexpensive, and the time needed for path inference is negligible. In total, segmenting the esophagus from a CT volume takes less than 16 s.

IV. CONCLUSION

We have presented a fully automatic method for esophagus segmentation from CT scans. An ROI is detected by finding salient anatomical landmarks. A powerful detector that learned a discriminative model of the appearance and an explicit model of the distribution of air is combined with prior knowledge about the esophagus shape. It is used to infer the approximate contour of the esophagus by finding the maximum a posteriori estimate. Two alternative methods for shape knowledge representation and inference are compared: A "detect and connect" approach using a Markov chain model, and a particle filter. Finally, a surface is generated and further adapted to better fit the boundary, again using discriminative learning.

The accuracy was measured using cross validation on 144 datasets. We found that the Markov chain based "detect and connect" approach can well handle difficult regions and resolve ambiguities. It clearly outperformed the particle filter, which is much more prone to tracking loss. Explicitly modelling respiratory and esophageal air to support the appearance based detector improves the mean error by 7.2%. Our proposed method segments the esophagus from a CT scan without user interaction with a mean error of 1.80 mm in less than 16 s, which is only 1 mm above the inter observer variability.

Apart from the ROI detection and explicitly modelling air, the method is not specific to the esophagus and can easily be adapted to other tubular structures like the spinal chord or larger vessels.

ACKNOWLEDGMENTS

This work originated in the German research project THESEUS Medico that was initiated and is financially supported by the German Federal Ministry of Economics and Technology.

The authors would furthermore like to thank Andreas Fieselmann for sharing the image data he used for evaluation in [5]. This helped to improve the comparability.

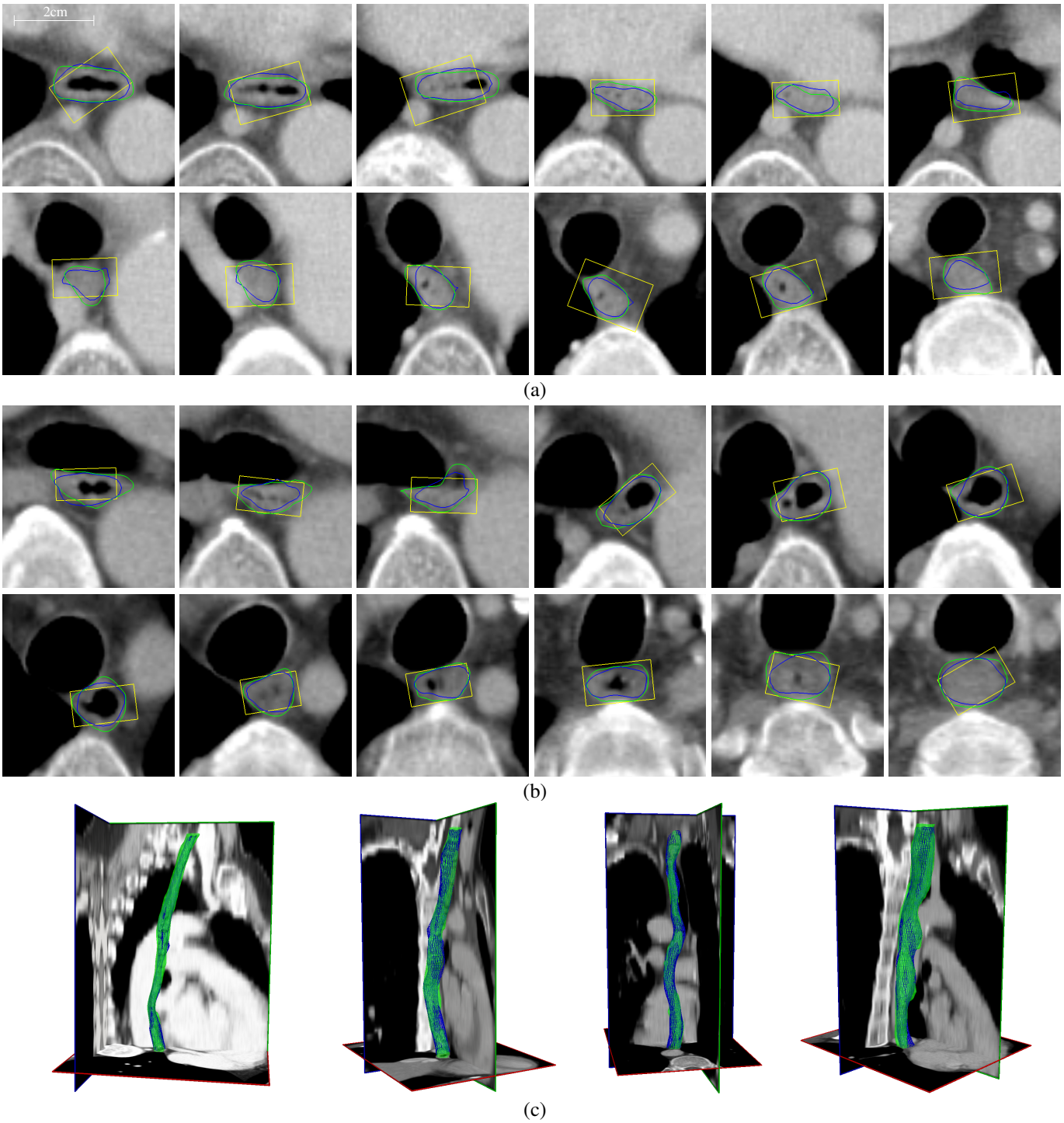


Fig. 12. Examples of segmentation results on unseen data. Axial slices are shown for two datasets (a), (b), and 3-D visualizations for four datasets (c). Blue is the automatic segmentation, green is the ground truth, and the yellow boxes show the inferred path. The mean errors of the segmentations are 0.95 mm (a), 0.88 mm (b). In subfigure (c) from left to right: 0.95 mm (same datasets as (a)), 1.15 mm, 0.99 mm and 0.95 mm. The bar in the top left slice indicates the scale.

REFERENCES

[1] B. V. Duwe, D. H. Sterman, and A. I. Musani, "Tumors of the mediastinum," *Chest*, vol. 128, no. 4, pp. 2893–2909, 2005. 1

[2] C. Pappone, H. Oral, V. Santinelli, G. Vicedomini, C. C. Lang, F. Manguso, L. Torracca, S. Benussi, O. Alfieri, R. Hong, W. Lau, K. Hirata, N. Shikuma, B. Hall, and F. Morady, "Atrio-Esophageal Fistula as a Complication of Percutaneous Transcatheter Ablation of Atrial Fibrillation," *Circulation*, vol. 109, no. 22, pp. 2724–2726, 2004. [Online]. Available: http://circ.ahajournals.org/cgi/content/abstract/109/22/2724 1

[3] M. Rousson, Y. Bai, C. Xu, and F. Sauer, "Probabilistic minimal path for automated esophagus segmentation," *Proceedings of the SPIE*, vol. 6144, pp. 1361–1369, 2006. 1, 11, 12

[4] S. Kurugol, N. Ozay, J. G. Dy, G. C. Sharp, and D. H. Brooks, "Locally deformable shape model to improve 3d level set based esophagus segmentation," *Pattern Recognition, International Conference on*, vol. 0, pp. 3955–3958, 2010. 1, 11, 12

[5] A. Fieselmann, S. Lautenschläger, F. Deinzer, M. John, and B. Poppe, "Esophagus Segmentation by Spatially-Constrained Shape Interpolation," *Bildverarbeitung für die Medizin*, pp. 247–+, 2008. 1, 11, 12

[6] T.-C. Huang, G. Zhang, T. Guerrero, G. Starkschall, K.-P. Lin, and K. Forster, "Semi-automated ct segmentation using optic flow and fourier interpolation techniques," *Comput. Methods Prog. Biomed.*, vol. 84, no. 2-3, pp. 124–134, 2006. 1

[7] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *CVPR*, vol. 1, p. 511, 2001. 1, 3

[8] Z. Tu, "Probabilistic boosting-tree: learning discriminative models for classification, recognition, and clustering," *ICCV*, vol. 2, pp. 1589–1596, 2005. 1, 3

[9] Y. Zheng, A. Barbu, B. Georgescu, M. Scheuering, and D. Comaniciu, "Fast automatic heart chamber segmentation from 3d ct data using marginal space learning and steerable features," *ICCV*, pp. 1–8, 2007. 1, 3, 7

[10] J. Feulner, S. K. Zhou, A. Cavallaro, S. Seifert, J. Hornegger, and D. Comaniciu, "Fast Automatic Segmentation of the Esophagus from 3D CT Data Using a Probabilistic Model," in *Proceedings of the 12th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2009), Part I, LNCS 5761*, ser. Lecture Notes on Computer Science 5761, Berlin, 2009, pp. 255–262. 1, 2

[11] J. Feulner, S. K. Zhou, M. Huber, A. Cavallaro, J. Hornegger, and D. Comaniciu, "Model-based esophagus segmentation from ct scans using a spatial probability map," in *MICCAI 2010, Lecture Notes in Computer Science, LNCS*, vol. 6361, no. 1, 2010, pp. 95–102. 2, 3

[12] S. Seifert, A. Barbu, S. K. Zhou, D. Liu, J. Feulner, M. Huber, M. Suehling, A. Cavallaro, and D. Comaniciu, "Hierarchical parsing and semantic navigation of full body ct data," in *Medical Imaging 2009: Image Processing*, J. P. W. Pluim and B. M. Dawant, Eds., vol. 7259, no. 1. SPIE, 2009, p. 725902. 2

[13] A. Fieselmann, S. Lautenschläger, F. Deinzer, and B. Poppe, "Automatic Detection of Air Holes Inside the Esophagus in CT Images," in *Bildverarbeitung für die Medizin 2008, Informatik aktuell, Volume . ISBN 978-3-540-78639-9. Springer Berlin Heidelberg, 2008*, 2008, pp. 397–401. 3

[14] R. Kindermann and J. L. Snell, *Markov Random Fields and Their Applications*. AMS, 1980. 4

[15] F. Kschischang, B. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, 2001. 4

[16] M. Fitzgibbon, A. W.and Pilu and R. B. Fisher, "Direct least-squares fitting of ellipses," *PAMI*, vol. 21, no. 5, pp. 476–480, May 1999. 5

[17] H. Burkhardt and B. Neumann, Eds., *A Smoothing Filter for CONDENSATION*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 1998, vol. 1406, pp. 767–781. [Online]. Available: http://dx.doi.org/ 6, 7

[18] C. Florin, N. Paragios, and J. Williams, "Particle filters, a quasi-monte-carlo-solution for segmentation of coronaries," *MICCAI*, pp. 246–253, 2005. 6

[19] M. Schaap, I. Smal, C. Metz, T. van Walsum, and W. Niessen, "Bayesian tracking of elongated structures in 3d images," *International Conference on Information Processing in Medical Imaging, IPMI*, 2007. 6

[20] M. Isard and A. Blake, "Condensation: Conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, pp. 5–28, 1998. 6, 7