Robust Segmentation of Various Anatomies in 3D Ultrasound Using Hough Forests and Learned Data Representations

Fausto Milletari[1], Seyed-Ahmad Ahmadi[3], Christine Kroll[1], Christoph Hennersperger[1], Federico Tombari[1], Amit Shah[1], Annika Plate[3], Kai Boetzel[3], and Nassir Navab[1,2]

[1] Computer Aided Medical Procedures, Technische Universität München, Germany
[2] Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA
[3] Department of Neurology, Klinikum Grosshadern, Ludwig-Maximilians-Universität München, Germany

Abstract. 3D ultrasound segmentation is a challenging task due to image artefacts, low signal-to-noise ratio and lack of contrast at anatomical boundaries. Current solutions usually rely on complex, anatomy-specific regularization methods to improve segmentation accuracy. In this work, we propose a highly adaptive learning-based method for fully automatic segmentation of ultrasound volumes. During training, anatomy-specific features are obtained through a sparse auto-encoder. The extracted features are employed in a Hough Forest based framework to retrieve the position of the target anatomy and its segmentation contour. The resulting method is fully automatic, i.e. it does not require any human interaction, and can robustly and automatically adapt to different anatomies yet enforcing appearance and shape constraints. We demonstrate the performance of the method for three different applications: segmentation of midbrain, left ventricle of the heart and prostate.

1 Introduction and Related Work

Manual segmentation of ultrasound volumes is tedious, time consuming and subjective. In the attempt to produce results that are invariant to the presence of noise, drop-out regions and poorly distinguishable boundaries, current computer-aided approaches either use complex cost functions, often regularized by statistical prior models, or require extensive user interaction. Many optimization-based methods utilize cost functions based on local gradients, texture, region intensities or speckle statistics [10]. Methods employing shape and appearance models often require a difficult and time-consuming training stage where the annotated data must be carefully aligned to establish correspondence across shapes in order to ensure the correctness of the extracted statistics. Learning approaches have been successfully proposed to solve localization and segmentation tasks both in computer vision [7,12] and medical image analysis [6]. Handcrafted features which exhibit robustness towards the presence of noise and artefacts have been often

© Springer International Publishing Switzerland 2015
N. Navab et al. (Eds.): MICCAI 2015, Part II, LNCS 9350, pp. 111–118, 2015.
DOI: 10.1007/978-3-319-24571-3_14

employed to deliver automatic segmentations [8]. Recent work [3] in the machine learning community focused on approaches leveraging single [5] or multi-layer [9] auto-encoders to discover features from large amount of data. In particular, sparse auto-encoders with a single-layer have been proven to learn more discriminative features compared to multi-layer ones [5] when a sufficiently large number of hidden units is chosen. In the medical community, recent approaches [4] have employed deep neural networks to solve segmentation tasks, despite their computational burden due to the presence of cascaded 3D convolutions when dealing with volumes.

The segmentation method proposed in this paper is (i) fully automatic, (ii) highly adaptive to different kinds of anatomies, and (iii) capable of enforcing shape and appearance constraints to ensure sufficient robustness. A sparse auto-encoder is trained from a set of ultrasound volumes in order to create a bank of 3D features, which are specific and discriminative to the anatomy at hand. Through a voting strategy, the position of the region to be segmented is assessed and a contour is obtained by patch-wise projection of appropriate portions of a multi-atlas. Differently from [12], each contribution to the contour is weighted by a factor dependent on the appearance pattern of the region that it was collected from. In this way, we effectively enforce shape *and* appearance constraints.

We demonstrate the performance of our method by segmenting three different and challenging anatomies: left ventricle of the heart, prostate and midbrain. The experiments show that our approach is competitive compared to state-of-the-art anatomy specific methods and that, in most cases, the quality of our segmentations lies within the expected inter-expert variability for the particular dataset.

2 Method

Our approach comprises a training and a test phase. During training, we discover anatomy-specific features that are employed to learn a Hough Forest. During testing, we perform simultaneous object localization and segmentation.

2.1 Feature Learning

Sparse Auto-Encoders are feed-forward neural networks designed to produce close approximations of the input signals as output (Fig. 1 - a). By employing a limited number of neurons in the hidden layer and imposing a sparsity constraint through the Kullback-Leibler (KL) divergence, the network is forced to learn a sparse lower-dimensional representation of the training signals [5,9].

The network has N inputs, K neurons in the hidden layer and N outputs. The biases $b_i^{(1,2)}$ are integrated in the network through the presence of two additional neurons in the input and hidden layer having a constant value of 1. The weights of the connections between the j-th neuron in one layer and the i-th neuron in the next are represented by $w_{ij}^{(1,2)} \in \mathbb{R}$, that are grouped in the matrices $\mathbf{W}_1 \in \mathbb{R}^{K\times(N+1)}$ and $\mathbf{W}_2^\top \in \mathbb{R}^{N\times(K+1)}$. Network outputs can

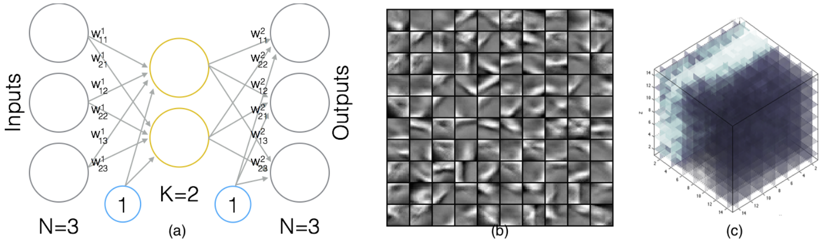

Fig. 1. a) Schematic Illustration of a Sparse Auto-Encoder (SAE); b) Bank of filter obtained from 2D ultrasound images of the midbrain; c) One filter obtained from 3D echocardiographical data through the SAE.

be written as $h_{\mathbf{W}^{(1,2)}}(\mathbf{X}) = f\left(\mathbf{W}_2^\top f\left(\mathbf{W}_1\mathbf{X}\right)\right)$, where $f(z) = \frac{1}{1+exp(-z)}$ is the sigmoid activation function.

The matrix $\mathbf{X}$ is filled with M unlabeled ultrasound training patches arranged column-wise. After a normalization step to compensate for illumination variations through the dataset, the network is trained via back-propagation. The network weights are initialized with random values. The objective function to be minimized comprises of three terms, enforcing the fidelity of the reconstructions, small weights magnitude and sparsity respectively:

$$C(\mathbf{X}, \mathbf{W}_{1,2}) = \frac{1}{2}\left\|h_{\mathbf{W}^{(\mathbf{1},\mathbf{2})}}(\mathbf{X}) - \mathbf{X}\right\|^2 + \frac{\lambda}{2}\sum_{l=1}^{2}\sum_{k=1}^{K}\sum_{n=1}^{N}\left(w_{nk}^{(l)}\right)^2 + \beta\sum_{j=1}^{K} KL(\rho\|\rho_j). \quad (1)$$

In the third term, we indicate as $\rho_j = \frac{1}{M}\left(\mathbf{1}^\top f\left(\mathbf{W}_1\mathbf{X}\right)\right)$ the average firing rate of the j-th hidden neuron, and we define the KL divergence, which enforces the sparsity constraint by penalizing deviations of ρ_j from ρ, as:

$$KL(\rho\|\rho_j) = \rho\, log\left(\frac{\rho}{\rho_j}\right) + (1-\rho) log\left(\frac{1-\rho}{1-\rho_j}\right). \quad (2)$$

The parameter ρ represents the desired firing rate of the neurons of the hidden layer, and must be set prior to training together with λ, β and K, which control the weight of the two regularization terms and the number of neurons of the hidden layer respectively.

After optimization, the rows of the weight matrix $\mathbf{W}_1$ can be re-arranged to form a set of 3D filters $\varXi = \{\xi_1 ... \xi_K\}$ having the same size as the ultrasound patches collected during training (Fig. 1 - b,c).

2.2 Training the Hough Forest

Our implementation of Hough Forests (HF) combines the classification performance of Random Forests (RF) with the capability of carrying out organ localization and segmentation. Differently from the classical Hough Forest framework [7], our method retrieves segmentations enforcing shape and appearance constraints.

We consider a training set composed of N data samples, $\mathbf{d}_{1...N}$, where each sample $\mathbf{d}_i = [d_x, d_y, d_z]^\top$ corresponds to a voxel element of an annotated volume V_t belonging to the training set T. Each data point is described by K features $F_{1...K}$, which are computed by applying one of the filters ξ from the set Ξ obtained via Sparse Auto-Encoders as described in the previous step. Specifically, we write

$$F_k(\mathbf{d}) = \sum_{i=-r_x}^{r_x} \sum_{j=-r_y}^{r_y} \sum_{k=-r_z}^{r_z} V_t(d_x + i, d_y + j, d_z + k) * \xi(i, j, k). \tag{3}$$

The annotation G_t, obtained in the form of a 3D binary mask associated with the volume V_t, determines the binary labels $l_i = \{f, b\}$ that characterize each data-point as belonging to the foreground or to the background. Foreground data-points are associated with a vote $\mathbf{v}_i = \mathbf{c}_t - \mathbf{d}_i$, which is expressed as a displacement vector connecting $\mathbf{d}_i$ to the centroid of the annotated anatomy $\mathbf{c}_t = [c_x, c_y, c_z]^\top$, obtained from G_t.

During training, the best binary decision is selected in each node of the Hough Forest, either maximizing the Information Gain (IG) or maximizing the Vote Uniformity (VU). In each node, we compute M random features and we determine S candidate splits through the thresholds $\tau_{1...S}$. Each split determines a partitioning of the data D_p reaching the parent node in two subsets $D_l = \{\mathbf{d}_i \in D_p : F_k(\mathbf{d}_i) \leq \tau_s\}$ and $D_r = \{\mathbf{d}_i \in D_p : F_k(\mathbf{d}_i) > \tau_s\}$ reaching the left and right child nodes, respectively. The Information Gain is obtained as:

$$IG(D_p, D_l, D_r) = H(D_p) - \sum_{i \in \{l,r\}} \frac{|D_i|}{|\bar{D}|} H(D_i), \tag{4}$$

where the Shannon entropy $H(D) = \sum_{c \in \{f,b\}} -p_c log\,(p_c)$ is obtained through the empirical probability $p_c = \frac{|D_c|}{|D|}$ using $D_c = \{d_i \in D : l_i = c\}$.

The Vote Uniformity criterion requires the votes $\mathbf{v}_j^{\{l,r\}}$ contained in D_l and D_r to be optimally clustered around their respective means $\bar{\mathbf{v}}^{\{l,r\}}$:

$$VU(D_l, D_r) = \sum_{i \in \{l,r\}} \sum_{\mathbf{v}_j} \left\| \mathbf{v}_j^i - \bar{\mathbf{v}}^i \right\|. \tag{5}$$

Once (i) the maximum tree depth has been reached or (ii) the number of data points reaching the node is below a certain threshold or (iii) the Information Gain is zero, the recursion terminates and a leaf is instantiated. The proportion of foreground versus background points $p_{\{f,b\}}$ is stored together with the votes $\mathbf{v}_i$ and the associated original positions $\mathbf{d}_i$. The coordinates $\mathbf{d}_i$, in particular, refer to training volumes which will be used as atlases during segmentation.

2.3 Segmentation via Hough Forests

Given an ultrasound volume I of the test set, we first classify it into foreground and background, then we allow foreground data-points to cast votes in order to

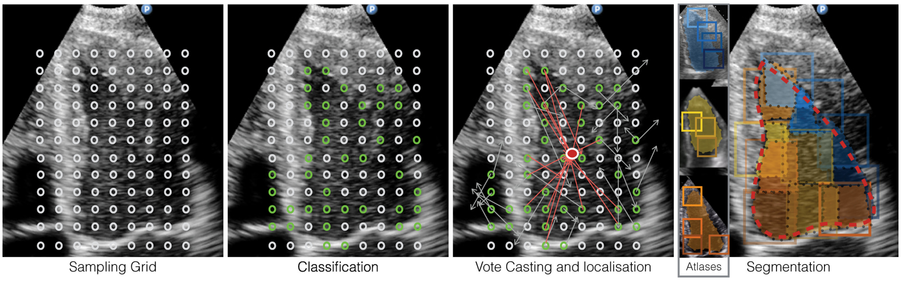

Fig. 2. Schematic representation of our segmentation approach shown in 2D.

localize the target anatomy, and finally, we obtain the contour by projecting 3D segmentation patches from the atlases associated with each vote that correctly contributed to localization (Fig. 2).

The data-points processed in the Hough Forest are obtained through a regular grid of sampling coordinates $S = \{\mathbf{s}_1 ... \mathbf{s}_{N_d}\}$. In this way, we can reduce the computational load during testing without significantly deteriorating the results. Each data-point $\mathbf{s}_i$ classified as foreground in a specific leaf l of the Hough trees is allowed to cast the n_v^l votes $\mathbf{v}_{1...n_v^l}$ stored in that leaf during training.

Each vote determines a contribution, weighted by the classification confidence, at the location $\mathbf{s}_i + \mathbf{v}_j$ of a volume C having the same size as I and whose content is initially set to zero.

The target anatomy is localized retrieving the position of the highest peak in the vote map. All the votes $\hat{\mathbf{v}}_j$ falling within a radius r around the peak are traced back to the coordinates $\hat{\mathbf{s}}_i$ of the data-points that cast them. Each vote $\hat{\mathbf{v}}_j$ is associated with the coordinates $\hat{\mathbf{d}}_j$ of a specific annotated training volume.

We retrieve the 3D appearance patch A_j and the segmentation patch P_j associated to each vote by using the coordinates $\hat{\mathbf{d}}_j$ to sample the appropriate training volume and its annotation. The segmentation patches are projected at the positions $\hat{\mathbf{s}}_i$ after being weighted by the Normalized Cross Correlation (NCC) between the patch A_j and the corresponding intensity patterns around $\hat{\mathbf{s}}_i$ in the test volume. The fusion of all the reprojected segmentation patches forms the final contour, which implicitly enforces shape and appearance constraints.

3 Results

We demonstrate the segmentation accuracy and the flexibility of our algorithm using three datasets of different anatomies comprising in total 87 ultrasound volumes. A brief description of each dataset and the relative state-of-the-art segmentation approach being used for comparison is provided below.

1. The left ventricle of the heart is segmented and traced in [2] using an elliptical shape constraint and a B-Spline Explicit Active Surface model. The dataset employed for our tests, comprising 60 cases, was published during

the MICCAI 2014 "CETUS" challenge. Evaluations were performed using the MIDAS platform[1].
2. The prostate segmentation method proposed in [11] requires manual initialization. Its contour is retrieved using a min-cut formulation on intensity profiles regularized with a volumetric size-preserving prior. We test on a self-acquired trans-rectal ultrasound (TRUS) dataset comprising 15 subjects. All the volumes were manually segmented by one expert clinician via 'TurtleSeg'. Our results are obtained via cross-validation.
3. Segmentation of the midbrain in transcranial ultrasound (TCUS) is valuable for Parkinson's Disease diagnosis. In [1], the authors employed a discrete active surface method enforcing shape and local appearance constraints. We test the methods on 12 ultrasound volumes annotated by one expert using 'ITK snap' and acquired through one of the skull bones of the patients. Our results are obtained via cross-validation.

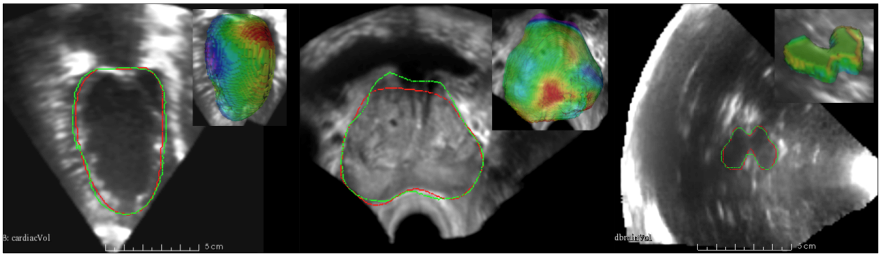

Fig. 3. Exemplary segmentation results (green curves) Vs. ground-truth (red curves). Mesh color encodes distances from ground truth in the range −3mm (red) to +3mm (blue), with green indicating perfect overlap.

Table 1 shows the performance of our method in comparison to the other state-of-the-art approaches on the three datasets. Results are expressed in terms of Dice coefficients and mean absolute distance (MAD) from ground truth annotation. Typical inter-expert annotation variability is also shown for each anatomy.

3.1 Parameters of the Model

The Sparse Auto-Encoder was trained to obtain $K = 300$ 3D filters having size $15 \times 15 \times 15$ pixels, with parameters $\lambda = 10^{-4}$, $\beta = 10$ and $\rho = 10^{-3}$. The Hough Forest includes 12 trees with at most 35 decision levels and leafs that contain at least 25 data-points. During testing, the images were uniformly sampled every 3 voxels. All the votes accumulating in a radius of 3 voxels from the object centroid were reprojected. The size of the segmentation and intensity patches employed for reprojection during segmentation was different for the three datasets due to

[1] Documentation under: `http://www.creatis.insa-lyon.fr/Challenge/CETUS`

Table 1. Overview of Dice coefficients and mean absolute distance (MAD) achieved during testing. Inter-expert-variabilities (IEV) are also reported. MAD was not provided by the authors of the algorithms used for comparison.

Dataset	Avg. our approach	MAD	Avg.state-of-the-art	IEV(Dice)
Left Ventricle	0.87 ± 0.08 *Dice*	2.90 ± 1.87 *mm*	0.89 ± 0.03 *Dice*	86.1%[2]
Prostate	0.83 ± 0.06 *Dice*	2.37 ± 0.95 *mm*	0.89 ± 0.02 *Dice*	83.8%[11]
Midbrain	0.85 ± 0.03 *Dice*	1.18 ± 0.24 *mm*	0.83 ± 0.06 *Dice*	85.0%[1]

the variable size the object of interest. Values for left ventricle, prostate and midbrain were 35 × 35 × 35, 30 × 30 × 30 and 15 × 15 × 15 pixels respectively.

Training time for the Auto-Encoder was approximately 24 hours per dataset, with 500,000 patches. The training time for the forest ranged from 20 minutes to 5 hours. The processing time during testing was always below 40sec. per volume.

3.2 Experimental Evaluation

In Fig. 4 we show the histogram of Dice scores observed during our tests. Its resolution is 0.05 Dice. Additional results can be found in Table 1 and Fig. 3.

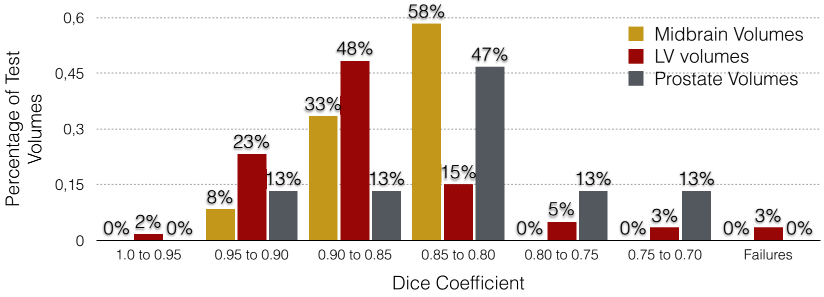

Fig. 4. Percentage of test volumes vs. Dice coefficient. This histogram shows the percentage of test volumes falling in each Dice bin on the horizontal axis.

3.3 Discussion

Localization of the target anatomy through a voting strategy, removes the need for user interaction while being very efficient in rejecting false positive datapoints, whose votes could not accumulate in the vicinity of the true anatomy centroid. During our tests, only one out of 87 localizations failed, resulting in a wrong contour. A trade-off between appearance and shape constraints can be set choosing the size of the segmentation patches. Bigger patches force smoother contours, while smaller ones lead to more adaptation to local volume contents. The method is not suited for segmentation of elongated structures (eg. vessels).

4 Conclusion and Acknowledgement

In this work, we presented a learning-based method for fully automatic localization and segmentation of various anatomies in 3D ultrasound. The method learns an optimal representation of the data and implicitly encodes a prior on shape and appearance. We apply the method on three clinical 3D ultrasound datasets of challenging anatomies, with results comparable to the state-of-the-art. This work was founded by DFG BO 1895/4-1 and ICT FP7 270460 (ACTIVE).

References

1. Ahmadi, S.A., Baust, M., Karamalis, A., Plate, A., Boetzel, K., Klein, T., Navab, N.: Midbrain segmentation in transcranial 3D ultrasound for Parkinson diagnosis. Med. Image Comput. Comput. Assist. Interv. 14(Pt 3), 362–369 (2011)
2. Barbosa, D., Heyde, B., Dietenbeck, T., Houle, H., Friboulet, D., Bernard, O., D'hooge, J.: Quantification of left ventricular volume and global function using a fast automated segmentation tool: validation in a clinical setting. Int. J.Cardiovasc. Imaging 29(2), 309–316 (2013)
3. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(8), 1798–1828 (2013)
4. Carneiro, G., Nascimento, J., Freitas, A.: Robust left ventricle segmentation from ultrasound data using deep neural networks and efficient search methods. In: 2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1085–1088. IEEE (2010)
5. Coates, A., Ng, A.Y., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: International Conference on Artificial Intelligence and Statistics, pp. 215–223 (2011)
6. Donner, R., Menze, B.H., Bischof, H., Langs, G.: Global localization of 3D anatomical structures by pre-filtered hough forests and discrete optimization. Medical Image Analysis 17(8), 1304–1314 (2013)
7. Gall, J., Yao, A., Razavi, N., Van Gool, L., Lempitsky, V.: Hough forests for object detection, tracking, and action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(11), 2188–2202 (2011)
8. Ionasec, R.I., Voigt, I., Georgescu, B., Wang, Y., Houle, H., Vega-Higuera, F., Navab, N., Comaniciu, D.: Patient-specific modeling and quantification of the aortic and mitral valves from 4-D cardiac CT and TEE. IEEE Trans. Med. Imaging 29(9), 1636–1651 (2010)
9. Le, Q.V.: Building high-level features using large scale unsupervised learning. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8595–8598. IEEE (2013)
10. Noble, J.A., Boukerroui, D.: Ultrasound image segmentation: a survey. IEEE Trans. Med. Imaging 25(8), 987–1010 (2006)
11. Qiu, W., Rajchl, M., Guo, F., Sun, Y., Ukwatta, E., Fenster, A., Yuan, J.: 3D prostate TRUS segmentation using globally optimized volume-preserving prior. Med. Image Comput. Comput. Assist. Interv. 17(Pt 1), 796–803 (2014)
12. Rematas, K., Leibe, B.: Efficient object detection and segmentation with a cascaded hough forest ISM. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 966–973, November 2011