Shape-Aware Multi-Atlas Segmentation

Jennifer Alvén*, Fredrik Kahl*†, Matilda Landgren†, Viktor Larsson† and Johannes Ulén†
*Department of Signals and Systems, Chalmers University of Technology, Sweden
Email: {alven,fredrik.kahl}@chalmers.se
†Centre for Mathematical Sciences, Lund University, Sweden
Email: {fredrik,matilda,viktorl,ulen}@maths.lth.se

***Abstract*—Despite of having no explicit shape model, multi-atlas approaches to image segmentation have proved to be a top-performer for several diverse datasets and imaging modalities. In this paper, we show how one can directly incorporate shape regularization into the multi-atlas framework. Unlike traditional methods, our proposed approach does not rely on label fusion on the voxel level. Instead, each registered atlas is viewed as an estimate of the position of a shape model.**

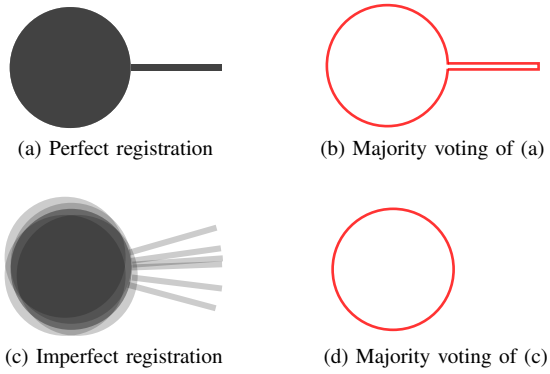
We evaluate and compare our method on two public benchmarks: (i) the VISCERAL Grand Challenge on multi-organ segmentation of whole-body CT images and (ii) the Hammers brain atlas of MR images for segmenting the hippocampus and the amygdala. For this wide spectrum of both easy and hard segmentation tasks, our experimental quantitative results are on par or better than state-of-the-art. More importantly, we obtain qualitatively better segmentation boundaries, for instance, preserving fine structures.

I. Introduction

Segmentation is an essential task in computer vision and has numerous applications. Our motivation stems from medical imaging where one is interested in obtaining accurate segmentation boundaries of organs and tissues for diagnostic and visualization purposes. In this paper, we concentrate on problems where the qualitative shape of the segmented object is of importance.

The research area can be divided into two classes of approaches. The first class uses an explicit shape model, where the prime examples are the Active Shape Model (ASM) and the Active Appearance Models (AAM) introduced by Cootes *et al.* [1], [2]. One clear advantage is that the possible solutions are constrained to lie close to a pre-defined shape space, but at the risk of losing accuracy if the shape model fails to capture all the possible object variations. The second class of approaches uses an implicit shape model. One such example is a low-order Markov Random Field model [3] where the resulting segmentation is not directly constrained by a global shape model. Another example of this class is the multi-atlas segmentation framework which has gained popularity in recent years due to its simplicity and notable performance in practical applications across several different modalities, see the recent survey [4]. This approach will be the focus of the present paper.

In multi-atlas approaches, pairwise registrations between the target image and each atlas image are computed. Thereafter, the given segmentation labels of the atlas images are transferred to the target image where the labels are combined into one segmentation proposal. The simplest fusion scheme is to perform a majority vote for each voxel (or pixel) in the target image. Several more sophisticated fusion schemes have been developed, *e.g.* based on statistical principles or machine learning [5]–[8]. In all these approaches, there is no explicit shape model or global shape regularization in the underlying model; the fusion process is performed at the voxel level and hence no awareness of the overall shape is obtained at this stage.



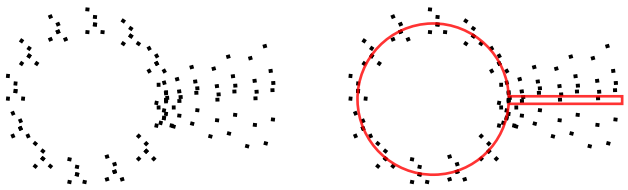
(a) Perfect registration (b) Majority voting of (a)

(c) Imperfect registration (d) Majority voting of (c)

Fig. 1. Illustration of why multi-atlas approaches may eliminate fine structures in the final segmentation when the registrations are inaccurate. (a) Result of label transfer when the atlas registrations are perfect. (b) The corresponding segmentation obtained by a majority vote. (c) Result of label transfer when there are inaccuracies in the registrations. (d) The resulting segmentation obtained via majority voting. Note that the thin structure present in the ideal segmentation (b) is lost when the registrations are not perfectly aligned (d).

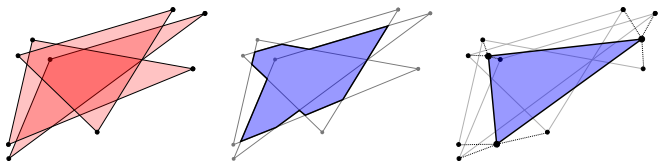
The absence of explicit shape constraints is the main drawback of multi-atlas approaches since it may yield implausible segmentation results. For example, fine structures may be lost due to the averaging of less accurate pairwise registrations, see Fig. 1, and topology may change resulting in (erroneous) disjoint organ shapes. This problem has been previously recognized in the literature and corrected for by using an explicit shape model as an additional refinement step, *e.g.* [9].

On the contrary to multi-atlas segmentation, shape-based methods have the advantage of favoring plausible shapes via constrained optimization. Typically, an intensity model is used in order to match the shape model to a target image while a statistical shape model ensures that the solution lies closely to a subspace spanned by the training examples. Notable disadvantages are the sensitivity to initialization and the dependency of correctly established one-to-one landmark correspondences. Most of these approaches rely on producing a single, accurate



(a) Transferred landmarks (b) SHAPEMAP

Fig. 2. Illustration of the SHAPEMAP-technique for the imperfect registration example in Fig. 1. (a) Each atlas registration gives a proposal of the different landmark positions (*black dots*) in the target image. (b) A robust average shape (*red curve*) is computed from these positions. Note that the thin structure is preserved despite inaccurate registrations.



(a) Triangles (b) Majority voting (c) Mesh averaging

Fig. 3. The concept behind the mesh averaging. (a) Three triangles have been registered to roughly the same position. (b) For each pixel, every triangle votes on whether the pixel is part of the object or not, resulting in an oddly shaped object. (c) The mesh points are averaged which ensures that the output is still a triangle.

guess of the initial shape position without taking advantage of the robustness and the information that several registrations may provide as in the multi-atlas approach.

A. Our approach

In this paper, we propose a framework that maintains the advantages of multi-atlas segmentation while eliminating the drawback of having a weak shape prior. It is achieved by directly incorporating a shape model into the multi-atlas segmentation framework. The core idea is that instead of voting on the voxel level, each atlas registration "votes" for the coordinates of a shape model. The shape model consists of landmark correspondences between all atlas images established in a pre-processing step. Each atlas registration provides a transferred shape and thus, an estimate of all the landmark coordinates in the target image. In contrast to ordinary multi-atlas voting which obtains a (pseudo) probability map, we obtain a map over the landmark positions. The segmentation is obtained by computing a robust average over the landmark coordinates given by the transferred shapes. Such a robust average shape estimate will be called a SHAPEMAP. In particular, the approach will be less sensitive to outlier registrations and it will ensure an overall feasible shape. An illustration is given in Fig. 2.

The main contribution of this paper is that we show how one can easily incorporate shape regularization into the multi-atlas segmentation framework. We compare our basic approach with standard majority voting using the same registration methods for both approaches on two publicly available datasets. Further, similar to many other label fusion schemes, we also perform a refinement step of the SHAPEMAP using standard techniques. We experimentally compare our results to two baselines: The winner of 13 of 20 organs in the VISCERAL Grand Challenge on multi-organ segmentation at ISBI 2015 [10] and the winner of the MICCAI 2012 Grand Challenge on multi-atlas labeling [7]. Both baselines rely on label fusion on a voxel level and refine the solution using machine learning techniques.

II. SHAPE-AWARE MULTI-ATLAS SEGMENTATION

In order to segment a new target image, each atlas image is aligned to the target image based on a standard intensity-based non-linear registration procedure (Section III-A). In addition to the intensity image, each atlas image has a manual segmentation which is considered to be the ground truth. In our system, the ground truth surfaces are represented by triangular meshes. The topology (*i.e.* the faces) of the mesh is the same for all atlases, but the coordinates of the mesh points change from atlas to atlas. Hence, the mesh points act as landmarks and these landmark correspondences are pre-computed for the atlas set (Section III-B).

A. Landmark fusion

From the pairwise registrations, we get non-rigid transformations from each atlas object into the target volume. Using these transformations we map the triangular meshes to the target image. Thereafter, the transferred meshes are combined into a single mesh by performing averaging on the mesh points. Fig. 3 shows the mesh averaging and illustrates the benefits compared to label fusion based on majority voting.

If some of the registrations were unsuccessful, then simply forming the mean shape might distort the solution. Instead we consider a *robust average* which can be thought of as the median shape. Let $\tilde{X}_{ki}$ denote the ith mesh point transferred from the kth atlas. The minimization problem we consider is then

$$\min_X \sum_{k=1}^{K}\sum_{i=1}^{N} \|X_i - \tilde{X}_{ki}\|_2 + \gamma \sum_{i=1}^{N}\sum_{j\in\mathcal{N}(i)} \|X_i - X_j\|_2^2, \quad (1)$$

where K is the number of atlases, N is the number of landmarks (for each triangular mesh), $\mathcal{N}(i)$ denotes the neighborhood (with respect to the triangular mesh) of the ith mesh point and X is the vector of concatenated mesh point coordinates X_i, $i = 1, \ldots, N$.

The first term is essentially the ℓ_1-norm of the residuals which will favor solutions close to the median point which gives robustness to outlier registrations. The second term promotes a smooth mesh grid by penalizing solutions with discontinuous jumps via the ℓ_2-norm. Note that the ℓ_1-norm would not be a good choice for the second term as it allows for jumps. The parameter γ is typically chosen small to reduce the shrinking bias of the regularizer.

To solve the minimization problem in (1) we use the Alternating Direction Method of Multipliers (ADMM) [11].

The problem in (1) can be reformulated as

$$\begin{aligned} \min_{X,Z} \; & \sum_{k=1}^{K}\sum_{i=1}^{N} \|Z_{ki}\|_2 + \gamma \sum_{i=1}^{N} \sum_{j\in\mathcal{N}(i)} \|X_i - X_j\|_2^2 \qquad (2) \\ \text{s.t.} \quad & Z_{ki} = X_i - \tilde{X}_{ki}, \end{aligned}$$

which has the augmented Lagrangian

$$\begin{aligned} L(X,Z,\Lambda) = & \sum_{k=1}^{K}\sum_{i=1}^{N} \|Z_{ki}\|_2 + \gamma \sum_{i=1}^{N} \sum_{j\in\mathcal{N}(i)} \|X_i - X_j\|_2^2 \\ & + \sum_{i=1}^{N}\sum_{k=1}^{K} \|Z_{ki} - X_i + \tilde{X}_{ki} + \Lambda_{ki}\|_2^2 - \|\Lambda_{ki}\|_2^2. \end{aligned} \qquad (3)$$

The updates in the ADMM procedure are given by

$$\begin{aligned} X^{t+1} &= \operatorname*{argmin}_{X} \; L(X, Z^t, \Lambda^t), \\ Z_{ki}^{t+1} &= \operatorname*{argmin}_{Z_{ki}} \; \|Z_{ki}\|_2 + \|Z_{ki} - X_i^{t+1} + \tilde{X}_{ki} + \Lambda_{ki}^t\|_2^2, \\ \Lambda_{ki}^{t+1} &= \Lambda_{ki}^t + Z_{ki}^{t+1} - X_i^{t+1} + \tilde{X}_{ki}. \end{aligned} \qquad (4)$$

The update in X is a simple least squares problem and the update in Z is given by the closed form expression

$$Z_{ki}^{t+1} = \begin{cases} 0 & \text{if} \quad \|p\| \leq \frac{1}{2} \\ p - \frac{1}{2}\frac{p}{\|p\|} & \text{otherwise} \end{cases}, \qquad (5)$$

where

$$p = X_i^{t+1} - \tilde{X}_{ki} - \Lambda_{ki}^t.$$

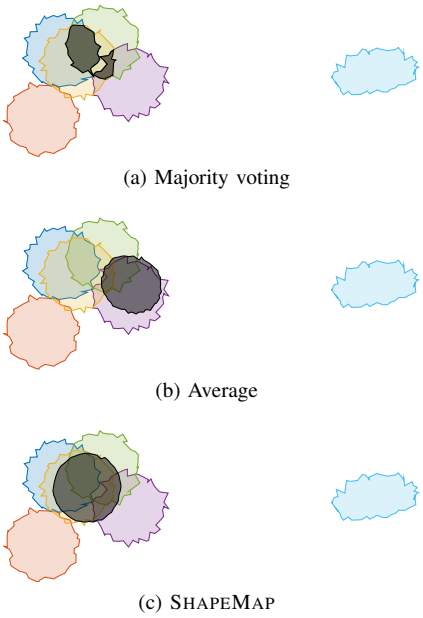
As the problem in (1) is convex, ADMM is guaranteed to converge to the global optimum [11].

1) Example: We now illustrate the potential gains with the robust averaging. In Fig. 4 we have simulated a rather poor and noisy registration of $K = 6$ circles into a two-dimensional target image. Using mean averaging results in a registration heavily influenced by the single outlier registration. The standard method, majority-vote, results in a disjoint and erroneous shape. Robust averaging, *i.e.* the solution to (1) which we refer to as a SHAPEMAP, on the other hand gives a coherent and plausible segmentation.

B. Refining the solution

The robust average solution gives us a rough idea where the object is located, however, it largely ignores the local appearance of each voxel. In order to fuse information from the shape model and the image volume itself we use a random forest classifier which will give us a likelihood estimate for each voxel belonging to the object or not. To obtain a smooth segmentation boundary, regularization is enforced by using a Markov Random Field model [3].

The random forest is applied for each voxel independently. Three types of features are calculated; (1) the signed distance to the SHAPEMAP segmentation boundary, where voxels inside the object segmentation have negative distance, (2) the intensity with and without Gaussian smoothing with standard deviations $\sigma = 1, 3$ and (3) the average of the warped, binary



(a) Majority voting

(b) Average

(c) SHAPEMAP

Fig. 4. Circles registered into a target image. The final segmentation is shown in grey. (a) Standard majority voting. This results in a disjoint and inconceivable object boundary. (b) Mean averaging of the circles' landmark positions. The resulting registration is both deformed and shifted to the right due to the outlier. (c) Robust averaging using ADMM. Note that the shape is plausible and coherent.

segmentation masks from the atlas registrations. Each feature is normalized before training to have zero mean and unit standard deviation with respect to the training set. The framework of Sherwood [12] is used both to train and test the random forest instances efficiently. We use random hyperplanes as weak learners and entropy as the splitting criterion.

For each object the random forest classifier generates an estimated probability, $P(i)$, for voxel i to be part of the object. Note that each decision in the classifier is taken independently of the output of the neighboring voxels of i. This has a tendency of producing noisy boundary estimates and therefore we will regularize the solution by using a special instance of the Potts model, *c.f.* [3].

In summary the final segmentation, $\boldsymbol{x}^\star$, is given by the solution to the optimization problem:

$$\begin{aligned} \boldsymbol{x}^\star = & \operatorname*{argmin}_{\boldsymbol{x}\in\{0,1\}^n} \sum_{i=1}^{n} x_i \Big(\frac{1}{2} - P(i)\Big) \\ & + \lambda \sum_{i=1}^{n} \sum_{j\in\mathcal{N}(i)} \mu_{ij} x_i (1 - x_j), \end{aligned} \qquad (6)$$

where λ is a regularization weight and μ_{ij} compensates for anisotropic resolution, see [13]. For all objects we use a 6-connected neighborhood $\mathcal{N}$. In order to save memory and speed-up calculations we only process a volume around the initial registration with a 20 voxels margin. The function in (6) is submodular and is minimized efficiently using [14].

III. Implementation details

In this section, we describe in more detail two important (but standard) subroutines in our system: (i) how the image registration is performed and (ii) how landmark correspondences are established.

A. Atlas registration

In order to propagate the landmark correspondences from our atlas images to a new target image, we estimate a standard intensity-based non-rigid coordinate transformation using the method in [10]. In our experimental evaluation, the compared label fusion methods use exactly the same registrations in order to obtain a fair baseline comparison.

B. Landmark correspondences

In order to establish landmark correspondences for an object between all atlases, we use a registration-based approach. We have found this method to work satisfactorily, but there are several more sophisticated methods [15] that may improve the performance. We estimate a set of accurate non-rigid coordinate transformations between a reference atlas and each of the remaining atlases with the method in [16]. These displacement fields are used in order to transfer triangular meshes of the ground truth surfaces to the coordinate system of the reference atlas. For each mesh point in the reference mesh we find the closest point on each transformed mesh using an algorithm based on [17].

IV. Experimental evaluation

In this section we will evaluate our method on two publicly available datasets and show both quantitative and qualitative comparisons with current state-of-the-art methods.

The first dataset we consider is the whole-body CT images used for the VISCERAL Grand Challenge on multi-organ segmentation [18] and is denoted VISCERAL. The difficulty of this dataset significantly changes from organ to organ. Only the training set, consisting of 20 whole-body CT scans, is publicly available. We split the dataset into three groups: the training volumes 1–10, the cross-validation volumes 11–15, and the test volumes 16–20. The random forest classifiers are trained for volumes 1–15. The cross-validation set is used to tune any parameter needed by the model.

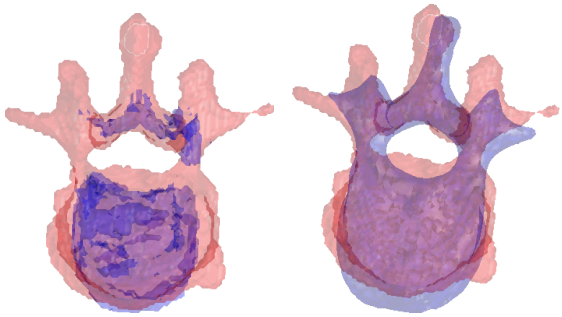
The second dataset we consider consists of 30 MR images of the brain [19]. We will denote this dataset HAMMERS. For this dataset we only consider the most commonly segmented regions of the brain, the hippocampus and the amygdala. This is a very difficult dataset as the boundary of the different organs often lack image contrast.

A. Evaluation of ShapeMap

In order to evaluate the usefulness of the SHAPEMAP we first consider the performance without the refinement step detailed in Section II-B. We compare it to a standard majority voting approach. Note that the same registrations are used for the two competing methods. Also, note that there are no tuning parameters for the two fusion methods, except for γ in (1) which we set to a small number for *all* experiments ($\gamma = 0.001$) and perform no tuning.

TABLE I
MEAN DICE COEFFICIENTS FOR THE VISCERAL DATASET, COMPARING SHAPEMAP TO THE RESULT GIVEN BY A STANDARD MULTI-ATLAS APPROACH WITH MAJORITY VOTING. THE SAME REGISTRATIONS ARE USED FOR BOTH METHODS.

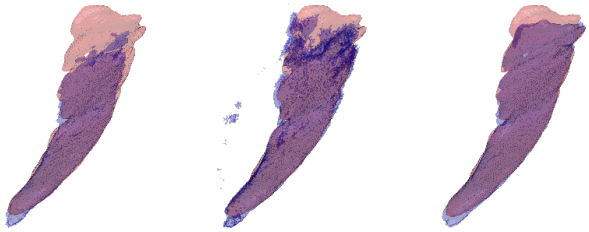
Organ	SHAPEMAP	Majority voting
Trachea	0.83	**0.85**
Pancreas	**0.54**	0.41
Gallbladder	**0.23**	0.07
Urinary Bladder	0.53	**0.65**
Sternum	**0.85**	0.84
Lumbar Vertebra 1	**0.83**	0.82
Right Kidney	0.75	**0.79**
Left Kidney	**0.85**	**0.85**
Right Adrenal Gland	**0.25**	0.13
Left Adrenal Gland	**0.43**	0.37
Right Psoas Major Muscle	**0.83**	**0.83**
Left Psoas Major Muscle	**0.82**	0.81
Right Rectus Abdominis	**0.54**	0.43
Left Rectus Abdominis	**0.58**	0.56
Aorta	**0.80**	0.77
Thyroid	**0.48**	0.43
Spleen	0.69	**0.75**
Average	**0.64**	0.61



(a) Majority vote (b) SHAPEMAP

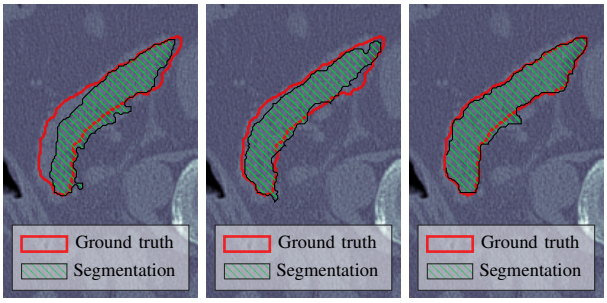
Fig. 5. Ground truth (red) and segmentation proposal (blue) of Lumbar Vertebra 1 with (a) majority voting and (b) SHAPEMAP. Both segmentations have a similar Dice coefficient, but the qualitative shape is superior for SHAPEMAP.

The segmentation proposals are compared to the ground truth using the Dice coefficient (a similarity metric measuring the overlap of two sets, see [20]). Table I shows mean Dice coefficients for the organs in the VISCERAL dataset. The SHAPEMAP approach produces better or on par results for a large majority of the organs (13 out of 17 organs) and achieves the best average result. In Fig. 5, a qualitative comparison of the two methods on a lumbar vertebra from the VISCERAL dataset is shown. In Fig. 5(a) voxel-based majority voting is used, resulting in a disjoint and implausible shape. In Fig. 5(b) the same input data is segmented using our proposed shape-aware method resulting in a coherent and plausible shape. Even though both solutions have similar Dice coefficients the results are qualitatively very different.



(a) Baseline I (b) Baseline II (c) SHAPEMAP

Fig. 6. Ground truth (red) and refined segmentation (blue) of the Right Rectus Abdominis done by (a) Baseline I [10], (b) Baseline II [7] and (c) Refined SHAPEMAP. The segmentations have similar Dice coefficients, but the qualitative shape is superior for SHAPEMAP.



(a) Baseline I (b) Baseline II (c) SHAPEMAP

Fig. 7. Comparison of our approach with the state-of-the-art methods of [10] and [7] on a slice of the pancreas. (a) The segmentation produced by Baseline I. (b) The segmentation produced by Baseline II. (c) The refined segmentation result of this paper.

B. Evaluation of the full framework

We also evaluate the performance of the complete framework including the refinement step from Section II-B. We again consider the VISCERAL dataset and we compare our method with [10] (Baseline I), which achieved the best average results in the whole-body CT segmentation challenge[1] at ISBI 2015 [18], and with [7] (Baseline II), which won the first place of the multi-atlas labeling challenge at MICCAI 2012 [21] and was one of the top performers in the SATA challenge[2] at MICCAI 2013 [22]. Both baselines utilize a pipeline similar to ours; first, an initialization is computed by label fusion (modified majority voting and joint label fusion respectively) and thereafter, the segmentation proposal is refined using machine learning techniques (random forest classification and corrective learning respectively). We use the authors' own implementations for both baselines. Note that the same registrations are used for all methods (including the proposed one).

The segmentation proposals are compared to the ground truth using i) the Dice coefficient and ii) the modified Hausdorff distance. The Dice coefficient is a standard metric in medical image applications while the modified Hausdorff distance (measuring the distance between two sets, see [23]) should capture qualitative improvements better. The similarity metrics are shown in Table II. A qualitative comparison between our approach and the two baselines for segmentation of Right Rectus Abdominis is shown in Fig. 6. An example of a slice of the pancreas in Fig. 7 highlights the performance gain achieved by SHAPEMAP compared to the two baselines.

C. Evaluation on the HAMMERS dataset

Table III shows the segmentation results for the HAMMERS dataset using majority voting and SHAPEMAP (without refinement) respectively. For completeness we also include the method MAPER, presented in [24], which has produced the best published results for this particular dataset. All the presented results are determined using leave-one-out testing. Similarly to our approach, MAPER is also based on multi-atlas segmentation, but has several additional components to improve performance: adaptive atlas selection, organ tissue classification, skull stripping, etc. Still, we can conclude that without any special modifications we can apply our framework to brain MRI segmentation and achieve results on par with the best published results on this dataset.

[1]The winner, measured in Dice, of 13 out of 20 organ classes.

[2]Segmentation: Algorithms, Theory and Applications challenge.

D. Running times

The running times of the different steps in our pipeline were measured on a PC with Intel i7-3770 CPU. Each pair-wise registration (for $512 \times 512 \times 800$ volumes) takes around 50s for large objects (*e.g.* kidneys, aorta) and less than 25s for small ones (*e.g.* pancreas, hippocampus). The transfer of the triangular mesh (typically between 500 – 5,000 triangles per atlas mesh) is negligible. The refinement step takes less than 1 minute for large objects and less than 5s for small ones.

V. CONCLUSIONS

This paper introduces a framework for incorporating prior information of plausible shapes into the standard multi-atlas pipeline by (i) propagating landmarks from a set of atlases and optimizing a robust shape average and (ii) incorporating properties of the robust shape average into a random forest classifier. As far as we are aware, this gives a novel way of regarding the multi-atlas voting procedure, where the standard label decision on a voxel level is replaced with the more robust decision of landmark positions. When evaluating the segmentation results with standard similarity metrics, our approach outperforms recent results for multi-organ segmentation in whole-body CT images. Moreover, our approach gives a significant qualitative improvement in the form of more realistic organ shapes than the compared methods.

A natural continuation of the SHAPEMAP would be to incorporate a full statistical shape model. In this paper, the algorithm for establishing landmark correspondences is not evaluated thoroughly, and may leave room for improvement. Though, the robust average decreases the risk of impairment of the results due to inaccurate landmarks. Possibly, introducing a complete point distribution model could increase the demand for more accurate landmark correspondences. Finding new ways of computing landmarks suitable for our purposes will be an interesting avenue of future research.

TABLE II
MEAN DICE COEFFICIENTS AND HAUSDORFF DISTANCES FOR THE VISCERAL DATASET, COMPARING OUR FINAL SEGMENTATION RESULTS TO BASELINE I [10] AND BASELINE II [7].

Organ	Mean Dice coefficients			Mean Hausdorff distance (voxels)		
	Refined SHAPEMAP	Baseline I	Baseline II	Refined SHAPEMAP	Baseline I	Baseline II
Trachea	**0.92**	**0.92**	0.91	**0.83**	0.86	1.72
Pancreas	**0.60**	0.45	0.50	**6.34**	9.76	12.94
Gallbladder	**0.23**	0.00	0.05	**15.62**	*	*
Urinary Bladder	**0.78**	0.76	0.74	**4.30**	4.62	5.93
Sternum	0.87	0.86	**0.89**	1.97	1.98	**1.13**
Lumbar Vertebra 1	**0.86**	0.84	0.80	**1.69**	2.01	3.61
Right Kidney	**0.84**	0.76	0.83	**3.28**	5.09	5.31
Left Kidney	**0.89**	**0.89**	**0.89**	**2.31**	2.47	3.09
Right Adrenal Gland	**0.25**	0.18	0.24	**5.73**	8.88	*
Left Adrenal Gland	**0.51**	0.34	0.46	5.53	**5.43**	21.94
Right Psoas Major Muscle	**0.84**	0.83	**0.84**	**3.62**	3.91	4.50
Left Psoas Major Muscle	**0.84**	0.83	**0.84**	**2.95**	2.97	3.08
Right Rectus Abdominis	**0.61**	0.58	0.52	**8.06**	11.32	*
Left Rectus Abdominis	0.70	0.56	**0.73**	4.36	12.09	**3.68**
Aorta	**0.82**	0.81	**0.82**	2.41	2.49	**2.28**
Thyroid	0.50	0.46	**0.51**	**4.92**	5.10	5.06
Spleen	**0.83**	0.8	**0.83**	5.99	5.94	**4.63**

*The method failed to produce a segmentation proposal for one or more of the test images.

TABLE III
MEAN DICE COEFFICIENTS FOR THE HIPPOCAMPUS AND THE AMYGDALA FROM THE HAMMERS DATASET.

	Hippocampus		Amygdala	
	Left	Right	Left	Right
SHAPEMAP	0.806	0.826	**0.805**	**0.804**
Majority voting	0.756	0.780	0.758	0.748
MAPER [24]	**0.822**	**0.835**	0.784	0.794

REFERENCES

[1] T. Cootes *et al.*, "Active shape models - their training and applications," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[2] T. F. Cootes *et al.*, "Active appearance models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[3] A. Blake *et al.*, Eds., *Markov Random Fields for Vision and Image Processing*. MIT Press, 2011.

[4] J. Iglesias and M. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Medical Image Analysis*, vol. 24, no. 1, pp. 205–219, 2015.

[5] T. Rohlfing and C. Maurer, "Shape-based averaging," *IEEE Trans. on Image Processing*, vol. 16, no. 1, pp. 153–161, 2007.

[6] G. Sanroma *et al.*, "Learning-based atlas selection for multiple-atlas segmentation," in *Conf. on Computer Vision and Pattern Recognition*, 2014.

[7] H. Wang *et al.*, "Multi-atlas segmentation with joint label fusion," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 611–623, 2013.

[8] S. Warfield *et al.*, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. on Medical Imaging*, vol. 27, no. 3, pp. 903–921, 2004.

[9] Z. Xu *et al.*, "Shape-constrained multi-atlas segmentation of spleen in CT," in *SPIE Medical Imaging*, 2014.

[10] F. Kahl *et al.*, "Good features for reliable registration in multi-atlas segmentation," in *Proceedings of the VISCERAL Challenge at ISBI*, 2015.

[11] S. Boyd *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[12] A. Criminisi *et al.*, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Foundations and Trends® in Computer Graphics and Vision*, vol. 7, pp. 81–227, 2011.

[13] V. Kolmogorov and Y. Boykov, "What metrics can be approximated by geo-cuts, or global optimization of length/area and flux," in *IEEE Int. Conf. on Computer Vision*, 2005.

[14] O. Jamriška *et al.*, "Cache-efficient graph cuts on structured grids," in *Conf. on Computer Vision and Pattern Recognition*, 2012.

[15] T. Cootes *et al.*, "Computing accurate correspondences across groups of images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1994–2005, 2010.

[16] J. Alvén *et al.*, *Überatlas: Robust Speed-Up of Feature-Based Registration and Multi-Atlas Segmentation*, ser. Lecture Notes in Computer Science. Berlin: Springer, 2015, pp. 92–102.

[17] D. Eberly, "Distance between point and triangle in 3D," 2008, geometric Tools, LLC, Available from: http://www.geometrictools.com/.

[18] O. del Toro *et al.*, "VISCERAL - VISual Concept Extraction challenge in RAdioLogy : ISBI 2015 challenge organization," in *Proceedings of the VISCERAL Challenge at ISBI*, ser. CEUR Workshop Proceedings, 2015.

[19] I. Gousias *et al.*, "Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest," *Neuroimage*, vol. 40, no. 2, pp. 672–684, 2008.

[20] L. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26(3), pp. 297–302, 1945.

[21] "MICCAI 2012 multi-atlas labeling challenge," in *MICCAI 2012 Workshop on Multi-Atlas Labeling*, B. Landman and S. Warfield, Eds., 2012.

[22] A. Asman *et al.*, "MICCAI 2013 segmentation: Algorithms. theory and applications (SATA) challenge results summary," in *MICCAI 2013 Challenge Workshop on Segmentation: Algorithms. Theory and Applications*, 2013.

[23] M. P. Dubuisson and A. K. Jain, "A modified Hausdorff distance for object matching," in *Proceedings of the Int. Conf. on Pattern Recognition*, vol. 1, 1994, pp. 566–568.

[24] R. Heckemann *et al.*, "Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation," *Neuroimage*, vol. 51, no. 1, pp. 221–227, 2010.