Creating a Large-Scale Silver Corpus from Multiple Algorithmic Segmentations

Markus Krenn[1], Matthias Dorfer[2], Oscar Alfonso Jiménez del Toro[3], Henning Müller[3], Bjoern Menze[4], Marc-André Weber[5], Allan Hanbury[6], and Georg Langs[1]

[1] Computational Imaging Research (CIR) Lab, Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Vienna, Austria
markus.krenn@meduniwien.ac.at
[2] Department of Computational Perception, Johannes Kepler University (JKU), Linz, Austria
[3] University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
[4] Institute for Advanced Study and Department of Computer Science, Technische Universität München, Munich, Germany
[5] Department of Diagnostic and Interventional Radiology, University of Heidelberg, Heidelberg, Germany
[6] Institute of Software Technology and Interactive Systems, TU Wien, Vienna, Austria

Abstract. Currently, increasingly large medical imaging data sets become available for research and are analysed by a range of algorithms segmenting anatomical structures automatically and interactively. While they provide segmentations on a much larger scale than possible to achieve with expert annotators, they are typically less accurate than experts. We present and compare approaches to estimate segmentations on large imaging data sets based on a small number of expert annotated examples, and algorithmic segmentations on a much larger data set. Results demonstrate that combining algorithmic segmentations is reliably outperforming the average individual algorithm. Furthermore, injecting organ specific reliability assessments of algorithms based on expert annotations improves accuracy compared to standard label fusion algorithms. The proposed methods are particularly relevant in putting the results of large image analysis algorithm benchmarks to long-term use.

Keywords: segmentation, label fusion, silver corpus

1 Introduction

Annotations are an important basis when developing algorithms that segment anatomical structures in medical imaging data. For relatively small sets, they can be manually generated by experts, and serve as means for the training, and evaluation of algorithms. If multiple annotations are available for the same target, label fusion algorithms such as STAPLE [26] provide improved estimates for true segmentations. Recently, increasingly large data sets have become available

to the research community [11]. Such datasets are often part of challenges, where a large number of state of the art algorithms are applied to localize or segment anatomical structures. In this paper we propose and compare approaches to estimate true segmentations on large medical imaging data, if expert annotations are available for only a small sub-set and less reliable algorithmic annotations are available for all data.

Label fusion approaches in medical image segmentation aim at finding the true (hidden) segmentation of a structure in an image by estimating a *consensus* of multiple segmentation estimates. Independent noise in different annotations causes the consensus to correct this variability and yield significantly more accurate segmentations than those derived from a single source [18, 22]. For instance, *majority voting* assigns the label with most *votes* among annotators to each voxel [8, 15, 16]. In the context of multi-atlas label fusion additional weighting or *weighted voting* based on image similarity can further improve results [2]. Further improvement can be gained by fusing labels of a well chosen sub-set of atlases [1] Fusing multiple atlas segmentations significantly outperforms single atlas segmentations [21]. One downside of weighting schemes, is that image similarity is a limited predictor of registration accuracy [2, 18]. Therefore, a range of approaches takes multiple annotations into account to assess consistency as a basis for estimating their reliability. Label fusion via *Simultaneous Truth And Performance Level Estimation (STAPLE)* [26] estimates performances and weights of contributing segmentations based on an expectation maximization. STAPLE simultaneously computes performance estimates based on sensitivity and specificity of contributing segmentations during maximization and establishes an estimate of the hidden ground truth segmentation, given performance estimates in the expectation step [26]. It outperforms majority voting, and competes with weighted voting approaches [1]. A *Selective and Iterative Method for Performance Level Estimation (SIMPLE)* proposed in [18] is another iterative label fusion algorithm. Similar to STAPLE it simultaneously computes an estimate of the hidden true segmentation and estimates performances of each contributing segmentation. SIMPLE additionally discards poorly performing segmentations during the fusion process and computes segmentation performances based on a spatial overlap measure of involved segmentations to the current estimate of the hidden ground true segmentation. In case of fusing a small number of expert annotations SIMPLE and STAPLE are reported to perform equally [18].

While these insights are important, existing approaches assume that we face a set of annotators with initially equal competency estimates. In this paper we extend these approaches to cases, where we have a small set of high-quality, or 'expert' annotations (gold corpus), and a large set of algorithmic and possibly less accurate annotations. We show how to use this for estimating segmentations of anatomical structures on a large data set, resulting in a so-called *silver corpus*.

We calculate a silver corpus annotation of anatomical structures in medical images by fusing gold corpus annotations from a limited number of templates by multi-atlas label fusion, and additional algorithmic estimates of the annotations.

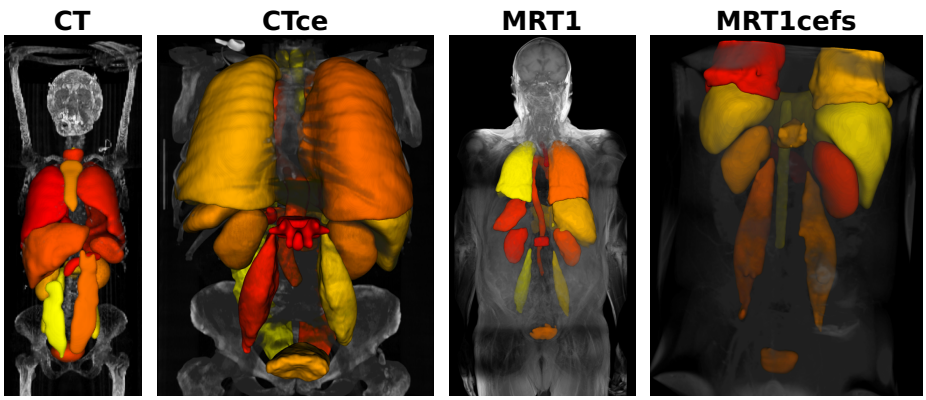

Fig. 1. Illustrations of generated silver corpus annotations of one volume in each modality. Volumes and annotations will be publicly available for the research community.

We estimate the reliability of each algorithm and corresponding weights based on gold corpus annotations on a limited number of cases.

We evaluate different strategies of fusing algorithmic segmentations, and estimate their reliability either by consensus, or by comparison across the annotation sets (gold corpus and algorithmic). Finally we demonstrate that the proposed approach consistently outperforms the average algorithm. Injecting reliability estimates further improves accuracy. We detail the benefit of specific fusion strategies, and provide a comprehensive evaluation of all approaches on 20 anatomical structures in 40 volumes of four different modalities.

We finally apply the best performing fusion method to 264 cases of the evaluated modalities that were part of the VISCERAL challenges [19]. We make the resulting silver corpus that contains 4323 annotated anatomical structures available for the research community. Visualizations of generated silver corpus annotations of one volume of each evaluated modality are shown in Figure 1.

2 Method

We start with a formal definition of our problem setting. Then, we explain the algorithm for silver corpus label fusion from expert annotations on a small set of cases, and multiple algorithmic annotations for each case. To facilitate reading we explain the segmentation of one anatomical structure in one target image of the silver corpus. This generalizes to arbitrary structures since organs, and target images are treated independently in the proposed approach.

2.1 Problem Setting

Given a segmentation gold corpus as the set of N expert annotated atlases $\mathfrak{A} = \langle \mathbf{I}_1, \ldots, \mathbf{I}_N; \mathbf{L}_1, \ldots, \mathbf{L}_N \rangle$ where an atlas $(\mathbf{I}_n, \mathbf{L}_n)$ is defined as a tuple, containing

image $\mathbf{I}_n$ and binary label image (annotation) $\mathbf{L}_n$. We aim to compute a silver standard annotation $\mathbf{L}'_T$ for each organ in a target image (image or volume) $\mathbf{I}_T$ by fusing data that is typically available during benchmarks or competitions. Benchmark data addresses the scenario where the gold corpus $\mathfrak{A}$ is complemented by a set of P automatic segmentation estimates $\mathcal{P} = \langle \hat{\mathbf{L}}_T^1, \dots \hat{\mathbf{L}}_T^P \rangle$ obtained by algorithms that are applied to target image $\mathbf{I}_T$. Each binary label image $\hat{\mathbf{L}}_T^p$ holds the segmentation estimate of algorithm p in target image $\mathbf{I}_T$. The atlas annotations are non-linearly mapped to $\mathbf{I}_T$ using image registration. We compute transformations $T_n(\mathbf{x})$ so that $\mathbf{I}_T(\mathbf{x}) \sim \mathbf{I}_n(T_n(\mathbf{x}))$ and transfer the corresponding annotations $\mathbf{L}_n$ towards the target volume, by applying the same transformation $\tilde{\mathbf{L}}_n(\mathbf{x}) = \mathbf{L}_n(T_n(\mathbf{x}))$. Our goal is to compute a segmentation silver corpus by introducing a label fusion approach taking both, gold corpus $\mathfrak{A}$ as well as algorithmic label estimates $\mathcal{P}$ into account.

2.2 Atlas Registration and Selection

We first map atlas annotations $\mathbf{L}_n$ to the target image in a two step procedure. Inspired by [1] we first evaluate the *Normalized Mutual Information* (NMI) for each pair of an atlas image $\mathbf{I}_n$ and the target image $\mathbf{I}_T$ [24].

After ranking all atlases based on their NMI to $\mathbf{I}_T$, the framework selects the set of top ranked J atlases and computes segmentation estimates of the desired structure in $\mathbf{I}_T$ for each selected atlas by a two step registration process: **1. Registration:** A non-linear alignment between $\mathbf{I}_T$ and each atlas image $\mathbf{I}_n$ is established (for details see Section 3), resulting in a transformation T_n, that maps $\mathbf{I}_T$ to $\mathbf{I}_n$ so that $\mathbf{I}_T(x) \approx \mathbf{I}_n(T_n(x))$. **2. Label propagation:** The computed transformation is used to compute a segmentation estimation of atlas n in the target image, denoted as $\tilde{\mathbf{L}}_T(x) = \mathbf{L}_n(T_n(x))$. We denote the set of J atlas based labelings of the target image $\mathbf{I}_T$ as $\mathcal{A} = \langle \tilde{\mathbf{L}}_1, \dots, \tilde{\mathbf{L}}_J \rangle$.

2.3 Computing Weights of Segmentation Estimates

The previous steps result in two sets of labelings for the target image: mapped gold corpus atlas annotations $\mathcal{A}$ and annotations by algorithms $\mathcal{P}$. Inspired by the work in [2], we perform weighted label fusion to generate $\mathbf{L}'_T$. In the following we explain how to determine the weights of individual segmentations.

Obtaining Atlas Weights We derive atlas segmentation performance estimates by leave-one-out cross validation on the test set. We propagate each annotation $\mathbf{L}_n$ to all remaining atlases based on non-rigid pair-wise registration of the corresponding images. We evaluate the overlap of propagated annotations by comparing them to the native expert annotation in the atlas image using the Dice coefficient [3]. The weight v_n of an atlas annotator $\tilde{\mathbf{L}}_n$ is the average Dice [3] coefficient of its segmentations propagated to all other atlas images, and the corresponding expert annotation. This results in weights $\mathbf{v} = v_1, \dots v_N$.

Obtaining Algorithm Weights To derive performance estimates $\mathbf{w} = w_1, \dots w_P$ of the algorithms, each algorithm is applied to all test set atlas volumes. Similar to atlas weights, the weights are calculated by averaging the Dice coefficents [3] of computed segmentations and atlas annotations.

2.4 Fusion

As final step we fuse propagated atlas segmentations $\mathcal{A}$ and algorithmic segmentations $\mathcal{P}$ to the resulting silver corpus annotation $\mathbf{L}'_T$ for the targeted structure. In the following we explain four label fusion approaches and compare them in the experiment section. After calculating weights of atlas- and algorithm annotations individually, we treat them equally during fusion, we denote $\mathcal{L} = \langle \mathbf{L}_1, \dots, \mathbf{L}_M \rangle$ as a set of M binary label images and $\mathbf{u} = u_1, \dots u_M$ as a vector of corresponding segmentation weights. $\mathbf{L}'_T(\mathbf{x})$ is the final segmentation estimate in $\mathbf{I}_T$.

1. **Majority Vote (MV):** The computed segmentation performance estimates are not considered during fusion and all contributing segmentations are weighted equally. Each voxel $\mathbf{x}$ of $\mathbf{L}'_T$ is assigned with the label that is most frequent in corresponding voxels of all contributing segmentation estimates:

$$\mathbf{L}'_T(\mathbf{x}) = \left\{ \begin{array}{l} 1, \left(\sum_{i=1}^{M} \mathbf{L}_i(\mathbf{x}) \right) \geq \frac{M}{2} \\ 0, \left(\sum_{i=1}^{M} \mathbf{L}_i(\mathbf{x}) \right) < \frac{M}{2} \end{array} \right\} \tag{1}$$

2. **Organ Level Weighted Voting:** $\mathbf{L}'_T$ is derived by a majority vote where the impact of each $\mathbf{L}_i \in \mathcal{L}$ is weighted by u_i. Since weights are determined for each organ independently, we call this algorithm Organ Level Weighted Voting (OLWV).

$$\mathbf{L}'_T(\mathbf{x}) = \left\{ \begin{array}{l} 1, \left(\sum_{i=1}^{M} \mathbf{L}_i(\mathbf{x}) \cdot u_i \right) \geq \frac{\sum u}{2} \\ 0, \left(\sum_{i=1}^{M} \mathbf{L}_i(\mathbf{x}) \cdot u_i \right) < \frac{\sum u}{2} \end{array} \right\} \tag{2}$$

3. **STAPLE:** For evaluation in this work we use the binary version of the STAPLE algorithm [26]. STAPLE takes a set of binary label images $\mathcal{L}$ as input and computes an estimation of the hidden ground truth $\mathbf{L}'_T$ based on expectation maximization [26]. We refer to STAPLE segmentations as $\mathbf{L}'_T = \mathit{STAPLE}(\mathcal{L})$.
4. **SIMPLE:** An implementation of SIMPLE as proposed in [18] is used in our work. Besides $\mathcal{L}$, the algorithm is parametrized by k and α, where α influences the performance level that contributing segmentations must exceed to remain in the set of contributing segmentation and k defines the number of iterations in which segmentations are kept for fusion even though the performance threshold is not reached. Optional, SIMPLE takes initial segmentation weights into account. We refer to SIMPLE segmentations computed without initialization as $\mathbf{L}'_T = \mathit{SIMPLE}(\mathcal{L}, k, \alpha)$ and as $\mathbf{L}'_T = \mathit{SIMPLE}^*(\mathcal{L}, k, \alpha, \mathbf{u})$ when pre-computed segmentation weights $\mathbf{u}$ are used for initialization.

3 Experiments

Data and Validation We evaluate the proposed framework and compare fusion methods using a set of 120 atlases with manual expert reference annotations. They cover four modalities (30 T1-weighted magnetic resonance (MRT1) images, 30 T1-weighted contrast enhanced fat saturated magnetic resonance images (MRT1cefs), 30 computed tomography scans (CT) and 30 contrast enhanced computed tomography (CTce) scans) with up to 20 annotated structures in each volume[7] MRT1 volumes have a field–of–view of the whole human body (voxels: $1.1 - 1.3 \times 1.1 - 1.3 \times 6 - 7$mm), MRT1cefs volumes of the abdomen ($1.2 - 1.3 \times 1.2 - 1.3 \times 3$mm), CT scans of the whole human body ($0.8 - 0.9 \times 0.8 - 0.9 \times 1.5$mm) and CTce scans include the chest and the abdomen ($0.6 - 0.7 \times 0.6 - 0.7 \times 1.2 - 1.5$mm).

Algorithmic segmentations are available from participants of the *VISCERAL Anatomy* 2 & 3 challenges[8], where 9 groups contributed algorithms for structures in CT and CTce and 2 in MRT1 and MRT1cefs volumes [6, 11]. Each participant has been able to submit up to 5 different parameter configurations, resulting in 20 independent algorithmic estimates in CT and CTce and 2 in MRT1 and MRT1cefs volumes. Most algorithms incorporate atlas based segmentation approaches [5, 9, 10, 12, 13], but are also based on shape and appearance modelling [7, 20, 25], anatomy based reasoning [4, 23] or use graph cuts and spatial relations [14]. All methods are trained on 20 atlases of each modality which are excluded from the test set [6, 11].

Since only $\mathcal{P}$ are independent for each target volume, but a sub-set of atlases has to be used for generating $\mathcal{A}$ and for estimating the performance of each algorithmic annotator, we performed leave-one-out cross validation on 10 atlases of each modality resulting in a test set of 40 volumes. Each atlas of the test set was selected once as target image and held out of the source atlas set and from the performance estimation process. Accuracy for all structures is reported as the Dice between segmentation estimate and gold corpus annotation [3].

Registration Annotations are propagated from atlases to target images based on NMI driven multiresolution affine- and non-rigid registration. We use the NiftyReg toolbox[9], with a CUDA based implementation for affine alignment and B-spline based non-rigid registration. For CTce volumes a spline grid with 7 mm, for CT with 9 mm, for MRT1cefs with 9 mm and for MRT1 a grid with 7 mm spacing is applied.

[7] Structures and RadlexIDs: r./l. lungs - RID 1302/1326, liver 58, r./l. kidneys 29662/29663, gallbladder 187, trachea 1247, aorta 480, first lumbar vertebra 29193, r./l. adrenal gland 30324/30325, r./l. psoas major 32248/32249, muscle body of r./l. rectus abdominis 40357/40358, pancreas 170, spleen 86, sternum 2473, urinary bladder 237 and thyroid gland 7578. For Radlex terminology refer to http://www.radlex.org/.

[8] Organized by the EU FP7 funded project VISCERAL: http://www.visceral.eu

[9] http://www.nitrc.org/projects/niftyreg/

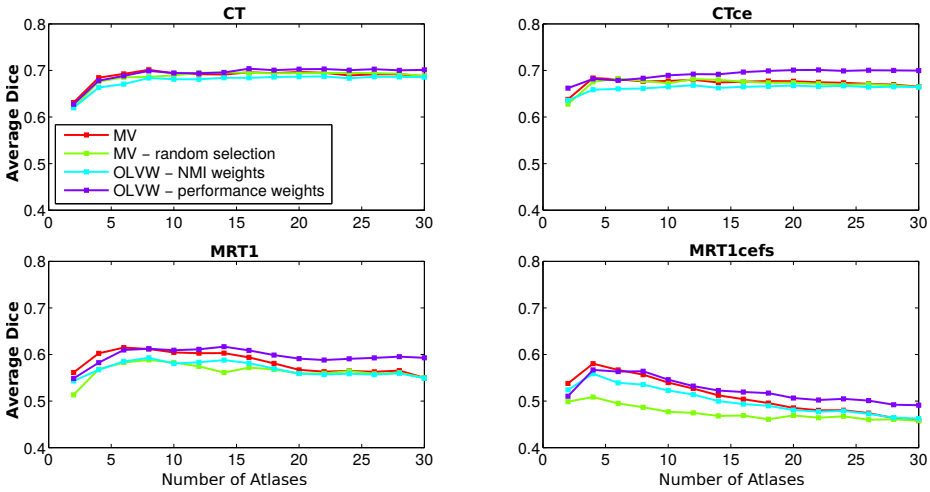

Fig. 2. Segmentation performances of *MV*, *MV - random selection*, *OLVW - NMI weights* and *OLVW - performance weights* averaged over all structures of each modalitiy evaluated for different numbers (x-axis) of selected atlases. *OLVW - performance weights* yields best results in all four modalities.

Fusion Approaches Method identifiers (*MV, OLWV, STAPLE, SIMPLE*) and input parameters follow those in Section 2. Depending on the input data, the method identifier is appended by letter A (atlases) if $\mathcal{L} = \mathcal{A}$, by P (participating algorithms) if $\mathcal{L} = \mathcal{P}$ and AP if $\mathcal{L} = \langle \mathcal{A}, \mathcal{P} \rangle$ (atlases and algorithms). Corresponding weights are calculated following Section 2.3. For SIMPLE, the number of iterations in which no segmentation is discarded is $k = 3$, similar to [18] and we set $\alpha = 1.25$ [17].

4 Results

Atlas Selection and Weighting First, we evaluate the effect of different numbers of weighted atlases without using algorithmic segmentations. Figure 2 compares majority voting of pre-registration selected atlases (*MV*), majority voting of randomly selected atlases (*MV - random selection*), OLVW with weights derived by NMI of the transformed atlas image and the target image (*OLVW - NMI weights*) and OLVW with performance weights calculated according to Section 2.3 (*OLVW - performance weights*). We show the Dice coefficients averaged over all structures for each modality and increasing numbers of atlases. While all methods yield comparable results on CT volumes, differences become visible in CTce and especially in MRT1 and MRT1cefs volumes. Results show that *MV* slightly outperforms *MV - random selection* as well as *OLWV - NMI weights*. Best results in all modalities are obtained by a weighted vote that incorporates performances weights (*OLWV - performance weights*). Results furthermore show

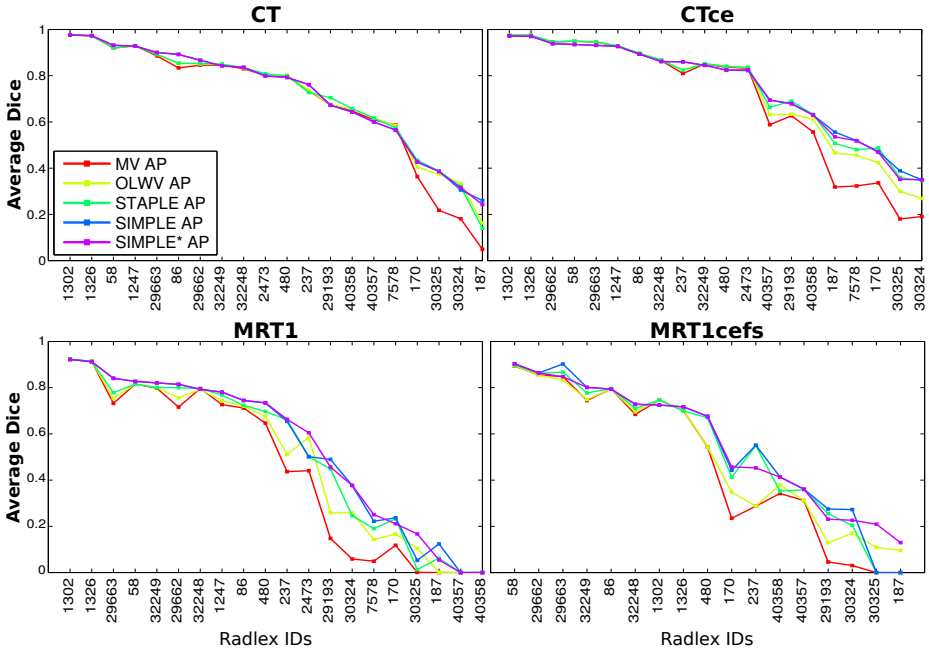

Fig. 3. Average segmentation performances of different fusion approaches that incorporate propagated atlases and algorithmic segmentation estimates on all modalities evaluated. *SIMPLE* segmentation yields in the best overall segmentation performances. Injecting reliability estimates (*SIMPLE**) results in similar performances compared to *SIMPLE* without initial segmentation weights.

an initial performance increase with an increasing amount of considered atlases followed by a constant (CT, CTce) or even decreasing segmentation performance when considering many poor atlases (MRT1, MRT1cefs).

Comparing Label Fusion Methods Integrating Atlases and Algorithms Figure 3 compares different label fusion approaches (*MV*, *OLWV*, *STAPLE* and *SIMPLE*) and the impact of performance based weights on the SIMPLE approach (indicated by $*$ in the legend) on test set segmentation accuracy. Here, all methods take atlases and segmentations of participants into account. Based on the results shown in Figure 2, we select the top 16 ranked atlases for CT, 20 for CTce, 14 for MRT1 and 8 for MRT1cefs volumes. The y-axis depicts average Dice [3] coefficients, the x-axis identifies anatomical structures, ordered by accuracy of the best performing approach.

All methods perform comparably for structures with high overall segmentation accuracy (e.g., the lungs in CT, CTce and MRT1). The benefit of weighting becomes visible for structures with lower overall segmentation accuracy, and is highest in MRT1 and MRT1cefs. Here, *OLWV AP* outperforms *MV AP* in the

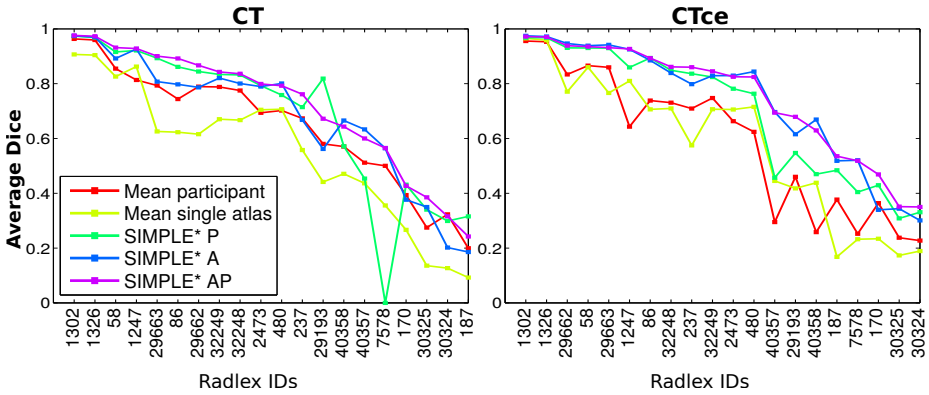

Fig. 4. Average accuracies of single algorithmic segmentations, atlases and different fusion approaches. Fusing segmentation estimates of both, atlases and algorithms (*SIMPLE* AP*) outperforms segmentations obtained by fusing estimates of one of both components (*SIMPLE* A*, *SIMPLE* P*).

majority of tested structures. That is, taking the split of annotators into experts and algorithms into account does improve accuracy. Fusing segmentations with *STAPLE AP* and *SIMPLE AP* improves segmentation accuracy for all investigated structures. Excluding poorly performing segmentations (*SIMPLE AP*) results in higher segmentation accuracy, again especially in structures with low overall segmentation quality. Using reliability estimates of each contributing segmentation to initialize SIMPLE (*SIMPLE* AP*) and SIMPLE without applying initial weights (*SIMPLE AP*) reach similar segmentation performances, which is plausible since SIMPLE is reported to be independent to initial segmentation weights [18]. Small performance gains are observed in some structures of MRT1 and MRT1cefs volumes.

Multiple Atlases vs. Multiple Algorithms Figure 4 compares average segmentation accuracies of individual atlases or individual algorithms with segmentations obtained by fusing atlases (*SIMPLE* A*), algorithms (*SIMPLE* P*) or both (*SIMPLE* AP*). As expected, all fusion approaches consistently outperform individual segmentations. Using atlases and algorithms jointly improves segmentation accuracies in the majority of structures compared to fusing algorithms or atlases only. Figure 5 illustrates the effect of fusing atlases and algorithms on liver segmentations. True positive, false negative and false positive segmented voxels of two algorithms, two propagated atlases and the resulting fused segmentation are shown in a CTce and a MRT1cefs volume.

The Resulting Silver Corpus Finally we apply the best performing fusion method *SIMPLE* AP* to 264 additional volumes of the modalities (62 CT, 65 CTce, 66 MRT1, 71 MRT1cefs) for which segmentation estimates could be generated using the algorithms submitted to *VISCERAL Anatomy* 2 & 3, resulting

in the **VISCERAL Anatomy Silver Corpus** that consists a set of 4323 segmentations of anatomical structures which will be available as a resource for the research community at www.visceral.eu.

For reference, Table 1 lists the number of computed segmentations (#) of all target structures in each modality as well as average segmentation performances (μ) and corresponding standard deviations (σ) which serve as structure and modality specific segmentation performance estimates of generated silver corpus annotations.

5 Conclusion

Algorithmic segmentation of anatomical structures is essential for computer aided diagnosis, since large scale manual annotation is infeasible. Benchmarks that evaluate multiple algorithms on medical imaging data contribute critically to assessing their accuracy, and thereby advancing method research. At the same time the variety and number of these algorithms can be leveraged to create large scale *silver corpora* of imaging data annotated by some sort of consensus of these contributions. Here, we propose and evaluate a framework for creating silver corpus annotations of large scale data. We first apply a number of different segmentation algorithms that were entries to an anatomy segmentation challenge to segment the data. Then, we fuse labels transferred from manually annotated atlases and the labels obtained by the algorithmic segmentations. The results demonstrate that adding algorithmic estimators improves accuracy compared to baseline segmentations obtained by mere expert atlas fusion. Furthermore, informing the label fusion by weighting algorithmic and atlas segmentations based on their accuracy in comparison to expert annotations improves accuracy over standard label fusion techniques. The accuracy gained by fusion is highest for anatomical structures on which algorithms perform poorly. Analogously, even though fusion of algorithmic segmentations is already beneficial across the entire range of organs, adding atlases furthermore improves accuracy for structures where algorithms have low accuracy. We applied the best performing method, *SIMPLE* AP* (fusing atlases and algorithmic segmentations, initialized with performance estimate weights), to a dataset of 264 volumes of four modalities (CT, CTce, MRT1, MRT1cefs). This resulted in a set of 4323 silver corpus annotations, which will be available for the research community at www.visceral.eu.

Acknowledgements The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements 318068 (VISCERAL) and 257528 (KHRESMOI). We furthermore acknowledge the support of NVIDIA Corporation with the donation of a Tesla K40 GPU used for this work and would like to thank all research groups contributing to this work by participating in the *VISCERAL Anatomy* 2 & 3 benchmarks [4, 5, 7, 9, 10, 12–14, 20, 23, 25].

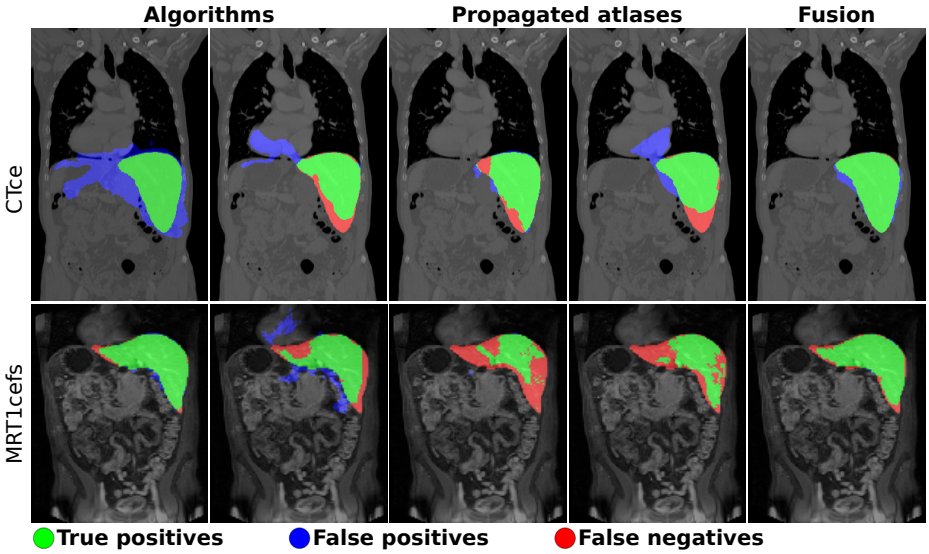

Fig. 5. Consensus plots of two algorithmic segmentation estimates, two mapped atlases and the resulting fusion of two liver segmentations obtained by *SIMPLE* AP*.

Table 1. Segmentation performances (μ, σ) obtained by *SIMPLE* AP*, evaluated on 10 annotated test set volumes (i.e. which were not included in algorithm training) per modality and number of silver corpus annotations (#) computed on additional volumes that will be available as a resource for the research community at www.visceral.eu.

Radlex ID	**Name**	**CT**			**CTce**			**MRT1**			**MRT1cefs**		
		#	μ	σ	#	μ	σ	#	μ	σ	#	μ	σ
58	liver	59	0.93	0.01	63	0.94	0.01	66	0.83	0.07	71	0.90	0.03
86	spleen	55	0.89	0.06	63	0.89	0.07	65	0.74	0.11	71	0.79	0.18
170	pancreas	57	0.43	0.19	60	0.47	0.18	63	0.21	0.21	71	0.46	0.13
187	gallbladder	40	0.24	0.19	49	0.54	0.15	46	0.05	0.05	61	0.13	0.20
237	urinary bladder	58	0.76	0.15	64	0.86	0.06	59	0.66	0.28	70	0.45	0.25
480	aorta	58	0.79	0.04	63	0.82	0.05	65	0.73	0.07	71	0.68	0.02
1247	trachea	57	0.93	0.02	62	0.93	0.02	63	0.78	0.10	-	-	-
1302	r. lung	60	0.98	0.01	64	0.97	0.01	66	0.92	0.02	-	-	-
1326	l. lung	61	0.97	0.01	63	0.97	0.01	66	0.91	0.03	-	-	-
2473	sternum	55	0.80	0.04	63	0.83	0.07	64	0.60	0.00	-	-	-
7578	thyroid gland	57	0.57	0.10	62	0.52	0.13	64	0.25	0.15	-	-	-
29193	first lumbar vertebra	57	0.67	0.36	63	0.68	0.34	58	0.46	0.25	71	0.23	0.12
29662	r. kidneys	57	0.87	0.12	63	0.94	0.01	65	0.81	0.11	71	0.86	0.18
29663	l. kidneys	58	0.90	0.03	63	0.93	0.02	64	0.84	0.06	71	0.85	0.20
30324	r. adrenal gland	54	0.32	0.20	56	0.35	0.14	50	0.38	0.14	60	0.23	0.11
30325	l. adrenal gland	54	0.36	0.19	53	0.35	0.17	41	0.17	0.22	49	0.21	0.12
32248	r. psoas major	58	0.84	0.02	63	0.86	0.02	65	0.79	0.06	71	0.73	0.12
32249	l. psoas major	56	0.84	0.02	63	0.85	0.05	65	0.82	0.06	71	0.80	0.05
40357	r. rectus abdominis	56	0.60	0.21	63	0.69	0.16	-	-	-	-	-	-
40358	l. rectus abdominis	55	0.64	0.14	64	0.63	0.17	-	-	-	-	-	-
$\sum$		**1122**			**1227**			**1095**			**879**		

References

1. P. Aljabar, R.A. Heckemann, A. Hammers, J.V. Hajnal, and D. Rueckert. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *NeuroImage*, 46(3):726 – 738, 2009.
2. X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de Solórzano. Combination strategies in multi-atlas image segmentation: Application to brain mr data. *IEEE Transactions on Medical Imaging*, 28(8):1266–1277, 2009.
3. L.R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.
4. Y. Dicente Cid, A. Depeursinge, O.A. Jiménez del Toro, and H. Müller. Efficient and fully automatic segmentation of the lungs in ct volumes. In O. Goksel, O.A. Jiménez-del Toro, A. Foncubierta-Rodríguez, and H. Müller, editors, *Proceedings of the VISCERAL Challenge at ISBI*, volume 1390, page 31, Apr 2015.
5. T. Gass, G. Szekely, and O. Goksel. Multi-atlas segmentation and landmark localization in images with large field of view. In B. Menze, G. Langs, A. Montillo, M. Kelm, S. Müller, H.and Zhang, W.T. Cai, and D. Metaxas, editors, *Medical Computer Vision: Algorithms for Big Data*, volume 8848 of *Lecture Notes in Computer Science*, pages 171–180. Springer International Publishing, 2014.
6. O. Göksel, O. A. Jiménez-del Toro, A. Foncubierta-Rodríguez, and H. Muller. Overview of the VISCERAL challenge at ISBI 2015. In Orçun Göksel, Oscar Alfonso Jiménez-del Toro, Antonio Foncubierta-Rodríguez, and Henning Müller, editors, *Proceedings of the VISCERAL Challenge at ISBI)*, New York, NY, May 2015.
7. B. He, C. Huang, and F. Jia. Fully automatic multi-organ segmentation based on multi-boost learning and statistical shape model search. In O. Goksel, O.A. Jiménez-del Toro, A. Foncubierta-Rodríguez, and H. Müller, editors, *Proceedings of the VISCERAL Challenge at ISBI*, volume 1390, pages 18–21, Apr 2015.
8. R.A. Heckemann, J.V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers. Automatic anatomical brain mri segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–126, 2006.
9. M.P. Heinrich, O. Maier, and H. Handels. Multi-modal multi-atlas segmentation using discrete optimisation and self-similarities. In O. Goksel, O.A. Jiménez-del Toro, A. Foncubierta-Rodríguez, and H. Müller, editors, *Proceedings of the VISCERAL Challenge at ISBI*, volume 1390, page 27, Apr 2015.
10. O.A. Jiménez del Toro, Y. Dicente Cid, A. Depeursinge, and H. Müller. Hierarchic anatomical structure segmentation guided by spatial correlations (anatseggspac): Visceral anatomy3. In O. Goksel, O.A. Jiménez-del Toro, A. Foncubierta-Rodríguez, and H. Müller, editors, *Proceedings of the VISCERAL Challenge at ISBI*, volume 1390, pages 22–66, http://ceur-ws.org, Apr 2015. CEUR-WS.
11. O.A. Jiménez del Toro, O. Goksel, B. Menze, H. Müller, G. Langs, M.A. Weber, I. Eggel, K. Gruenberg, M. Holzer, A. Jakab, G. Kotsios-Kontokotsios, M. Krenn, T. Salas Fernandez, R. Schaer, T. Abdel Aziz, M. Winterstein, and A. Hanbury. Visceral-visual concept extraction challenge in radiology: {ISBI} 2014 challenge organization. In O. Göksel, editor, *Proceedings of the VISCERAL Challenge at ISBI, CEUR Workshop Proceedings*, pages 6–15, 2014.
12. O.A. Jiménez del Toro and H. Müller. Hierarchic multi–atlas based segmentation for anatomical structures: Evaluation in the visceral anatomy benchmarks. In B. Menze, G. Langs, A. Montillo, M. Kelm, S. Müller, H.and Zhang, W.T. Cai, and D. Metaxas, editors, *Medical Computer Vision: Algorithms for Big Data*, volume 8848 of *Lecture Notes in Computer Science*, pages 189–200. Springer International Publishing, 2014.

13. F. Kahl, J. Alvén, O. Enqvist, F. Fejne, J. Ulén, J. Fredriksson, M. Landgren, and V. Larsson. Good features for reliable registration in multi-atlas segmentation. In O. Goksel, O.A. Jiménez-del Toro, A. Foncubierta-Rodríguez, and H. Müller, editors, *Proceedings of the VISCERAL Challenge at ISBI*, volume 1390, pages 12–17, Apr 2015.
14. R. Kéchichian, S. Valette, M. Sdika, and M. Desvignes. Automatic 3d multiorgan segmentation via clustering and graph cut using spatial relations and hierarchically-registered atlases. In B. Menze, G. Langs, A. Montillo, M. Kelm, S. Müller, H.and Zhang, W.T. Cai, and D. Metaxas, editors, *Medical Computer Vision: Algorithms for Big Data*, volume 8848 of *Lecture Notes in Computer Science*, pages 201–209. Springer International Publishing, 2014.
15. J. Kittler and F.M. Alkoot. Sum versus vote fusion in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(1):110–115, 2003.
16. J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.
17. S. Klein, M. Staring, K. Murphy, M. Viergever, and J.P.W. Pluim. Elastix: a toolbox for intensity-based medical image registration. *IEEE Transactions on Medical Imaging*, 29(1):196–205, 2010.
18. T.R. Langerak, U.A. Van der Heide, A.N.T.J. Kotte, M.A. Viergever, M. Van Vulpen, and J.P.W. Pluim. Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (simple). *IEEE Transactions on Medical Imaging*, 29(12):2000–2008, 2010.
19. G. Langs, A. Hanbury, B. Menze, and H. Müller. Visceral: Towards large data in medical imaging—challenges and directions. In *Medical Content-Based Retrieval for Clinical Decision Support*, pages 92–98. Springer Berlin Heidelberg, 2013.
20. X. Li, C. Huang, F. Jia, Z. Li, C. Fang, and Y. Fan. Automatic liver segmentation using statistical prior models and free-form deformation. In B. Menze, G. Langs, A. Montillo, M. Kelm, S. Müller, H.and Zhang, W.T. Cai, and D. Metaxas, editors, *Medical Computer Vision: Algorithms for Big Data*, volume 8848 of *Lecture Notes in Computer Science*, pages 181–188. Springer International Publishing, 2014.
21. T. Rohlfing, R. Brandt, R. Menzel, and C.R. Maurer. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *NeuroImage*, 21(4):1428–1442, 2004.
22. F. Roli, J. Kittler, G. Fumera, and D. Muntoni. An experimental comparison of classifier fusion rules for multimodal personal identity verification systems. In *Multiple Classifier Systems*, pages 325–335. Springer, 2002.
23. A.B. Spanier and L. Joskowicz. Rule-based ventral cavity multi-organ automatic segmentation in ct scans. In B. Menze, G. Langs, A. Montillo, M. Kelm, S. Müller, H.and Zhang, W.T. Cai, and D. Metaxas, editors, *Medical Computer Vision: Algorithms for Big Data*, volume 8848 of *Lecture Notes in Computer Science*, pages 163–170. Springer International Publishing, 2014.
24. C. Studholme, D.L.G. Hill, and D.J. Hawkes. An overlap invariant entropy measure of 3d medical image alignment. *Pattern recognition*, 32(1):71–86, 1999.
25. C. Wang and O. Smedby. Automatic multi–organ segmentation using fast model based level set method and hierarchical shape priors. In O. Goksel, O.A. Jiménez-del Toro, A. Foncubierta-Rodríguez, and H. Müller, editors, *Proceedings of the VISCERAL Challenge at ISBI*, volume 1194, pages 25–31, 2014.
26. S.K. Warfield, K.H. Zou, and W.M. Wells. Validation of image segmentation and expert quality with an expectation-maximization algorithm. In *Medical Image Computing and Computer-Assisted Intervention*, pages 298–306. Springer, 2002.