www.visceral.eu

Prototype of Silver Corpus Merging Framework

Deliverable number	*D3.3*
Dissemination level	*Public*
Delivery data	*30.4.2014*
Status	*Final*
Authors	*Markus Krenn, Allan Hanbury, Georg Langs*

This project is supported by the European Commission under the Information and Communication Technologies (ICT) Theme of the 7th Framework Programme for Research and Technological Development.

Executive Summary

The VISCERAL project organizes algorithm benchmarks for researchers who develop localization-, segmentation-, detection, or retrieval algorithms in the context of medical imaging. In the first line of benchmarks, named *VISCERALanatomy* we are evaluating algorithms that localize and segment anatomical structures. A small part of the available medical imaging data comprising computed tomography, and magnetic resonance imaging data is annotated by experts. This set forms the *Gold Corpus*. Part of it is used for training, and a remaining part not available to participants is used for evaluation. The majority of the data is not annotated. Instead all participants' algorithms are applied to this data, and a *Silver Corpus* is generated by merging the estimates collected from all algorithms. This deliverable describes and evaluates different merging strategies based on data and algorithm results gathered during the first *VISCERALanatomy* benchmark.

We present results for different approaches of deriving a silver corpus labeling from algorithmic labelings, and evaluate their accuracy compared to ground-truth annotations.

Table of Contents

List of Figures

Notation

$\mathbf{I}_i$	Image or volume with index i. If it is 2D or 3D data will become clear from the context.
$\mathbf{I}_i \in \mathbb{R}^2$	2D data such as images.
$\mathbf{L}^j_{i,k}$	Label image of structure k in volume with index i from participant j.
$\mathbf{L}^A_{i,k}$	Label image of a manual annotation of structure k in volume i.
$\mathbf{L}'_{i,k}$	Label image of a silver corpus segmentation of structure k in volume i.
$\mathbf{A}_p$	Atlas with index p, the gold corpus volumes serve as atlases for atlas based label fusion.
GC	VISCERAL Gold Corpus: A set of 120 volumes from four modalities, together with 20 manually annotated anatomical structures in each volume.
SC	VISCERAL Silver Corpus: A set of 800 volumes from four modalities, together with segmentations of 20 structures which are computed from the silver corpus merging framework.
Ct-Wb	Computed tomography whole body image.
Ctce-ThAb	Contrast enhanced computed tomography image image covering Thorax and Abdomen.
OLWV	Organ Level Weighted Voting

Abbreviations

SIMPLE	Selective and Iterative Method for Performance Level Estimation [3]
STAPLE	Simultaneous Truth And Performance Level Estimation [4]

1 Introduction

Reference annotations by experts are an important part of developing and evaluating algorithms that segment anatomical structures, or localize landmarks in medical imaging data. Typically annotations are time consuming, costly, and are therefore only available for relatively small data sets. Sometimes multiple annotations are acquired for the same data, in order to estimate the reliability and accuracy if the annotations are used as *standard of reference*. Approaches exist to obtain a joint estimate of the true labeling from a set of labelings by independent experts. Different annotators might have different levels of expertise specific to an organ. Methods such as STAPLE [4], or SIMPLE [3] calculate a joint labeling, by fusing multiple labelings, while at the same time estimating their accuracy, and taking this into account when weighting the contributions of multiple annotators to the final estimate.

In VISCERAL we have collected a substantial amount of medical imaging data comprising both MR, and CT images. We split the data into a Gold Corpus, for which expert annotations are performed, and a Silver Corpus for which only algorithmic annotations are done. Each participant algorithm is applied to the Gold Corpus data, and the comparison with the expert annotations is the basis for the quantitative evaluation of the algorithms. In addition, all algorithms are applied to data, for which we do not have expert annotations. For these data, we build a Silver Corpus, by fusing the labelings resulting from all participant algorithms. In this deliverable we evaluate different methods to perform this label fusion as part of the Silver Corpus Merging Framework.

In Section 2 we explain the framwork, and its components, in Section 3 we detail the methods to obtain a joint label estimates from multiple participants entries. In Section 4 we outline the design of the underlying database, and in Section 5 we report an initial quantitative evaluation of the fusion methods.

2 Silver Corpus Merging Framework

The silver corpus merging framework contains six main components which are illustrated in Figure 2. The *Silver Corpus Label Fusion Server* implements and evaluates different label fusion techniques and computes a silver standard segmentation for each structure of each volume in the silver corpus data set. We use three sources of information for the label fusion:

1. **Participant segmentations:** All participants' algorithms are applied to all data in the silver corpus set. The resulting segmentations (in the present evaluation, we use algorithms submitted to *VISCERALanatomy* Benchmark 1) form the input data of the silver corpus server. Each segmentation of each organ is a voxel wise labeling of the entire image volume into foreground, and background. We will refer to one specific participant segmentation (label) image as $\mathbf{L}^{j}_{i,k}$, where $j = 1,...,M$ indicates the participant, $i = 1,...N$ denotes the volume ID and $k = 1,...S$ is the index of the anatomical structure (e.g. lungs, liver,...) of the segmentation.

2. **Participant evaluation results:** In addition to the algorithmic labelings, the Gold Corpus carries expert annotations. We use these annotations to evaluate the algorithms, and to assign corresponding structure-specific weights to each algorithm. These weights are

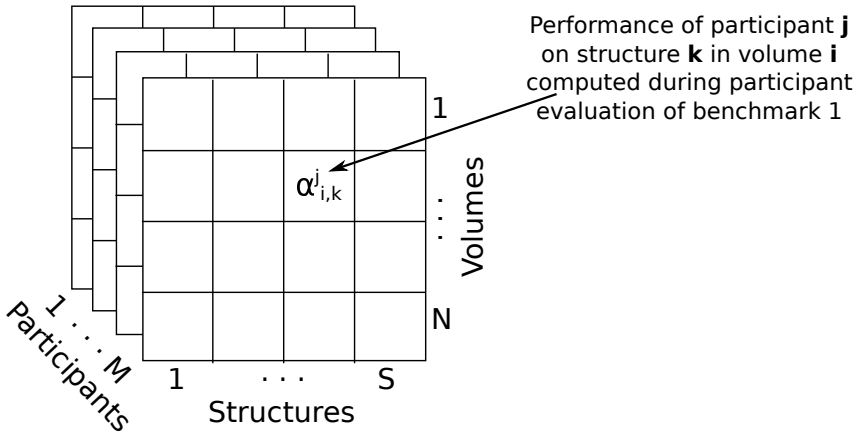

Figure 1: 3D Matrix of participant performances

then used during fusion. The results of Benchmark 1 are stored in the 3 dimensional array $\alpha^j_{i,k}$ (see Figure 1) and are used to estimate the performance of one participant for each structure in one modality (e.g. lung in CT's, heart in MR's,...). This estimate serves as basis for weighting the impact of a participants segmentation during the label fusion process.

3. **Manual annotations:** Since the benchmark is performed on a confined image domain, we can transfer information across examples to some extent. That is, in addition to using the participant algorithms on each volume, we can use the gold corpus volumes as atlases $\mathbf{A}_P$, and transfer their labels via multi-atlas label fusion to unlabeled cases. After registration of the gold corpus volumes to the silver corpus volumes, we use the transformed annotations as additional segmentation estimates. This is particularly relevant for structures where participant algorithms perform poorly. In the quantitative evaluation reported in Section 5 we use these annotations to evaluate the performances of the implemented label fusion approaches in a cross-validation scheme.

We have implemented different approaches to perform label fusion in the *Silver Corpus Label Fusion Sever*. They include:

1. **Majority vote:** For each volume, and each anatomical structure, each participant algorithm contributes one vote for the labelling of each voxel into fore- or background. This approach does not take any performance estimates into account, but treats all algorithms equally.

2. **Organ level weighted votes:** Similar to majority voting, each algorithm casts one vote for the labelling of each voxel. However, here we take the accuracy evaluation results of the algorithms on the gold corpus into account. The weight of each vote is determined by the algorithm accuracy on the gold corpus (of participants based on performance results of benchmark 1).

3. **SIMPLE:** Here we fuse the algorithm segmentations by using the SIMPLE algorithm [3] and do not take the performance of the algorithms on the gold corpus into account.

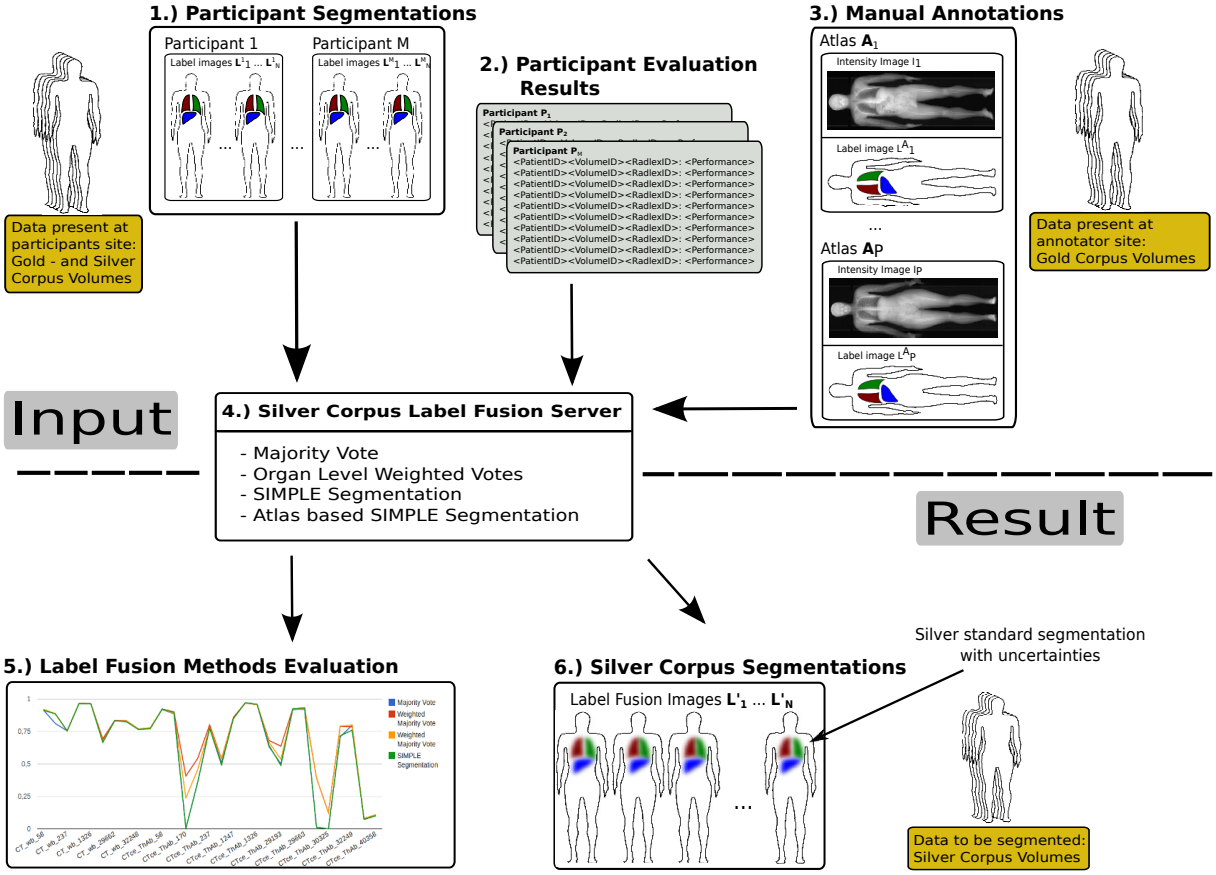

Figure 2: Components and Workflow of the Silver Corpus Merging Framework

4. **Atlas based SIMPLE:** Similar to SIMPLE, but we take both algorithmic segmentations on the volume, as well as labels transferred from gold corpus volumes into account, when estimating the segmentation. Again we use the SIMPLE algorithm to fuse the segmentations.

After the evaluation of the implemented label fusion techniques on the gold corpus volumes, we will use the best performing approach to compute the silver standard segmentation (label) images $\mathbf{L}'_{i,k}$ of all structures in all modalities in the silver corpus data set.

In the following we describe the two kinds of label fusion processes relevant in the framework: (1) fusing multiple segmentations in one volume, and (2) fusing segmentations across different volumes. The result of the label fusion process is the silver corpus segmentation of each volume.

2.1 Label fusion within a volume

Given a target image $\mathbf{I}_i$, a target structure k and the set of M participant segmentations $\mathbf{L}^j_{i,k}$ of the the structure in the image where $j = 1...M$, the aim of label fusion is to estimate a more accurate segmentation $\mathbf{L}'_{i,k}$, of the hidden ground truth segmentation by fusing the M participant

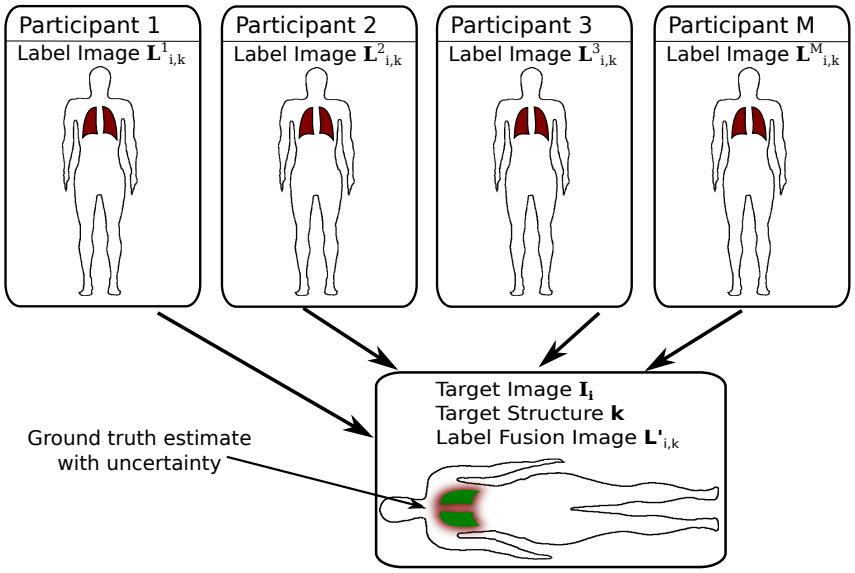

Figure 3: Illustration of label fusion within one image

segmentations $\{\mathbf{L}^1_{i,k},\ldots,\mathbf{L}^M_{i,k}\}$. Figure 3 shows the principle of the label fusion process where each participating segmentation is a segmentation of the same image, and the same organ. For each voxel in $\mathbf{I}_i$ we have to estimate if it is foreground, or background based on the set of estimates of the algorithms, and our quality assessment of each algorithm, that can influence the weight we assign to its voxel labels.

2.2 Atlas based Label Fusion

Again, we want to find the segmentation $\mathbf{L}'_{i,k}$ of structure k in a target image $\mathbf{I}_i$. Instead of using participant annotations in the target image, we perform atlas based label fusion to estimate the hidden ground truth segmentation. We fuse multiple segmentations of the same structure that are available in other images. We refer to the segmentations in other images as atlases. Here we have to solve three tasks.

1. **Registration** First, we establish correspondence between the target image $\mathbf{I}_i$, and all atlas images $\{\mathbf{I}^A_1,\ldots,\mathbf{I}^A_P\}$. This is done by performing non-rigid image registration.

2. **Label transfer:** After correspondences have been established in the form of an image transform $\mathbf{T}_{i\to p}$, we can transfer the labels from the atlas images to the target images, by assigning each voxel $\mathbf{x}$ in the target image the label value in the atlas image at the position $\mathbf{T}_{i\to p}(x)$.

3. **Label fusion:** Finally, after we have label estimates for all voxels, from all atlas images, we fuse the labels analogously to the label fusion within the volume.

An efficient approach of pre-registration atlas selection and label propagation has been introduced in Section 3 of the VISCERAL Deliverable 3.1. After selecting P relevant atlases $\mathbf{A}_P$

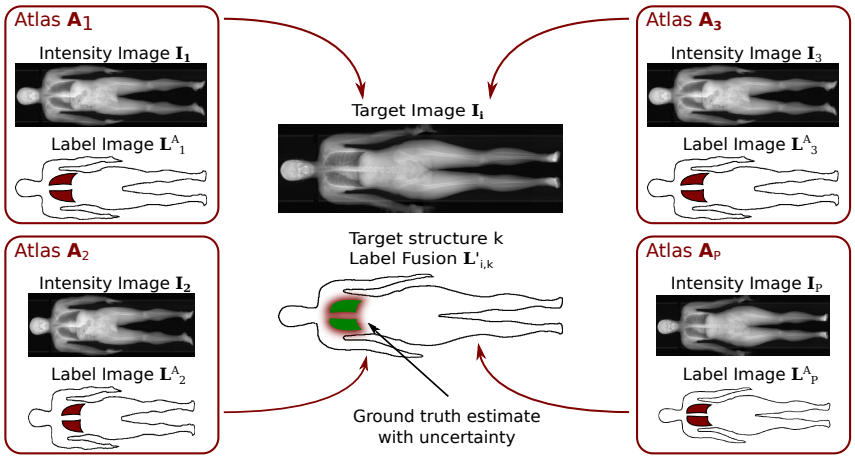

Figure 4: Atlas based label fusion, see also Deliverable D3.1

we transform their label images to the target image and perform label fusion to estimate the hidden ground truth segmentation $\mathbf{L}'_{i,k}$. The process of atlas based label fusion is illustrated in Figure 4. While one can use both participant segmentations, and ground truth annotations as atlas labels, in our experiments, we only evaluate ground-truth annotations in the gold-corpus as atlas labels.

3 Label fusion methods

The following section describes and discusses different label fusion methods that have been implemented and evaluated from the visceral silver corpus server. As described in the previous section, a label fusion algorithm seeks to merge M segmentation estimates or binary label images $\mathbf{L}^j_{i,k}, j = 1...M$ of a structure k in a target image i in order to compute an improved estimation of the hidden ground truth segmentation $\mathbf{L}'_{i,k}$.

3.1 Majority vote

Majority vote counts for each voxel $\mathbf{x}$ how many segmentation estimates vote for the presence (value 1) or absence (value 0) of the structure that has to be segmented. Label fusion is done by assigning a voxel $\mathbf{x}$ of the ground truth estimate $\mathbf{L}'_{i,k}(\mathbf{x})$ with value 1 if the majority of the participating segmentation estimates have voted for $\mathbf{x}$ to be foreground [1].
The label fusion server collects all participant segmentations $\mathbf{L}^j_{i,k}$ available for image i, and computes the estimated hidden ground truth as follows:

$$\mathbf{L}'_{i,k}(\mathbf{x}) = \left\{ \begin{array}{ll} 1, & \left(\sum_{j=1}^{M} \mathbf{L}^j_{t,s}(\mathbf{x})\right) \geq \frac{M}{2} \\ 0, & \left(\sum_{j=1}^{M} \mathbf{L}^j_{t,s}(\mathbf{x})\right) < \frac{M}{2} \end{array} \right\} \quad (1)$$

3.2 Organ level weighted voting

Here, we are given M binary label estimates $\mathbf{L}^j_{t,s}$ and a vector $\phi = (\phi_1, ..., \phi_M)$ holding weights $\phi_j \in (0,1)$ which specify the influence of the single estimates to the resulting ground truth fusion $\mathbf{L}'_{i,k}$. Even though the method is originally called global weighted voting because the weights are assigned to the entire region of one label volume, we will refer to it as Organ Level Weighted Voting (OLWV) since one label volume covers one and only one anatomical structure of the human body. Taking the additional information into account, label fusion assigns the fused ground truth estimates $\mathbf{L}'_{i,k}(\mathbf{x})$ as follows [1]:

$$\mathbf{L}'_{i,k}(\mathbf{x}) = \left\{ \begin{array}{ll} 1, & \left(\sum_{j=1}^{M} \mathbf{L}^j_{i,k}(\mathbf{x}) \cdot \phi_j\right) \geq \frac{\sum \phi_j}{2} \\ 0, & \left(\sum_{j=1}^{M} \mathbf{L}^j_{i,k}(\mathbf{x}) \cdot \phi_j\right) < \frac{\sum \phi_j}{2} \end{array} \right\} \tag{2}$$

This means each ground truth estimate $\mathbf{L}^j_{i,k}$ is weighted by the performance parameter ϕ_j holding a value that defines the confidence in an annotator's decision.

The silver corpus label fusion server implements and evaluates three different weighting functions which are applied to global weighted voting:

1. **Based on segmentation performance:** Estimates the weight ϕ_j for each segmentation $\mathbf{L}^j_{i,k}$ by computing the Dice coefficient [2] with the manual annotated ground truth segmentation $\mathbf{L}^A_{i,k}$.

$$\phi_j = \frac{2|\mathbf{L}^j_T \cap \mathbf{L}^A_{i,k}|}{|\mathbf{L}^j_T| + |\mathbf{L}^A_{i,k}|} \tag{3}$$

2. **Based on average structure performance:** Estimating ϕ_j for each segmentation $\mathbf{L}^j_{i,k}$ by computing the mean performance of participant j on structure k over all volumes.

$$\phi_j = \frac{\sum_{i=1}^{N} \alpha^j_{i,k}}{N} \tag{4}$$

3. **On top k-ranked segmentations** Computes ϕ_j as described in Equation 4 and takes only the top k-ranked segmentations and their weights into account, where

$$k = max(2, \frac{M}{3}) \tag{5}$$

3.3 SIMPLE Segmentation

Selective and Iterative Method for Performance Level Estimation (SIMPLE) proposed by Langerak et. al. [3] is based on an iterative strategy that alternates on

1. Estimating the hidden ground truth segmentation of the target image
2. Estimating the performances of the contributing segmentations

The actual label fusion component is an interchangeable component in the SIMPLE method. This allows for a comparison of different methods and if necessary a fine tuning of label fusion to specific tasks. Another advantage of the method is that it can be applied for atlas based segmentation label fusion as well as the combination of multiple segmentations of the same image [3].

Algorithm description of SIMPLE

- **Input:** A set of M estimates $\mathbf{L}_{i,k}^{j}$ for the true segmentation $\mathbf{L}_{i,k}$
- **Compute performance estimates and label fusion:**
  1. Combine all available segmentations $\mathbf{L}_{i,k}^{j}$ to get an initial estimate for the ground truth $\mathbf{L}_{i,k}'$ using for instance majority voting or weighted voting.
  2. Estimate the performance ϕ_j for each segmentation $\mathbf{L}_{i,k}^{j}$ by computing a binary overlap measure (e. g. Dice coefficient [2], see equation 4) with the initial ground truth estimation $\mathbf{L}_{i,k}'$.
  3. Identify badly performing segmentations based on the estimated performances by applying a performance level threshold θ that is chosen a priori.
  4. Exclude all badly performing segmentations ($\phi_j < \theta$) and compute an update of the fused ground truth $\mathbf{L}_{i,k}'$ based on the remaining segmentation estimates and their estimated performances. For this purpose the performance weighted label fusion described in Section 3.2 is applied.
  5. Re-estimate the performances ϕ_j of the single segmentations based on the updated ground truth estimate.
  6. Re-consider early discarded segmentations in the first k iterations of the procedure. This allows discarded segmentations to get back in the label fusion process after updating the ground truth estimate.
  7. Iterate ground truth and performance estimation until convergence (no changes in the considered atlases and their performances).
- **Output:** A set of M estimated performances ϕ_j, a set of selected segmentations $\mathbf{L}_{i,k}^{j}$, as well as their fusion $\mathbf{L}_{i,k}'$. This fusion is defined as estimate for the real hidden ground truth segmentation $\mathbf{L}_{i,k}$.

The silver corpus server implements and evaluates two different versions of SIMPLE segmentation as follows:

1. **Participant segmentation based SIMPLE segmentations**
   This configuration takes all participant segmentation $\mathbf{L}_{i,k}^{j}$ of structure k in target image t as input for the SIMPLE algorithm. The system is designed to run the SIMPLE segmentation with two different label fusion configurations:
   (a) *Majority Vote:* Each segmentation has the same weight, see Section 3.1

(b) *Organ Level Weighted Voting:* Each segmentation is weighted according to its participants average performance on a structure, see Section 3.2

2. **Atlas and participant segmentations based SIMPLE segmentation**
This configuration merges all participant segmentations $\mathbf{L}^{j}_{i,k}$ of structure k in target image i as well as manual annotations of all available atlas images $\mathbf{A}_P$ as input for the SIMPLE algorithm. Please note that we do not have an estimate performance of one atlas, which leads to Majority Voting as label fusion type.

4 Database Design

This section describes the underlying database design and its key components in the silver corpus merging framework. Figure 5 illustrates the enhanced entity-relationship model of the data base.

- **Volume** holds both silver and gold corpus Volumes. A volume is defined by its *VolumeID* and *PatientID*. *Modality* and *Bodyregion* are stored with the purpose of evaluating performances of both annotators and participants.
- **ManualAnnotation** contains manual segmentations of the gold corpus volumes. One annotation is defined by *VolumeID, PatientID, RadlexID, AnnotatorID*.
- **Structure** stores the anatomical structures which are annotated by annotators or segmented by participants. The *RadlexID* identifies an anatomical structure.
- **ParticipantSegmentation**
is defined by *VolumeID,PatientID,RadlexID,ParticipantID,configuration*. Each participant has been able to upload five different configurations of his algorithm, which makes it necessary to differentiate between those configurations in order to evaluate the best performing configuration for each structure. The field *Performance* holds information about the performance of the segmentation in relation to the ground truth manual annotation of the segmented structure. This value is used to compute the weight of a segmentation during the label fusion process.
- **RegistrationResults** is defined by its source and target volume. Points to a file containing the results of affine and non-rigid image alignment. Registration results are required for the purpose of atlas based label fusion.
- **LabelFusionType** holds the different types of label fusion methods that are implemented from the framework. This table is required to evaluate the performance of the different label fusion types.
- **SilverCorpusSegmentation** a silver corpus segmentation is defined by its *VolumeID, PatientID, RadlexID, LabelFusionType*. The field performance is used to store the performance of the segmentation in relation to the manual annotation (gold standard). Filename points to the segmentation file, whereas LabelFusionResults holds information about the

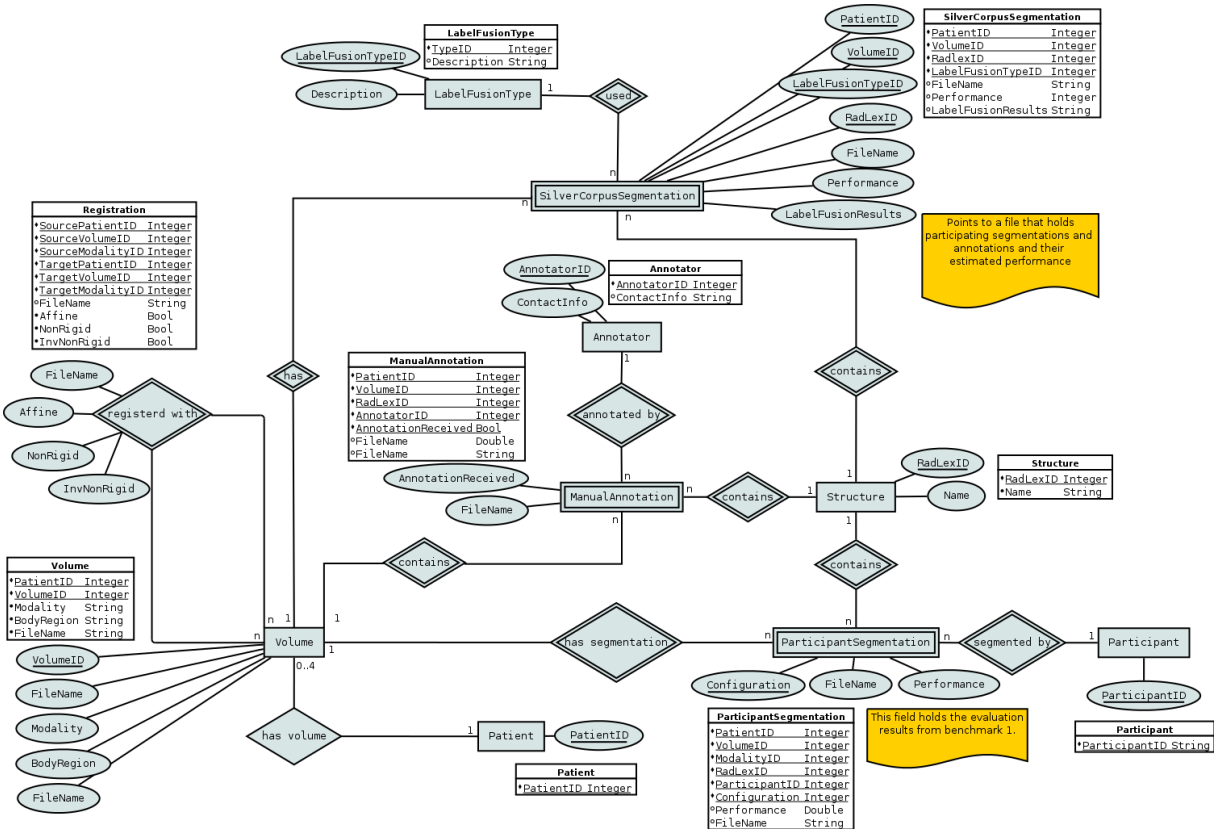

Figure 5: EER Diagram of the data base

participating segmentations (i.e. which participants, which atlas volumes have been involved in the label fusion process).

5 Evaluation and results

The following sections describe the data used in order to evaluate the described label fusion approaches as well as the evaluation process and shows initial results.

5.1 Data in use

Table 1 gives an overview of the submitted participant segmentations that have been available to evaluate the label fusion methods. Please note that only structures with at least two submitted participant segmentations of one volume have been used to evaluate the label fusion performances of the described methods.

5.2 Evaluation of label fusion algorithms

For the evaluation purpose we computed 331 segmentations with each of the proposed label fusion methods covering 20 structures in 2 modalities (CT-Wb and Ctce-ThAb). Segmentation

Table 1: Submitted participant segmentations

RadlexID	CT-wb	CTce-ThAb	MRT1-wb	MRT1cefs-Ab	∑
58	167	339	13	9	**528**
86	76	239	14	9	**338**
170	15	36		6	**84**
187	10	109		4	**123**
237	75	226	14	10	**325**
480	15	79	14		**108**
1247	15	146	14		**175**
1302	77	243	14		**334**
1326	77	243	14		**334**
2473	77	140			**217**
7578	12	19	7		**38**
29193	15	141	12	6	**174**
29662	77	240	14	8	**339**
29663	75	240	13	9	**337**
30324	7	24	4		**35**
30325	9	33	4		**46**
32248	66	139			**205**
32249	77	151	14	8	**250**
40357		60			**60**
40358		67			**67**
∑	**942**	**2941**	**165**	**69**	**4417**

performance is measured using the Dice coefficient [2] which measures the overlap of a computed segmentation and its ground truth, the manual expert annotation. Figures 6 and 7 show segmentation performances of a representative subset of structures in both modalities of the following segmentation methods:

1. **Best performing participant:** Includes all results of the best performing participant of the target structure. Note that this is for comparison only, since this information is not available in practice.

2. **All participants:** Includes all participant results of each structure.

3. **Majority Vote:** Label fusion method as described in Section 3.1.

4. **OLWV - GT Performance:** Organ Level Weighted Voting, weights are derived from a segmentations overlap with its ground truth annotation. This is not feasible in practice since it uses information from the ground truth on the cases segmented. It is only included as a reference. See Section 3.2.

5. **OLWV - Structure Performance:** Organ Level Weighted Voting, weights are derived from the participants mean performance on the target structure, but only other volumes are taken into account. This is a correct simulation of what is feasible during silver corpus generation, since there we can use performance measures on the gold corpus, but not on the silver corpus. See Section 3.2.

6. **OLWV - Top k-ranked:** Organ Level Weighted Voting as in item 5, only considering the top k-ranked participants instead of all participants when building the weighted consensus, see Section 3.2.

7. **SIMPLE - MajVote:** SIMPLE segmentation where the initial ground truth estimate is computed using Majority Vote, see Section 3.3.

8. **SIMPLE - OLWV:** SIMPLE segmentation where the initial ground truth estimate is derived from Organ Level Weighed Majority vote, see also Section 3.3.

9. **Atlas based SIMPLE:** SIMPLE segmentation combining participant segmentations as well as manual annotations from the gold corpus volumes, see also Section 3.3.

Please note that the groups 1 and 2 depict performances from single participant segmentations, which are included to illustrate the benefit of using label fusion methods while segmenting a single structure. Also note that label fusion method 4 *OLWV - GT Performance*, is only available if there exists ground truth annotation for each single structure and will thus not be used for the segmentation of volumes from the silver corpus.

5.2.1 Results

The results shown in Figure 6 indicate that the proposed label fusion methods perform similar in structures where the overall performance of participant segmentations is consistent. Label fusion methods can decrease the influence of poorly performing outliers.

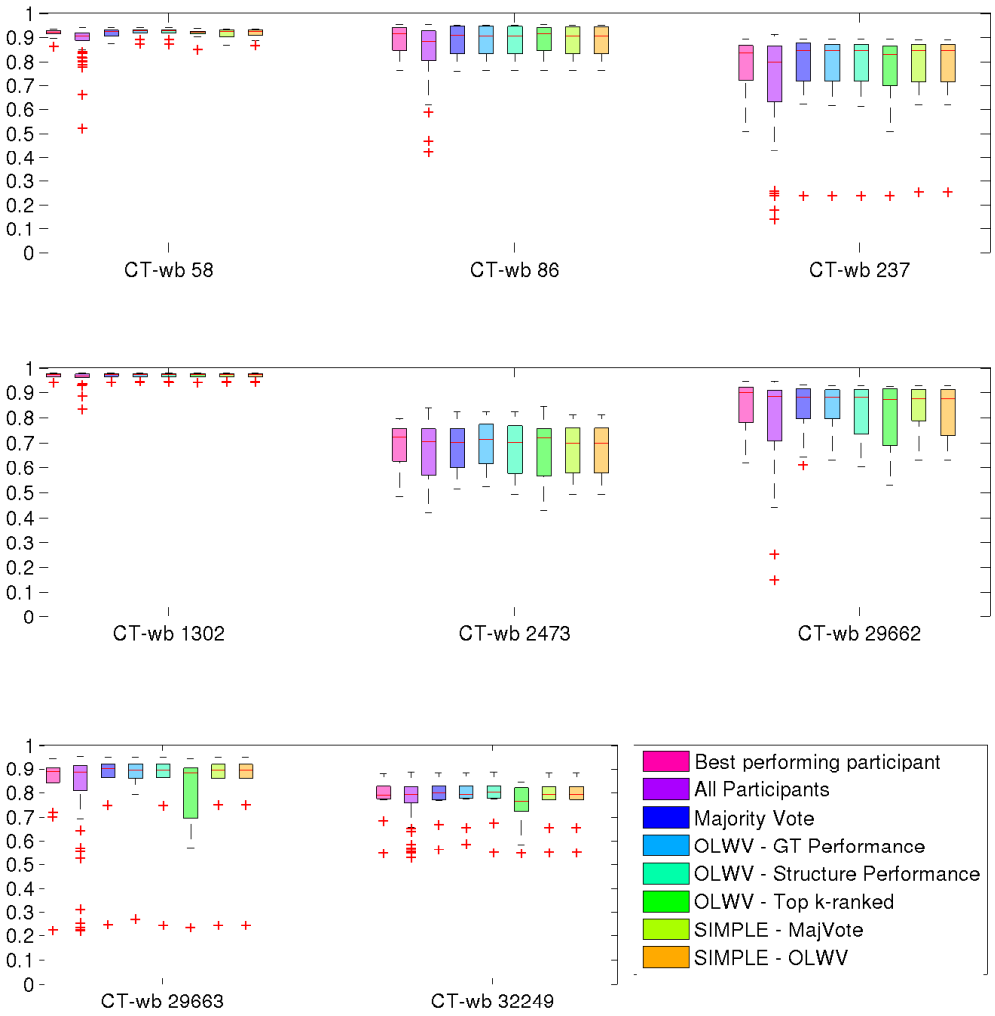

Figure 6: Label fusion performances of structures in whole body CT volumes

The results in Figure 7 show the differences of the approached methods on structures where the overall segmentation performance is not as good (structures 29193, 40357, 480). Top k-ranked OLWV not only outperforms the average participants performance and global weighted voting, but as well the approaches that implement SIMPLE segmentation. This indicates that the accuracy scores of one participant on one specific structure (e.g. lungs) in one specific modality (e.g. Ctce-ThAb) are predictive for the scores on the same structure in other volumes. Additionally top-k ranked OLWV discards poorly performing segmentations whereas the implemented SIMPLE algorithms take those into account to estimate the initial ground truth performance. Overall, building the consensus from top ranked participants typically outperforms other approaches, and the performance of the method on the gold-corpus is a good estimator for the performance on volumes for which no ground-truth annotation is available.

We also studied how label transfer across volumes can be used to estimate segmentations, and how this can be combined with native algorithm segmentations. To this end we evaluated one method that combines manual annotations from images other than the target image with the participant segmentations (*Atlas based SIMPLE)*. Please note that this method has been evaluated on only two images and on a subset of all structures. The results indicate that the combination of atlas based segmentation and the fusion of participant segmentations can increase the performance of non atlas based approaches especially in structures where the overall participant segmentation performance is poor, see also Figure 7, structures (237, 29193, 40357).

Figure 8 illustrates the average performance of the label fusion methods over all structures. Top k-ranked OLWV label fusion (green line) shows the most promising results on structures with good overall performances and is as well robust against outliers (see structure CTce-ThAb-480). The Figure shows as well that Atlas based SIMPLE segmentation is capable to outperform top k-ranked OLWV on structures with bad overall performance.

Figure 9 provides an example of a computed silver corpus segmentation. It shows four selected liver segmentation submissions for one contrast enhanced CT volume, together with its ground truth manual annotation (top right corner) and its label fusion segmentation (method in use: Top k-ranked OLWV).

6 Conclusion

This document describes the VISCERAL Silver Corpus Merging Framework Prototype and presents initial quantitative evaluations of different label fusion algorithms, that will be used in building the silver corpus. This framework calculates label estimates for images from participant algorithm results in those cases, where no expert annotation is available. The images together with the resulting segmentations form the Silver Corpus. In this deliverable, we detail the algorithms available in the framework, and report initial evaluation results for different label fusion approaches. The framework uses three sources of information to estimate the labels. First the participant algorithm segmentations, second the accuracies of these algorithms when compared to Gold Corpus annotations, and finally the Gold Corpus annotations themselves. Future work will improve the methods based on including the full data set labeled in the VISCERAL anatomy challenge.

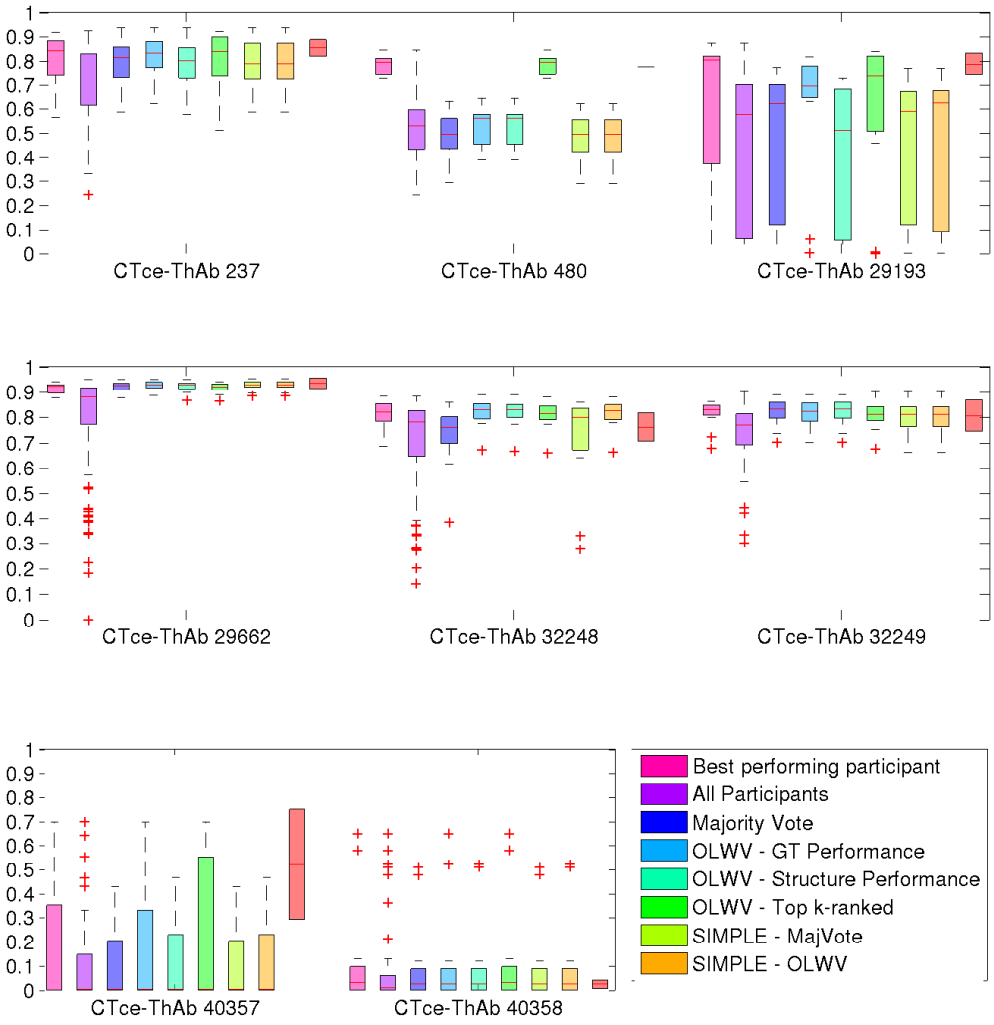

Figure 7: Label fusion performances of a representative set of structures in CTce-ThAb volumes.

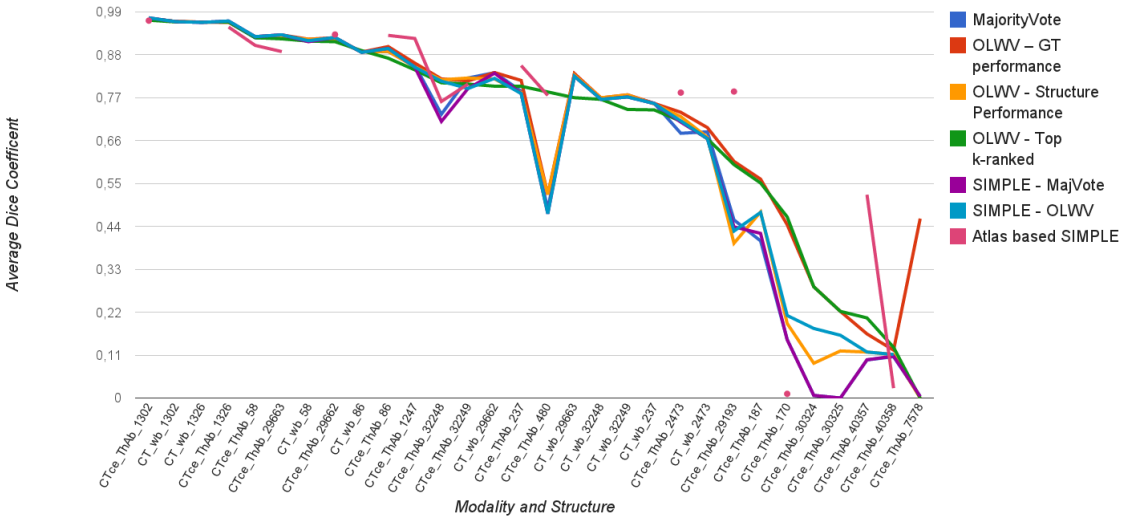

Figure 8: Average label fusion performances

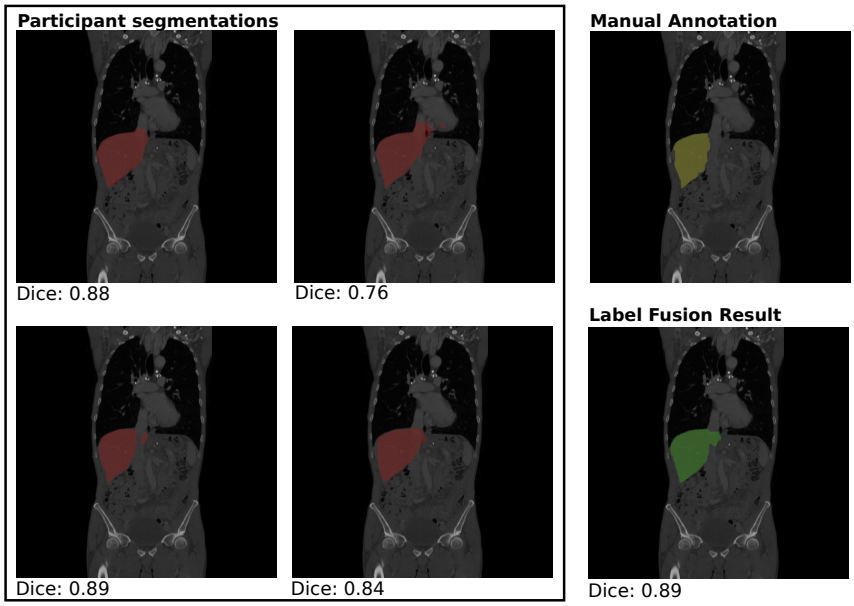

Figure 9: Participant segmentation and label fusion results.

7 References

[1] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de Solorzano. Combination strategies in multi-atlas image segmentation: Application to brain mr data. *Medical Imaging, IEEE Transactions on*, 28(8):1266–1277, 2009.

[2] Lee R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302, July 1945.

[3] 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). *Medical Imaging, IEEE Transactions on*, 29(12):2000–2008, 2010.

[4] Simon K. Warfield, Kelly H. Zou, and William M. Wells. Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.