www.visceral.eu

Prototype of gold corpus active annotation framework

Deliverable number	*D3.1*
Dissemination level	*Public*
Delivery data	*30.4.2013*
Status	*Final*
Authors	*Matthias Dorfer, Georg Langs, Oscar Jimenez, Henning Mueller*

SEVENTH FRAMEWORK PROGRAMME

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

This deliverable document describes the gold corpus annotation framework for the first VISCERAL benchmark in detail. VISCERAL aims at distributing a substantial amount of medical imaging data together with high quality expert annotations, the *gold corpus*. Part of the gold corpus is distributed to the participantes for algorithm development and training. The other part of the gold corpus is held back, and serves as basis for evaluation of participant algorithms. The objective of the gold corpus annotation framework is to ensure high quality annotations, optimal use of annotation resources, and an volume- and organ specific estimate of annotation reliability. In this document we describe the main components of the system in detail.

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.

Abbreviations

PACS	Picture archiving and communication system
STAPLE	Simultaneous Truth and Performance Level Estimation [10]
MDS	Multi Dimensional Scaling
SIMPLE	Selective and Iterative Method for Performance Level Estimation [5]

1 Introduction

VISCERAL aims at releasing a substantial amount of medical imaging data together with expert annotations to foster research, and comparability of methodology. Part of the available data is annotated by experts forming the *gold corpus*. The second, and far larger part is only assessed by algorithms developed by the participants, The best estimate of the truth based on those algorithm results forms the *silver corpus*. In this deliverable we describe the framework, algorithms and procedure for gold corpus annotation.

A key to optimal use of the data is the accurate annotation, quality control and choice of annotated examples. The framework ensures the following aspects during gold corpus annotation:

1. **Estimate annotation confidence of annotated volumes** The system estimates the confidence of annotations for each organ, and each example volume that has been already annotated.

2. **Estimate confidence of labels propagated to non-annotated examples** The system uses label fusion to propagate labels to volumes, not yet annotated. It uses the transferred labels to estimate how well the not yet annotated example is represented by the already annotated volumes.

3. **Choice of examples to annotate** Based on the previous confidence estimate, the system ranks volumes, and organs. The ranking indicates the data for which annotation would add maximum information to the annotated set. This ensures optimal use of the annotation resources, since only a small part of the overall data can be annotated. Note that the volumes and organs that are marked for future annotation, can include volumes and organs that have been annotated already, if the reliability estimate of the existing annotation is not satisfactory.

4. **Matching of annotators and organs according to their previous annotation accuracy** Based on confidence estimates that are organ-, volume- and annotator specific, annotation requests are optimized by assigning those annotators to specific organs, that obtained maximum reliability score in the already annotated corpus.

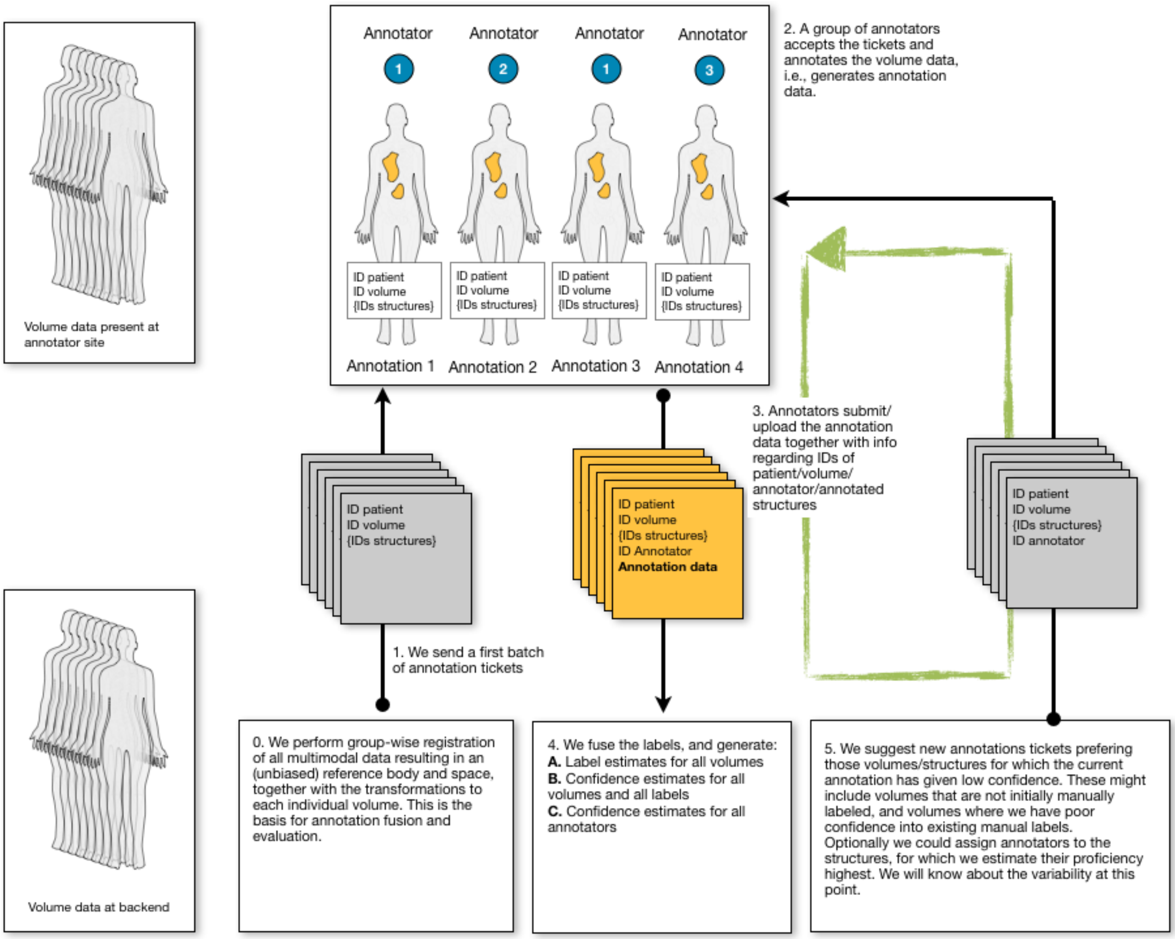
5. **Quality control** Reliability estimates also serve as a quality control for the annotations.

In this document we explain the framework, and underlying algorithms in detail.

2 Annotation Framework Overview

The framework scheme is illustrated in Figure 1. Data to annotate is available at both the annotator site, and the site of the annotation backend. Annotators and Backend are connected by an online interface that transfers annotation requests (*tickets*) from the backend to the annotators and annotations from the annotators to the backend.

In the first phase the backend analyses all image data, and non-linearly aligns the data to a common reference space, i.e., performs group-wise registration of the data set. After this initial phase, the system chooses an initial set of volumes for which annotations of all organs are



2. A group of annotators accepts the tickets and annotates the volume data, i.e., generates annotation data.

3. Annotators submit/ upload the annotation data together with info regarding IDs of patient/volume/ annotator/annotated structures

1. We send a first batch of annotation tickets

0. We perform group-wise registration of all multimodal data resulting in an (unbiased) reference body and space, together with the transformations to each individual volume. This is the basis for annotation fusion and evaluation.

4. We fuse the labels, and generate:
A. Label estimates for all volumes
B. Confidence estimates for all volumes and all labels
C. Confidence estimates for all annotators

5. We suggest new annotations tickets prefering those volumes/structures for which the current annotation has given low confidence. These might include volumes that are not initially manually labeled, and volumes where we have poor confidence into existing manual labels. Optionally we could assign annotators to the structures, for which we estimate their proficiency highest. We will know about the variability at this point.

Figure 1: Overview of the annotation framework. Annotation tickets are generated by the backend based on already annotated data, and the estimated information likely to be gained by a specific annotated organ in a specific case. Annotators annotate organs, and send by the annotation files via a web interface, triggering a regeneration of new tickets.

requested. Annotators obtain the tickets corresponding to these requests. Each ticket contains the specification of three IDs: the volume, the organ, and the annotator. At this stage annotators are assigned randomly.

The annotators annotate the data corresponding to the tickets, and upload the annotation files (nii.gz files) via the web-interface. They are sent to the backend together with corresponding volume-, organ-, and annotator IDs.

The backend collects the annotations, and propagates them to all volumes in the entire data set. Then, it estimates the reliability of both the expert annotations, and the propagated labels. Finally it ranks the volumes and organs corresponding to this reliability, and generates annotation tickets for those with the lowest reliability.

In the following we will describe the annotation backend (Section 3), the transfer framework (Section 4) in detail. The annotation interface will be described briefly in Section 5. However it is subject of deliverable D1.1.

3 Annotation Backend

Given a set of multi-modal medical images or volumes $\mathbf{I}_i$ with $i = 1, ..., N$ either in $\mathbb{R}^2$ or $\mathbb{R}^3$. In the reminder of the deliverable both images and volumes are denoted as images for a consistent naming. The global goal of the active annotation backend is to provide in all N images the correct annotations (labels) $\mathbf{L}_{i,s}$ for all $s = 1, ..., S$ selected anatomical structures (e.g. lung, liver, ...). To achieve this ground truth (gold standard) segmentation, we have $j = 1, ..., M$ annotators available, who provide us with manual segmentations $\mathbf{L}^j_{i,s}$ (estimates for the real ground truth) for each structure contained in the images $\mathbf{I}_i$.

The following subsections deal with the single processing steps of the active annotation backend proposed in the previous section in detail:

1. Group-wise registration and reference space (template) computation
2. Selection of an initial batch of volumes for manual annotation
3. Annotation propagation, label fusion, and performance estimation
4. Reliability estimation and annotator ranking

3.1 Group-wise registration

For the purpose of initial set selection (Subsection 3.2) and pre-registration atlas selection (Subsection 3.3.1) a central template $\mathbf{T}$ is required in the annotation backend. The template has to represent the entire population of images $\mathbf{I}_1, ..., \mathbf{I}_N$ and has to be as unbiased as possible. To achieve this goal we first select the subject best representing all remaining images [7] in the set. Based on this initialization we use a method described in [4] to iteratively reduce bias. The output of this procedure is a template $\mathbf{T}$ in the center of the population with respect to tissue intensity as well as anatomical shape [4].

3.2 Selection of an initial volume set for annotation

The purpose of this subsection is the selection of an initial set of volumes which is sent to the annotators for manual segmentation. This subset of all available images has to represent the entire population of images (subjects) to ensure optimal label propagation in the remaining steps of the annotation framework. Based on the number N of available images we suggest two different methods for the choice of the initial set. Both methods are supported by available meta data such as age, gender or pathologies [1]. The result of the selection methods is a subset of the entire population of available images to be submitted to the annotators for manual segmentation.

3.2.1 All to all registration selection

This selection method is based on the work of Park et. al. [7] who published a method for least biased target selection for anatomical atlas construction. They compute the pair-wise registrations ($O(N^2)$) between all N available images in the set. Since image registrations is computationally expensive and time consuming [6] the number of images that are concerned in this approach is limited [1]. Based on the pair-wise registration costs $d_{i,j}$ between all Images $\mathbf{I}_i$ and $\mathbf{I}_j$ a distance matrix $\mathbf{D} = [(d_{i,j})]$ is composed. This distance matrix is transformed towards a Multi Dimensional Scaling (MDS) embedding space. The embedding illustrates the similarity structures in the underlying image set. The subject located closest to the center of the embedding space is expected to have the lowest bias with respect to the remaining population and is selected as template (atlas).

We reuse this concept in the annotation backend. However, we do not only select a single central template subject, but use all the information on anatomical shape variation in the set. We select multiple volumes by equally sub-sampling the entire MDS embedding space. The principal of this approach is depicted as Method 1 in Figure 2. Points (coordinates) and their respective Euclidean distances in the embedding space reflect the pair-wise similarities of the original input images. The closer two coordinates are placed to each other in the embedding space, the closer are their anatomies. The large green circle shows the lowest distance to all remaining images and is expected to be the least biased subject in the set [7]. The blue circles are an example for a possible subset gathered by equally sub-sampling the embedding space to cover the entire shape variation in the population. This subset is sent to the annotation experts for manual segmentation.

3.2.2 Affine template registration selection

If the number of available candidate images is to large, the pair-wise registrations can not be computed to form the distance matrix $\mathbf{D}$. In this case a method for pre-registration atlas selection published by Aljabar et. al. [1] is reused for pair-wise similarity computation. Instead of registering all N images in the set with each other, we compute for each subject a coarse affine alignment towards a single central template ($O(N)$). After template alignment all subjects live in a common central space where we compute the pair-wise similarities d_{ij} as in Method 1. This similarity information is again used to form the distance matrix $\mathbf{D}$ as well as the MDS embedding space. The remaining selection procedure is performed as described above. This principle is illustrated as Method 2 in Figure 2.

3.3 Annotation propagation and label fusion

This subsection deals with the selection of representative atlas volumes and the fusion of multiple segmentations for an anatomical structure in the same subject. The result of label fusion is taken as the estimated ground truth of the segmentation and used as basis for a ranking of the contributing annotations and the respective annotators.

3.3.1 Pre-registration atlas selection (optional)

For the propagation of existing labels in the atlases towards new subjects the pair-wise registration of the atlas images with the test subject is required. This registration yield transformations which are applied to the annotations to propagate the labels towards the respective target subjects. Since image registration is computationally expensive a pre-selection of atlas candidates is required [1]. The atlas selection in this framework is based on the available meta data such as gender, age, or pathologies and by the following approach proposed by Aljabar et. al. [1].

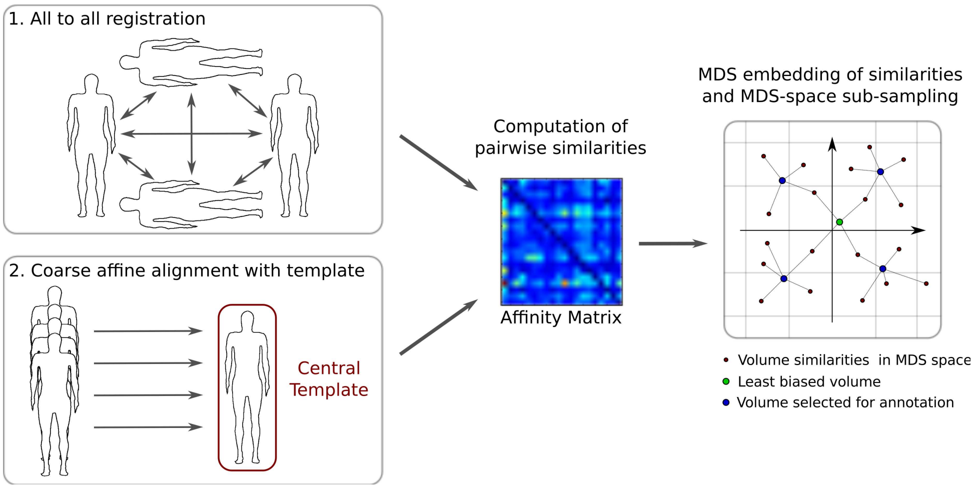

1. All to all registration
2. Coarse affine alignment with template
Central
Template
Computation of
pairwise similarities
Affinity Matrix
MDS embedding of similarities
and MDS-space sub-sampling
• Volume similarities in MDS space
● Least biased volume
● Volume selected for annotation

Figure 2: Overview of the selection of an initial set of volumes for annotation.

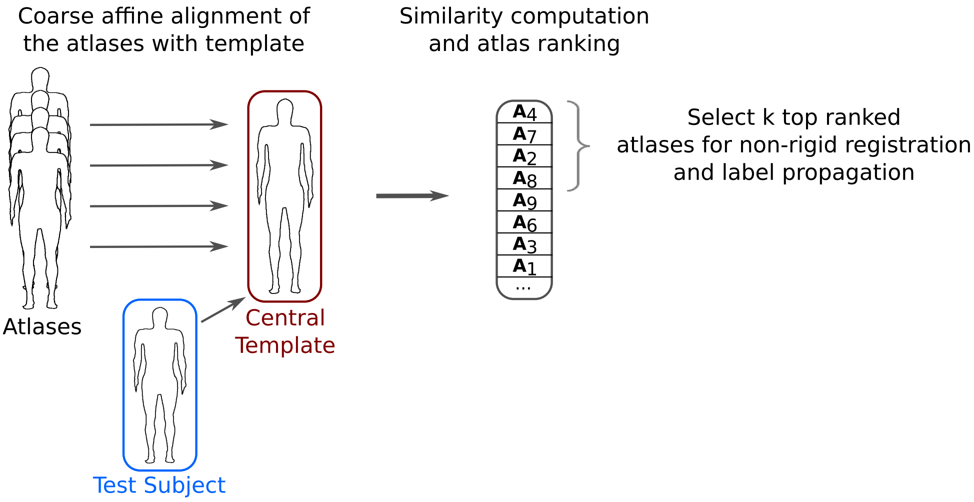


Figure 3: Overview of pre-registration atlas selection.

We take an initial template with minimal bias **T** and compute an coarse affine alignment of all available atlas images with this template. This is performed analogously to the selection of an initial set of volumes for annotation. The number of affine registrations required is linear in the number of available atlases and is computed offline. If a novel test image has to be annotated, we compute as with the atlas volumes its affine alignment with the central un-biased template. After this alignment step all available atlases and the test image live in the same space and we compute the pair-wise similarities of the affine aligned atlases with the test subject. Based on the resulting similarities we select the k top ranked atlases for an accurate non-rigid registration for label propagation, and finally for label fusion.

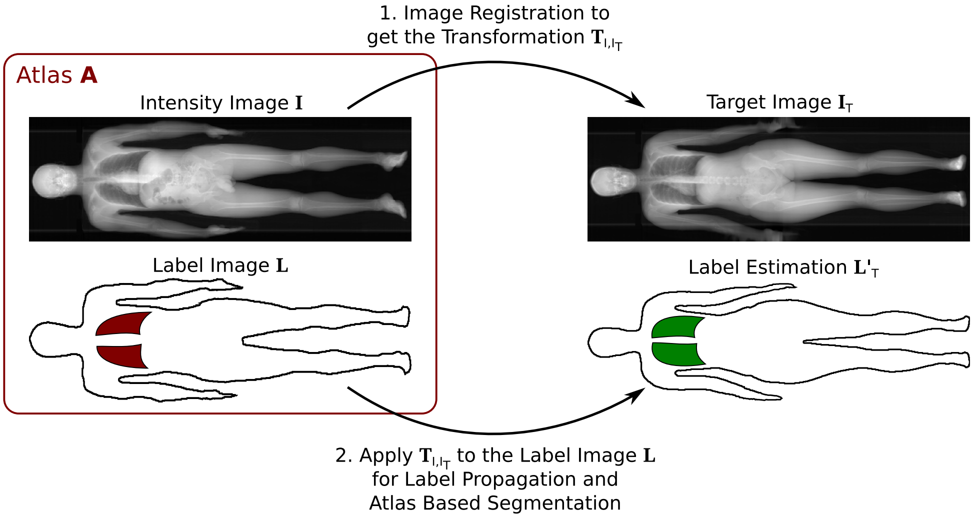

1. Image Registration to
get the Transformation $\mathbf{T}_{I,I_T}$
Atlas **A**
Intensity Image **I**
Target Image $\mathbf{I}_T$
Label Image **L**
Label Estimation $\mathbf{L}'_T$
2. Apply $\mathbf{T}_{I,I_T}$ to the Label Image **L**
for Label Propagation and
Atlas Based Segmentation

Figure 4: Overview of atlas based segmentation.

3.3.2 Annotation propagation for atlas based segmentation

Given an atlas **A** defined as a pair $(\mathbf{I},\mathbf{L})$ consisting of an intensity image **I** and its corresponding label image **L** (e.g. by manual annotation). The aim of annotation propagation is the automatic segmentation of an novel test image $\mathbf{I}_T$. For this purpose the atlas intensity image **I** is registered with the test image $\mathbf{I}_T$ yielding the transformation $\mathbf{T}_{I,I_T}$ from the atlas towards the test image. An estimation for the true segmentation $\mathbf{L}_T$ of the test image is computed by also applying the registration transformation $\mathbf{T}_{I,I_T}$ to the atlas label image **L**.

$$\mathbf{L}'_T = \mathbf{T}_{I,I_T}(\mathbf{L}) \qquad (1)$$

Figure 4 summarizes all components contributing to atlas based segmentation using image registration and label propagation.

3.3.3 Label fusion and segmentation performance estimation

Given a target image $\mathbf{I}_T$ for annotation and a set of estimations $\mathbf{L}_T^j$ with $j = 1,...,M$ for its true (unknown) segmentation $\mathbf{L}_T$. The segmentations are defined to cover a single structure and are represented by a binary label image with the same dimensions as the target image. This means in particular that the voxels containing the structure of interest have label 1 and the background label 0 assigned. The estimated segmentations can either originate from the manual annotation by human experts or from automatic segmentation algorithms. The goal of label fusion is to combine the M segmentation estimates in a single fused label image $\mathbf{L}'_T$ to provide an improved estimation of the real hidden ground truth segmentation $\mathbf{L}_T$ of the target image. The principle of label fusion based segmentation is illustrated in Figure 5. In addition to the estimated ground truth $\mathbf{L}'_T$ we are interested in the contributions of the single estimates $\mathbf{L}_T^j$ to the resulting fused segmentation, as well as their performances (consensus) ϕ_j with respect to the estimated ground truth.

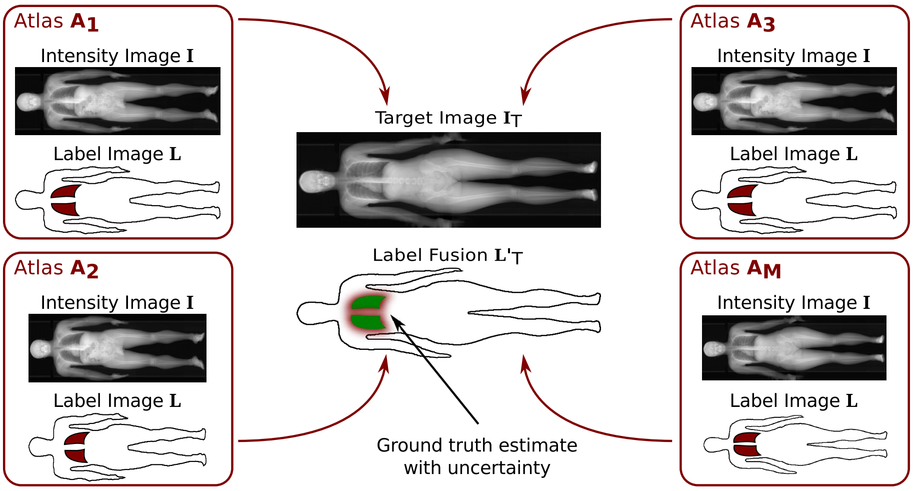


Figure 5: General principal of label fusion for atlas based segmentation.

There already exist methods which provide a solution to the problem addressed. In the following we describe a Selective and Iterative Method for Performance Level Estimation (SIMPLE) proposed by Langerak et. al. [5]. SIMPLE is based on an iterative strategy and alternates in:

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 expert annotations of the same subject [5]. For describing the SIMPLE algorithm we use *majority voting* as well as *global weighted voting* two label fusion methods which are introduced in the following paragraphs. Although majority voting is used at this point, it can be replaced by methods such as Simultaneous Truth and Performance Level Estimation (STAPLE) [10] or the generative model for label fusion based image segmentation published in [9]. The latter allows for a parameterization of the label fusion concept to vary between local weighted voting, majority voting, global weighted fusion, as well as semi-local weighted fusion.

Majority voting [2] Given M binary label estimates for point $\mathbf{x}$, $\mathbf{L}_T^j(\mathbf{x})$, majority voting counts for each voxel, how many annotators vote for the presence (value 1) or absence (value 0) of the structure that has to be annotated. Label fusion is done by assigning each voxel at coordinate $\mathbf{x}$ of the ground truth estimate $\mathbf{L}'_T(\mathbf{x})$ the value having the maximum number of votes.

$$\mathbf{L}'_T(\mathbf{x}) = \left(\sum_{j=1}^{M} \mathbf{L}_T^j(\mathbf{x})\right) > \frac{M}{2} \tag{2}$$

Global weighted voting [2] Given M binary label estimates $\mathbf{L}_T^j$ 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 $\mathbf{L}'_T$. The method is called global weighted voting since the weights are assigned to the entire region of the volumes. Taking the additional information into account, label fusion assigns the fused ground truth estimates $\mathbf{L}'_T(\mathbf{x})$ as follows:

$$\mathbf{L}'_T(\mathbf{x}) = \left(\sum_{j=1}^{M} \mathbf{L}_T^j(\mathbf{x}) \cdot \phi_j \right) > \frac{\sum \phi_j}{2} \tag{3}$$

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

Algorithm description of SIMPLE

- **Input:** A set of M estimates $\mathbf{L}_T^j$ for the true segmentation $\mathbf{L}_T$
- **Compute performance estimates and label fusion:**
  1. Combine all available segmentations $\mathbf{L}_T^j$ to get an initial estimate for the ground truth $\mathbf{L}'_T$ using majority voting (note that any other fusion method can be used at this point as well).
  2. Estimate the performance ϕ_j for each segmentation $\mathbf{L}_T^j$ by computing a binary overlap measure (e. g. Dice coefficient [3]) with the initial ground truth estimation $\mathbf{L}'_T$.

     $$\phi_j = \frac{2|\mathbf{L}_T^j \cap \mathbf{L}'_T|}{|\mathbf{L}_T^j| + |\mathbf{L}'_T|} \tag{4}$$

  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 $\mathbf{L}'_T$ based on the remaining segmentation estimates and their estimated performances. For this purpose performance weighted label fusion described above 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 $\mathbf{L}_T^j$, as well as their fusion $\mathbf{L}'_T$. This fusion is defined as estimate for the real hidden ground truth segmentation $\mathbf{L}_T$.

3.4 Annotator reliability estimation and ranking

The aim of this subsection is the ranking of the $j = 1,...,M$ annotators based on all annotations they provided. The data available for computing this ranking consists of:

1. A set of images $\mathbf{I}_i$ with $i = 1,...,N$
2. A set of $s = 1,...,S$ structures selected for annotation
3. A set of binary label images $\mathbf{L}_{i,s}^j$ for each structures s in each image $\mathbf{I}_i$ provided by the $j = 1,...,M$ annotators.
4. A consensus fusion of all label estimates provided by the annotators (output of SIMPLE). This fusion is defined as the ground truth segmentation estimate $\mathbf{L}_{i,s}^\prime$ for structure s in image $\mathbf{I}_i$.
5. The performances $\phi_{i,s}^j$ of the segmentations (output of SIMPLE). $\phi_{i,s}^j$ is the performance of annotator j for structure s in image $\mathbf{I}_i$

3.4.1 Hierarchical ranking of annotators

Based on this information we suggest the following hierarchical ranking model.

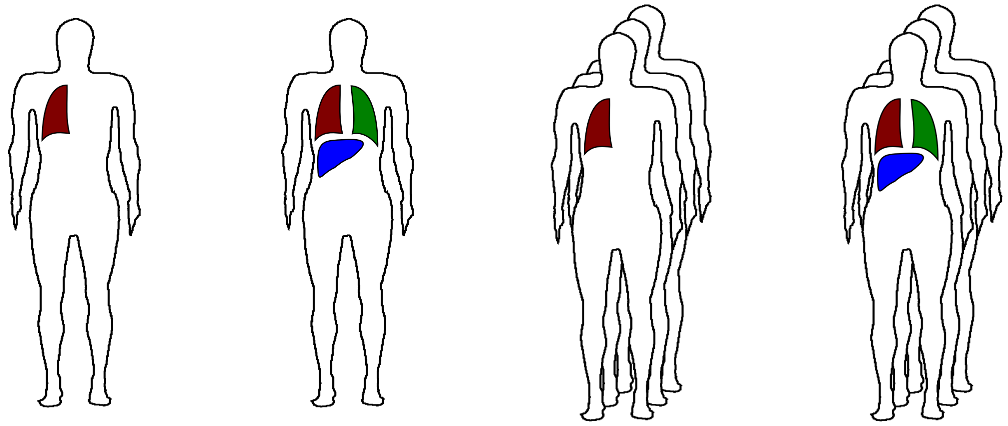
1. Ranking of annotations for a single image and a single structure: This is the basis for label fusion and the lowest level of annotator ranking.
2. Ranking of annotators based on a single image and multiple structures: This helps to find the best annotator for a single image:
3. Ranking of annotators based on multiple images for a single structure: This helps to determine annotation experts for a certain structure.
4. Ranking of annotators based on multiple images and multiple structures: This helps to find the best general annotator.

Figure 6 illustrates the four levels of performance estimation. The hierarchical model helps to ensure in Subsection 3.5 an optimum coverage of the entire population by providing three different strategies for assigning the annotation tickets to the appropriate annotators. In the following the single ranking levels are discussed in detail.

3.4.2 Ranking of annotations for a single image and a single structure

The ranking measures at this level are computed for each annotator j and each structure s in all images $\mathbf{I}_i$. This is the lowest level of performance estimation and serves as a basis for more general (averaged) ranking measures at higher levels.

SIMPLE performance The first measure of this category is the performance estimate $\phi_{i,s}^j$ of the SIMPLE method. As described above, $\phi_{i,s}^j$ is computed as the dice coefficient of the annotation estimations $\mathbf{L}_{i,s}^j$ and the estimated ground truth $\mathbf{L}_{i,s}^\prime$. This binary overlap measure reflects the consensus of segmentations $\mathbf{L}_{i,s}^j$ with the estimated ground truth $\mathbf{L}_{i,s}^\prime$.



Ranking of annotations for a single image and a single structure

Ranking of annotators based on a single image and multiple structures

Ranking of annotators basedon multiple images for a single structure

Ranking of annotators based on multiple images and multiple structures

Figure 6: Overview of annotator ranking hierarchy.

Voxel-wise performance A second class of measures is computed on a per voxel basis. Each voxel is treated as a binary decision between foreground (presence of the structure) and background (absence of the structure). This principle is derived from a crowd sourced labelling approach for annotator ranking published in [8]. The sensitivity $\alpha = Pr[1|1]$ is defined as the true positive rate and represents the probability that a foreground voxel is correctly classified as foreground. The true negative rate $\beta = Pr[0|0]$ represents the probability that a background voxel is correctly classified as background. Both measures are combined into the $accuracy_{i,s}^{j}$ of segmentation $\mathbf{L}_{i,s}^{j}$ with respect to the ground truth estimate $\mathbf{L}_{i,s}'$. The accuracy is the ratio of correctly classified voxels and computed as:

$$accuracy_{i,s}^{j} = \alpha_{i,s}^{j} p_{i,s} + \beta_{i,s}^{j}(1 - p_{i,s}) = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

with the following notation:

- TP / TN: Number of voxels which got segmented correctly as foreground / background.
- FP / FN: Number of voxels which got segmented incorrectly as foreground / background.

The measures at this level enable the ranking of annotators based on their annotations of a single structure in a single image and are used as basis for the measures in the higher levels.

3.4.3 Ranking of annotators based on a single image and multiple structures

The ranking measures in this level are applied if multiple structures (≥ 2) get annotated in a single image. For this purpose we first fuse all binary label estimates $\mathbf{L}_{i,s}^{j}$ of image $\mathbf{I}_i$ for each structure s into one categorical (multi-)label image $\mathbf{L}_i^j$ where each voxel carries the value s of its originating binary label image $\mathbf{L}_{i,s}^{j}$. We assume at this point that the label estimates for the single structures are distinct.

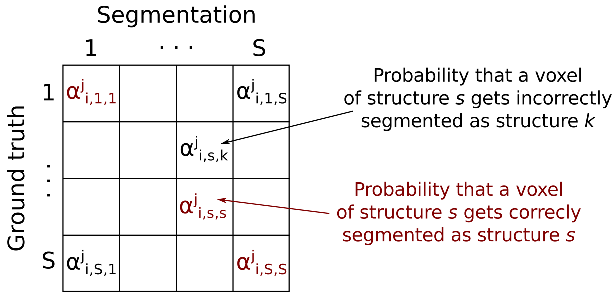

Segmentation
1
· · ·
S
Ground truth
1
· · ·
S
α^j_{i,1,1}
α^j_{i,1,S}
α^j_{i,s,k}
α^j_{i,s,s}
α^j_{i,S,1}
α^j_{i,S,S}
Probability that a voxel of structure s gets incorrectly segmented as structure k
Probability that a voxel of structure s gets correcly segmented as structure s

Figure 7: Confusion matrix of annotator j for the segmentation of image $\mathbf{I}_i$.

SIMPLE performance The first measure which reflects the reliability of an annotator over different structures in a single image is the average SIMPLE performance ϕ_i^j and defined as:

$$\phi_i^j = \frac{1}{S}\sum_{s=1}^{S}\phi_{i,s}^j \tag{6}$$

Voxel-wise performance Such as the SIMPLE performance we extend the *accuracy* to a multi-structure (label) version. Instead of the true and false positive rate we introduce for each annotator j and each image $\mathbf{I}_i$ the multi modal parameter $\alpha_{i,s,k}^j$ [8] as:

$$\alpha_{i,s,k}^j = Pr[\mathbf{L}(\mathbf{x})_i^j = k|\mathbf{L}_i'(\mathbf{x}) = s] = \frac{c_{k,s}}{\sum_{l=1}^{S} c_{l,s}} \tag{7}$$

with

$$c_{k,s} = \sum_{\mathbf{x}\in\mathbf{x}_s}\left(\mathbf{L}_i^j(\mathbf{x}) == k\right) \tag{8}$$

$c_{k,s}$ is described as the number (count) of voxels of structure s that are classified as structure k. $\mathbf{x}_s$ is the set of voxels which belong to structure s in the estimated ground truth $\mathbf{L}_j'$.

$$\mathbf{x}_s = \{\mathbf{x}|\mathbf{L}_i'(\mathbf{x}) = s\} \tag{9}$$

$\alpha_{i,s,k}^j$ reflects the probability that annotator j assigns a voxel belonging to structure s the label k. This general formulation reduces for a binary segmentation problem to sensitivity α and specificity β as defined above [8]. $\alpha_{i,s,s}^j$ is described as the true positive rate of structure s.

Having the parameters for all possible combinations of labels we combine them into the *confusion matrix* $\mathbf{C}_i^j = [(\alpha_{i,s,k}^j)]$ with $s = 1,\ldots,S$ and $k = 1,\ldots,S$. Figure 7 illustrates the confusion matrix of annotator j for image $\mathbf{I}_i$. Using the entries of the confusion matrix we define the *multi label accuracy* of annotator j as follows:

$$accuracy_i^j = \frac{1}{S}\sum_{k=1}^{S}\alpha_{i,k,k}^j \tag{10}$$

Both, the average SIMPLE performance as well as the multi label accuracy reflect the annotation performance of an annotator in a single image over all structures annotated.

3.4.4 Ranking of annotators based on multiple images for a single structure

The measures in this section reflect the ability of an annotator to annotate a certain structure over multiple images. This is useful if we have a 'difficult' case of the targeted structure to annotate. We use the measures of this category to find the highest ranked annotator for the structure over all annotated images and assign the structure for annotation.

SIMPLE performance The first measure which reflects the reliability of an annotator for a specific structure s is the average SIMPLE performance $\phi^j_{\cdot,s}$ over all images:

$$\phi^j_{\cdot,s} = \frac{1}{N}\sum_{i=1}^{N}\phi^j_{i,s} \tag{11}$$

Voxel-wise performance We also adopt the accuracy of an annotator to an average version over all images for structure s:

$$accuracy^j_{\cdot,s} = \frac{1}{N}\sum_{i=1}^{N} accuracy^j_{i,s} \tag{12}$$

The averaging is computed over the number of images N and not on a per voxel basis. A voxel-wise averaging would be biased by the resolution (number of voxel per structure) of the contributing images.

3.4.5 Ranking of annotators based on multiple images and multiple structures

The measures at this level provide an overall performance value for the annotators based on all their annotations concerning all structures. This is the highest (most general) level of performance estimation and helps to determine the best *'all-round'* annotator.

SIMPLE performance The first measure is the overall average over all SIMPLE performances. This allows a ranking of the annotators based on their ability to correctly annotate a range of different structures in the images:

$$\phi^j = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{s=1}^{S}\phi^j_{i,s}\right) \tag{13}$$

Voxel-wise performance Further we compute the average $\alpha^j_{\cdot,s,k}$ of the multi modal parameters $\alpha^j_{i,s,k}$ and compose an average confusion matrix $\mathbf{C}^j = [(\alpha^j_{\cdot,s,k})]$ for all annotators j concerning all their annotations.

$$\alpha^j_{\cdot,s,k} = \frac{1}{N}\sum_{i=1}^{N}\alpha^j_{i,s,k} \quad \forall j = 1,\ldots,M \ \text{ and } \ \forall s,k = 1,\ldots,S \tag{14}$$

The global average accuracy $accuracy^j$ of an annotator is computed as:

$$accuracy^j = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{S}\sum_{s=1}^{S}\alpha^j_{i,s,s}\right) \tag{15}$$

Table 1: Summary of SIMPLE performance measures

Structure	Annotation 1 ... Annotation N	Number of Annotations	AVG Performance
Structure 1	$\phi^j_{1,1} \quad \cdots \quad \phi^j_{N^j_1,1}$	N^j_1	$\phi^j_{\cdot,1}$
...			
Structure S	$\phi^j_{1,S} \quad \cdots \quad \phi^j_{N^j_S,1}$	N^j_S	$\phi^j_{\cdot,S}$
	$\phi^j_{1,\cdot} \quad \cdots \quad \phi^j_{N_S,\cdot}$	N^j	ϕ^j

3.5 Ensuring Optimal Coverage of Dataset

In this section we make use of the ranking measures introduced to assign images to annotators for manual segmentation. There are different reasons why a manual annotation is required and we distinguish between three different cases. This distinction helps to provide an ideal strategy for annotation ticket assignment and to ensures optimal coverage of the entire set of images. Before we discuss the three cases in detail, a summary of the ranking measures introduced is provided.

3.5.1 Summary of ranking measures

SIMPLE performance Table 1 summarizes all measures computed based on the performance output ϕ of the SIMPLE algorithm. For ranking we have the following performance values for each of the M annotators available:

- Performance $\phi^j_{i,s}$ for the annotation of a single structure in a single image.
- Average performance $\phi^j_{\cdot,s}$ for the annotation of a single structure over all image.
- Average performance ϕ^j_i for the annotation of multiple structures in a single image.
- Average performance ϕ^j overall annotations.

Since not every annotator will annotate each structure in each image we introduce the values N^j_s which count the number of annotations of annotator j for the respective structures.

Voxel-wise performance The accuracy measures computed based on a per voxel decision basis are computed at the same hierarchical levels as the simple performance measures summarized above. Hence SIMPLE performance and accuracy are either interchangeable components or are used in parallel to evaluate the results.

In the following sections the SIMPLE performance measures are used to describe how the annotators for the three different annotation tasks are determined.

3.5.2 Case 1: Poor confidence in existing manual segmentation

The quality check described in this subsection focuses on the manual annotation of a single structure. This enables the consulting of annotation experts for the targeted structure if required.

For the purpose of evaluation and annotation quality assurance we fuse the labels of the volumes which already have manual annotations towards the target image selected for evaluation. After label fusion we estimate the confidence in the existing manual segmentation by computing its consensus with the estimated ground truth (result of label fusion). If a threshold τ is not reached, we assign the image and the respective structure for manual re-annotation. To ensure a correct annotation we select a group of top ranked annotators (experts) for the structure of interest.

Annotator selection The reliability of an annotator for a specific structure is reflected by its average structure performance $\phi^{j}_{\cdot,s}$. In addition to the performance we take for a decision the experience of the annotators (number N^j_s of annotations for structure s) into account. This is done by applying a sigmoid shaped weighting to the annotator performances. To allow the experience threshold to evolve during the annotation process, we define an experienced annotator as someone who has annotated at least the same number of structures s as the average of all annotators. Using this definition, the experience weighted performance is computed as:

$$\phi^{j*}_{\cdot,s} = \phi^{j}_{\cdot,s} \cdot \frac{1}{1+e^{(N^j_s - \overline{N}_s)}} \quad \text{with } \overline{N}_s = \frac{1}{M}\sum_{j=1}^{M} N^j_s \tag{16}$$

The principal of annotator performance weighting is exemplary illustrated in Figure 8. Annotator one and three have less annotations than the average of all annotators and are not taken into account as structure-experts. Annotator four has the highest performance and enough experience to annotate the problematic structure and gets the ticket assigned.

3.5.3 Case 2: Poor confidence in label fusion for a novel image

Given a novel test image for automatic segmentation using image registration and label fusion. If the fused labels of the existing atlases show a low agreement for the new image and the respective structure, we suggest it for manual annotation. The level of agreement is derived by:

1. The number of segmentations discarded by SIMPLE (the less segmentations are discarded the better the agreement) → threshold

2. The average performance $\phi_{\cdot,s} = \frac{1}{M}\sum_{j=1}^{M} \phi^{j}_{\cdot,s}$ of all segmentations not discarded by SIMPLE (the higher the better) → threshold

Annotator selection If we decide to manually annotate the structure s concerned in the new image we select the k annotators having the highest average performances $\phi^{j}_{\cdot,s}$ for structure s. Annotators with a low number of annotations are favoured to distribute the annotation tasks equally to the entire group of annotators.

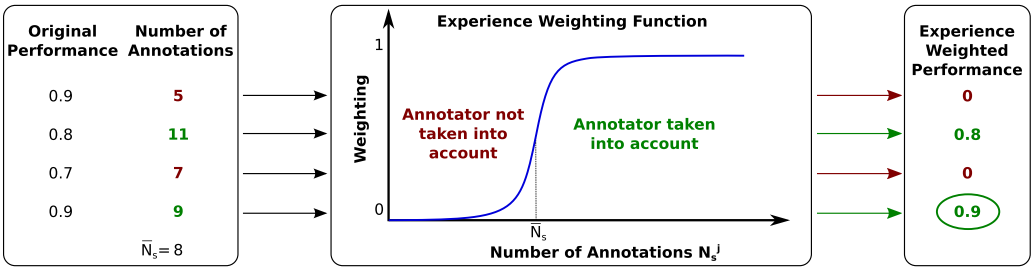

Original Performance
Number of Annotations
0.9
5
0.8
11
0.7
7
0.9
9
N̄s = 8
Experience Weighting Function
Weighting
1
0
Annotator not taken into account
Annotator taken into account
N̄s
Number of Annotations N_s^j
Experience Weighted Performance
0
0.8
0
0.9

Figure 8: Weighting of annotator performances based on the number of annotations.

3.5.4 Case 3: Introduction of a novel structure for annotation

If a new structure is selected for annotation the following two step approach is used for annotation ticket assignment:

1. Select a representative set of images analogously to the selection of the initial set in Subsection 3.2 to ensure a representative set of subjects covering the entire population.

2. Select the k annotators having the highest average ranking ϕ^j over all annotated images and structures and assign them the images. The overall average annotator performance is taken into account, since we have no information on who of the annotators performs best on the new structure at this point.

The result of this procedure is a subset of images having the novel structure annotated. Once this is done we proceed with image registration and label fusion to get the entire set annotated. Problems occurring in the label fusion process are again covered by Case 1 and Case 2.

4 Annotation Transfer Framework

For the implementation of the annotation process to the annotators, a web–interface is created to distribute and retrieve the manual annotations. The interface is designed using Java language programming. The framework includes the following steps:

1. Creation of an annotation list with the annotations that need to be done ('tickets') and its upload to the web interface.

2. The web interface has a login user name and password for each of the annotators.

3. The annotators ID is used in the naming of the tickets:

$$subjectXX\_acquisitionZZ[modalityYY]\_RadLexID\_annotatorID.nii$$

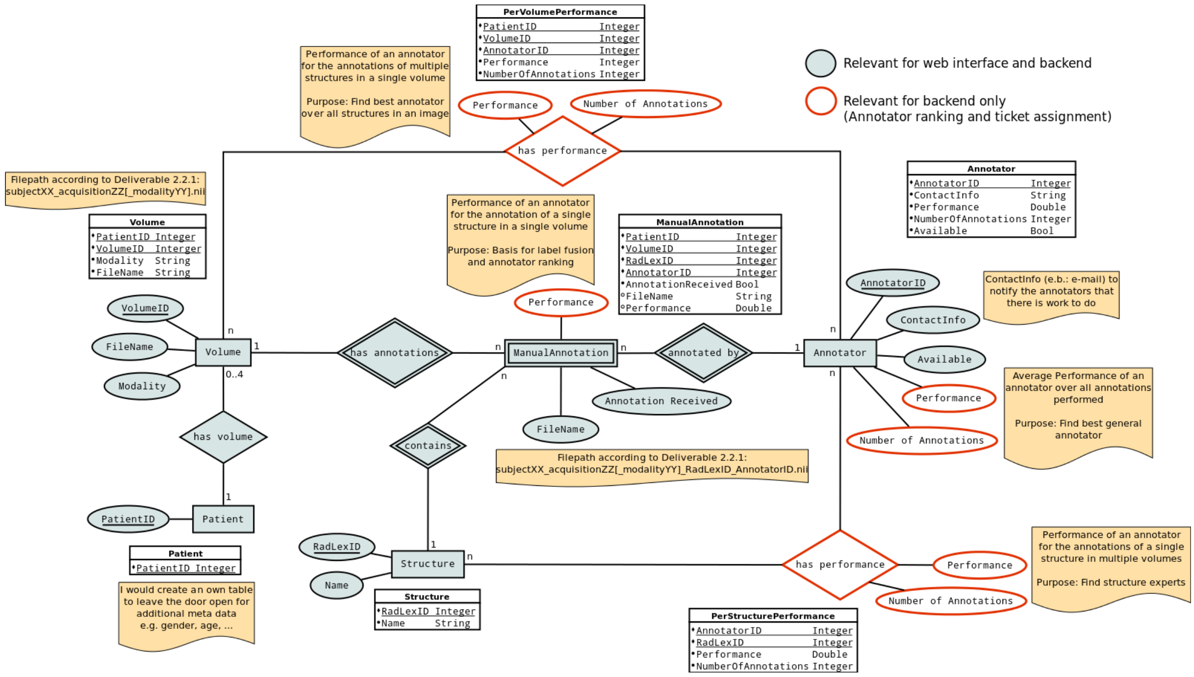
4. The annotators upload their files next to the ticket and the name of the file is implemented to be the same as the ticket for their back end analysis.

5. All the annotations are saved in the same folder to download them and use them in the analysis.

6. The annotation back end produces a new list of tickets for the new annotations needed and from which annotators.

7. The list in the interface is updated.

8. The annotators receive their new set of tickets when they login.



○ Relevant for web interface and backend

○ Relevant for backend only
(Annotator ranking and ticket assignment)

Performance of an annotator for the annotations of multiple structures in a single volume

Purpose: Find best annotator over all structures in an image

Filepath according to Deliverable 2.2.1:
subjectXX_acquisitionZZ[_modalityYY].nii

Performance of an annotator for the annotation of a single structure in a single volume

Purpose: Basis for label fusion and annotator ranking

ContactInfo (e.b.: e-mail) to notify the annotators that there is work to do

Average Performance of an annotator over all annotations performed

Purpose: Find best general annotator

Filepath according to Deliverable 2.2.1:
subjectXX_acquisitionZZ[_modalityYY]_RadLexID_AnnotatorID.nii

I would create an own table to leave the door open for additional meta data e.g. gender, age, ...

Performance of an annotator for the annotations of a single structure in multiple volumes

Purpose: Find structure experts

Figure 9: Common database for web-interface and annotation backend.

5 Annotation Interface

Microsoft GeoS annotation tool prototype has been selected for the VISCERAL gold corpus creation. It was selected after reviewing the different image analysis frameworks available that provide semi–automatic segmentation methods that facilitate the manual annotations of 3D medical images. The annotation tool comparison and overview of GeoS is included in the Deliverable 1.1.

6 Conclusion

In this document we described the gold corpus annotation framework for VISCERAL. The system serves as the central hub for gold corpus annotation. It manages the already annotated data, continuously estimates annotation reliability, and ensures that the annotation resources are used optimally.

7 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 Solorzano. Combination strategies in multi-atlas image segmentation: Application to brain mr data. *Medical Imaging, IEEE Transactions on*, 28(8):1266–1277, 2009.

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

[4] A. Guimond, J. Meunier, and J. P. Thirion. Average brain models: A convergence study. *Computer Vision and Image Understanding*, 77(2):192 – 210, 2000.

[5] 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.

[6] M. Modat, G. R. Ridgway, Z. A. Taylor, M. Lehmann, J. Barnes, D. J. Hawkes, N. C. Fox, and S. Ourselin. Fast free-form deformation using graphics processing units. *Comput. Methods Prog. Biomed.*, 98(3):278–284, June 2010.

[7] H. Park, Peyton H. Bland, Alfred O. Hero, and Charles R. Meyer. Least biased target selection in probabilistic atlas construction. In *Proceedings of the 8th international conference on Medical image computing and computer-assisted intervention - Volume Part II*, MICCAI'05, pages 419–426, Berlin, Heidelberg, 2005.

[8] Vikas C. Raykar and Shipeng Yu. Ranking annotators for crowdsourced labeling tasks. In J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1809–1817. 2011.

[9] M. R. Sabuncu, B. T. T. Yeo, K. Van Leemput, B. Fischl, and P Golland. A Generative Model for Image Segmentation Based on Label Fusion. *IEEE Transactions on Medical Imaging*, 29(10):1714–1729, October 2010.

[10] 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.