www.visceral.eu

Definition of the evaluation protocol and goals for Competition 1

Deliverable number	*D4.1*
Dissemination level	*Public*
Delivery data	*30.4.2013*
Status	*Final*
Authors	*Georg Langs, Abdel Aziz Taha, Bjoern Menze, Allan Hanbury*

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

VISCERAL is a support action that aims at distributing a substantial amount of medical imaging data together with expert annotations to the research community. The distribution will occur in the context of two large benchmarking campaigns, that allow researchers, to evaluate their algorithms on test data. This document outlines the goals of the first benchmark (*competition 1*), and describes the protocols that will be used during evaluation. The benchmark is aimed at making data and evaluation useful to a wide variety of algorithms, in the context of medical image analysis. On one hand we will evaluate algorithms that can localize and segment specific anatomical structures, on the other we evaluate algorithms that learn segmentation- or localization models for arbitrary data offering a challenge task during which algorithms have to localize and segment a previously unseen organ.

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

LBP	Local Binary Patterns
PACS	Picture archiving and communication system

1 Introduction

In the course of VISCERAL, two substantial data sets of medical imaging data will be distributed to the research community. The data will be distributed together with expert annotations of organs and anatomical landmarks on part of the data. In addition to the data release, VISCERAL will coordinate and host two benchmarking campaigns, or *competitions* that allow researchers to evaluate their algorithms on test data, not available. Not making the second data set publicly available ensures its usability for continuous evaluation in the future.

In this document we describe the goal of the first competition, that focuses on whole body labelling in medical imaging data.

2 Benchmark Goals

The data release together with a multi-layered competition serves several purposes:

1. **Access to large scale annotated medical imaging data** It makes data available to a large number of research groups, who might have advanced technology, and methodology, but only limited access to large scale medical imaging data.

2. **Fostering basic research** Research groups focusing on basic science, and methodology development, who are not directly connected to application in the clinical domain, often have no access to medical imaging data, or precise knowledge of the relevant tasks for which their methods might be relevant. VISCERAL aims at including these groups in the scientific discourse, and offering them fine grained tasks by which they can evaluate also methodology that only addresses part of an entire image processing pipeline.

3. **Comparability** In the past the availability of large representative data sets together with evaluation frameworks had tremendous impact on the advance in fields ranging from computer vision to neuroimaging. The comparability of methods is essential to assess progress, to pinpoint difficulties shared across many methods, and to identify promising methodology approaches.

4. **Fine grained benchmarks to allow for participation of generalizable algorithms tackling specific aspects of the analysis** The evaluation is structured, so that algorithms, that perform only part of the analysis pipeline such as only segmentation, or localization can be included in the benchmark.

In the following we will first outline the tasks in the first VISCERAL competition (Section 3), then we explain the evaluation protocol in detail in Section 4.

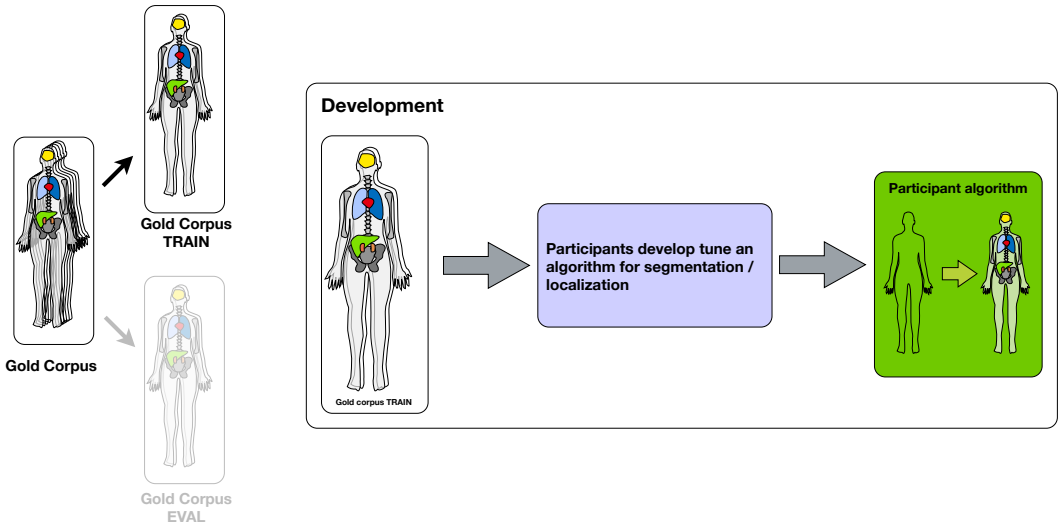

Figure 1: During the development phase, annotated data is available to the participants. They can develop, and train their algorithms.

3 Benchmark Tasks

3.1 Multi layered benchmark tasks

The benchmark tasks are (1) segmentation of anatomical structures (lung, liver, kidney) in non annotated whole body MR- and CT- volumes, and (2) the identification of anatomical landmarks in this data. To ensure that algorithms that for instance are only able to segment organs, but not able to localized them in a large volume, we will provide additional initialization information, if participants desire. The tasks that the participants algorithms have to perform in the evaluation phase are

1. Full run segmentation: Given a whole body volume, locate and segment a specified list of organs.

2. Full run landmarks: Given a whole body volume, locate a specified list of anatomical landmarks.

3. Half run segmentation: Given a whole body volume, and the centroid of a specified list of organs, segment the organs

During the the development phase (Figure 1) a part of the image data together with annotations corresponding to the benchmark tasks is available to all participants. During the evaluation, the participant algorithms run on virtual machines provided by VISCERAL, and configured by the participants. During evaluation the algorithms access the test data not available during benchmark preparation (Figure 2).

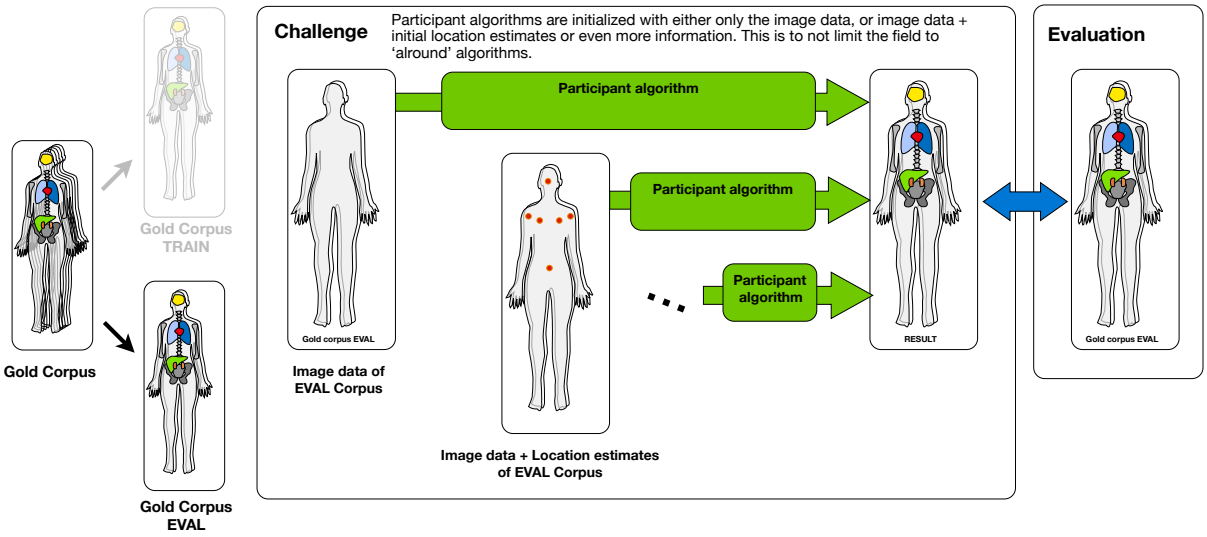

Figure 2: During evaluation, the participants algorithm perform localization, or segmentation tasks, and are evaluated against a part of the gold corpus not publicly available.

3.2 The surprise organ: evaluating learning algorithms

In addition to evaluating algorithms that are developed for specific organs, we also evaluate algorithms, that aim at generalizing their learning capabilities to arbitrary structures. Avoiding the *overfitting* of methods to data, or specific problems is a hard to solve and essential problem in medical imaging. This part of the benchmark aims to evaluate algorithms that are not tuned to a specific organ, but instead can learn to segment, or localize any structure, given sufficient training data.

During the development phase the data distributed is the same as for the standard challenge. However instead of developing algorithms only for the given organs, the participants use the data to train and develop algorithms, that learn localization- and segmentation models, that can be transferred to structures different from those included in the development data set.

During the evaluation phase (Figure 4) the algorithms have to solve the following task:

1. Full run segmentation learner: You are given image data together with annotations of an organ not previously seen during development for part of the image data. Learn a segmenter that segments this organ in the non annotated part of the evaluation data.

2. Full run localization learner: You are given image data together annotations of an organ not previously seen during development for part of the image data. Learn a localizer that localizes the anatomical landmarks in the non annotated part of the evaluation data.

3. Half run segmentation learner: You are given image data together annotations of an organ not previously seen during development for part of the image data. For the remaining part you are given the centroids of the organ in question. Learn a segmenter that segments this organ in the non annotated part of the evaluation data.

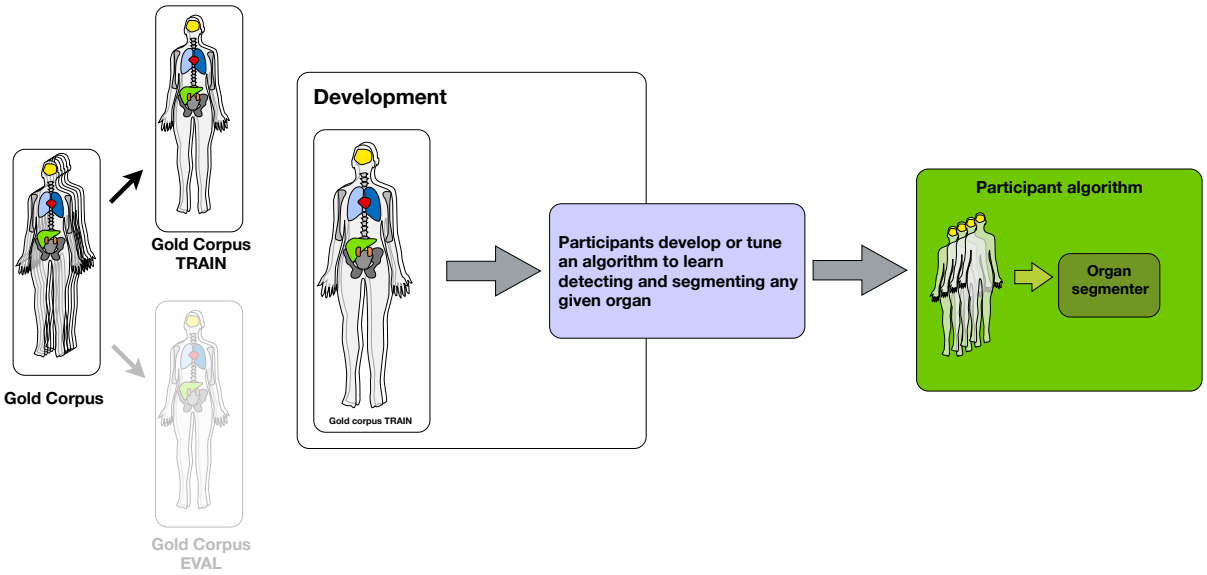

Figure 3: Similar to the development in the standard challenge annotated data is distributed to the participants in the surprise organ challenge. However, instead of building segmention algorithms or localization algorithms for the given organs, they develop learners that can adapt to new organs, or anatomical landmarks.

4 Evaluation Protocol

4.1 Introduction

Because large data sets are crucial in medical image analysis and retrieval research, one of the objectives of VISCERAL is to innovate through the use of a cloud infrastructure to provide participants with a huge amount of data with variability reflecting what is encountered in everyday practice in a hospital. It is a challenging task to make available such a large data set. VISCERAL will master this difficulty by providing an evaluation infrastructure in the cloud: the data will be available in the cloud, and computing instances will be provided in the cloud, so that participants can deploy and test their algorithms in the cloud. This means moving the software to the data, thereby avoiding problems associated with moving the data, like long download times and sending physical disks. Computing instances will also provide evaluation scripts that can be performed by participant to test their algorithms in the development phase. Computing instances will be entirely financed by VICSERAL, which means that participant shouldnt́ pay any fees. Generally, the challenge consists of three steps: registration, training phase and evaluation phase. Each of the steps is described below.

4.2 Registration

To use the training data and take part in the benchmark, participants should register and accept the terms of use for the data. This will be done through a registration system online, together with a signed *terms of use* document.

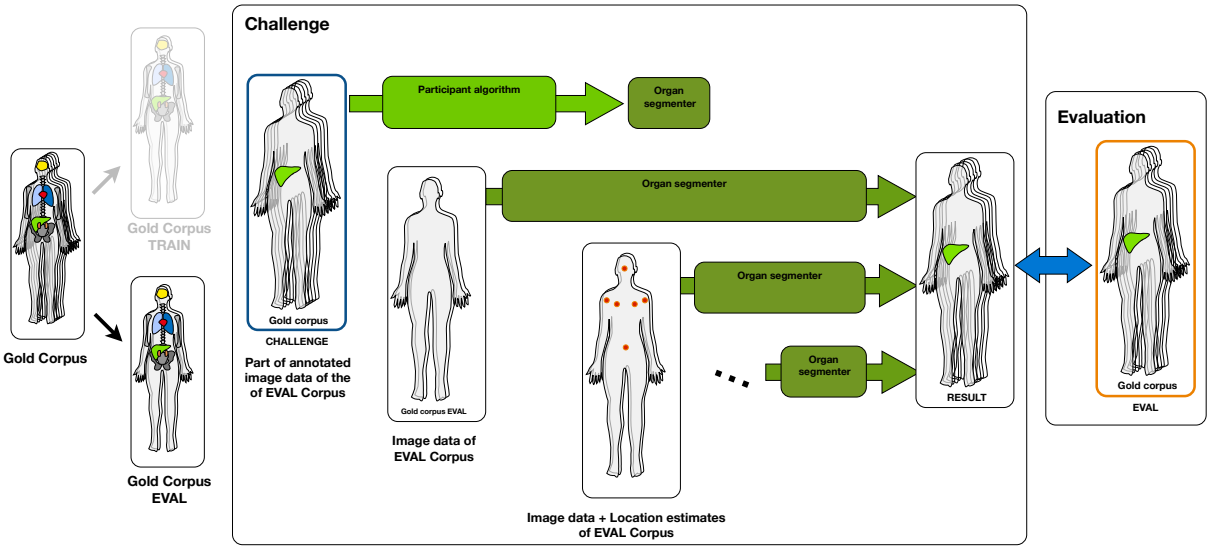

Figure 4: During surprise organ evaluation, the algorithms in the surprise organ challenge, have to segment or localize an organ not previously seen, after learning the segmenter or localizer on a small annotated data set only available during evaluation.

4.3 Training phase

Participants will be provided with a large data set of full body 3D image data together with training annotations on a subset of the data. Each participant should use his/her own computing instance to freely deploy and test algorithms in the cloud. Costs of using these computing instances are covered by the organizers up to a specified limit.

Information about how to use the computing instances and about the exact formats and communication protocol (annotation guidelines, name convention for delivered participant programs, parameter, specification, used datasets, etc.) will be provided when the benchmark is opened.

On the benchmark deadline, control of the computing instances will pass to the organizers. It will only be necessary to leave executable programs (satisfying the specifications) in the computing instance. Participants should ensure that any information that they wish to keep confidential (such as source code or confidential data) is removed from the computing instance by the submission deadline. Figure 5 illustrates the use of cloud computing instances in the training phase of the benchmark.

To ensure that participants master the computing instances, a test submission should be performed short time after the benchmark (deadline for test submission will be determined on benchmark). The goal of the test submission is to ensure submissions satisfying the specification at challenge deadline.

4.4 Evaluation phase

On the challenge deadline, participants are disconnected from their instances which are then connected to the organizer to perform the evaluation. Before participants are disconnected, they

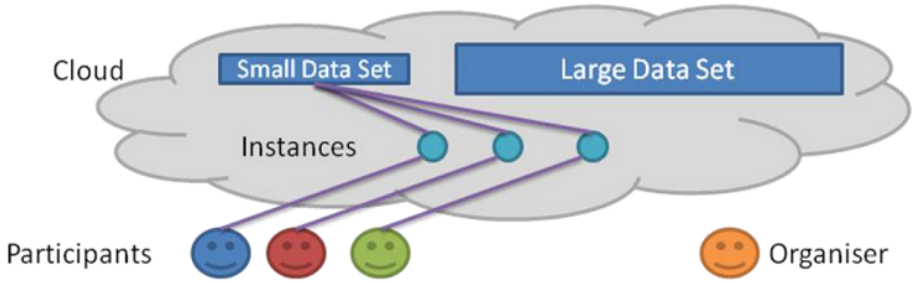

Figure 5: Use of computing instances in the cloud during the training phase.

should have deployed their algorithms in form of executable software that is ready to be used by the organizer. Figure 6 shows the use of instances in the cloud to perform evaluation. The

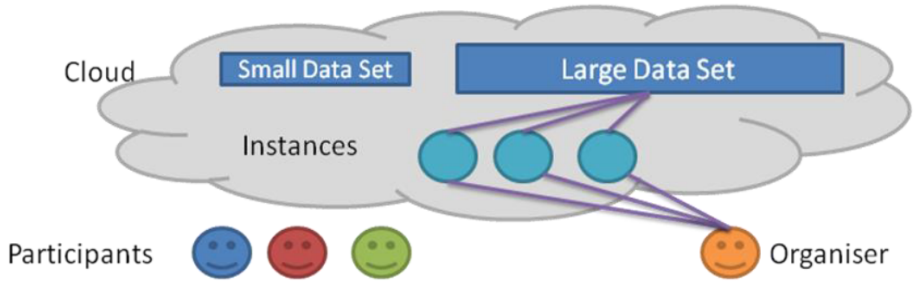

Figure 6: Use of computing instances in the cloud in the evaluation phase.

participant algorithms are then applied by the organizer on a large dataset of 3D volumes to get the participant annotations (runs). Two tasks will be considered in the benchmark: organ identification (localization) and segmentation of anatomical structures present in the data like bones, inner organs and relevant substructures. These two tasks are described in more detail in section 3. The benchmark test data will consist of two parts: a small part with gold annotation (manual annotation) and a very large part without annotation. We will evaluate both with regard to

- comprehensive identification
- subset localization, in order to be able to include algorithms developed for specific organs as a secondary task within the competition

The evaluation will be done in four steps:

1. Evaluating only participant annotations for which gold manual annotation exists to get $score_{gold}$.
2. Generating a silver corpus from participant annotations based on the results of step 1 ($score_{gold}$)

3. Evaluating the rest of the participant annotations against the silver corpus to get $score_{silver}$

In the next sub sections we will describe each of these steps as well as how to obtain scores from metrics and end scores from partial scores.

4.4.1 Evaluation stage 1

Participant annotations of data for which gold annotations exist, will be evaluated against these gold annotations. As a result of this step, each participant algorithm will be assigned a score ($score_{gold}$). Scores will be at both participant and structure level: scores on participant level have the goal to determine the best algorithm all-around while score at structure level determine the best algorithm in annotating a particular organ. Scores at structure level will be also used in generating the silver corpus considering that some algorithms work better on certain anatomical structures than others, and thus each participant will be given a score at structure level which will then be used as a reliability measure of a particular participant and a particular structure that is considered in the process of silver corpus fusion. If necessary, several structures may be merged to a single group (structure), which will be then additionally evaluated as a single structure. This will be for example the case, if two or more structures are difficult to separate by most of the participant algorithms.

4.4.2 Generating the Silver corpus

The portion of test data for which no annotation exists will be automatically annotated based on the participant entries to generate a large silver corpus: A heuristic will be applied on entries provided by participants to automatically merge them through a probabilistic metric taking the following into account:

1. agreement between the different participant algorithms: the higher the agreement on particular annotations, the more likely is that this annotation will be incorporated into the silver corpus

2. initial scores ($score_{gold}$) from the first evaluation stage: when merging annotations belonging to a particular structure, annotations of participants will be weighted in voting according to their score regarding this structure

A similar approach for generating a silver standard corpus from participant annotations was used by CALBC [17] with the difference that CALBC used a consensus model as a reference (annotations with high participant agreement) to measure the precision of each participant because CALBC silver standard corpus was built from pure automatic annotations. As we will have a gold portion of the test data, we will use this as a reference to measure participant reliability which is then used to generate the silver corpus. The details of silver corpus generation are subject of deliverable D3.3.

4.4.3 Evaluation stage 2

Once the silver corpus is generated, participant data are evaluated against the silver annotations to get new scores ($score_{silver}$). These scores will be also published as additional information on the performance of the participant algorithms and also on the performance of the silver corpus.

4.5 Evaluation metrics

There are many metrics that have been used in domains with similar tasks. We will consider these metrics as a pool of candidate metrics and we will select a subset of them. This metrics selection will be published by the time of opening the benchmark. Selecting the metrics will be based on the following criteria:

- At least the silver annotations will be expressed as probability maps which means that image segmentation will have fuzzy membership to structures at voxel level. There will be at two type of metrics.
  Type-I: metrics that directly deal with fuzzy segmentations, and
  Type-II: metrics that are calculated relative to an estimated hard truth, which is obtained by averaging the probabilistic segmentations and extracting the level-set at a determined optimal threshold.

- Metrics from various categories will be selected, that is metrics from the following categories will be included (1) distance-based metrics, (2) spatial overlap metrics, and (3) probabilistic and information theoretic metrics

- The most suitable metrics from the pool will be selected: The suitability will be decided according to the results of an analysis that is being performed on participant results from a similar domain, that is the brain segmentations from the BRATS12 challenge [2] in relation to the metric candidates

4.5.1 Metric candidates

Metrics in Table 1 will be the candidates from which we will select a subset to be used for evaluation in the challenge. The selection will be according to the guidelines mentioned previously. The selection of the candidates in the Table was according to a literature research: a metric was only included if there are at least two papers that describe its usage in the same domain, i.e. segmentation of medical volume images. Metrics with low use (less than two papers) were not included.

5 Conclusion

This document describes the goals of the first challenge in VISCERAL, and explains the evaluation protocol for the participant algorithms. The goal of the challenge is to provide data and evaluation for a wide range of algorithms that perform segmentation and localisation in medical imaging data. The evaluation tasks are designed so that both groups specializing on individual organs, as well as groups, who aim at developing basic methodology can make use of the data and evaluation.

metric	reference	category
Dice	[21] [22] [13] [3] [7] [11] [20] [1]	2
Hausdorff distance	[2] [6] [15] [7] [1] [12]	1
Jaccard	[2] [7] [18] [19]	2
Sensitivity	[2] [22] [11]	2,3
Specificity	[2] [22] [11]	2,3
Consistency Error	[14] [18]	2
Volumetric Similarity	[18] [19] [1] [4]	2
Mutual Information (MI) / Variation of Information (VOI)	[22] [8] [18]	3
Probabilistic Distance	[6] [7]	3
Cohens kappa	[2] [21]	3
ROC curve (AUC)	[22] [9]	2,3
Average distance	[2][12]	1
Rand Index RI/probabilistic RI	[18] [19]	2,3
interclass correlation coefficient	[6] [5]	3
Mahalanobis Distance	[16] [4]	3
coefficient of variation	[7] [10]	3

Table 1: Metric Candidates: the column reference shows papers where the metric has been used in evaluation of medical volume segmentation, category assigns the metric to (1) distance-based metrics, (2) spatial overlap metric, or (3) probabilistic and information theoretic metrics

6 References

[1] K. O. Babalola, B. Patenaude, P. Aljabar, J. Schnabel, D. Kennedy, W. Crum, S. Smith, T. F. Cootes, M. Jenkinson, and D. Rueckert. Comparison and evaluation of segmentation techniques for subcortical structures in brain mri. *Medical image computing and computer-assisted intervention*, 2008.

[2] Menze Bjoern, Jakab Andras, Debrecen, Bauer Stefan, Reyes Mauricio, Prastawa Marcel, and Van Leemput Koen. Brats multimodal brain tumor segmentation (http://www2.imm.dtu.dk/projects/brats2012), 2012.

[3] X. Cai, Y. Hou, C. Li, J. Lee, and W.G. Wee. Evaluation of two segmentation methods on mri brain tissue structures. *Conf Proc IEEE Eng Med Biol Soc*, 2006.

[4] Ruben Cardenes, Rodrigo de Luis-Garcia, and Meritxell Bach-Cuadra. A multidimensional segmentation evaluation for medical image data. *Comput. Methods Prog. Biomed.*, 96(2):108–124, 2009.

[5] Thomas M. Doring, Tadeu T.A. Kubo, L. Celso H. Cruz, Mario F. Juruena, Jiosef Fainberg, Romeu C. Domingues, and Emerson L. Gasparetto. Evaluation of hippocampal volume

based on mr imaging in patients with bipolar affective disorder applying manual and automatic segmentation techniques. *Journal of Magnetic Resonance Imaging*, 33(3):565–572, 2011.

[6] Guido Gerig, Matthieu Jomier, and Miranda Chakos. A new validation tool for assessing and improving 3d object segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2001*, pages 516–523. 2008.

[7] Sylvain Gouttard, Martin Styner, Marcel Prastawa, Joseph Piven, and Guido Gerig. Assessment of reliability of multi-site neuroimaging via traveling phantom study. In *Proceedings of the 11th International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 263–270, 2008.

[8] S. Gupta, K.P. Ramesh, and E.P. Blasch. Mutual information metric evaluation for pet/mri image fusion. In *Aerospace and Electronics Conference*, 2008.

[9] Jin Huang and Charles X. Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17:299–310, 2005.

[10] Moonis Gand Liu J, Udupa JK, and Hackney DB. Estimation of tumor volume with fuzzy-connectedness segmentation of mr images. *AJNR Am J Neuroradiol*, pages 356–63, 2002.

[11] Kasiri Keyvan, Dehghani Mohammad Javad, Kazemi Kamran, Helfroush Mohammad Sadegh, and Shaghayegh Kafshgari. Comparison evaluation of three brain mri segmentation methods in software tools. In *Biomedical Engineering (ICBME)*, 2010.

[12] Hassan Khotanlou, Olivier Colliot, Jamal Atif, and Isabelle Bloch. 3d brain tumor segmentation in mri using fuzzy classification, symmetry analysis and spatially constrained deformable models. *Fuzzy Sets Syst.*, 160(10):1457–1473, may 2009.

[13] Stefan Klein, Uulke A. van der Heide, Bas W. Raaymakers, Alexis N. T. J. Kotte, Marius Staring, and Josien P. W. Pluim. Segmentation of the prostate in mr images by atlas matching. In *ISBI*, 2007.

[14] Kevin McGuinness, Gordon Keenan, Tomasz Adamek, and O Connor. Image segmentation evaluation using an integrated framework. In *4th International Conference on Visual Information Engineering*, 2007.

[15] Fredric Morain-Nicolier, Stephane Lebonvallet, Etienne Baudrier, and Su Ruan. Hausdorff distance based 3d quantification of brain tumor evolution from mri images. In *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2007.

[16] W.J. Niessen, K.L. Vincken, and M.A. Viergever. Evaluation of mr segmentation algorithms. In *International Society Magnetic Resonance in Medicine*, 1999.

[17] Dietrich Rebholz-Schuhmann, Antonio Jose Jimeno J. Yepes, Erik M. Van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. Calbc silver standard corpus. *Journal of bioinformatics and computational biology*, 2010.

[18] A. Ramaswamy Reddy, E. V. Prasad, and L. S. S. Reddy. Abnormality detection of brain mr image segmentation using iterative conditional mode algorithm. *International Journal of Applied Information Systems*, 5(2):56–66, 2013.

[19] Nagesh Vadaparthi, Srinivas Yarramalle, Suresh Varma Penumatsa, and P.S.R.Murthy. Segmentation of brain mr images based on finite skew gaussian mixture model with fuzzy c-means clustering and em algorithm. *International Journal of Computer Applications*, 28(10):18–26, 2011.

[20] Kelly H. Zou, Simon K. Warfield, Aditya Baharatha, Clare Tempany, Michael R. Kaus, Steven J. Haker, William M. Wells, Ferenc A. Jolesz, and Ron Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index. *Academic Radiology*, 11:178–189, 2004.

[21] K.H. Zou, S.K. Warfield, A. Bharatha, C.M. Tempany, M.R. Kaus, S.J. Haker, W.M. Wells III, F.A. Jolesz, and R. Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index. *Acad Radiol*, 11(2), 2 2004.

[22] Kelly H. Zou1, William M. Wells, Ron Kikinis, and Simon K. Warfield. Three validation metrics for automated probabilistic image segmentation of brain tumours. *Statistics in Medicine*, 23, 2004.