Valmet: A New Validation Tool for Assessing and Improving 3D Object Segmentation

[1,2]Guido Gerig, [1]Matthieu Jomier, [2]Miranda Chakos

[1]Department of Computer Science, UNC, Chapel Hill, NC 27599, USA
[2]Department of Psychiatry, UNC, Chapel Hill, NC 27599, USA
Software: http://www.ia.unc.edu/public/valmet
email: gerig@cs.unc.edu

Abstract. Extracting 3D structures from volumetric images like MRI or CT is becoming a routine process for diagnosis based on quantitation, for radiotherapy planning, for surgical planning and image-guided intervention, for studying neurodevelopmental and neurodegenerative aspects of brain diseases, and for clinical drug trials. Key issues for segmenting anatomical objects from 3D medical images are validity and reliability. We have developed VALMET, a new tool for validation and comparison of object segmentation. New features not available in commercial and public-domain image processing packages are the choice between different metrics to describe differences between segmentations and the use of graphical overlay and 3D display for visual assessment of the locality and magnitude of segmentation variability. Input to the tool are an original 3D image (MRI, CT, ultrasound), and a series of segmentations either generated by several human raters and/or by automatic methods (machine). Quantitative evaluation includes intra-class correlation of resulting volumes and four different shape distance metrics, a) percentage overlap of segmented structures (R intersect S)/(R union S), b) probabilistic overlap measure for non-binary segmentations, c) mean/median absolute distances between object surfaces, and maximum (Hausdorff) distance. All these measures are calculated for arbitrarily selected 2D cross-sections and full 3D segmentations. Segmentation results are overlaid onto the original image data for visual comparison. A 3D graphical display of the segmented organ is color-coded depending on the selected metric for measuring segmentation difference. The new tool is in routine use for intra- and inter-rater reliability studies and for testing novel automatic machine-segmentation versus a gold standard established by human experts. Preliminary studies showed that the new tool could significantly improve intra- and inter-rater reliability of hippocampus segmentation to achieve intra-class correlation coefficients significantly higher than published elsewhere.

1. Scoring Measurement Methods

The computer vision community started several efforts with a number of workshops, conferences, and special issues of journals on the topic of empirical evaluation technique, see for example [Bowyer 1998], [Niessen 2000], [Vincken 2000], [Chalana 1997], and [Remiejer 1999]. Measuring performance of algorithms or human raters in image segmentation requires an appropriate metric, a “goodness” index that gives us a

W. Niessen and M. Viergever (Eds.): MICCAI 2001, LNCS 2208, pp. 516-523, 2001.

valid measure of the quality of a segmentation result. A good source and discussion of techniques is found in the most recent book, *Performance Characterization in Computer Vision* [Klette 2000]. Typical procedures for validation of computer-assisted segmentation are listed in [Kapur 1996]: Segmentation results are validated by a) visual inspection, b) comparison with manual segmentation, c) tests with synthetic data, d) use of fiducials on patients, and e) use of fiducials and/or cadavers. The problem is not that there is no ground truth for medical data, but that the ground truth is not typically available to the segmentation validation system in any form that can be readily used. For some structures or parts of structures the boundary can only be known with non-negligible tolerance. For other structures, the ground truth/gold standard is in reality barely fuzzy.

The following list covers metrics to measure the differences of segmentation results for measuring and comparing the reliability of intra-rater, inter-rater and machine-to-rater segmentations. We have developed a new tool called VALMET [freely avaible at http://www.ia.unc.edu/public/valmet] that reads an original 3D image to be segmented and a series of 3D segmentation results. Valmet calculates different metrics to assess pairwise segmentation differences and differences between groups. It further displays volumetric images with overlaid segmentations as 3 orthogonal sections with coupled cursors and as 3D renderings (see Fig. 2).

1.1 Volumes

A feature most easily accessible is the total volume of a structure. This is the simplest morphologic measure and often used in reliability studies in neuroimaging applications. For binary segmentations, we calculate the number of voxels adjusted by the voxel volume. More precise volumetric measurements can be obtained by fitting a surface (marching cubes, e.g.) with sub-voxel accuracy and calculating the volume by integration. Comparing volumes of segmented structures does not take into account any regional differences and does not give an answer to the question where differences occur. Further, over- and underestimation along boundaries or surfaces cancel and can give excellent agreement even if the boundary segmentation is poor [Niessen 2000].

1.2 Volumetric Overlap (True and False Positives, True and False Negatives)

One approach for taking in to account the spatial properties of structures is a pair-wise comparison of two binary segmentations by relative overlap. Assuming spatial registration, images are analyzed voxel by voxel to calculate false positives, false negative, true positive and true negative voxels. Well accepted measures are the intersection of subject and reference divided by the union, $(S \cap R)/(S \cup R)$, or intersection divided by reference $(S \cap R)/R$. Both measures give a score of 1 for perfect agreement and 0 for complete disagreement. The first is more sensitive to differences since both denominator and numerator change with increasing or decreasing overlap. The measure gives comparable results if applied at different institutions if structures and resolution of image data are standardized. However, the overlap measure depends on the size and the shape complexity of the object and is related to the image sampling. Assuming that most of the error occurs at the boundary of objects, small objects are penalized and get a much lower score than large objects.

1.3 Probabilistic Distances between Segmentations

In a lot of medical image segmentation tasks there are no clear boundaries between anatomical structures. Absolute ground truth by manual segmentation does not exist and only a `fuzzy' probabilistic segmentation is possible. Manual probabilistic segmentations can be generated by aggregating repeated multiple segmentations of the same structure done either by a trained individual rater or by multiple raters. We have developed a probabilistic overlap measure between two fuzzy segmentations derived from the normalized L^1 distance between two probability distributions. The probabilistic overlap is defined as

$$POV(A,B) = 1 - \frac{\int | P_A - P_B |}{2\int P_{AB}},$$

where P_A and P_B are the probability distributions representing the two fuzzy segmentations and P_{AB} is the pooled joint probability distribution.

1.4 Maximum Surface Distance (Hausdorff Distance)

The Hausdorff-Chebyshev metric defines the largest difference between two contours. Given two contours C and D, we first calculate for each point *c* on *C* the minimal distance to all the points on contour D, $d_c(c,D)$, $d_C(c,D) = \min\{d_{ps}(c,s), s \subset D\}$. We calculate this minimal distance for each boundary point and take the maximum minimal distance as the "worst case distance", $h_C(C,D) = \max\{d_C(c,D), c \in C\}$. The Hausdorff metric is not symmetric and $h_c(C,D)$ is not equal to $h_c(D,C)$ (see drawn figure), which is accounted for by finally calculating $H_C(C,D) = \max\{h_C(C,D), h_C(D,C)\}$. The Hausdorff metric calculation is computationally very expensive, as we need to compare each contour point to all the other ones. A comparison of complex 3D surfaces would require huge number of calculations. The VALMET implementation uses 3D Euclidean distance transform calculation on one object and overlay of the second object to efficiently calculate the measure. The measure is extremely sensitive to outliers and does not reflect properties integrated along the whole boundary or surface. In certain cases, however, where a procedure does have to stay within certain limits, this measure would be the metrics of choice.

1.5 Mean Absolute Surface Distance

The mean absolute surface distance tells us how much on average the two surfaces differ. This measure integrates over both over- and under-estimation of a contour, and results in an L^1 norm with intuitive explanation [Chalana 1997]. The calculation is not straightforward if point to point correspondence on two surfaces is not available. We use a similar strategy as for the Hausdorff metric calculation, namely signed Euclidean distance transforms on one object and overlay of the second object surface. We then trace the surface and integrate the distance values. This calculation is not

symmetric, since distances from A to B are not the same as B to A (see discussion Hausdorff distance above). We therefore derive a common average by combining the two averages. The mean absolute distance, as opposed to binary overlap, does not depend on the object size. As a prerequisite, however, it requires existing surfaces and is therefore only suitable for single object comparison.

1.6 Interclass Correlation Coefficient for Assessing Intra-, Inter-rater and Rater-Machine Reliability

A common measure of reliability of segmentation tasks is the intraclass correlation coefficient. The measure calculates the ratio between the variance of a normally population and the "population of measurements", i.e. the variance of the population σ_b^2 plus the variance of the rater σ_0^2. The intraclass correlation is thus defined as $\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_0^2}$. If the rater variance is small relative to the total, then the variation in measurements among different cases will be due largely to natural variation in the population and thus close to 1. Hence we can be confident in the rater's reliability.

In neuroimaging applications, inter- and intra-rater reliability studies based on volumetric measurements have become standard. Commonly accepted values range from 0.9 to 0.99 for volume assessments with manual tracing of simply shaped subcortical structures or organs like kidneys and liver, for example.

2 Visualization of Intra- and Inter-rater Reliability

Visualization of intra- and inter-rater reliability on 2D cross-sections with label overlay and by 3D surface renderings is shown in Fig. 1. The concept is as follows: We load a 3D volumetric gray level image dataset and a series of segmentation results either by different raters or as repeated measurements of one rater into the tool. The labels are overlaid onto the original image with variable opacity. The 3D rendering reconstructs the 3D surfaces and displays either intra-rater or inter-rater variability as color overlays. This tool has shown its usefulness to act as a training tool for manual rater's segmentation. The new capability to visually assess rater differences on 2D slices and 3D views is new and not available by other packages.

Fig. 2 shows the screen of VALMET applied to hippocampus segmentation. Repeated 3D manual segmentations of the hippocampus provided by several experts are compared to qualitatively and quantitatively assess the intra- and inter-rater reliability. The tool displays three orthogonal cuts with overlay of labeled regions and a 3D surface rendering of the object boundaries. The hue indicates the local surface direction either inwards (blue, see color bar in Fig. 2) and outwards (red, again see color bar in Fig. 2) relative to the reference object and the distance between the surfaces, according to the metric chosen.

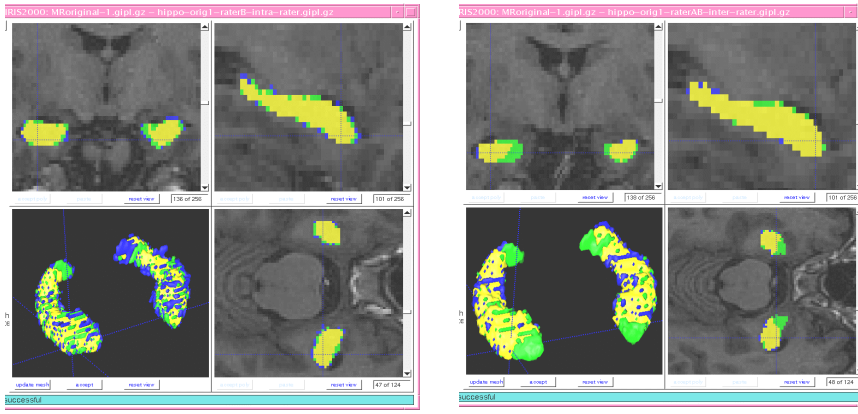

Fig. 1: Qualitative assessment of intra- and inter-rater reliability. The images show 2D orthogonal sections of a region of interest of the interior brain with segmentation of the left and the right hippocampal structures. Left: Intra-rater variability of 3 segmentations (observations) by one rater with yellow=3, green=2, and blue=1 votes per voxel. The 3D rendering displays the regional fuzziness of the boundary. Right: Inter-rater variability between two raters by comparison of two average segmentations. Yellow marks the region segmented by both, and blue and green regions segmented by only one of them. This displays clearly illustrates the agreement/disagreement between the raters, which is dominant in the hippocampus amygdala transition area (HATA) and the region of the hippocampal tail.

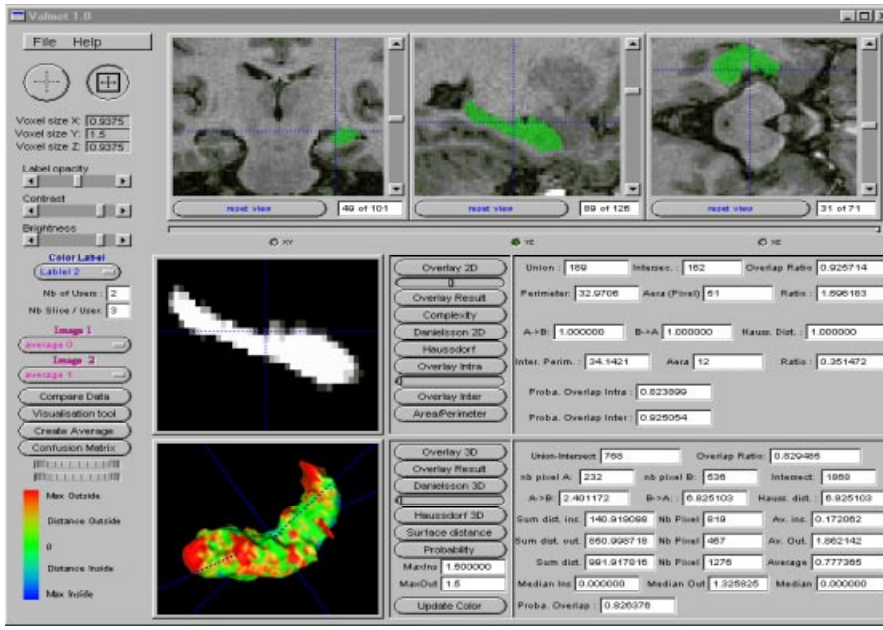

Fig. 2: User interface of VALMET. The tool calculates overlap measures, Hausdorff distance, mean absolute (and signed) surface distances, and probabilistic overlap. The 3D rendering provides a color display of both intersecting surfaces (green and red), showing regional differences between two surfaces. The application shows the result of a inter- and intra-rater hippocampus segmention study.

3. Segmentation Validation: Manual Hippocampus Segmentation

Unlike some anatomical structures the hippocampus as imaged through MRI has no clear boundaries, and it is very difficult to establish ground truth by manual segmentation. Hence it is very important to quantify variability in manual segmentations done by trained raters. As part of a large schizophrenia neuroimaging study, intra- and inter-rater reliability were tested with blind studies of series of 3D image data. For each series, we randomly selected 5 cases from an ongoing schizophrenia. The 5 cases were replicated 3 times and numbered randomly resulting in 15 image datasets, numbered differently for each rater. Trained raters go through all theses cases and segment left and right hippocampal structures using a new 3D segmentation tool IRIS [IRIS, 1999] developed by our group. The tool allows triplanar region editing and graphical 2D/3D interaction between image planes and segmented objects. We used an intraclass correlation program written in SAS (SAS Institute Inc.) to calculate intra- and inter-rater reliability.

Table 1 shows the reliability of two of the raters. We tested the reliability in two series, a first series after raters have been trained with the tools and became familiar with the instructions for hippocampus, and a second series after they evaluated and compared their results using the new tools described above. The results of the first series show that the reliability of raters A and B differs significantly between right and left hippocampus, each achieving a high reliability for one of the structures. The inter-rater reliability of 0.75 for the right and 0.62 for the left

Table 1: Reliability of manual hippocampus segmentation.

Intraclass Correlation: Manual Hippocampus Segmentation

Study design: 2 raters, 5 cases, 3 observations each
Analysis: Individual and pooled analysis

First reliability series

	individual analysis		pooled analysis	
	intra-rater rater A	intra-rater rater B	intra-rater A and B	inter-rater A vs. B
right hippocampus	0.89067	0.66422	0.77241	0.75062
left hippocampus	0.69061	0.85157	0.81391	0.61923

Second reliability series

	individual analysis		pooled analysis	
	intra-rater rater A	intra-rater rater B	intra-rater A and B	inter-rater A vs. B
right hippocampus	0.96073	0.88145	0.93229	0.67325
left hippocampus	0.95416	0.94822	0.96094	0.48218

hippocampus suggest that the left hippocampus is more difficult to segment than the right hippocampus. The second series was measured after the two raters visualized their segmentations using VALMET and revised the protocol. Interestingly, they both are becoming very reliable. This is reflected in reliabilities up to 0.95 and in the pooled intra-rater reliability of 0.93 and 0.96. However, the reliability between raters (inter-rater) became worse and dropped significantly from 0.75 to 0.67 for the right and from 0.61 to 0.48 for the left hippocampal structures. The second series used 5 different cases with 3 replications. In conclusion, we find that the intra-rater reliability for manual hippocampus segmentation was very high in comparison to studies done at other sites (Hogan 2000). A reliability of 0.95 for the manual segmentation of a structure as difficult as the hippocampus has to be considered excellent. We attribute this performance to the 2D/3D capabilities of the IRIS segmentation tool and VALMET. The inter-rater reliability is insufficient and reflects that both raters do excellent but different segmentations.

4. Discussion

No consensus exists regarding a necessary and sufficient set of measures to characterize segmentation performance. We plan to provide a suite comprising a reasonable variety of geometric and statistical methods. In addition to the measures already implemented in the prototype validation tool VALMET, we will consider providing a number of others including moments and volume of error voxels normalized by the surface area. Measures implemented in VALMET and other geometric measures reported in the literature tend to favor least squares measures. Measures in this class are intuitive and work well for noise-free data. However real medical images have structure noise and random noise that can lead to high spatial frequencies in segmented surfaces. Methods based on least-squares measures are very sensitive to even a small number of extreme data values in the sense that a small number of outlier voxels can disproportionately bias a measure and make an otherwise good segmentation appear to compare poorly with truth. Statistically robust methods include quantiles of distance, which are robust to extreme values. A next version of VALMET will include the calculation of a surface distance histogram and choice of arbitrary quantiles.

Bibliography

Bowyer, K.W., Phillips, P., Empirical Evaluation Techniques in Computer Vision, IEEE Computer Society, 1998

Chalana V and Kim Y: A methodology for evaluation of boundary detection algorithms on medical images, IEEE Trans. Med. Imaging 16: 642-652 (1997)

Hogan, R.E., Mark, K.E., Wang, L., Joshi, S., Miller, M.I. and Bucholz, R.D., Mesial Temporal Sclerosis and Temporal Lobe Epilepsy: MR Imaging Deformation-based Segmentation of the Hippocampus in Five Patients, Radiology 216, pp. 291-297, July 2000

IRIS (1999): Interactive Rendering and Image Segmentation, UNC student project spring 1999, Gregg, D., Larsen, E., Neelamkavil, A., Sthapit, S. and Wynn, Chris, Dave Stotts and Guido Gerig, supervisors, http://www.cs.unc.edu/~stotts/COMP145/homes/iris/

Kapur, T., Grimson, E.L., Wells, W.M., and Kikinis, R., Segmentation of brain tissue from magnetic resonance images, Medical Image Analysis, 1(2);109-127, 1996

Klette R, Stiehl SH, Viergever MA, and Vincken KL, eds: Performance Characterization in Computer Vision, Kluwer Academic Publishers (2000)

Niessen, W.J., Bouma, C.J., Vincken, K.L., Viergever, M.A., Error Metrics for Quantitative Evaluation of Medical Image Segmentation, in Performance Characterization in Computer Vision, Kluwer Academic Publishers, pp. 299-311, 2000

Remiejer P, Rasch C, Lebesque JV, and van Herk M: A general methodology for three-dimensional analysis of variation in target volume delineation. Med. Phys. 27: 1961-1970 (1999)

Vincken, K.L., Koster, A.S.E., De Graaf, C.N. and Viergever, M.A., Model-based evaluation of image segmentation methods, in Performance Characterization in Computer Vision, Kluwer Academic Publishers, pp. 299-311, 2000