Methodology of Evaluation in Medical Image Computing

March 11 – 16, 2001

About one decade ago, Yannis Aloimonos complained that "Unfortunately, there is a disconcerting lack of visual systems which perform well in real-world environments, particularly when compared to the amount of mathematical theory published on the subject."—a complaint which not only holds for computational vision in general but in particular also for the safety-critical case of medical image computing (a terminus technicus which commonly subsumes medical image formation, processing, analysis, interpretation, and visualization). One reason for this unfortunate situation is clearly the fact that the experimental basis of computational vision as a scientific discipline is still rather weak. As a down-to-earth-consequence, e.g., it is by no means clear for an industrial system designer, on which grounds she/he should rely on a particular algorithm, method, or proposed tool once she/he is faced with the problem of putting academic research to work. Neither it seems to be clear for a clinician, what kind of as well as what degree or quality of support in his routine work she/he can expect from proffered medical image computing (MIC) tools claimed to support routine work. Put in other words, MIC seen as a coin has a shiny and scientifically rewarding theory side but a rather rusty, not to say puny, practice side.

Meanwhile in the MIC community a growing awareness of the fact can be observed that evaluation aiming at performance characterization is a critical issue. In a complementing way, a strong need from both clinical and industrial actors for tackling theoretical as well as experimental problems associated with this issue has to be stated, since dissemination of theoretical advances into practical settings requires a deep understanding of assets, limitations, application scope, etc. of MIC algorithms, methods, and tools. Moreover it is safe to state that without such a deep understanding gained from a scientific approach the design of interactive MIC systems will be severely hampered, since human-centered efficient interaction should take place on the basis of results of computational processes which are trustworthy—ideally results, which are consistent with theoretical proofs of a computational theory. In contrast to other application domains drawing upon visual data, MIC stands out for reasons of required safety, accuracy, robustness, ergonomy, etc. Apart from that, MIC is seen as a major future high-tech market also, hence the development of successful products strongly depends on bridging the gap between theory, experiment, and practice. Obviously, solutions to these problems reside in a space composed of multiple dimensions to name a few: MIC theory, practice of MIC (incl. design of algorithms and visual data structures), clinical requirement analysis, and industrial platform constraints.

Due to the lack of well-grounded, internationally accepted, and standardized methods for evaluation and given the specifity of MIC as briefly sketched above, it was high-time to bring together leading experts from the MIC community in the inspiring atmosphere of Schloss Dagstuhl to discuss the state-of-the-art/technology as well as routes to be jointly

taken in the near future. After the successful first seminar on more general issues of performance characterization in computational vision in 1998 and given the most recent publications on domain-unspecific topics of evaluation in computational vision (see, e.g., "Empirical Evaluation Techniques in Computer Vision", edited by K.W. Bowyer and J. Phillips, IEEE Computer Society Press, 1998; "Proc. Workshop on Performance Characterisation and Benchmarking of Vision Systems (Las Palmas de Gran Canaria, Canary Islands, January 1999)", edited by A. Clark and P. Courtney; IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 21, No. 4, April 1999, Special Section on "Empirical Evaluation of Computer Vision Algorithms", edited by P.J. Phillips and K.W. Bowyer; "Performance Characterization and Evaluation of Computer Vision Algorithms", edited by R. Klette, H.S. Stiehl, M.A. Viergever, and K.L. Vincken, Kluwer Academic Publishers, 2000; "Tutorial on Performance Characterisation of Computer Vision Techniques (European Conf. on Computer Vision, Dublin, Ireland, June 2000)" by P. Courtney and N. Tacker; "Proc. 2. Workshop on Empirical Evaluation Methods in Computer Vision (Dublin, Ireland, June 2000)", edited by H.I. Christensen and P.J. Phillips, published as CPAV technical report no. 243, Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, Sweden, June 2000), this seminar will focus on particular domain-specific issues as related to medical imagery, e.g. performance characterization of computational processes for segmentation, analysis, registration, and real-time visualization of multi-dimensional and multi-modal images.

In terms of priority, the focus w.r.t. presentations and discussions has been set on the following concrete topics:

1. validation and evaluation of accuracy, robustness, etc. of algorithms for interactive/semi-automatic/automatic segmentation, analysis, registration, and visualization of medical imagery

2. theoretical/methodological issues such as definition of ground truth and gold standards, value of phantoms, imaging simulators, and synthetic test data

3. selection of a representative set of clinical routine images related to specific domains and tasks (certified clinical reference cases and test image data base)

4. identification of open questions (see appendix)

One of the main goals of the successful seminar was to contribute towards a more seamless methodology of validation, evaluation, and performance characterization across various levels—thus to contribute also to bridge the gap between MIC theory and the end user and to provoke fruitful discussions beyond the ivory tower. Despite the recent progress and achievements reported by the seminarians, it became quite clear during the week we had at our disposal that the MIC community has to retain the seminar topic on its research agenda for the years to come.

Acknowledgment

We are grateful to the administration of the Dagstuhl enterprise for creating such an inspiring environment as well as for providing the excellent facilities which significantly con-

tributed to the success of our meeting. We are also grateful to Sönke Frantz for compiling the abstract booklet.

Kevin W. Bowyer Murray H. Loew H. Siegfried Stiehl Max A. Viergever

Hamburg, June 2001

List of Speakers

Monday

P. Courtney, K. Rohr, S. Frantz, G. Prause, J.Z. Chen, G. Gerig, T. Netsch

Tuesday

E.B. Dam, M. Nielsen, S. Delgado Olabarriaga, A. Pommert, F. Pernus, C. Barillot, X. Pennec, J.A. Schnabel, J.C. Gee (evening session)

Wednesday

C. Rodríguez-Carranza, M. Mitschke, M.A. Viergever, M.H. Loew (evening session)

Thursday

A. Kuba, B. Likar, N. Kiryati, S.M. Pizer, K. van Bemmel, N. Karssemeijer, N. Thacker

Friday

Round table discussion with all participants (moderated by M.H. Loew and H.S. Stiehl)

Contents

Session I: Image Segmentation

Evaluation of a Medical Image Segmentation Algorithm

Patrick Courtney and Neil Thacker

Division of Imaging Science & Biomedical Engineering,
The University of Manchester, United Kingdom

This talk focussed on the importance of (1) understanding the assumptions underlying an algorithm and (2) matching the statistical properties of the data to those assumptions.

We examine an existing segmentation algorithm from Rusinek for determining partial volume estimate in a three-tissue brain image acquired using MRI. The assumptions underlying the algorithm are described and reviewed in the light of alternative imaging sequences. A revised more general form of the algorithm is proposed and its improved error properties (uniform errors) are presented. Experimental results are shown for a range of image sequencing protocols.

The ideas presented are extended in the later talk by Neil Thacker.

Accuracy of Landmark Localization

Karl Rohr

School of Information Technology,
International University in Germany, Bruchsal, Germany

This work analyses the accuracy of estimating the location of landmarks and characteristic image structures. Based on nonlinear estimation theory we here study the minimal stochastic errors of the position estimate of landmarks caused by noisy data. Given analytic models of the image intensities we derive closed-form expressions for the Cramér-Rao bound for different 2D and 3D structures, e.g., edges, lines, corners, circular symmetric landmarks, and blobs. It turns out, that the precision of localization depends on the noise level, the size of the observation window, the width of the intensity transitions, as well as on other parameters describing the considered image structure. The derived lower bounds can serve as benchmarks and the performance of existing algorithms can be compared with them. Medical applications of landmarks are, for example, measurement tasks or (nonrigid) image registration.

References

[1] C. Drewniok and K. Rohr. Model-Based Detection and Localization of Circular Landmarks in Aerial Images. *Internat. Journal of Computer Vision*, 24(3):187–217, 1997.

[2] K. Rohr. On the Precision in Estimating the Location of Edges and Corners. *Journal of Mathematical Imaging and Vision*, 7(1):7–22, 1997.

[3] K. Rohr. *Landmark-Based Image Analysis Using Geometric and Intensity Models.* Computational Imaging and Vision Series, Vol. 21. Kluwer Academic Publishers, Dordrecht Boston London, 2001.

Validation of Semi-Automatic Differential Approaches to 3D Point Landmark Extraction

Sönke Frantz[1], Karl Rohr[2], and H. Siegfried Stiehl[1]

[1]Arbeitsbereich Kognitive Systeme, Fachbereich Informatik,
Universität Hamburg, Germany

[2]School of Information Technology,
International University, Bruchsal, Germany

Three-dimensional (3D) anatomical point landmarks are useful image features for a variety of medical image analysis tasks, in particular for 3D image registration. Apart from manually extracting such landmarks from 3D images, which is usually time-consuming and difficult to reproduce, only a few, mainly differential approaches have been developed [1],[2] (for a recent alternative approach to landmark extraction based on deformable models, see [3]). While differential approaches have rather low computational costs, a problem is that often a rather large number of false detections is obtained. In [4], we developed an improved semi-automatic differential approach to landmark detection whose key features are i) the automatic selection of a suitable region-of-interest (ROI) size to exclude disturbing neighboring structures, ii) the application of a robust 3D differential detection operator using only first order image derivatives ([2]), and iii) the incorporation of additional a priori knowledge of the intensity structure at a landmark in terms of the surface curvature to automatically reject potential false detections.

In our experimental studies, we first analyzed the detection performance of our new approach for different anatomical landmarks in 3D MR and CT images of the human head, considering, e.g., the number and loci of detections, the significance and the separability of the obtained detections w.r.t. the detection operator response. The new approach significantly improves the detection performance, as compared to applying the detection operator alone within a ROI of fixed size and without utilizing a priori landmark knowledge.

The next step was a validation study in which we compared the performance of our new semi-automatic approach with that of a purely manual (standard) procedure for landmark extraction [5]. As application, we considered rigid registration of 3D MR and CT images. The main result of our study is that compared to a purely manual procedure, a) the elapsed time for landmark extraction can be significantly reduced with the semi-automatic procedure, b) the registration results based on semi-automatic landmark extraction generally show similar quality, and c) the inter-observer variability of the localized landmark posi-

tions as an indicator for reproducibility is significantly smaller with the semi-automatic procedure.

This work was supported by Philips Research Hamburg, Project IMAGINE (IMage- and Atlas-Guided Interventions in NEurosurgery).

References

[1] J.-P. Thirion. New Feature Points based on Geometric Invariants for 3D Image Registration. *Internat. Journal of Computer Vision*, 18(2):121–137, 1996.

[2] K. Rohr. On 3D differential operators for detecting point landmarks. *Image and Vision Computing*, 15(3):219–233, 1997.

[3] S. Frantz, K. Rohr, and H.S. Stiehl. Localization of 3D Anatomical Point Landmarks in 3D Tomographic Images Using Deformable Models. In S.L. Delp, A.M. DiGioia, and B. Jaramaz (eds.), *Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI 2000)*, Lecture Notes in Computer Science 1935, pp. 492–501. Springer-Verlag, Berlin, 2000.

[4] S. Frantz, K. Rohr, and H.S. Stiehl. Improving the Detection Performance in Semi-automatic Landmark Extraction. In C.J. Taylor and A.C.F. Colchester (eds.), *Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI 1999)*, Lecture Notes in Computer Science 1679, pp. 253–262. Springer-Verlag, Berlin, 1999.

[5] S. Frantz, K. Rohr, H.S. Stiehl, S.-I. Kim, and J. Weese. Validating Point-based MR/CT Registration Based on Semi-automatic Landmark Extraction. In H.U. Lemke, M.W. Vannier, K. Inamura, and A.G. Farman (eds.), *Proc. Computer Assisted Radiology and Surgery (CARS 1999)*, pp. 233–237. Elsevier Science, Amsterdam, 1999.

[6] S. Frantz. *Local and Semi-Global Approaches to the Extraction of 3D Anatomical Landmarks from 3D Tomographic Images.* Doctoral dissertation, Fachbereich Informatik, Universität Hamburg, May 2001.

Evaluation of Edge-Based Segmentation Approaches

Guido Prause, Andrea Schenk, and Heinz-Otto Peitgen

MeVis - Center for Medical Diagnostic Systems and Visualization, Bremen, Germany

Fully-automated segmentation is still an unsolved problem for most clinical applications due to the large variety of image modalities, imaging protocols, and biological variability. Hence there is a strong need for interactive approaches allowing clinicians an efficient and reproducible segmentation of volumetric medical images. The evaluation of interactive segmentation methods requires not only well-defined quantitative measures such as contour

or region metrics but also has to take into account the clinical task and the aspect of usability.

As an example of a user-steered edge-based approach for the segmentation of the liver parenchyma in three-dimensional CT and MRT data sets, we present a combination of live-wire and shape-based interpolation [1, 2]. First evaluation results of the method's application for pre-operative planning of living-related liver transplants and oncologic liver surgery are reported.

Finally, a recently started interdisciplinary network (VICORA - Virtual Institute for COmputer assistance in clinical RAdiology [3]) is presented, bringing together clinical radiologists, medical imaging scientists, and industry in Germany. This network will serve as a platform for a consensual specification, development, and evaluation of software tools as well as for clinical multi-center studies.

References

[1] A. Schenk, G. Prause, H.-O. Peitgen. Efficient Semiautomatic Segmentation of 3D Objects in Medical Images. In S.L. Delp, A.M. DiGioia, and B. Jaramaz (eds.), *Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI 2000)*, Lecture Notes in Computer Science 1935, pp. 186–195. Springer-Verlag, Berlin, 2000.

[2] A. Schenk, G. Prause, H.-O. Peitgen. Local Cost Computation for Efficient Segmentation of 3D Objects with Live Wire. In M. Sonka and K.M. Hanson (eds.), *Proc. SPIE's Internat. Symposium on Medical Imaging 2001: Image Processing,* Vol. 4322. SPIE Press, Bellingham, WA, in press.

[3] http://www.vicora.de/

Segmentation Validation via Monte Carlo Simulation

James Chen, Stephen Pizer, and Edward Chaney

Medical Image Display & Analysis Group,
University of North Carolina, Chapel Hill, NC, USA

For the many currently available image segmentation methods, it is important to evaluate their performance (accuracy and efficiency) and make comparison in different applications. As a large set of test images with known ground truth segmentation will be desired for this purpose, we propose to generate realistic synthetic images via Monte Carlo simulation.

Given a pair of images of interest collected from the normal population, we can regard one image as the geometric transformation of the other. Therefore, if one particular image is taken as the master reference image (R), an arbitrary image of the same type (R') can be transformed from R by a warp function W: R'=W(R). Similarly, segmentation of R' (denoted as S') can be derived from the ground truth segmentation of R (denoted as S, assuming it is known) by the same warp function: S'=W(S).

Since a particular segmentation method needs to be tested on many images if any statistical conclusion can be reached regarding its performance, the synthetic image generation proceeds in two steps: training and sampling. First, the reference image R is warped into a collection of clinical images, which represents the biological variation among the normal population. Then the probability distribution of the variation among the family of warp functions W can be studied by Principal Components Analysis (PCA) or Markov Random Field Analysis (MRF) technique. After we have learned the variation probability distribution p(W) within the training images, an infinite number of realistic synthetic images of interest and the corresponding "ground truth" segmentations can be generated by sampling this probability distribution p(W).

When the synthetic data set becomes available, it will first be used to validate the M-rep model based image segmentation method. The ultimate goal of this project is to make this validation tool available via web to the medical imaging research community.

VALMET: A new tool for assessing and improving reliability of object segmentation from 3D medical images

Guido Gerig and Matthieu Jomier

Departments of Computer Science and Psychiatry,
University of North Carolina, Chapel Hill, NC, USA

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 variations. Input to the tool are an original 3D image (MRI, CT, ultrasound) and a series of segmentations either generated by several human raters 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 Hausdorff (maximum). 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. Current driving applications are organ segmentation (kidney, bladder, prostate) from the pelvic region in radiotherapy planning and segmentation of subcortical brain structures in neuroimaging studies, with focus to the hippocampus, amygdala and caudate nucleus. We can demonstrate that the use of our new tool could significantly improve intra- and inter-rater reliability of hippocampus segmentation to achieve intra-class correlation coefficients significantly higher than published elsewhere. In conclusion, we expect the new tool with its advanced validation methodology to play an important role in a variety of quantitative medical imaging projects including clinical studies and the development of new segmentation techniques.

Validation of Medical Registration and Segmentation Algorithms at Philips Research Hamburg

Thomas Netsch

Philips Research Laboratories Hamburg, Germany

The presentation focuses on the technical evaluation of image processing algorithms developed within the Digital Imaging Group at Philips Research in Hamburg. After a brief description of the framework for development and evaluation, two examples of current projects with respect to their technical validation are explained in more detail: the spatial alignment of functional MR brain images and the segmentation of 3D images with shape models. The latter example also describes the validation of the shape model. Finally, the results of a clinical trial for the automated detection of shutter edges in X-ray images are presented.

Session II: Interactive Image Segmentation

Objective Evaluation of an Interactive Segmentation Tool

Erik B. Dam and Mads Nielsen

IT University of Copenhagen, Denmark

For an evaluation of the clinical value of an interactive segmentation tool the user is indispensable. However, for a preliminary development-phase evaluation an automatic, quantitative method can be used for the numerous test runs needed for optimisation of the internal parameters. This talk described an evaluation method for a building block based interactive segmentation method and showed how the evaluation was used to optimise the program. Specifically, the effect of the choice of diffusion scheme in the underlying multi-scale watershed segmentation method was evaluated. The evaluation used real and simulated ground truth segmentations of the white and grey matter tissues from MR brain scans.

Some Thoughts on the Evaluation of Interactive Methods for Medical Image Segmentation

Sílvia Delgado Olabarriaga

Informatics Institute,
Federal University of Rio Grande do Sul, Porto Alegre, Brazil

Interactive methods for the segmentation of medical images are composed of basically two parts: the user and the computer. When evaluating the behaviour of such methods, both parts should be taken into account in a combined fashion. From the user perspective, some of the important aspects are safety, output quality, ease of use, required effort to obtain the desired segmentation, aids offered by the program and clinical benefit [1]. From the computer perspective, it is important to characterise the method's behaviour with respect to parameter values, classes of admitted inputs (images, user input), output quality, processing effort (time and complexity) and implementation quality [2, 3]. When both perspectives are combined, it becomes clear that the evaluation of such methods is far from a trivial task.

Only a few criteria extracted from the list above are discussed in the talk, namely accuracy (the result corresponds to the truth), precision (the result is repeatable), and efficiency (the amount of effort/time needed to complete the task to a satisfactory quality degree) [4]. Although some of these criteria also apply for image segmentation methods in general, the talk focuses on their interpretation in the context of interactive methods. Some thoughts are presented on the needs for an evaluation setting for the criteria. Examples

are presented to provoke discussion on how experiments could be organised is a manner to obtain clear answers about a method's performance with respect to this set of evaluation criteria.

References

[1] S. Graham, R.H. Taylor, and M. Vannier. Needs Assessment for Computer-Integrated Surgery Systems. In S.L. Delp, A.M. DiGioia, and B. Jaramaz (eds.), *Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI 2000)*, Lecture Notes in Computer Science 1935, pp. 931–939. Springer-Verlag, Berlin, 2000.

[2] K.W. Bowyer. Validation of Medical Image Analysis Techniques. In M. Sonka and J.M. Fitzpatrick (eds.), *Handbook of Medical Imaging*, Vol. 2: Medical Image Processing and Analysis, Chapter 10, pp. 567–607. SPIE Press, Bellingham, WA, 2000.

[3] T.S. Yoo, M.J. Ackerman, and M. Vannier. Toward a common validation methodology for segmentation and registration Algorithms. In S.L. Delp, A.M. DiGioia, and B. Jaramaz (eds.), *Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI 2000)*, Lecture Notes in Computer Science 1935, pp. 422–431. Springer-Verlag, Berlin, 2000.

[4] S.D. Olabarriaga and A.W.M. Smeulders. Interaction in the Segmentation of Medical Images: a Survey. *To appear in Medical Image Analysis.*

Visual Feedback for Interactive Segmentation in Volume Visualization

Andreas Pommert

Institute of Mathematics and Computer Science in Medicine,
University Hospital Eppendorf, Hamburg, Germany

Modern volume visualization methods are capable of rendering the data with arbitrary subvoxel precision. However, as regards the accuracy of the visualized surfaces, limitations are implied by the underlying data. Assuming an object with a high contrast step edge, we are investigating surface location error as a function of point spread function (PSF) of the imaging device, as well as interpolation method and threshold value used for rendering. Provided that a suitable threshold and a width (full width at half maximum, FWHM) of the PSF equal or larger than the voxel spacing are used, it could be shown that the surface location error is about one order of magnitude better than the voxel spacing. Furthermore, it is shown that the artifacts appearing on the surface provide a visual feedback for optimizing the threshold value.

Session III: Image Registration

Registration of 3D MR and 2D x-ray images

Franjo Pernus

Department of Electrical Engineering,
University of Ljubljana, Slovenia

There is a growing need for non-invasive intraoperative localisation of the patient in the operating or radiotherapy room. Common imaging modalities for guiding certain interventions are x-ray fluoroscopy, digital radiography, and ultrasound. Although helpful and real time, these modalities are only two-dimensional (2D), so they lack the 3D spatial information contained in computed tomography (CT) or magnetic resonance (MR) images, which are usually taken preoperatively for surgery or radiotherapy planning. One method of allowing 3D information from CT or MR images to be used during interventional procedures is to register the CT or MR scan to one or more 2D intraoperative images. Basically, one must perform a 2D to 3D registration by which that transformation is found, which brings a 3D image of an anatomy into the best spatial correspondence with respect to 2D images of the same anatomy. The registration of 3D MR and 2D x-ray images is especially difficult because images of different modalities and different domains need to be registered. In the present paper we address this problem by finding the best match between 2D projection images of the 3D MR gradient image and the x-ray gradient image. Validation of a registration method is a crucial and difficult issue. We have validated the proposed method by using the Visible Human CT and MR data of the pelvis and hips. By projecting CT data onto a plane, simulated x-ray images were created which were then registered with MR data. The experimental results show that in 60% of registrations a target registration error smaller than 10 mm was achieved.

Linear Versus Non-Linear Registration of Brainscans - What can we Expect? Some Evaluation Issues

Christian Barillot

IRISA, Projet VISTA, Rennes, France

Within the scope of three-dimensional brain imaging and brain functional mapping, I have presented a retrospective evaluation framework of inter-individual fusion scheme to register brain cortical anatomy and functional activations. We have evaluated local, global, linear, piecewise affine and non-linear registration approaches, coming from more than five international research groups. The evaluation procedure relies on global and local criteria such as gray/white matter segmentation map overlap, correlation ratio of the mean Lvv intensities (i.e. 3D measure of the curvature of the cortex from the MRI volumes), distance

and shape variation of cortical sulcal landmarks (segmented using the "active ribbon" approach).

The data set was made of 18 MRI volumes from healthy right handed men volunteers. MEG somatosensory activation of fingers have been performed on these subjects. Visual and global measures seem to advantage the non-linear global registration methods while local measures, based on sulci, did not show any significant differences between all global methods. In addition, it has been shown that a new local non-linear registration method based on active shape models and thin-plate splines is able to better retrieve the inter-individual variability of the functional somatosensory activities.

Evaluation of the Uncertainty in Various Registration Problems

Xavier Pennec

Epidaure, INRIA Sophia Antipolis, France

When analysing automatic registration algorithms, one can distinguish between gross errors (convergence to wrong local minima) and small errors around the exact transformation. The *robustness* is quantified by the size of the basin of attraction of the right solution or the probability of false positives. The small errors may be sorted into *systematic biases*, *repeatability* and *accuracy*. The repeatability accounts for the errors due to internal parameters of the algorithm, mainly the initial transformation, and to the finite numerical accuracy of the optimisation algorithm, while the external error accounts for the propagation of the error in the data into the optimisation result.

In the first part, we present the theoretical methods to propagate the uncertainty from the data to the resulting transformation in the case of mono-modal feature based algorithms [2]. On standard MR or CT images of the same patient, our method typically predict an accuracy of less than 0.1 mm in the area of interest. To verify this prediction, we used a series of multiple echo-images. The comparison of intra-echo transformations validates the uncertainty prediction while the comparison of inter-echo transformations exhibit a bias of about 0.2 mm in translation. After the correction of the bias, the accuracy predictions are fully validated on several patient series [1].

Generalising this kind of theoretical approaches to curves, surfaces and intensity-based registrations is difficult because the homology assumption between matched features is false. Moreover, there is a spatial correlation between neighbouring features. To still perform an accuracy evaluation in these cases, we propose in the second part an *a posteriori* study on a multimodal (3D MR/ 3D US) intensity-based registration method. In a first step, we show how to evaluate the robustness and the repeatability using a Monte-Carlo sampling of the initial transformation. In a second step, we perform an accuracy study using registration loops. Typically, the variability of a point in MR_1 transformed into MR_2, then US_i and back to MR_1 will be $\sigma^2_{loop} = 2 * \sigma^2_{MR/US} + \sigma^2_{MR/MR}$. In this scheme, the influence of the intra-modality registration is evaluated and minimised using a multiple registration. Using many of these loops, we show that the expected MR/US registration

accuracy of our method is of the order of the MR voxel size [3].

In the future, we plan to extend the theoretical framework to point-surface rigid registration and validate it using an a posteriori evaluation method. Another work in progress is the analysis of registration errors in fMRI time-sequences registrations.

References

[1] X. Pennec, C.R.G. Guttmann, and J.P. Thirion. Feature-based registration of medical images: Estimation and validation of the pose accuracy. In W.M. Wells, A.C.F. Colchester, and S.L. Delp (eds.), *Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI 1998)*, Lecture Notes in Computer Science 1496, pp. 1107–1114. Springer Verlag, Berlin, 1998.

[2] X. Pennec and J.P. Thirion. A framework for uncertainty and validation of 3D registration methods based on points and frames. *Internat. Journal of Computer Vision*, 25(3):203–229, 1997.

[3] A. Roche, X. Pennec *et al.* Generalized Correlation Ratio for Rigid Registration of 3D Ultrasound with MR Images. In S.L. Delp, A.M. DiGioia, and B. Jaramaz (eds.), *Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI 2000)*, Lecture Notes in Computer Science 1935, pp. 567–577. Springer-Verlag, Berlin, 2000.

Visual Assessment and Quantitative Validation of Non-Rigid Registration Techniques

Julia A. Schnabel[1,a], Derek L.G. Hill[1], David J. Hawkes[1], and Daniel Rueckert[2,b]

[1]Computational Imaging Science Group,
Division of Radiological Sciences and Medical Engineering,
Guy's, King's and St. Thomas' School of Medicine, London, United Kingdom

[2]Visual Information Processing, Department of Computing,
Imperial College of Science, Technology and Medicine, London, United Kingdom

Validation of non-rigid registration is becoming a highly important topic in medical image processing. Traditionally, registration validation relies on measurements of robustness, consistency, visual assessment, and gold standard comparisons, e.g. when using intrinsic or extrinsic markers. Most of these approaches however are not applicable for non-rigid registration. In this presentation we discuss general registration aims and validation techniques, followed by summarizing the non-rigid registration algorithm by Rueckert et al. [1] which is based on a combined global and local motion model, where global motion is modelling a rigid or affine transformation, and local motion is modelling residual local deformations using free-form deformations based on B-Splines. The registration cost function consists of

normalised mutual information as a voxel similarity measure, and a regularizing smoothness constraint. This method has been developed for registration of contrast-enhanced MR mammography, for the application of which we present quantitative measures and visual assessment using expert ranking in order to show the registration performance [2]. We have investigated local volume changes in enhanced lesions which are caused by the algorithm [3], and furthermore, we have applied the registration method to inter-subject brain MRI registration of which we present visual assessment as well as landmark- and intensity based quantitative results. Finally, we discuss novel non-rigid registration validation approaches using tagged cardiac MRI [4], and simulation of biomechanical deformations [5].

[a]Funded by EasyVision Advanced Development, Philips Medical Systems, Best, NL.

[b]Received funding from EPSRC GR/L08519 whilst at CISG-GKT.

References

[1] D. Rueckert, L.I. Sonoda, C. Hayes, D.L.G. Hill, M.O. Leach, and D.J. Hawkes. Non-rigid registration using Free-Form Deformations: Application to breast MR images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999.

[2] E.R.E. Denton, L.I. Sonoda, D. Rueckert, S.C. Rankin, C. Hayes, M. Leach, D.L.G. Hill, and D.J. Hawkes. Comparison and evaluation of rigid and non-rigid registration of breast MR images. *Journal of Computer Assisted Tomography*, 23(5):800–805, 1999.

[3] C. Tanner, J. A. Schnabel, D. Chung, M. J. Clarkson, D. Rueckert, D.L.G. Hill, and D.J. Hawkes. Volume and shape preservation of enhancing lesions when applying non-rigid registration to a time series of contrast enhancing MR breast images. In S.L. Delp, A.M. DiGioia, and B. Jaramaz (eds.), *Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI 2000)*, Lecture Notes in Computer Science 1935, pp. 327–337. Springer Verlag, Berlin, 2000.

[4] R. Razavi, D. Rueckert, P. Summers, J.A. Schnabel, M. Miquel, E. Rosenthal, and E. Baker. Detection of arrythmogenic substrate by Magnetic Resonance Imaging. In *Society of Cardiovascular Magnetic Resonance (SCMR'01)*, 2001.

[5] J.A. Schnabel, C. Tanner, A. Castellano Smith, M.O. Leach, R. Hose, D.L.G. Hill, and D.J. Hawkes. Validation of non-rigid registration using Finite Element Methods. In *Proc. Information Processing in Medical Imaging (IPMI 2001)*, Lecture Notes in Computer Science, pp. 345–358. Springer Verlag, Berlin, in press.

Automatic detection of large misregistrations of multimodality medical images

Claudia E. Rodríguez-Carranza[1,3] and Murray H. Loew[2,3]

[1]Department of Computer Science,
[2]Department of Electrical and Computer Engineering,
[3]Institute for Medical Imaging and Image Analysis,
The George Washington University, Washington DC, USA

Before a retrospective registration algorithm can be used routinely in the clinic, methods must be provided for distinguishing between registration solutions that are clinically satisfactory and those that are not [1]. One approach is to rely on a human observer to inspect the registration results and reject images that have been registered with insufficient accuracy. As alternative, in this paper we present an algorithmic procedure that discriminates between large or "definitively bad" registrations and those that are close to correct alignment or "possibly good" registrations. We found that a new measure of distance between brain contours appears to identify misregistrations on the order of 15mm or more of root mean squared error at the corners of the volumes.

References

[1] J.M. Fitzpatrick, D.L.G. Hill, Y. Shyr, J. West, C. Studholme, and C.R. Maurer. Visual assessment of the accuracy of retrospective registration of MR and CT images of the brain. *IEEE Transactions on Medical Imaging*, 17(4):571–585, 1998.

Evaluation of Automated Reduction of Patient Motion Artefacts in Digital Subtraction Angiography

Erik H.W. Meijering, Wiro J. Niessen, and Max A. Viergever

Image Sciences Institute,
University Medical Center Utrecht, The Netherlands

The purpose of this study was to evaluate the performance of the automatic registration technique for motion artefact reduction in digital subtraction angiography (DSA) images, developed by the authors [1]. A total of 104 cerebral DSA images were processed both manually by means of pixel shifting, the current clinical standard, and automatically by using the automated technique. Four observers assessed the quality of the resulting corrected images, by comparing them mutually, and by comparing each type of corrected image to the corresponding original uncorrected image. The results of the evaluation indicated that the automatic technique is not only considerably faster, but also statistically significantly better than manual pixel shifting.

References

[1] E.H.W. Meijering, K.J. Zuiderveld, and M.A. Viergever. Image registration for digital subtraction angiography. *Internat. Journal of Computer Vision*, 31(2/3):227–246, 1999.

Session IV: General Issues

Performance evaluation of medical image processing algorithms: Report on the 2000 SPIE Workshop

James C. Gee

Department of Radiology,
University of Pennsylvania, Philadelphia, PA, USA

Modern imaging techniques in medicine have revolutionized the study of anatomy and physiology in man. A central factor in the success and increasingly widespread application of imaging-based approaches in clinical and basic research has been the emergence of sophisticated computational methods for extracting salient information from image data. The utility of image processing has prompted the development of numerous algorithms for medical data, but these have largely remained research tools and few have been incorporated into a clinical workflow. A primary cause of this poor track record is the lack of validation of these methods. A workshop, [1], was held at the Image Processing Conference of the 2000 SPIE International Symposium on Medical Imaging to discuss and stimulate developments in performance characterization research for medical image processing algorithms. This talk presents highlights from the workshop presentations and from the panel discussion with the audience.

References

[1] J.C. Gee. Performance evaluation of medical image processing algorithms. In K.M. Hanson (ed.), *Proc. SPIE's Internat. Symposium on Medical Imaging 2000: Image Processing,* Vol. 3979, pp. 19-27. SPIE Press, Bellingham, WA, 2000.

Session V: Image Generation, Processing, and Visualization

Online recovery of projection geometry of a mobile X-ray C-arm system

Matthias Mitschke[1,2], Oliver Schütz[1], and Nassir Navab[2]

[1]Siemens Medical Solutions, Erlangen, Germany
[2]Siemens Corporate Research, Princeton, CT, USA

Tomographic three-dimensional reconstruction of high-contrast objects such as bones or contrast-agent enhanced blood vessels from two-dimensional planar X-ray images has been the focus of research over the last years. Our goal is to use mobile C-arm systems, which are commonly used as intra-operative imaging device, for such an application. The challenging task is the recovery of X-ray projection geometry. Due to the mechanical design of many available C-arms the reproducibility of its motion can not be assumed accurate enough. This then requires an online approach for the recovery of X-ray projection geometry.

We have evaluated two different methods and sensor systems. The first uses a CCD camera attached to the housing of the X-ray source and an optical marker system in order to compute the motion of the C-arm. The second approach uses a commercially available pose tracking system to determine the motion of the X-ray source. Both methods require some offline calibration in order to relate their measurements to the X-ray projection geometry.

The two methods have been evaluated and compared both qualitatively and quantitatively to the gold standard of recovering the X-ray projection geometry using an X-ray calibration phantom for an online calibration.

According to the absolute error of motion estimation the second approach is better, but the first approach is superior when an application relevant quantitative evaluation and the qualitative evaluation are applied.

This demonstrates the importance of a carefully selected error measure in order to describe the error relevant for the particular application, see [1].

References

[1] M. Mitschke and N. Navab. Recovering projection geometry: How a cheap camera can outperform an expensive stereo system. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2000)*, Vol. 1, pp. 193–200. IEEE Computer Society, Los Alamitos, CA, 2000.

Discrete tomography and its application in medical imaging

Attila Kuba

Department of Applied Informatics,
University of Szeged, Hungary

Discrete tomography (DT) deals with the reconstruction of functions from their projections when the functions have discrete known range (e.g., binary functions with values 0 and 1). DT has many interesting mathematical problems (uniqueness, existence, reconstruction, complexity. etc.). These problems can be solved in polynomial time in the case of 2D discrete sets and two projections, but they are NP-hard in higher dimensions and also if we make reconstruction from more than 2 projections. In order to find the most suitable solution among the many feasible solutions, some a priori information should be included into the reconstruction. In the medical applications some heuristic reconstruction method is used mostly or the reconstruction problem is considered and solved as an optimization problem. In the talk some of these methods and results of medical imaging (DSA, generation of attenuation map for SPECT) were shown. Further medical imaging applications of DT are to be found.

Evaluation of Shading Correction Methods

Bostjan Likar

Department of Electrical Engineering,
University of Ljubljana, Slovenia

Shading is an adverse phenomenon in microscopical and magnetic resonance images, manifesting itself as spurious intensity variations that are not present in original scenes. Correction of shading is often required for many tasks in image analysis, such as segmentation or registration, and especially if quantitative analysis is the final goal. We outline major sources of shading, review the existing methods for shading correction, and stress the importance of proper definition of shading for deriving a criterion that will allow objective evaluation of shading correction methods. The so-called coefficient of joint variations, which is a sum of standard deviations of two distinct tissue classes divided by the absolute difference of corresponding means, was proposed as a good evaluation criterion that is invariant to both global multiplicative and additive intensity transformation. We performed the evaluation of many shading correction methods on real and simulated microscopical and magnetic resonance images for which either full or partial segmentations were used to compute the coefficients of joint variations. The results show high inconsistencies between the shading correction components found by different methods, most likely due to the fundamentally different assumptions about the shading effect, indicating that the problem of shading correction is not yet well understood. On the experimental sets of images the information minimization shading correction method performed the best in terms of reducing the coefficient of joint variations.

Voxel-based Surface Flattening

Nahum Kiryati

Tel Aviv University, Israel

A voxel-based method for flattening a surface in 3-D space into 2-D while best preserving distances is presented. Triangulation or polyhedral approximation of the voxel data are not required. The problem is divided into two main parts: Voxel-based calculation of the minimal geodesic distances between points on the surface, and finding a configuration of points in 2-D that has Euclidean distances as close as possible to these distances. The method suggested combines an efficient voxel-based hybrid distance estimation method, that takes the continuity of the underlying surface into account, with classical multi-dimensional scaling (MDS) for finding the 2-D point configuration. The proposed algorithm is efficient, simple, and can be applied to surfaces that are not functions. Experimental results are shown.

Joint work with Ruth Grossmann and Ron Kimmel.

Session VI: Image Analysis

Intrinsic Object-Based Geometric Correspondence in the Validation of Image Analysis Methods

Stephen M. Pizer

Medical Image Display & Analysis Group,
University of North Carolina, Chapel Hill, NC, USA

Just as the segmentation of anatomic objects from 3D medical images must involve intrinsic geometry at multiple scales of deformable objects, so must the validation of segmentation. A medially based non-Euclidean geometry is peculiarly capable of compactly and effectively representing spatial relationships with respect to deformable single figures, objects made from multiple attached figures, and multiple non-overlapping objects. This geometry provides a correspondence of positions, orientations, and distance metrics that is useful in measuring the geometric difference between the true form of an object and the form extracted by a segmentation method under test.

Blood Pool Agent CE-MRA: Improved Arterial Visualization of the Aortoiliac Vasculature in the Steady-State using First-Pass Data

Kees van Bemmel, Wiro J. Niessen, Onno Wink, Bert Verdonck, and Max A. Viergever

Image Sciences Institute,
University Medical Center Utrecht, The Netherlands

Blood pool agents (bpa's) for Contrast Enhanced Magnetic Resonance Angiography (CE-MRA) have a prolonged intravascular halflife and provide strong T_1-relaxation even at low resolution. Therefore, these agents allow imaging in the steady-state, thus providing longer time windows for image acquisition, which can be advantageous if high contrast and/or resolution is required. However, an important drawback is the simultaneous enhancement of arteries and veins, which hampers the interpretation of the steady-state data.

It is investigated whether information from the arteriogram acquired in the first-pass of the contrast agent can be used for improved arterial visualization of the steady-state images. Hereto a path tracking tool is utilized, which automatically outlines the central arterial axes (CAA) of the aortoiliac region in the first-pass images based on four user defined points. Subsequently, by registering the first-pass and steady-state data, the path is transformed to the steady-state. A number of visualization techniques can be utilized once the CAA is known in the steady-state images.

Acquiring the CAA in the steady state by an observer is a tedious procedure which can not be used in clinical practice. Moreover, for improved arterial visualization and further post-processing it is more important that the method is sufficiently accurate.

Evaluation of computer aided diagnosis in breast cancer screening

Nico Karssemeijer

Department of Radiology,
University Medical Center Nijmegen, The Netherlands

Screening errors due to inadequate human interpretation frequently occur in breast cancer screening. Even when two radiologists read screening mammograms independently, breast cancer can often be identified retrospectively on prior mammograms taken years before detection. To avoid such errors, computer aided detection methods are being developed that mark suspicious abnormal areas in mammograms. It is assumed that radiologists can use such methods to focus attention to areas they might otherwise overlook. To evaluate the benefit of computer aided detection a large observer study was conducted to analyse which type of errors radiologists make in practice. It turned out that most problems are related to interpretation of mammographic regions identified as suspicious, and not to incomplete search of the images. This means that the maximum benefit of computer aided diagnosis can be expected when methods are aimed at helping radiologists to improve interpretation. A method was developed to evaluate the potential benefit of computer aided interpretation. In this approach data collected in an observer study is independently combined with the output of a pattern classification algorithm. It was shown that by implementing computer aided detection in this way the performance of experienced screening radiologists can be significantly improved. The advantage of this approach is that the potential benefit of a new computer aided detection scheme can be assessed using existing observer data, thus avoiding the need to repeat expensive experiments with observers.

The Role of Statistical Principles in the Design of Medical Image Segmentation Algorithms

Neil Thacker

Division of Imaging Science & Biomedical Engineering,
The University of Manchester, United Kingdom

Many segmentation methods work by applying calculations which at first sight appear to have nothing in common with either probability theory or statistics, but if we wish to explain the best approaches to MR image segmentation we must be able to relate these techniques to the statistical models that they are based upon. The most direct piece of information which we can obtain from data is in the form of a conditional probability. $P(C|D)$ is the probability of the interpretation C given the data D. Given such probabilities for each pixel in an image we can segment regions or locate the boundaries between tissues. If the method (or algorithm) used to determine these probabilities is appropriate, then the regions and boundaries determined in this manner will be optimal. That is, they will (by definition) have extracted all of the useful information relating to the problem from the data. Determining that an algorithm is appropriate amounts to being able to confirm that the assumptions underlying the statistical approach are valid. To do this we must first know what these assumptions are. Areas of algorithmic research which give rise to image processing algorithms are fundamentally linked to matching assumptions to data sets. Though it is possible to develop good algorithms blindly (by guesswork and testing) it is always better to apply a statistical methodology, systematically testing the effects of any assumption on the result.

The consequence of all of this is that there is no technique that can be guaranteed to work on any data set, that is a "magic bullet". For any method to work it must be applied to data which falls within the range of behaviour for which it has been designed. Different algorithms have varying ranges of applicability. Algorithms which make the most assumptions regarding the data often have very limited use in comparison to those which take into account a broader range of data characteristics. Though algorithms which make a large number of assumptions can often be simple it is not necessarily true that complex algorithms will always perform better. The extra complexity must be used wisely and for good reason. Extra complexity can just as easily result in unreliability as in improved results. These are the issues that algorithm designers consider when developing a new technique and algorithms can be grouped according to the underlying assumptions for the purposes of evaluation.

BBQs on Methodology of Evaluation in MIC: Result of the Round Table Group Discussion

H. Siegfried Stiehl

Arbeitsbereich Kognitive Systeme, Fachbereich Informatik,
Universität Hamburg, Germany

Throughout the week the moderators of the different sessions were asked by the organizers to collect hot topics and crucial open problems (coined "burning burning questions or, as acronym, BBQs) in written form on a flip-chart. The final list was made up of the following BBQs (in random order):

1. characterization and formalization of noise in imagery
2. modelling of shading effects in e.g. microscopy or MR images
3. effect of preprocessing (e.g. image protocol selection, noise reduction, shading correction, contrast enhancement) on subsequent processing stages
4. influence of pathology in images on design of computational processes
5. importance of volumetric nature of image data
6. impact of anisotropic (e.g. CT and MR) image data on design of computational processes
7. effectiveness of computer graphics visualization for validation purposes
8. importance of unique terminology of validation and evaluation
9. relation of validation and evaluation through visual assessment to characteristics of visual system
10. mathematical foundation of image registration metrics
11. importance of correspondence problem in image registration
12. analysis of effect of local and global shape on computational processes
13. analysis of discretization (sampling and quantization) effects given continuous theory of computational processes
14. importance of spatial scale(s) for the design of computational processes
15. appropriateness of morphing (e.g. elastic registration) approaches w.r.t. anatomical variability

16. availability of ground truth

17. availability of annotated test image databases

18. importance of certification and of sharing of code

19. necessity of revealing assumptions of algorithms

On the basis of the above list of BBQs, a final round table group discussion was organized on Friday morning. In order to stimulate the discussion, the seminarians (in random order) were asked to provide their personal prime three BBQs or, in other words, hot topics to be put on the agenda with priority. The moderators (M.H. Loew and H.S. Stiehl) took notes of the different BBQs on wallpapers (while J.Z. Chen transcribed key words in parallel on his laptop) and the seminarians were asked to group them into more general topics (a task which turned out to be achievable to a partial extent only; see below). After about two hours of brainstorming, the following list has been compiled which should be conceived of as ;-) "The 2001 Legacy of the Dagstuhl 01111 Seminarians:

1. need for synopsis of (ideally annotated) test image data bases available to the scientific MIC community

2. research on description and representation of shape variability and appearance variability (pathology, etc.)

3. compilation of built-in assumptions of algorithms

4. definition of groups of algorithms, methods, and tools

5. importance of (and necessity of increase of) open source

6. high value of theoretical studies and usefulness of synthetic data

7. availability of classes of test data according to different clinical tasks, different groups of algorithms, etc.

8. public availability of explicit assumptions for synthetic test images (and data bases)

9. establishment of an independent MIC validation and evaluation institution (e.g. in cooperation with national institutions such as TÜV in Germany, FDA/NIST in US, or NPL in UK)

10. development of novel test image data bases (providing also gold standards, clinical task description, and certified code for benchmarking as well as allowing for augmentation through MIC community in terms of two-way-communication)

11. need for standardized terminology, e.g. figures of merit, assumptions, metrics, quality of segmentation

12. profit from other sources, e.g. national standardization bodies, measurement theory, experimental physics, psychophysics, etc.

13. necessity for taxonomy of test image data
14. development of best-practice guides for experimental MIC
15. need for future (probably more focussed) meetings
16. How to encourage community to do evaluation? or: creation of a "cultural change (e.g. institutionalization through regular special conference sessions, committal call-for-papers, change of editorial policy w.r.t. journal sections, funding through industry)
17. increased consideration of clinical task in experimental MIC
18. improvement of free availability (and exchange) of code, benchmarks, and test image data
19. design of special test image data bases (e.g. w.r.t. selected clinical tasks, health care system priorities, etc.
20. encouragement of replication of results by journals
21. problem of practical and financial limits on clinical trials
22. input from other sources to hitherto weak terminology of experimental MIC
23. formal grounding of performance characterization
24. necessity of full understanding of problem domain (e.g. clinical task, work-flow, etc.)
25. probabilistic and geometrical (to be combined in general) modelling of pathology
26. methods for measuring differences from segmentation ground truth from different observers
27. necessity of routine use of multi-scale methods
28. definition of statistically valid sample size of test images in data bases
29. IAEA quality of software report / documentation
30. sophistication of algorithms vs. requirements from low-/mid-/high-tech medicine
31. MIC target: academic high-tech hospital vs. rural reality
32. scientific value of validation, or: How to make a PhD thesis out of a validation task?
33. establishment of an industry working/pressure group on validation (e.g. FDA requirement)
34. profit from rigor of other disciplines with strong experimental part (e.g. theory-experiment cycle in physics, statistical grounding of experimentation in psychophysics, application of standard statistics in clinical studies)

35. importance of discretization of continuous MIC theories (continuous-case vs. discrete-case validation)

36. scientific basis of comparability, e.g. common procedures, biostatistics methods, test methodologies, etc.

37. types of and relevance of measures of performance of MIC methods

38. definition of standards of test image data properties

39. issue of validation of validation, e.g. completeness of training/test image data set, biological realism of synthetic image data, etc.

40. two-way-access to test image data bases for MIC community, e.g. for refinement, augmentation, etc.

41. usefulness of probabilistic representations (e.g. MNIs probabilistic brain atlas) for ground truth segmentation

42. problem of truth of result vs. beauty of result (in eye of beholder)

43. sufficiency/appropriateness of math for modelling (of e.g. noise characteristics, digital data, anatomical variability, etc.) and trade-off between reality and accuracy

>>> REALITY >>>			
Math Model	Phantom	Cadaver	In Vivo
<<< ACCURACY <<<			

44. consideration of full scientific background of methodology (Greek/new-Latin terminus for theory of scientific methods), or: What is "methodology?

45. What is "is? (This question was spontaneously raised along with a sigh right after the latter deep question in 44. by a seminarian who kindly asked to refrain from making his name public in this report and shortly after all seminarians agreed upon that this question though of highest philosophical interest should be taken as a sure indicator of common mental exhaustion.)

46. design of curricula for teaching students and doctors about methodology of validation and evaluation

47. necessity of change of attitude towards validation and evaluation as a bonus/pursuit for members of MIC community

48. need for code of ethics in MIC validation and evaluation or, in general, experimental computer vision (see e.g. IEEE, GNU, and others for inspiration)

After a short discussion of several seconds the attempt of ranking and clustering the topics enumerated above was characterized to be an NP-hard or even worse problem and for the sake of both the participants of the subsequent seminar (who were observed to arrive) and the excellent chefs in the kitchen (who were hypothesized to be ready to serve lunch) the majority was in favour of lunch and espresso as demonstrated by unique body language. (Note that at least a partial solution could have been computed by an unknown seminarian: Topic 47 perfectly goes with topic 16.)

As result of a last mental and intellectual rearing-up it should be reported that a couple of action points have been collected:

1. creation of a web site for a forum on validation and evaluation in the context of MIC (Action Point: P. Courtney)
2. publication of editorial in IEEE Transactions on Medical Imaging (Action Point: The Seminar Organizers)
3. proposal on special issue of Medical Image Analysis (Action Point: The Seminar Organizers)
4. synopsis of public and private test image databases (Action Point: The Seminarians via web site from above)
5. mutual exchange of our own local test image data (Action Point: G. Gerig and The Seminarians)
6. national proposals for creation of annotated, task-specific etc. test image data bases (Action Points: The Seminarians)
7. collection and examination of existing codes of ethics (Action Point: The Seminarians via web site)
8. compilation of terms, definitions, etc. related to validation and evaluation in MIC (Action Point: S. Olabarriaga and The Seminarians via web site)
9. brief report on our Dagstuhl seminar to be published in newsletters of national societies (Action Points: The Seminarians)
10. regular exchange of upcoming national validation and evaluation events (Action Point: The Seminarians via web site)

Famous last words: It was a great seminar and all of us promised to keep the Dagstuhl spirit!