How Accurate Is Brain Volumetry?

A Methodological Evaluation

Horst K. Hahn[1], Benoît Jolly[1], Miriam Lee[1], Daniel Krastel[2], Jan Rexilius[1], Johann Drexl[1], Mathias Schlüter[1], Burckhard Terwey[2], and Heinz-Otto Peitgen[1]

[1] MeVis – Center for Medical Diagnostic Systems and Visualization,
Universitätsallee 29, 28359 Bremen, Germany, hahn@mevis.de
[2] Center for Magnetic Resonance Imaging, St.-Jürgen Str., Bremen, Germany

Abstract. We evaluate the accuracy and precision of different techniques for measuring brain volumes based on MRI. We compare two established software packages that offer an automated image analysis, EMS and SIENAX, and a third method, which we present. The latter is based on the Interactive Watershed Transform and a model based histogram analysis. All methods are evaluated with respect to noise, image inhomogeneity, and resolution as well as inter-examination and inter-scanner characteristics on 66 phantom and volunteer images. Furthermore, we evaluate the N3 nonuniformity correction for improving robustness and reproducibility. Despite the conceptual similarity of SIENAX and EMS, important differences are revealed. Finally, the volumetric accuracy of the methods is investigated using the ground truth of the BrainWeb phantom.

1 Introduction

Various indications exist for brain volume measurements. Major fields of application are diagnosis, disease monitoring, and evaluation of potential treatments in Multiple Sclerosis (MS) [1,2,3] and neurodegenerative diseases, most importantly Alzheimer's disease (AD) [4,5]. Rudick *et al.* propose the brain parenchymal fraction (BPF), which they define as the ratio of brain parenchymal volume to the total volume within the brain surface contour, as a marker for destructive pathologic processes in relapsing MS patients [1]. De Stefano *et al.* found substantial cortical gray matter (GM) volume loss in MS. They propose that neocortical GM pathology may occur early in the course of both relapsing-remitting and primary progressive forms of the disease and contribute significantly to neurologic impairment [2]. In addition to a process that is secondary to white matter (WM) inflammation, they also assume an independent neurodegenerative process, which mainly affects GM and raises the need for robust measures to independently quantify WM and GM volumes.

Fox *et al.* report a mean brain atrophy progression of approximately one percent (12.3 ml) per year in the AD group compared to less than 0.1 percent (0.3 ml) in the control group [4]. I. e., the relative precision of brain volume measurements must be approximately 0.5 percent in order to significantly measure atrophy within one year. Brunetti *et al.* conclude that GM and WM atrophy quantification could complement neuropsychological tests for the assessment of disease severity in AD, possibly having an impact on therapeutic decisions [5].

C. Barillot, D.R. Haynor, and P. Hellier (Eds.): MICCAI 2004, LNCS 3216, pp. 335–342, 2004.

In addition to global measurements, a regional analysis is important, for example to derive separate volumes for brain lobes, deep gray matter, hippocampus, cerebellum, etc. Fox *et al.* visualize regional atrophy by the differences between aligned and normalized serial scans, while computing tissue loss that takes into account partial volume effects (PVE) by the integral of the change over the whole brain [4]. Andreasen *et al.* derive volume measurements for twelve standardized regions based on the Talairach coordinate system [6].

In this paper, we concentrate on whole brain characteristics, for total brain volume (TBV=WM+GM), BPF (=TBV/(TBV+CSF)), as well as GM and WM volumes. These measures are expected to remain important parameters in clinical imaging and within clinical trials [1,2,5]. Our objectives are:

1. To evaluate the accuracy and precision of different methodologies for whole brain volumetry on a software phantom with known ground truth and on real MRI data. We aim toward contributing to a methodological comparison that we find poorly developed for image analysis in general and for brain volumetry in particular.
2. To investigate the importance of image nonuniformity, noise, and resolution in brain volumetry and the possibility of improving results using nonuniformity correction.
3. To propose a novel technique which is simple, fast, robust, and accurate. Robustness will be assessed with respect to nonuniformity, noise, and resolution, while the major criterion is reproducibility in terms of inter-examination (scan-rescan) and also inter-scanner characteristics.

2 Material and Methods

For the evaluation of brain volumetry, we used phantom, volunteer, and patient data. The phantom data was obtained from the BrainWeb database [7] at different noise (3%, 5%, 7%, 9%) and nonuniformity levels (0%, 20%, 40%) as well as axial resolutions (1 mm, 3 mm, 5 mm; cf. Fig. 5). The in-plane resolution of the phantom is $(1.0\,\text{mm})^2$ throughout. The exact volumetric ground truth is known a-priori and provided on the web site; the values for GM and WM are 902.9 ml and 680.8 ml, respectively, if glial matter (6.0 ml) is counted as WM; TBV sums up to 1583.7 ml.

To evaluate scan-rescan reproducibility, we used data from three healthy volunteers (subjects 1–3), which have been scanned each five times on the same day with independent repositioning and head rotation. Two of the subjects have also been repeatedly scanned twice on two different scanners on another day, such that inter-scanner characteristics are available. The two devices and acquisition protocols were: (A) Siemens Magnetom Symphony, T1 MPR 3D, $\text{TR} = 1900\,\text{ms}$, $\text{TE} = 4.15\,\text{ms}$, $\text{TI} = 1100\,\text{ms}$, and (B) an older Siemens Magnetom Vision Plus, T1 MPR 3D, $\text{TR} = 9.7\,\text{ms}$, $\text{TE} = 4.0\,\text{ms}$, $\text{TI} = 300\,\text{ms}$. We used an isotropic resolution of $(1.0\,\text{mm})^3$ for all volunteer images with an acquisition time of approximately nine minutes. Inter-examination images were acquired on scanner A for subjects 1 (M, 39 y) and 2 (F, 34 y) and on scanner B for subject 3 (F, 27 y). To limit scan times, we resampled the image data from subject 3 (all five independent acquisitions) at four different axial resolutions using a three-lobed Lanczos filter (1.8 mm, 3.0 mm, 4.8 mm, 7.2 mm).

Various techniques exist to address brain volumetry based on MRI. We have evaluated two of them, which have reached a certain popularity, namely the software packages

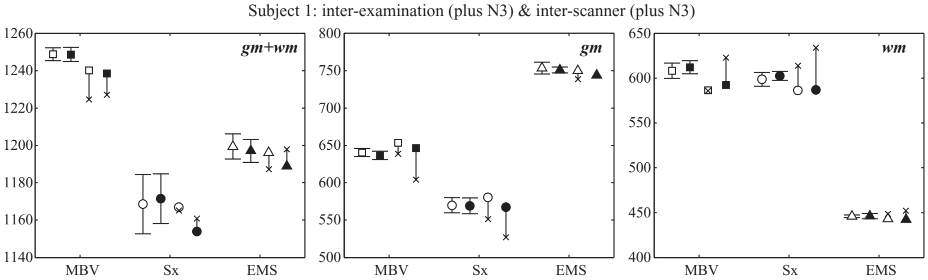

Fig. 1. Inter-examination and inter-scanner characteristics (subject 1) of total brain (***gm+wm***), GM, and WM volumes (mean and SD) for three methods: MBV ('square'), SIENAX (Sx, 'circle'), and EMS ('triangle'). The filled symbols represent the results after N3 nonuniformity correction. Error bars correspond to single SD. The candle plots show inter-scanner differences ('cross' for scanner B), while each symbol represents the mean of two independent acquisitions.

SIENAX [8] and EMS [9]. Both are well documented and available on the internet for research purposes [8,9]. Furthermore, we have evaluated a novel method that builds upon the image analysis platform MeVisLab [10], referred to as MeVisLab Brain Volumetry (MBV). We briefly describe the concepts of the three methods.

SIENAX is a fully automated command-line tool and part of FMRIB's FSL library (Steven M. Smith, University of Oxford, UK) [8]. It is based on a hidden Markov random field (MRF) model and an associated iterative Expectation-Maximization (EM) algorithm for estimating tissue intensity parameters and spatial intensity bias field. Before this, the images are registered to a standard space brain image, where a mask is used to remove non-brain tissue. SIENAX explicitly estimates PVE by evaluating the tissue intensity model on a particular voxel's neighborhood.

EMS was developed by Koen Van Leemput at the Medical Image Computing Group at KU Leuven, Belgium [9], and builds upon the SPM package (Wellcome Department of Imaging Neuroscience, University College London, UK). Much like SIENAX, EMS relies on unsupervised EM tissue classification, corrects for MR signal inhomogeneities, and incorporates contextual information through MRF modeling. Instead of using a brain mask, a digital brain atlas containing prior expectations about the spatial location of tissue classes is used after mutual information based affine registration to initialize the algorithm.

MBV is simpler than the two other methods in that it does not comprise EM classification, MRF or nonuniformity modeling. Rather, MBV relies on skull stripping and histogram analysis only. More precisely, the following four subsequent steps are performed: (i) Interactive definition of a cuboid ROI on three orthogonal and synchronized 2D views. (ii) Automatic resampling to an isotropic grid (spacing $0.9\,\text{mm}^3$) using a Mitchell filter for x, y, and z directions. (iii) Skull stripping using the marker-driven, three-dimensional Interactive Watershed Transform (IWT), which has been described in detail by Hahn and Peitgen [11,12]. (iv) Automatic histogram analysis for the 3D region defined by the IWT result [13]. The analysis is based on a model consisting of four Gaussian distributions for pure tissue types (WM, GM, CSF, and bone/air), as well as dedicated partial volume distributions for mixed tissue types. The histogram model

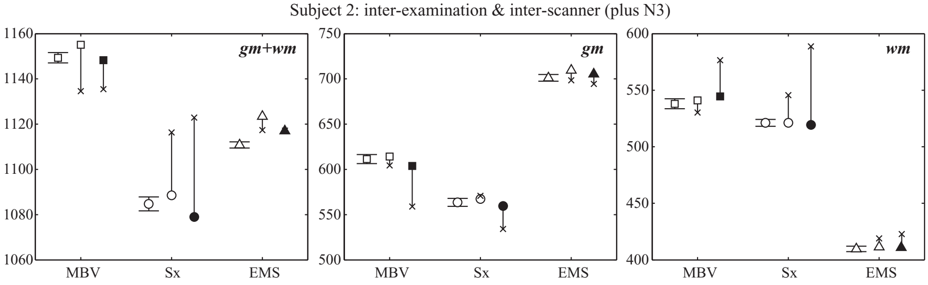

Fig. 2. Inter-examination and inter-scanner characteristics (subject 2). Same categories as Fig. 1, but without inter-examination N3 correction.

corresponds to the assumptions formulated by Santago and Gage [14], i. e. uniform PVE from each pair of two tissues, and is fitted to the histogram by minimizing least square deviations. We use a fourth class to cover air and bone tissue, which is required since the IWT when applied to the original data (interpreted as depth information) includes most of the partial volume regions at the outer brain surface (GM-CSF interface) and extracerebral CSF (CSF-bone interface) [12].

Finally, we investigated the N3 method by John Sled *et al.* [15] in order to improve the volumetric results in presence of a low-frequency image bias field caused by RF inhomogeneity and typical for MR images. N3 employs an iterative approach to estimate both the multiplicative bias field and the true intensity distribution. It requires only two parameters to be selected, one controlling the smoothness of the estimated nonuniformity, the other controlling the tradeoff between accuracy and convergence rate.

3 Results

Several results were acquired on a total of 66 images (phantom: 10, volunteer: 43, N3 corrected: 13). We computed inter-examination mean and standard deviations (SD) of TBV, BPF, GM, and WM volumes for the three subjects (error bars in Figs. 1–3). For subjects 1 and 2, we also assessed inter-scanner variability (candle plots in Figs. 1, 2, and 3 b). For the inter-scanner images and for all images of subject 1, we applied N3 nonuniformity correction using the parameters suggested by J. Sled *et al.* [15] (filled symbols in Figs. 1–3). Figure 4 shows the dependency of inter-examination mean values and variations on the axial image resolution. Figure 5 comprises measured volumes and ground truth for ten phantom images (different noise, nonuniformity, and resolution) in a combined graph. The characteristics for TBV and BPF over all subjects are summarized in Table 1.

Since it is an interactive technique, a second observer (H.K.H.) used MBV to analyze the five original images of subject 3 (diamonds in Fig. 3 a). We found inter-observer differences (mean ± SD) for TBV, GM, and WM volumes to be $+4.5 \pm 4.9$ ml, $+3.6 \pm 3.4$ ml, and $+1.0 \pm 1.6$ ml, respectively. From our experience with MBV, we did not observe significant variations between observers [3], so that we only chose the subject for inter-observer test that required most interaction.

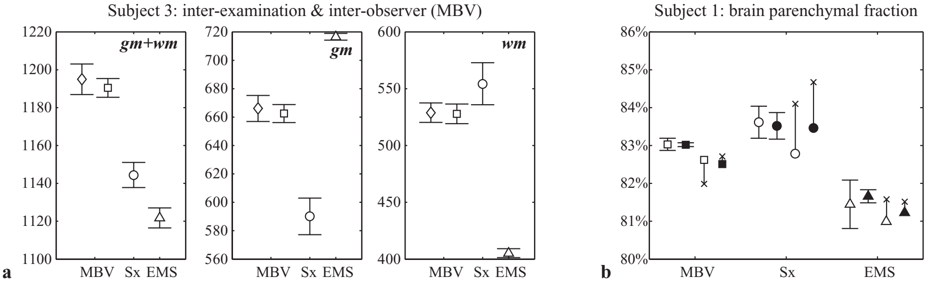

Fig. 3. a: Inter-examination characteristics (subject 3) with results from two observers for MBV (diamonds and squares). **b:** BPF characteristics for subject 1 (cf. Fig. 1).

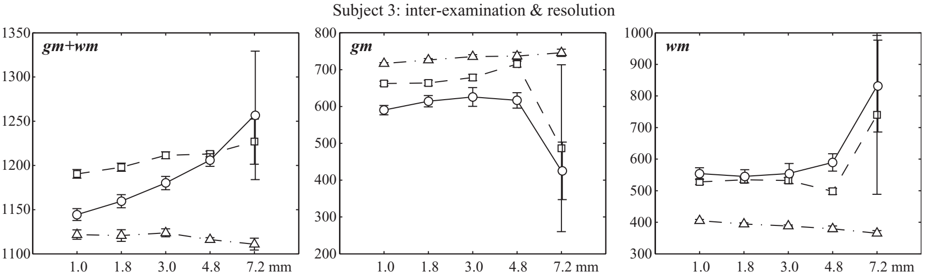

Fig. 4. Inter-examination characteristics for five different axial resolutions (1.0–7.2 mm, subject 3). Each element is computed from five independently acquired and resampled images (mean ± SD, symbols cf. Fig. 1).

We used all methods without modification or – except the N3 correction – image preprocessing (resampling, noise reduction, etc.). SIENAX (FSL 3.1, July 2003) was operated under Linux on a Dual 2.8 GHz Xeon with processing times of approximately eleven minutes per case. For our evaluation, we used the given brain volumes before normalization. EMS and MBV were operated under Windows 2000 on a 1.7 GHz Pentium III. For EMS, we used SPM99 and Matlab 6.1 with processing times of approximately 26 minutes per case (15 min registration plus 11 min segmentation), including modeling of MRF and fourth-order bias field. EMS volumes were calculated as weighted sum of tissue probability maps. For convergence, we had to manually provide the EMS registration with good initial parameter settings (e. g., 20° frontal rotation was a good starting point in many cases). For MBV, steps i, ii, plus iv required less than one minute, while the interactive skull stripping (step iii) required approximately 1–4 minutes for marker placement and visual inspection of the actual segmentation, resulting in an overall analysis time of less than five minutes.

For all image processing and evaluation, we paid attention to three important aspects: (a) our own method and its parameters were not altered after the first data set was analyzed; (b) the comparative evaluation of all three methods was conducted exactly once by a person (B.J.), which has not been involved in any method development; and (c) for the phantom, the operator was blinded with respect to the real tissue volumes.

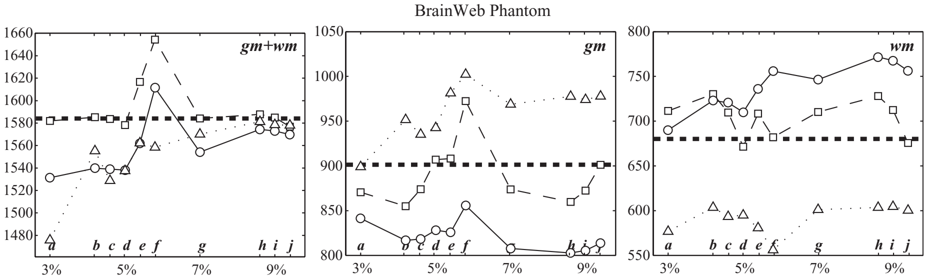

Fig. 5. Volumetric results on ten phantom images. The bold dashed lines indicate the ground truth (symbols cf. Fig. 1). Simulated noise levels are indicated on x-axis. Nonuniformity levels are 0% (b,h), 40% (d,j), and 20% (others). Axial slice thickness is 3 mm (e), 5 mm (f), and 1 mm (others).

4 Discussion

Within our study, each of the three methods revealed advantages and disadvantages. Even though similar in concept, we found considerable differences between SIENAX and EMS. The inter-examination characteristics were slightly better for EMS and MBV than for SIENAX, while EMS performed best for GM and WM (cf. Figs. 1–3) and MBV for TBV and BPF (cf. Table 1 left). In total, including inter-resolution (Fig. 4) and inter-scanner results (cf. Table 1 right), EMS was the most robust method.

One important question, to which a phantom can provide answers to a certain extent, is which method gets closest to the ground truth. Despite its stability, EMS yielded the highest GM volumes and lowest WM volumes on all images, which is consistent with a GM overestimation and WM underestimation compared to the phantom ground truth (Fig. 5). Conversely, SIENAX consistently underestimated GM and overestimated WM volumes compared to the ground truth, and yielded lowest GM volumes on all images. WM volumes were similar for SIENAX and MBV in some cases (Figs. 1 and 2). On the phantom data, MBV was closest to the ground truth on average and yielded values between the other two methods in most cases (cf. Fig. 5). We also analyzed images of AD patients and found similar results (cf. Fig. 6).

The systematic behavior of EMS could be influenced by the atlas based prior tissue probabilities; in Fig. 6 right, a comparably broad GM layer can be discerned. The GM

Table 1. Overall characteristics for TBV and BPF, from left to right: mean inter-examination SD (n=20 images at 1 mm slice thickness: S1, S2, S3, S1+N3); effect of N3 (n=13 images, mean and SD of pair-wise differences); effect of N3 on inter-scanner differences (n=16 images, mean difference $\langle V_A - V_B \rangle$, without → with N3).

	inter-examination SD		N3 effect		inter-scanner N3 effect	
	TBV (ml)	BPF (%)	TBV (ml)	BPF (%)	TBV (ml)	BPF (%)
MBV	3.6	0.17	-0.8 ± 4.2	0.00 ± 0.41	18.0 → 12.2	0.56 → 0.11
Sx	9.7	0.30	-2.0 ± 10.9	0.22 ± 0.52	14.8 → 25.4	1.17 → 1.61
EMS	4.9	0.33	-1.4 ± 9.6	0.09 ± 0.46	8.3 → 4.0	0.34 → 0.16

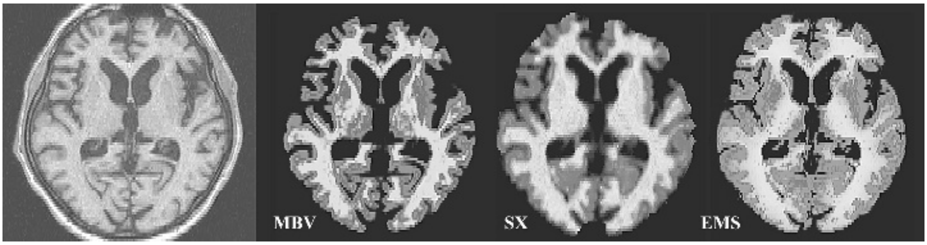

Fig. 6. AD patient (M, 60y). Note that for MBV, partial volume voxels are not shown. The results were (TBV/GM/WM/BPF) 1004.2 / 560.4 / 443.8 / 68.81% for MBV, 972.08 / 475.33 / 496.75 / 77.96% for SX, and 1096.5 / 671.7 / 424.8 / 79.40% for EMS.

underestimation of SIENAX could be caused by a too rigorous brain masking, which removes some of the partial volume voxels at the pial brain surface (cf. Fig. 6 center).

BPF is a normalized measure of brain atrophy, which can be used in cross-group studies. When computing the mean inter-examination SD for BPF, MBV performed slightly better than the other methods (cf. Fig. 3 b and Table 1 left).

Pair-wise comparison of results with and without N3 revealed the least changes of TBV and BPF for the MBV method, showing its high robustness to image nonuniformity (Table 1 middle). Note that this is despite the fact that MBV is the only method that comprises neither nonuniformity correction nor spatial (MRF) regularization, such that we expected it to benefit most from the N3 method. Moreover, N3 did not significantly reduce inter-scanner differences. It yielded some improvements for TBV and BPF, but only for EMS and MBV (Table 1 right); GM and WM differences were mostly worsened by N3 (cf. Figs. 1 and 2).

5 Conclusion and Future Work

In conclusion, EMS was most robust for very large slice thickness and between scanners, and with best inter-examination characteristics for WM and GM. SIENAX has a major advantage in its full and robust automation. MBV showed best inter-examination characteristics for TBV and BPF, and was least influenced by N3. Despite its interactivity, MBV was the fastest of the three methods, but only if a human operator is available. Schubert *et al.*, therefore, investigated a full automation of MBV's remaining interactive steps [16]. Still, we see an advantage in MBV's possibility to interactively control and refine the brain mask and that it does not depend on image registration or anatomical template.

Future work includes a more extensive investigation of these and other methods on a variety of patient images. In particular, the systematic behavior with respect to image resolution (cf. Figs. 4 and 5 e/f) requires further research on a larger set of MR images. Other methods include SIENA (also FSL) that is designed for longitudinal brain volume measurements. One issue that is important for clinical use was not explicitly addressed in this paper, namely sensitivity. While reproducibility is a measure for the robustness of a method, small changes, e. g. associated with GM atrophy, could remain undetected by a highly reproducible method. This needs to be addressed in future studies.

Acknowledgments. We kindly acknowledge the Medical Image Computing Group at KU Leuven (Paul Suetens), for providing their software EMS, the University of Oxford and the Image Analysis Group of FMRIB Oxford (Steven M. Smith) for providing their software FSL, the McCornell Brain Imaging Centre, MNI, McGill University (Alan C. Evans) for providing the BrainWeb phantom as well as their software N3, and Koen Van Leemput for assisting in the use of EMS. We thank our colleagues Florian Link and Tobias Boskamp for their groundbreaking work on MeVisLab.

References

1. Rudick, R.A., Fisher, E., Lee, J.C., et al.: Use of the brain parenchymal fraction to measure whole brain atrophy in relapsing-remitting MS. Neurology **53** (1999) 1698–1704
2. De Stefano, N., Matthews, P.M., Filippi, M., et al.: Evidence of early cortical atrophy in MS: Relevance to white matter changes and disability. Neurology **60** (2003) 1157–1162
3. Lukas, C., Hahn, H.K., Bellenberg, B., et al.: Sensitivity and reproducibility of a new fast 3D segmentation technique for clinical MR based brain volumetry in multiple sclerosis. Neuroradiology (2004) in print
4. Fox, N.C., Freeborough, P.A., Rossor, M.N.: Visualisation and quantification of rates of atrophy in Alzheimer's disease. Lancet **348** (1996) 94–97
5. Brunetti, A., Postiglione, A., Tedeschi, E., et al.: Measurement of global brain atrophy in Alzheimer's disease with unsupervised segmentation of spin-echo MRI studies. J Magn Reson Imaging **11** (2000) 260–266
6. Andreasen, N.C., Rajarethinam, R., Cizadlo, T., et al.: Automatic atlas-based volume estimation of human brain regions from MR images. J Comput Assist Tomogr **20** (1996) 98–106
7. Collins, D.L., Zijdenbos, A.P., Kollokian, V., et al.: Design and construction of a realistic digital brain phantom. IEEE Trans Med Imaging **17** (1998) 463–468, `//www.bic.mni.mcgill.ca/brainweb/`.
8. Smith, S., Zhang, Y., Jenkinson, M., et al.: Accurate, robust and automated longitudinal and cross-sectional brain change analysis. NeuroImage **17** (2002) 479–489, `//www.fmrib.ox.ac.uk/fsl/`.
9. Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P.: Automated model-based tissue classification of MR images of the brain. IEEE Trans Med Imaging **18** (1999) 897–908, `//bilbo.esat.kuleuven.ac.be/web-pages/downloads/ems/`.
10. MeVisLab development environment available at: `//www.mevislab.de/`.
11. Hahn, H.K., Peitgen, H.O.: IWT – Interactive Watershed Transform: A hierarchical method for efficient interactive and automated segmentation of multidimensional grayscale images. In: Med Imaging: Image Processing. Proc. SPIE 5032, San Diego (2003) 643–653
12. Hahn, H.K., Peitgen, H.O.: The skull stripping problem in MRI solved by a single 3D watershed transform. In: MICCAI – Medical Image Computing and Computer-Assisted Intervention. LNCS 1935, Springer, Berlin (2000) 134–143
13. Hahn, H.K., Millar, W.S., Klinghammer, O., et al.: A reliable and efficient method for cerebral ventricular volumetry in pediatric neuroimaging. Methods Inf Med 43 (2004) in print
14. Santago, P., Gage, H.D.: Quantification of MR brain images by mixture density and partial volume modeling. IEEE Trans Med Imaging **12** (1993) 566–574
15. Sled, J.G., Zijdenbos, A.P., Evans, A.C.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging **17** (1998) 87–97
16. Schubert, A., Hahn, H.K., Peitgen, H.O.: Robust fully automated brain segmentation based on a 3D watershed transform. In: Proc. BVM, Springer, Berlin (2002) 193–196, in German.