Comparison and Evaluation of Methods for Liver Segmentation From CT Datasets

Tobias Heimann*, Bram van Ginneken, *Member, IEEE*, Martin A. Styner, *Member, IEEE*, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, *Member, IEEE*, György Bekes, Fernando Bello, *Member, IEEE*, Gerd Binnig, Horst Bischof, *Member, IEEE*, Alexander Bornik, Peter M. M. Cashman, Ying Chi, Andrés Córdova, Benoit M. Dawant, Márta Fidrich, Jacob D. Furst, Daisuke Furukawa, Lars Grenacher, Joachim Hornegger, *Member, IEEE*, Dagmar Kainmüller, Richard I. Kitney, Hidefumi Kobatake, Hans Lamecker, Thomas Lange, Jeongjin Lee, Brian Lennon, Rui Li, Senhu Li, Hans-Peter Meinzer, *Member, IEEE*, Gábor Németh, Daniela S. Raicu, Anne-Mareike Rau, Eva M. van Rikxoort, Mikaël Rousson, László Ruskó, Kinda A. Saddi, Günter Schmidt, Dieter Seghers, Akinobu Shimizu, *Member, IEEE*, Pieter Slagmolen, Erich Sorantin, Grzegorz Soza, Ruchaneewan Susomboon, Jonathan M. Waite, Andreas Wimmer, and Ivo Wolf

***Abstract*—This paper presents a comparison study between 10 automatic and six interactive methods for liver segmentation from contrast-enhanced CT images. It is based on results from the "MICCAI 2007 Grand Challenge" workshop, where 16 teams evaluated their algorithms on a common database. A collection of 20 clinical images with reference segmentations was provided to train and tune algorithms in advance. Participants were also allowed to use additional proprietary training data for that purpose. All teams then had to apply their methods to 10 test datasets and submit the obtained results. Employed algorithms include statistical shape models, atlas registration, level-sets, graph-cuts and rule-based systems. All results were compared to reference segmentations five error measures that highlight different aspects of segmentation accuracy. All measures were combined according to a specific scoring system relating the obtained values to human expert variability. In general, interactive methods reached higher average scores than automatic approaches and featured a better consistency of segmentation quality. However, the best automatic methods (mainly based on statistical shape models with some additional free deformation) could compete well on the majority of test images. The study provides an insight in performance of different segmentation approaches under real-world conditions and highlights achievements and limitations of current image analysis techniques.**

***Index Terms*—Evaluation, liver, segmentation.**

I. Introduction

"ANYTHING you can do, I can do better (no you can't)..." is the title of a highly amusing paper by Price [1], in which he complained that "computer vision suffers from an overload of written information but a dearth of good evaluations and comparisons." Written more than 20 years ago, the key message of this text still holds true today, also for our field of medical image analysis. Each year, an increasing number of new methods and algorithms is published at conferences and in journals. Although paper commonly include an objective evaluation these days, most of the experiments exhibit two severe shortcomings. Firstly, new algorithms are generally not compared sufficiently against current state of the art methods. While in some rare cases a new technique allows to resolve a formerly unsolved problem, the vast majority of published works present gradual improvements or variations to existing solutions. Each variation may be evaluated against (and proven better than) the original solution, but how do different improvements of the same method compare to each other? And how do they compare to methods employing a completely different technique? As most algorithms are not freely available and re-implementation often is too tedious and time-consuming, comprehensive comparisons that would answer this kind of questions are still rare. Secondly, the data employed for evaluation is typically not representative of the real-world images used in the clinic. To be of practical value, algorithms have to cope with data from different sources, acquired with varying protocols, and featuring artifacts and pathology. As many research groups—especially the ones rooted in engineering or mathematics departments—do not have access to the variety of images required, too many new methods are tested on a restricted set of images and may fail in practical application. We believe this is a major cause for the gap between the published state of the art and the methods actually employed in the clinic (which are to a large extent still manual and cumbersome to operate).

Both shortcomings are already described in the above cited paper by Price. Since awareness of these issues has risen in the image analysis community, there is a growing amount of initiatives to amend the situation. On one side, open source toolkits as ITK[1] or OpenCV,[2] which offer a free collection of algorithms, are gaining more and more followers. On the other side, open data repositories for e.g., brain,[3] lung [2] or retinal data[4] provide a comprehensive set of clinical images to evaluate specific applications. A good possibility to further motivate the comparison of

Manuscript received November 27, 2008; revised January 09, 2009. First published February 10, 2009; current version published July 29, 2009. This work was supported by Siemens AG, Healthcare Sector (Forchheim, Germany). *Asterisk indicates corresponding author.*

Please see the Acknowledgment section of this paper for the author affiliations.

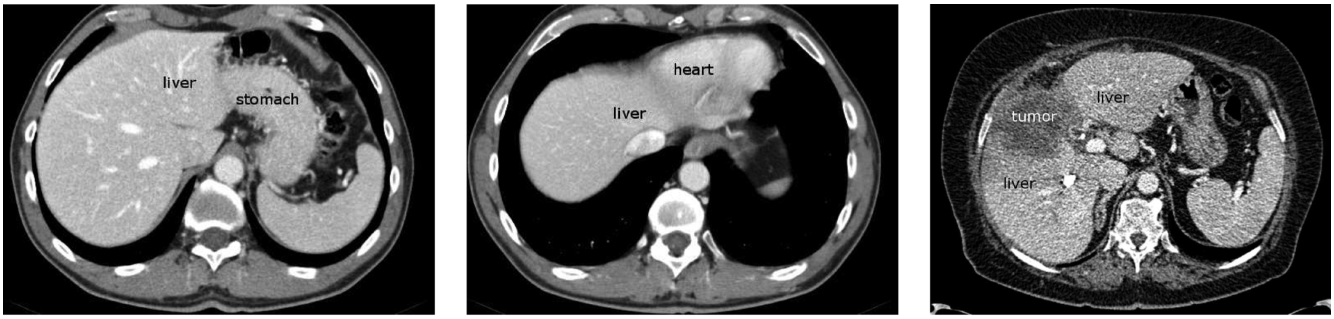
Digital Object Identifier 10.1109/TMI.2009.2013851

[1]See http://www.itk.org.

[2]See http://www.intel.com/technology/computing/opencv.

[3]See http://www.cma.mgh.harvard.edu/ibsr.

[4]See http://www.isi.uu.nl/Research/Databases/DRIVE.



(a) (b) (c)

Fig. 1. Examples why liver segmentation is a challenging task. In the first two images, liver tissue has to be separated from adjacent organs stomach (a) and heart (b). The gray-values in all structures are highly similar, which makes boundary detection difficult without a-priori information about the expected shape in these regions. In the third image (c), the tumor should be segmented as part of the liver. However, there is a considerable intensity difference between both structures, which often leads to misclassification of the tumor as nonliver tissue.

methods are collective efforts and events as the "Retrospective Image Registration Evaluation Project"[5] or the series of competitions for liver tissue and tumor segmentation that took place in Japan [3], [4]. Outside the medical domain, this kind of events is more common. The computer vision community e.g., has established competitions for stereo matching,[6] face recognition,[7] tracking,[8] and 3-D object retrieval,[9] to name just a few.

This paper is based on a competition for liver segmentation that took place during the "Grand Challenge" workshop at the Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2007 Conference in Brisbane, Australia. A second competition held at that workshop involved segmentation of the *Caudate Nucleus* and is described in a companion paper. Liver segmentation is the basis for computer-based surgery planning of interventions as tumor resection, living donor transplantations, or minimal invasive surgery [5], [6]. It can also be used for diagnosis and monitoring purposes [7], [8]. In computed tomography (CT) images usually acquired for these purposes, intensities of adjacent organs and tissue are very similar to liver tissue itself. This is often the case for boundaries to stomach and heart, but also for the boundary to the subcostal fat of the rib cage. In these problem regions, automatic segmentation of the liver based on gray-value alone is almost infeasible. Region-growing approaches e.g., leak into surrounding tissue and require subsequent manual corrections [9]. An additional issue is that most clinical images show severe pathologies like large tumors, cirrhosis, or partial liver resection with remaining scars. All these special cases have to be treated correctly by segmentation algorithms. Fig. 1 shows some examples of the various difficulties. Due to its highly varying shape, the liver is also a very challenging structure to describe with model-based approaches. Methods that have been examined for this task include deformable models [10], a probabilistic atlas [11], and statistical shape models [12], [13]. Due to the difficulties outlined above, systems commonly employed in clinical practice rely on manual segmentation, which is tedious, time-consuming, and generally not reproducible [14]–[17]. Main objectives of the workshop were to evaluate how far automatic segmentation methods have come in recent years and to provide a state-of-the-art overview of current techniques. We also wanted to provide a forum for interactive methods that allow a fast and accurate segmentation under full user control, which is a key requirement for clinical practice.

To evaluate the quality of segmentation methods, there exists a number of possibilities. Firstly, resulting segmentations can be compared to human expert delineations using a number of similarity measures [18], [19]. This approach faces the problem that human delineations of medical images are not a true gold standard (see Bioux *et al.* [20] for a comprehensive discussion), but it is the most objective solution. Secondly, resulting segmentations can be rated by a number of human experts [3]. Apart from the practical problem of finding enough experts and the time required to have every image analyzed, this approach is subjective and not guaranteed to deliver repeatable performance estimates. Thirdly, performance of algorithms can be estimated by their common agreement, e.g., using the STAPLE algorithm [21]. Although this approach provided good results when comparing brain tissue classifiers [20], there is the inherent danger that algorithms obtain good scores by simply producing the same errors as other methods (which also leads to a high common agreement). We decided to employ the first approach and compared all segmentations to human references in this study. Out of the three presented approaches, this is the only one that effortlessly allows adding more algorithms to the comparison at a later stage. When using the common agreement approach, all results have to be recalculated when adding new segmentations (possibly changing scores of older methods). For the human expert rating, adding new segmentations at a later stage requires the raters to go through a number of older results each time to recall their old ratings and avoid bias, a highly impractical procedure.

II. Materials and Methods

A. Image Data and Organization

The images employed in this study were provided by several clinical partners. A variety of different CT scanners was used for acquisition, including machines with 4, 16, and 64 detector

[5]See http://insight-journal.org/rire.

[6]See http://vision.middlebury.edu/stereo/.

[7]See http://www.ee.surrey.ac.uk/CVSSP/banca/icpr2004/.

[8]See www.pets2007.net.

[9]See http://www.aimatshape.net/event/SHREC/.

rows of different manufacturers. All datasets were acquired contrast-dye-enhanced in the central venous phase. Depending on machine and protocol used, pixel spacing varied from 0.55 to 0.8 mm in x/y-direction, and slice distance varied from 1 to 3 mm. One of the used protocols required patients to lie on their side, i.e., the entire anatomy is rotated around the z-axis. Most images in the study were pathologic and included tumors, metastasis and cysts of different sizes.

To generate reference segmentations for the study, radiological experts manually outlined liver contours for all images in transversal slice-by-slice fashion. Generally, the first tool used in this Process. was an intensity-based seeded region-grower. In case of leakage or inaccurate boundaries, segmentations were corrected by drawing freehand contours for the affected parts. The employed segmentation protocol defined the segmentation mask as the entire liver tissue including all internal structures like vessel systems, tumors etc. A vessel is considered internal if it is completely surrounded by liver tissue in the respective transversal slice. In case of partial enclosure (occurring where large vessels as *Vena Cava* and portal vein enter or leave the liver), the parts enclosed by liver tissue are included in the segmentation, thus forming the convex hull of the liver shape. A binary median filter of $3 \times 3 \times 3$ size was employed to emend inconsistencies between neighboring transversal slices. To ensure consistency of segmentations over different data sets, all results were inspected and if necessary corrected by a single expert at the end of the Process. Finally, all patient- and center-related information in the datasets was removed by converting them to a raw file format.

From a total of 40 images, 20 were randomly selected as training set and 10 as test set for this study. The remaining 10 images were used for the on-site competition at the Grand Challenge workshop, which is only briefly discussed in this report. Participants could download the training set including reference segmentations and employ this data to train, test and tune their algorithms. They were also allowed to use additional proprietary training data for that purpose. The 10 test images were made available without reference segmentations, in order to prevent participants from systematically tuning their algorithms for this set and biasing the evaluation. Before the given deadline, all participants had to upload their segmentation results for the test images to a central website. In the following months until the beginning of the Grand Challenge workshop, all teams had the opportunity to bug-fix and fine-tune their methods for the on-site competition, and to send in updated results. This offer was used by 8 out of 10 teams with automatic methods and three out of six teams with interactive methods. All results reported in this paper are produced by the final version of all algorithms, and results may therefore vary from the numbers published in the workshop proceedings [22].

B. Evaluation Measures and Scoring

To evaluate the quality of a given segmentation, we follow the empirical discrepancy approach [23]. Segmentations are compared to expert-generated references and rated according to detected deviations. There exists a multitude of different measures for this purpose, as described by Niessen *et al.* [18] and Gerig *et al.* [19], among others. Most commonly used are metrics based on volumetric overlap and surface distances. Both categories feature several variants, i.e., different mathematical definitions of how overlap and distances are measured. Depending on the application for which segmentation results are evaluated, there may be a clear preference for one category over the other. When e.g., measuring tumor volumes, an assessment of the volumetric error may be preferred over distance measures, since accurate determination of the volume is the single most important objective in that case. However, for a general evaluation of segmentation quality as in the presented study, a variety of different measures should be employed. The main advantage of using multiple measures is that each individual measure highlights different aspects of segmentation quality. A result might be very similar to the reference for the largest part of the boundary but show one large deviation at a small local area. In this case, the average error will be low, but the maximum error will be high. Calculating different measures for average and maximum error will convey more information than just using one measure. Similarly, measuring errors as volumetric differences will usually lead to different results than measuring errors as surface distances. Again, employing both kinds of measures will convey more information and a better estimate of the overall segmentation quality.

The problem when using multiple measures is how to combine different results to allow a quick assessment of general quality and to produce a ranking of segmentation methods. We solve this question by transforming the result ϵ_i of each error measure i to a gauged score $\phi_i \in [0, 100]$ and averaging all values to one final score

$$\phi = \frac{1}{N} \sum_{i=1}^{N} \phi_i. \tag{1}$$

To calibrate the scores, we let a second rater segment all test images manually and compared the results to the respective references. This yielded average user errors $\bar{\epsilon}_i$ for each measure. Defining the performance of this second rater as 75 out of 100 points, we calculate the corresponding score for measure i by

$$\phi_i = \max\left(100 - 25\frac{\epsilon_i}{\bar{\epsilon}_i}, 0\right). \tag{2}$$

Thus, a score of 100 points corresponds to a perfect match with the reference segmentation, and a score around 75 can be regarded as roughly equivalent to human performance.[10] Scores from different test images can be averaged to estimate the overall performance of the segmentation algorithm in question. To prevent a single segmentation with very large errors from biasing this final score too much, all scores ϕ_i are truncated at zero, as described by (2). This means that a deviation from the reference more than four times as high as the second observer is essentially considered as complete failure. Please note that ϵ_i and ϕ_i are computed for each individual segmentation, while $\bar{\epsilon}_i$ is computed once across all segmentations created by the second rater.

[10]The second rater received adequate training in liver segmentation, but was no radiological expert. An experienced user should be able to obtain more than 75 points.

The described scoring system was employed with the following five error measures.

1) Volumetric Overlap Error: The volumetric overlap error between two sets of voxels A and B is given in percent and defined as $100(1-(|A\cap B|/|A\cup B|))$. The ratio between intersection and union used in this term is also known as Tanimoto [24] or Jaccard [25] coefficient. The volumetric overlap error is 0 for a perfect segmentation and 100 if segmentation and reference do not overlap at all. This metric is one of the most popular methods to evaluate segmentation accuracy and was selected for that reason.

2) Relative Volume Difference: The relative volume difference between two sets of voxels A and B is given in percent and defined as $100(|A|-|B|/|B|)$, with A as segmentation and B as reference. Since the measure is not symmetric, it is no metric. A value of 0 means both volumes are identical. Note that this does not imply that A and B are identical, or actually overlap with each other. For this reason, the relative volume difference should never be used as the only measure of segmentation quality. In combination with other measures though, it reveals if a method tends to over- or undersegment. For this purpose, results of the relative volume difference are given as signed numbers. To calculate the corresponding score, the absolute value is used. Apart from its role in distinguishing over- from undersegmentation, this measure was selected because it directly evaluates volumetric information. For applications as liver surgery planning, volumetric information is the single most important number that segmentation provides.

3) Average Symmetric Surface Distance: The average symmetric surface distance is given in millimeters and based on the surface voxels of two segmentations A and B. Surface voxels are defined by having at least one non-object voxel within their 18-neighborhood. For each surface voxel of A, the Euclidean distance to the closest surface voxel of B is calculated using the approximate nearest neighbor technique [26] and stored. In order to provide symmetry, the same Process. is applied from the surface voxels of B to A. The average symmetric surface distance is then defined as the average of all stored distances, which is 0 for a perfect segmentation.

Formally: let $S(A)$ denote the set of surface voxels of A. The shortest distance of an arbitrary voxel v to $S(A)$ is defined as:

$$d(v,S(A)) = \min_{s_A\in S(A)} \|v-s_A\| \tag{3}$$

where $\|.\|$ denotes the Euclidean distance. The average symmetric surface distance is then given by:

$$\mathrm{ASD}(A,B) = \frac{1}{|S(A)|+|S(B)|}\left(\sum_{s_A\in S(A)} d(s_A,S(B)) + \sum_{s_B\in S(B)} d(s_B,S(A))\right). \tag{4}$$

Together with the volumetric overlap error, the average symmetric surface distance is one of the most commonly used measures. Therefore, it was also included in our selection.

4) Root Mean Square Symmetric Surface Distance: The root mean square (rms) symmetric surface distance is given in millimeters and is also based on surface distances. It is calculated as the average symmetric surface distance described above, except that Euclidean distances between surface voxels are squared before storing them. The root of averaged squared distances then yields the rms symmetric surface distance, which is 0 for a perfect segmentation

$$\mathrm{RMSD}(A,B) = \sqrt{\frac{1}{|S(A)|+|S(B)|}} \times \sqrt{\sum_{s_A\in S(A)} d^2(s_A,S(B)) + \sum_{s_B\in S(B)} d^2(s_B,S(A))}. \tag{5}$$

The rms distance is highly correlated with the average distance, but has the advantage that large deviations from the true contour are punished stronger. In our opinion, this is one of the single best choices to evaluate segmentation accuracy.

5) Maximum Symmetric Surface Distance: The maximum symmetric surface distance is given in millimeters and determined similar to the previous two metrics. It also known as Hausdorff distance [27]. Differences between both sets of surface voxels are determined using Euclidean distances, and the maximum value yields the maximum symmetric surface distance. For a perfect segmentation this distance is 0

$$\mathrm{MSD}(A,B) = \max\left\{\max_{s_A\in S(A)} d(s_A,S(B)), \max_{s_B\in S(B)} d(s_B,S(A))\right\}. \tag{6}$$

We included this metric to our set of error measures because it is sensitive to outliers and returns the true maximum error. This is required for applications as surgical planning, where the worst case error is more important than average errors.

TABLE I
NUMBER OF TRAINING SAMPLES USED IN EACH AUTOMATIC METHOD. VALUES >20 MEAN THAT ADDITIONAL PROPRIETARY DATA WAS USED TO TRAIN THE ALGORITHM

Method	Number of used training samples
Kainmüller *et al.*	112
Heimann *et al.*	35
Saddi *et al.*	50
Schmidt *et al.*	0
Chi *et al.*	20
Ruskó *et al.*	0
Seghers *et al.*	20
Furukawa *et al.*	20
Rikxoort *et al.*	12
Susomboon *et al.*	20

C. Automatic Segmentation Methods

This section describes all fully automated methods that participated in the comparison study. "Fully automated" implies that each algorithm had to use the same set of parameters for all test images. The amount of training images used in each method is summarized in Table I.

1) Shape-Constrained Segmentation With Heuristic Appearance Model: The method presented by Kainmüller *et al.* [28] matches a statistical shape model (SSM [29]) to the image data and refines the solution using a deformable mesh. The

SSM consists of around 7.000 landmarks and is built from an extensive training set of 112 liver shapes. The required correspondences are determined using a semi-automatic method where the principal ridges of each liver are specified manually. Thus, the surface is divided into several patches that are matched using surface parameterizations [30]. The employed appearance model is based on profiles running perpendicular to the surface at all landmarks. Depending on the phase of the segmentation, each profile consists of 10–60 intensity samples spaced at 0.2–1 mm distance. A heuristic method estimates the position of liver boundary within each profile, based on classifying intensity samples as liver tissue, tumor tissue, or background. This classification process employs specific intensity models for liver and tumor, each represented by an intensity interval $[i_{\min}, i_{\max}]$. The interval bounds are estimated by means of a histogram of the voxel intensities inside the current liver-segmentation. A sum of Gaussians is fitted to this histogram via the expectation-maximization (EM) algorithm [31], and the interval bounds are derived from the obtained means and standard deviations. Before starting the segmentation, images are smoothed using nonlinear isotropic diffusion [32]. To initialize the pose of the SSM, the right lung lobe is identified using thresholding and morphological operations [33]. The detected lobe is then projected downwards to the first slice that does not contain any lung voxels, and the center and orientation of the projection entail the initial transformation of the SSM. Subsequently, four iterations of intensity model estimation and SSM search are conducted with increasing number of modes for the SSM and varying parameters. After that, the resulting surface is remeshed to initialize the freely deformable mesh. Mesh evolution is guided by three vector fields: displacement towards best fit to data, smoothing, and shape preservation. The first one employs the above-described appearance model, the second one draws each vertex towards the center of its direct neighbors, and the last one ensures the mesh does not leave a narrow band defined around the SSM segmentation result. The entire segmentation process takes approximately 15 min per image on a 3.2 GHz personal computer (PC).

2) Shape-Constrained Segmentation With Nonlinear Appearance Model: Heimann *et al.* [34] employ a deformable mesh with internal forces based on an SSM and external forces based on image data. The underlying SSM consists of 2.500 landmarks and is built fully automated from 35 training shapes. Correspondences are determined by a population-based optimization approach that is based on the description length of the model [35]. The appearance model for each landmark consists of seven sample long intensity profiles perpendicular to the surface at varying image resolutions. During model construction, profiles are sampled at true boundary and displaced positions to train a moderated k-nearest-neighbor-classifier [36]. To improve performance of the classifier, landmarks with similar appearance are clustered to one model. For initializiation of the SSM, an evolutionary algorithm [37], [38] is employed. During evolution, each solution (i.e., individual of the population) consists of pose and shape parameters for the SSM. It is evaluated by multiplying probabilities of the most reliable appearance models [39] in a 4 times down-sampled image. The resulting weights are used in a random sampling process to assemble the population for the next iteration. Each chosen solution is mutated by adding Gaussian noise to all parameters. After a fixed number of iterations, pose and shape parameters of the solution with the best weight are used to initialize a deformable mesh in the image. Mesh evolution is guided by two vector fields: internal and external forces. Internal forces are based on differences in edge lengths and angles to the underlying SSM and draw all vertices towards the closest valid shape. External forces employ the appearance model to estimate boundary probabilities in a narrow band around the current surface. An optimal surface detection with hard constraints [40] is used to find the optimal target locations for all vertices given the image data. During the course of the segmentation, weights between internal and external forces are shifted so that the latter gain more influence toward the end. Moreover, the entire process runs in a multiresolution fashion, starting at a 4 times down-sampled image and ending at the original resolution. Segmentation of one image takes approximately 7 min on a 3 GHz PC.

3) Shape-Constrained Segmentation Using a Variational Framework: Saddi *et al.* [41] embed their algorithm in a variational framework [42], where an SSM segmentation step is followed by free deformation. The statistical model is built from 50 training samples which are represented by signed distance functions. As no landmarks are involved, determination of correpondences is not necessary. Instead of a Gaussian distribution of shape parameters which is commonly assumed, the authors use a nonparametrical shape distribution based on kernel density estimation [43]. The appearance model consists of intensity distributions for inside and outside of the liver, which are dynamically estimated during segmentation. To initialize the SSM, an intensity histogram of the image is analyzed using a Gaussian mixture model. The intensity of liver tissue deduced from this analysis is employed to threshold the image. The pixel yielding the highest response in a subsequent smoothing of the mask is assumed to lie inside the liver and is used as starting point for the level set evolution. A gradient descent algorithm is then employed to find the boundary which minimizes the segmentation energy. This energy is composed of a shape term (based on 30 SSM modes) and a data term (based on the appearance model). Subsequently, nonrigid registration is used to refine the segmentation. The new energy to be minimized contains the same data term as before, but instead of using the shape term, transformations are regularized by a variant of the fluid registration approach [44]. During optimization, small displacements are concatenated to find the optimal high-dimensional transformation. The whole process is conducted in a coarse-to-fine fashion with five levels of resolution. Segmentation of one image takes 5.5 min on a 2 GHz dual core PC.

4) Rule-Based Segmentation Using Scripting Language: Schmidt *et al.* [45] employ a specialized scripting language to define a set of rules which is used to successively extract different structures from the test images. The order of extraction is: background air, lungs and other intrabody air, subcutaneous fat and muscle layer, bones within muscle layer, aorta, spine, heart, and liver. In each extraction step, the system makes use of already detected structures to aid in the image analysis.

Rules can incorporate knowledge about intensity distributions, neighborhood relations, geometric features, etc. After having extracted the above listed structures up to the heart, a seed region for the liver is chosen by thresholding in the right side of the image below the heart, until an object fulfilling certain size criteria is detected. From this seed region, a process similar to region-growing (but including surface smoothness constraints) is started. The growing process is blocked by certain previously detected structures as fat, and attracted by others as lungs (letting the liver grow up to the lung boundary). All rules are defined without making use of the supplied training data, and parameters have not been optimized systematically. Computation time varies with the size of the image to be analyzed and ranges from 6 to 20 min on a standard 3 GHz PC.

5) Gradient Vector Flow Snake Using a Variational Framework: The approach of Chi *et al.* [46] is based on rotational template matching, clustering, and level sets [47]. The supplied training images are used to construct a 2-D template of the heart. Matching this template to different slices of an unseen image, the rough location of the liver and rotation of the patient can be inferred. This information is used to threshold the topmost slice of the liver and initialize a closed curve at its boundary. In order to make the curve evolution more robust, two techniques are employed. Firstly, the rib cage is detected by thresholding, and an edge enhancement in its vicinity aids to prevent leakage of the curve into subcostal fat. Secondly, a 2-D k-means clustering [48] is run in all relevant slices, using predefined area proportion rates to label pixels as liver, kidney, vessels/heart, or tumor. Nonliver labels are then employed to block the curve from entering irrelevant areas. Curve evolution is implemented in a variational framework and includes curvature and stopping terms. Image forces based on a distance transform are used for rough segmentation and subsequently refined by a narrow band gradient vector flow. Segmentation is started at the topmost slice and then continued downwards, initializing new slices with the results of previous ones. The required computation time depends on the number of slices in the dataset; for 200 slices the algorithm runs for 30 min.

6) Three-Dimensional Region-Growing: Ruskó *et al.* [49] employ an advanced region-growing technique [50] and subsequent postprocessing to segment the liver. To improve the runtime of the algorithm, the image to be analyzed is first resampled in z-direction to obtain a slice thickness between 2 and 3 mm. A histogram of intensities between -50 and 250 is then created for the right side of the image and analyzed to estimate the intensity distribution of the liver. The detected upper and lower values are employed to threshold the image to a binary mask, which is subsequently eroded with a large kernel [33]. This process results in an isolated region in the center of the liver that is used as seed region for subsequent region-growing. After reestimating the intensity distribution of liver tissue from the seed region, adjacent voxels are consecutively added if all voxels within 5 mm radius are within the estimated intensity interval. To prevent leakage of the region-growing into the heart, a surface connecting the bottom parts of both lung lobes is detected in advance and used as blocker. Postprocessing is started at the boundary to the right lung lobe: a new intensity distribution is estimated from not yet classified voxels and employed by a locally constrained second region-growing in this area. Subsequently, leakage through the *Vena Cava* is corrected by detecting circles of predefined radius in transversal slices and deleting them if they fulfill the required length criterion. Cavity filling is employed to label vessel trees and smaller tumors as liver tissue. Finally, the binary mask is converted to a mesh, smoothed, and converted back to voxel data of the original z-resolution. The entire method runs in half a minute on a standard PC.

7) Shape-Constrained Segmentation Using a Local Shape Model: Seghers *et al.* [51] propose a local shape model and dynamic programming [52] to segment the test datasets. The employed SSM is built from 20 training samples and consists of 2004 landmarks. Correspondences are determined by a point registration approach in two stages. First, an arbitrarily chosen training shape is registered to the remaining shapes, constituting a first set of correspondences. Second, bias is removed by generating a template from the mean of these correspondences and registering it to all training meshes to generate the final landmarks. Unlike the above described global shape models, the authors build a statistical model for each edge of the mesh, capturing mean and covariance of the edge vector. The appearance model is based on 24 feature images (intensity and first derivatives in various scales), which are sampled using "spherical intensity profiles:" one point is sampled at each landmark and six from a sphere of 10 mm radius surrounding the center (15 mm radius for a rougher resolution). The model is initialized by an affine registration of the new image to the template image used for landmarking. A set of candidate points is sampled in a grid around each landmark and evaluated with the specific appearance model. For each landmark, 100 points with the best matches are kept. Subsequently, a number of 1-D paths is generated by randomly joining adjacent edges. On each path, dynamic programming is used to find the combination of candidates that minimizes shape and appearance energies. The 20% least voted candidates are removed, until only one candidate per landmark remains and delivers the final segmentation. The algorithm is run in two resolutions with differently-spaced candidate grids and runs approximately half an hour on a 2.8 GHz PC.

8) Bayesian Voxel Classification With Probabilistic Atlas and Level-Set Refinement: Furukawa *et al.* built upon their workshop contribution [53] of maximum *a posteriori* (MAP) estimation and level sets. In preprocessing (applied to both training and test dataset), images are first resampled to isotropic voxels. Lungs and bones are then marked by thresholding in combination with morphological operations. Rotations around the z-axis are corrected by analyzing a maximum-intensity-projection of the rib cage. To build the probabilistic atlas, heart, right kidney, and other nonliver tissue are manually segmented in all training images. One training image is arbitrarily selected as template. All others are aligned to it by a translation matching the centers of mass of the lungs and subsequent x/y-scaling to reach the same lung size. Using the same transforms on the respective label masks yields the probabilistic atlas. To segment a new image, the same preprocessing and alignment is used. Voxels are classified as liver, heart, kidney, or other using MAP estimation based on intensity value, response to min–max filter, and the three coordinates. The required conditional probabilities for these features are estimated using the EM algorithm for an extended Gaussian mixture distribution [54]. The algorithm is initialized with values learned from training data and the probabilistic atlas as prior. It outputs maximum likelihood estimates of parameters as well as posterior probabilities of organs at each voxel, leading to a MAP-based rough segmentation that finds the organ label with the maximum probability. The resulting segmentation is employed to obtain a better alignment with the atlas, based on positions of right kidney and lung. After a second MAP estimation and morphological processing [33] to infer the position of the heart given the lung boundaries, lesions within the liver are detected by an additional voxel classification. Features in this step are based on different filter responses and evaluated by an AdaBoost-trained classifier cascade [55]. Candidate regions for lesions that fulfill certain criteria regarding their intensity distribution are added to the current liver mask. The final step consists of a geodesic level set algorithm with an additional term to prevent the liver from growing too close to the body surface (skin). On a standard PC, one segmentation requires around 36 min, or 15 min on a quad-core PC.

9) Non-Linear Voxel Classification With Multiatlas Segmentation: The approach proposed by van Rikxoort *et al.* [56] is based on voxel classification in combination with a multiatlas registration. A k-nearest-neighbor (kNN) classifier [57] is employed to label each voxel within an automatically detected candidate region as liver tissue or background. Used features are 44 Gaussian derivatives up to second order (in all three directions and at various scales) and three spatial features based on atlas registration. Sequential floating forward feature selection is employed on the training data to isolate the 10 strongest features from this set. In preprocessing, each image is resampled to isotropic voxels. To detect and correct rotations around the z-axis, bones are detected by thresholding and the spread of the resulting binary mask in x-direction is maximized by applying different rotations. After that, lungs are detected by thresholding and the potential liver region is limited to a fixed height around the lower lung rim. Twelve selected training scans are then registered to the new image using an affine transform followed by a *B*-splines approach [58] in multiple resolutions. For this purpose, a negative mutual information cost function [59] is optimized by a stochastic gradient optimizer. The resulting transformation fields are used to map individual training segmentations to the new image. This yields a probabilistic atlas segmentation [60] on which the three spatial features are based: they represent the percentage of the probabilistic segmentation above, left, and behind the voxel in question. After classifying each voxel in the vicinity of the mask with a 15-nearest-neighbor classifier, results are postprocessed by smoothing and morphological operations [33]. Segmentation of one image takes around 45 min on a standard PC.

10) Clustering, Voxel Classification and 2-D Region-Growing: The method proposed by Susomboon *et al.* [61] uses clustering, voxel classification and region-growing to segment the liver. The EM algorithm [31] is employed to define confidence intervals for the intensity distributions of air, fat, soft-tissue, and bone in the supplied training images. Using these intensity values, a quad-tree decomposition is employed to separate each slice of the image to be analyzed into tissue and nontissue regions. For all tissue regions, Haralick texture features [62] are calculated and forwarded to a classification and regression tree [63]. The tree then estimates the probability for liver tissue. The region with the highest probability of liver tissue in all slices is selected as seed region. A 2-D region-growing process is started from there, adding new voxels if all gray-values in the 9×9 neighborhood fall within the trained confidence interval. The process is repeated in neighboring slices. Newly detected regions are added to the segmentation if they feature an overlap of at least 85% with the current slice. Segmentation of one image takes around 25 min.

D. Interactive Segmentation Methods

This section describes segmentation methods that require a certain amount of user interaction to complete. The type of interaction employed varies from providing a single seed-point to extensive manual refinement of the binary segmentation mask. To allow a fair comparison of methods, we have classified all approaches according to the time required for interaction. Less than 1 min is regarded as *low interaction*, less than 5 min as *medium interaction*, and more than 5 min as *high interaction*.

1) Graph-Cut and Interactive Refinement: Beichel *et al.* [64] present an approach with high interaction, which is based on a graph-cut segmentation [65] and two refinement steps. To initialize the method, the user has to mark one or more seed regions inside the liver. Background seeds are set automatically, and a graph-cut algorithm [65] is employed to separate liver from background. Costs for the graph-cut are based on a region and a boundary term: the first one penalizes intensity differences between analyzed voxel and the distribution at the seed-points, while the latter one prefers cutting the graph at voxels with high "surfaceness" measure. Surfaceness is calculated based on image gradients, which are nonlinearly averaged for adjacent voxels and weighted to punish nonmaximal gradient magnitudes. In addition to forming a cost term for the graph-cut, surfaceness is also used to fragment the image volume to a collection of chunks. In a first refinement step, the user can add or remove individual chunks to or from the current segmentation result. Subsequently, the binary segmentation mask is converted to a simplex mesh and can be refined using a variety of different tools [66]. In the complete setup of the approach, both refinement steps are conducted within a virtual reality environment with stereoscopic vision [67]. The time required to segment one test image is approximately 36 min, consisting of 30 min of computation time for the graph-cut solver and chunk generation and 6 min for user interaction.

2) Region-Growing and Interactive Refinement: The method by Beck and Aurich [68] is another approach presented at the workshop with high interaction component. Core of the method is a 3-D region-grower with a nonlinear coupling criterion: new voxels are added to the seed region if the weighted intensity difference of an 11^3 neighborhood to the seed intensity is below a given threshold. Usually, the region-grower is employed iteratively at different locations inside the liver until the entire organ is segmented. Leaked regions or missing parts are corrected manually by a "virtual knife," which removes all labels on one side of a user-specified cutting plane. A postprocessing step extends the segmentation by calculating the convex hull inside limited local regions around the boundary. Segmentation of one image takes 3–15 min, with an average of 7 min for the supplied test datasets.

3) Two-Dimensional Level Sets With Transversal Contour Initialization: The method proposed by Dawant *et al.* [69] requires medium interaction and is based on level sets with a dynamic speed function. To initialize the algorithm, the user has to draw rough contours inside the liver tissue in a number of transversal slices. These contours are used as initialization for a 2-D level set evolution, which employs a novel definition for the speed function [70]. As opposed to other approaches in which the speed function is spatially dependent but independent of the propagation path, the employed function is dynamically modified at every iteration. The basic tenet in this algorithm is that one does not need to design a speed function that is identically equal to zero over the organ boundary to stop the contour. If one can keep track of the contour path and include this information in the definition of the speed function, it is sufficient to decelerate the contour over possible edge pixels. If the front passes over isolated points that are likely to be edges, it keeps propagating. If it passes over several such points in sequence it slows down exponentially until it is permanently stopped. To prevent contours from growing inside subcostal fat, skin and rib cage are previously extracted by thresholding and morphological operations [33]. Results of this step are used in an additional term in the speed function as blocking regions. Once all slices with user-drawn contours are segmented, results are interpolated to the remaining slices. The entire method features six free parameters, of which five are estimated from the user-drawn contours, and the sixth one (width of the narrowband) is claimed to be not critical for the segmentation result. Computation time is approximately 20 min on a 3.2 GHz PC.

4) Two-Dimensional Level Sets With Seed-Point Initialization: Lee *et al.* [71] present a segmentation method with low interaction based on level sets [72]. To initialize the algorithm, the user has to specify one seed-point at the top of the liver and one seed-point at the bottom of every lobe. The speed image for the level set evolution is generated by smoothing the input image using curvature diffusion filtering [73] and then calculating gradient magnitudes. From the user-placed seed-points, 2-D regions are grown on the speed image until reaching the high gradients at the liver boundary. These first results are refined using geodesic active contours in a narrow band around the initialization. After convergence, results are copied to the next slice toward the center of the liver and used as the initialization for a geometric active contour evolution. This process is repeated until the segmentations from top and bottom of the liver meet at the center of the liver. As a postprocessing step to fill vessels at the boundary of the liver, a rolling ball filter [74] is run on the resulting binary volume. In order to prevent filling correctly segmented gaps, the average intensity value of a newly added region is compared to that of the previously segmented liver region. Regions with an average intensity lower than the average liver intensity are considered as false positives and excluded. To process one image, this approach requires approximately 7 min on a 2.4 GHz quad core PC.

5) Three-Dimensional Level Sets With Orthogonal Contour Initialization: The approach described by Wimmer *et al.* [75] requires a medium amount of interaction for initialization and subsequently uses level sets [76] for the final segmentation. The user defines a number of 2-D contours in image planes resampled from various directions (preferably orthogonal to each other). Contours are defined by placing points on the liver boundary, which are interpolated using cubic splines. After the user has set 6–8 contours, radial basis functions are employed to generate a smooth surface passing through all contours and interpolating in between [77]. This surface is used as initialization for an edge-based geodesic active contour. As it is usually close to the true liver boundary, no constant speed forces are required during the level set evolution. The distance of the current level set surface to the user-specified input contours is used as additional shape-preserving term. Segmentation of one image takes 4–7 min on a 1.8 GHz PC.

6) Atlas Matching Using B-Splines: Slagmolen *et al.* [78] employ nonrigid atlas matching to segment the liver, requiring medium interaction to initialize their method. The atlas is built from the 20 provided training images. To register two images for atlas construction, the first step is an affine transform using the mutual information (MI) metric [79]. Subsequently, a *B*-spline nonrigid registration [58] is conducted using a combination of MI, regularizing costs, and surface distance penalty between the corresponding segmentations. A multiresolution approach [58] is employed for better robustness. Each training image is registered to all others, and the resulting deformation maps are averaged to transform each image into a common coordinate system. Averaging all images and segmentations transformed with this method yields the final atlas. To segment a test image, the user has to define an affine transform that roughly matches the atlas to the image. After optimization of the affine transform, the user has to define a region of interest around the liver in which the subsequent nonrigid registration is conducted. This step employs the same method as during atlas construction, but without using the surface distance penalty. The resulting deformation field is used to transform the probabilistic atlas segmentation to the test image. Thresholding this segmentation at 50% and performing a morphological opening [33] and subsequent removal of unconnected components yields the final result. The approach requires approximately 1 h per image, the largest part of the time is spent for the nonrigid registration.

III. RESULTS

A. Individual Results

All participating teams uploaded binary segmentation masks obtained with their method for the 10 test images to a central website. Beichel *et al.* submitted three results for different stages of their interactive approach, rated as low interaction (<30 s) for the graph-cut segmentation (GC), medium interaction (approximately 90 s) after chunk-based refinement (CBR) and high interaction (approximately 6 min) after mesh-based refinement (MBR). Using the measures and scoring system presented in Section II-B, all submissions were compared to the corresponding manual reference segmentations. Resulting scores per image are visualized in Fig. 2(a) for automatic methods and in Fig. 2(b) for interactive methods. Averaged results for individual measures are summarized in Tables II and III for automatic and interactive methods, respectively.

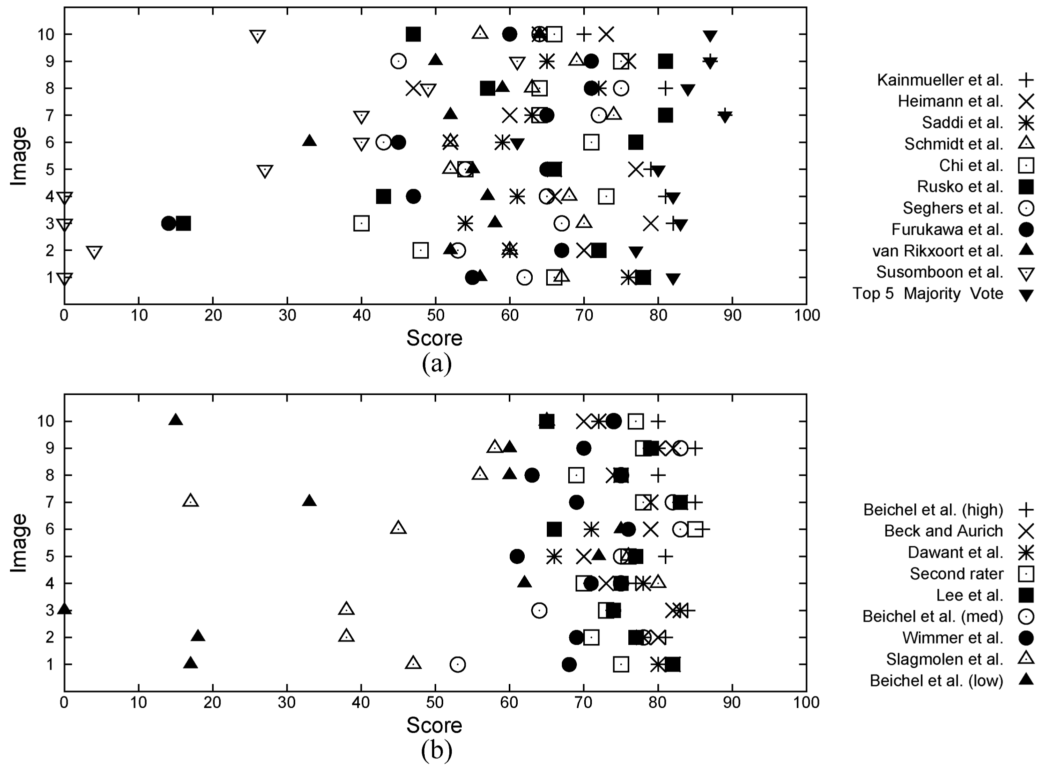


Fig. 2. Individual scores per image for all segmentation methods. All displayed scores are calculated according to (1). (a) Automatic segmentation methods. (b) Interactive segmentation methods.

TABLE II
OVERVIEW OF RESULTS FOR AUTOMATIC SEGMENTATION METHODS. RESULTS FOR EACH MEASURE ARE REPORTED AS MEAN AND STANDARD DEVIATION OVER ALL TEST IMAGES, TOGETHER WITH MEAN SCORE. ALL SCORES ARE AVERAGED TO A FINAL SCORE GIVEN AS MEAN AND STANDARD DEVIATION OVER ALL IMAGES

Method	Runtime	Overlap error		Volume difference		Avg. distance		RMS distance		Max. distance		Final
	[min]	[%]	Score	[%]	Score	[mm]	Score	[mm]	Score	[mm]	Score	Score
Kainmüller *et al.*	15	6.1 ± 2.1	76	-2.9 ± 2.9	85	0.9 ± 0.3	76	1.9 ± 0.8	74	18.7 ± 8.5	75	77 ± 9
Heimann *et al.*	7	7.7 ± 1.9	70	1.7 ± 3.2	88	1.4 ± 0.4	65	3.2 ± 1.3	55	30.1 ± 10.2	60	67 ± 11
Saddi *et al.*	5.5	8.9 ± 1.8	65	1.2 ± 4.4	80	1.5 ± 0.4	62	3.4 ± 0.8	52	29.3 ± 8.4	62	64 ± 6
Schmidt *et al.*	6–20	10.4 ± 1.9	59	-4.9 ± 3.0	74	1.7 ± 0.4	56	3.1 ± 1.1	57	24.0 ± 8.0	68	63 ± 8
Chi *et al.*	34	9.1 ± 2.8	65	2.6 ± 6.3	73	1.7 ± 0.6	58	3.3 ± 1.2	54	30.8 ± 9.2	60	62 ± 11
Ruskó *et al.*	0.5	10.1 ± 4.5	61	-3.8 ± 6.4	72	1.7 ± 0.9	58	3.5 ± 2.3	53	26.7 ± 11.7	65	61 ± 21
Seghers *et al.*	30	10.7 ± 2.5	58	-6.8 ± 2.3	64	1.8 ± 0.4	55	3.2 ± 1.1	56	25.2 ± 10.1	67	60 ± 10
Furukawa *et al.*	36	10.8 ± 3.7	58	-7.3 ± 4.7	61	1.9 ± 1.1	53	3.7 ± 1.9	49	31.6 ± 12.7	58	56 ± 17
van Rikxoort *et al.*	45	12.5 ± 1.8	51	1.8 ± 4.2	80	2.4 ± 0.3	40	4.4 ± 1.5	40	32.4 ± 13.7	57	53 ± 8
Susomboon *et al.*	25	26.4 ± 24	31	-11.5 ± 30	42	10.2 ± 13	15	17.1 ± 18	12	74.0 ± 41.5	23	24 ± 22
Top 5 majority vote	-	5.0 ± 1.3	81	-0.7 ± 1.7	93	0.8 ± 0.3	81	1.7 ± 0.8	77	19.1 ± 8.4	75	81 ± 8

All scores are rounded down to the nearest integer. Runtime is given as the average time to segment one image volume. For interactive methods (Table III), this comprises interaction time and computation time.

To provide an additional qualitative impression of the presented methods, coronal cross sections of results on one test image are displayed for all methods. Test image 10 was chosen for this purpose—as Fig. 2(a) and (b) reveal, a challenging dataset with lower than average scores but no complete failures. The selected cross section shows many of the typical "hazards" for medical image analysis, including large tumors and low contrast boundary in several areas. Results of automatic methods on this data are shown in Fig. 3, of interactive methods in Fig. 4. For a larger variety of images from different directions and for all test cases, please refer to the results section of the official website of the comparison study.[11]

B. Combined Automatic Results

To test the hypothesis that multiple independent raters reach better results than a single one, we employed majority voting to generate additional results. The top five automatic methods (according to their reached final score in Table II) were combined by comparing segmentation results on a per-voxel basis. In case the majority of methods labeled a voxel as liver tissue, it was also marked as liver in the final result. These combined results were evaluated using the same measures as the individual results and added to Table II and Fig. 2(a) and Fig. 3.

[11]See http://www.sliver07.org/.

TABLE III
OVERVIEW OF RESULTS FOR INTERACTIVE SEGMENTATION METHODS. RESULTS FOR EACH MEASURE ARE REPORTED AS MEAN AND STANDARD DEVIATION OVER ALL TEST IMAGES, TOGETHER WITH MEAN SCORE. ALL SCORES ARE AVERAGED TO A FINAL SCORE GIVEN AS MEAN AND STANDARD DEVIATION OVER ALL IMAGES. THE AMOUNT OF REQUIRED INTERACTION IS INDICATED IN PARENTHESES

Method	Runtime	Overlap error		Volume difference		Avg. distance		RMS distance		Max. distance		Final
	[min]	[%]	Score	[%]	Score	[mm]	Score	[mm]	Score	[mm]	Score	Score
Beichel *et al.* MBR (*high*)	36	5.2 ± 0.9	80	1.0 ± 1.7	91	0.8 ± 0.2	80	1.4 ± 0.4	80	15.7 ± 3.5	79	82 ± 2
Beck and Aurich (*high*)	7	6.6 ± 1.6	74	1.8 ± 2.5	88	1.0 ± 0.3	74	1.9 ± 0.4	73	18.5 ± 4.1	76	77 ± 4
Dawant *et al.* (*med*)	20	7.2 ± 1.2	72	2.5 ± 2.3	86	1.1 ± 0.2	73	1.9 ± 0.5	74	17.1 ± 5.4	77	76 ± 5
Second rater		6.4 ± 1.0	75	4.7 ± 1.8	75	1.0 ± 0.2	75	1.8 ± 0.5	75	19.3 ± 5.6	75	75 ± 4
Lee *et al.* (*low*)	7	6.9 ± 1.4	73	1.3 ± 2.9	88	1.1 ± 0.3	73	2.1 ± 0.5	71	21.3 ± 4.0	72	75 ± 5
Beichel *et al.* CBR (*med*)	31	6.5 ± 1.1	74	1.1 ± 1.9	90	1.1 ± 0.4	72	2.5 ± 1.2	66	23.4 ± 10.5	69	74 ± 9
Wimmer *et al.* (*med*)	4–7	8.1 ± 1.1	68	6.1 ± 2.6	68	1.3 ± 0.2	67	2.2 ± 0.4	69	18.7 ± 4.6	75	69 ± 5
Slagmolen *et al.* (*med*)	60	10.4 ± 3.1	59	3.7 ± 6.2	70	2.0 ± 0.7	50	5.0 ± 2.4	34	40.5 ± 18.2	47	52 ± 19
Beichel *et al.* GC (*low*)	30	14.3 ± 9.4	48	3.1 ± 10.7	62	3.6 ± 3.1	34	7.9 ± 5.9	24	49.2 ± 20.4	38	41 ± 27

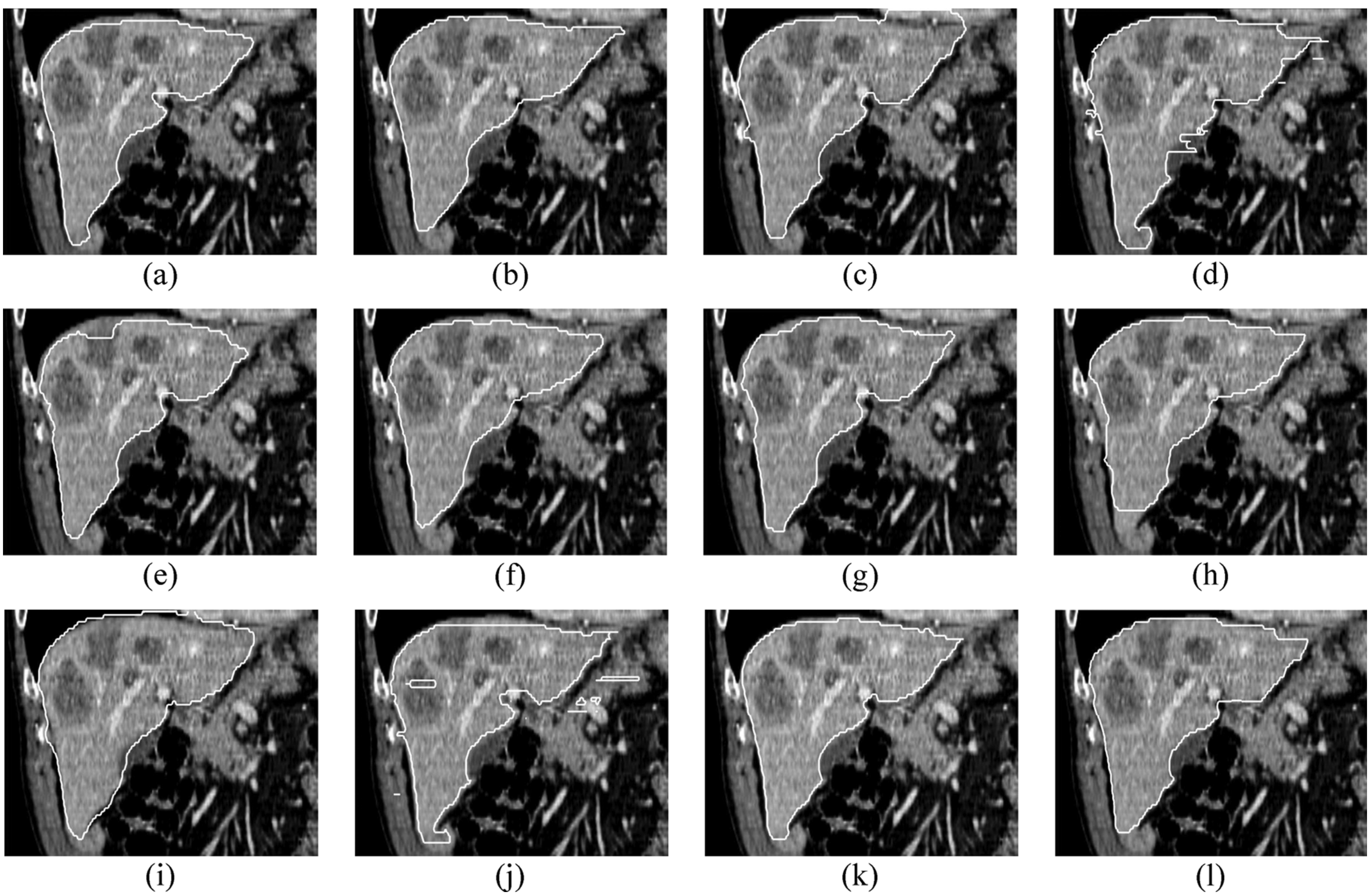


Fig. 3. Coronal view of segmentation results on one test image for all automatic methods. (a) Kainmüller *et al.* (b) Heimann *et al.* (c) Saddi *et al.* (d) Chi *et al.* (e) Ruskó *et al.* (f) Seghers *et al.* (g) Furukawa *et al.* (h) Schmidt *et al.* (i) van Rikxoort *et al.* (j) Susomboon *et al.* (k) Top five majority vote. (l) Reference.

C. On-Site Segmentation Contest

During the Grand Challenge workshop at the MICCAI 2007 conference, 10 new images that none of the participants had seen before were used in an on-site competition for the best segmentation algorithm. Participants were given 3 h of time to segment all images, either on notebook computers on-site or on remote machines in the respective laboratories. To guarantee fair conditions, teams affiliated with the workshop organizers (Heimann *et al.* and van Rikxoort *et al.*) were excluded from this competition. All remaining teams were invited to participate, and most of them accepted. Beichel *et al.* could not take part for technical reasons, as their system requires a sophisticated virtual reality environment which was not possible to bring along to the workshop. Chi *et al.* participated, but could not finish all segmentations due to technical problems. Results of all successfully participating teams are given in Tables IV and V. As this report focuses on the 10 publicly available images, only the final scores of the on-site contest are displayed.[12]

IV. DISCUSSION

In this section, we first briefly discuss individual methods and then elaborate on more general aspects of the comparison

[12]Participants of the contest explicitly demanded publication of these scores. Many deemed the on-site experiments to be fairer and more accurate than evaluation on public data, as the time constraint prevented teams from tuning their methods to deliver high scores on specific test images.

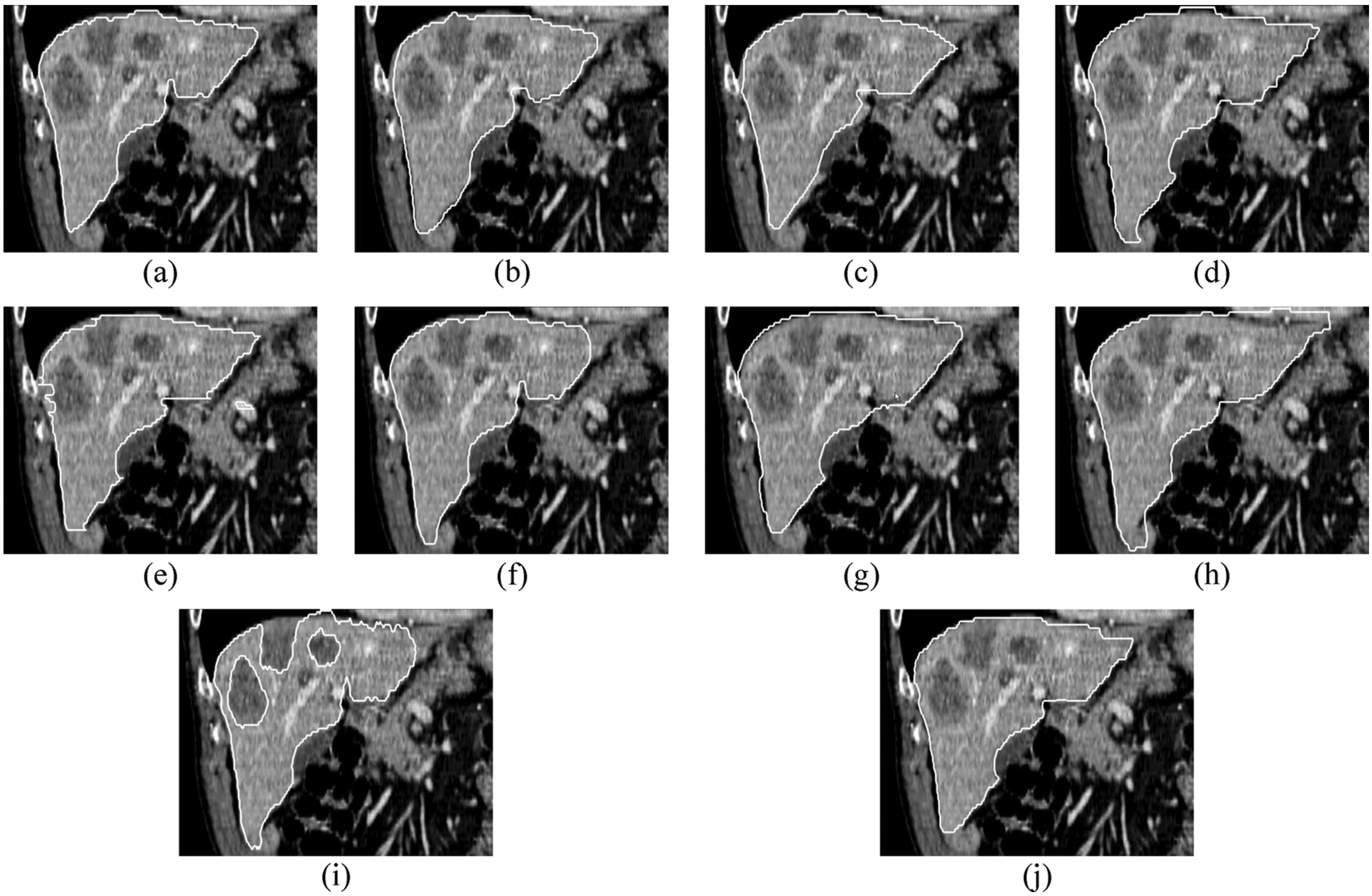


Fig. 4. Coronal view of segmentation results on one test image for all interactive methods. (a) Beichel *et al*. MBR (high). (b) Beck and Aurich. (c) Dawant *et al*. (d) Second rater. (e) Lee *et al*. (f) Beichel *et al*. CBR (med). (g) Wimmer *et al*. (h) Slagmolen *et al*. (i) Beichel *et al*. GC (low). (j) Reference.

TABLE IV
OBTAINED SCORES FOR AUTOMATIC SEGMENTATION METHODS DURING THE ON-SITE COMPETITION AT THE GRAND CHALLENGE WORKSHOP

Method	Final Score
Kainmüller *et al.*	68
Ruskó *et al.*	57
Schmidt *et al.*	53
Seghers *et al.*	51
Saddi *et al.*	51
Furukawa *et al.*	42
Susomboon *et al.*	5

TABLE V
OBTAINED SCORES FOR INTERACTIVE SEGMENTATION METHODS DURING THE ON-SITE COMPETITION AT THE GRAND CHALLENGE WORKSHOP. AMOUNT OF REQUIRED INTERACTION IS INDICATED IN PARENTHESES

Method	Final Score
Dawant *et al.* (*med*)	75
Beck and Aurich (*high*)	73
Lee *et al.* (*low*)	70
Wimmer *et al.* (*med*)	68

study. For an in-depth discussion of the individual segmentation approaches, we refer the reader to the respective workshop papers [22].

A. Results of Automatic Methods

The fully automated category produced clearly discernible performance differences between the different methods: final scores range from 24 to 77 points on a scale from 0 to 100. It is interesting to note that the three best-rated approaches are all based on statistical shape models with some form of additional deformation. Even the team of Kainmüller *et al.*, although employing more than 100 training shapes, does not rely on the shape-space alone and uses a subsequent free deformation step. This supports our initial assessment that the large variation of the liver cannot be captured completely by model-based approaches, but it also shows that utilizing shape information is still beneficial. The local shape model by Seghers *et al.* was applied to 3-D data for the first time in this comparison, thus there is still ample room for refinement of the method. The concept of guiding a deformable surface by statistical shape information without being restricted to a global model is arguably related to the first three methods. Ruskó *et al.* demonstrate that it is possible to produce competetive results with low-level image processing techniques, too. One of the major strengths of their approach is the speed of segmentation, which is with approximately half a minute a magnitude faster than all other methods. The approach by Schmidt *et al.*, although featuring more complex relationships between different objects, is similar in the respect that it does not make explicit use of the provided training data. Moreover, it is probably the most generic system that participated in the study, as the employed scripting language can be used to detect virtually all kinds of objects. The atlas-based methods (Furukawa *et al.* and Rikxoort *et al.*) apparently have more problems to adapt to the wide variety of liver shapes than the explicit shape models. To improve the results, both approaches employ the atlas only as one step among others, which leads to decoupling from the global model (as does the free deformation for shape models). Regarding the systems employing level sets without shape information (Chi *et al.* and Furukawa *et al.*), both utilize additional terms to prevent leakage into subcostal fat (as does the rule-based system by Schmidt *et al.*). Apparently, stopping the level-set evolution based on the standard terms is not reliable due to the low contrast between both tissue types. Most of the presented approaches work natively in 3-D, which has the advantage that all three directions are treated equally. Both methods utilizing slice-based segmentation and propagation in z-direction (Chi *et al.* and Susomboon *et al.*) show some form of inconsistencies in the 3-D view (Fig. 3).

B. Results of Interactive Methods

In the interactive segmentation category, half of the approaches reached scores around 75 points (see Table III). This supports the employed scoring system, as the second rater results were also produced using an interactive system. The highest score reached among all systems was produced by the team of Beichel *et al.* We attribute this excellent result also to their extensive experience with liver segmentation and the close collaboration with radiologists to achieve their goal of clinical applicability. The approach also requires a high amount of interaction, as does the second-best scoring system by Beck and Aurich. In general, there is a clear tendency that systems with high interaction reach better results than systems with medium interaction, which in turn score better than approaches with low interaction. The only exception to this rule is the method by Lee *et al.*, which reaches a good score in the middle field although it only uses low interaction. The approach by Slagmolen *et al.* is rated as *medium interaction* due to the reported timings. However, user interaction is only required for two initialization steps and the authors have already automated both by now. As alluded to earlier, it is very difficult to draw boundaries between different classes of interaction. Our approach by categorizing them according to the required user time is only one of many possibilities. For this reason, an objective comparison and rating as between the fully automatic methods is hardly possible in this case.

Another potential bias for the comparison is the variance between different users when operating the software. In our setup, the participants could freely chose the person generating the segmentations. In most cases, this person was one of the developers of the software, i.e., a computer expert. Designating this task to medical experts with less computer experience would probably slow down these methods and lead to less accurate results. However, some programs might be easier to operate than others, either because of simpler interaction schemes or more intuitive interfaces. For the goal of segmentation in the clinic, these are key points of a software, as medical experts have to accept the program as valuable tool. For the presented study, a different setup with medical experts as operators for all programs is much more difficult to organize and supervise. To sum up, the reader should take note that the presented scores of interactive methods do not reveal much about the usability of the software.

C. General Observations

Comparing automatic and interactive segmentation approaches, a much larger standard deviation of the final scores can be observed for automatic methods. This large standard deviation is mainly due to outlier errors, as can be seen when comparing Fig. 2(a) and (b). While many successful results of automatic methods reach similar scores as interactive methods, reliability of automatic methods is generally still inferior. Interestingly, the problems showing for different automatic methods often arise at different test images and regions. Although some areas (as the ones presented in the introduction) cause more errors than others on average, there is no single region where all methods fail. This observation together with the large variation of results over different test images strengthens the call for a sufficiently large and diverse collection of test images when evaluating performance. Although the image database used in this study features a large diversity, a higher number of training and test images would be beneficial. Evidence for this is the fact that most teams could improve their results noticeably between the first submission and the workshop by small code changes and parameter tuning. In the on-site contest on a new training set, performance broke down considerably for many approaches (compare the scores in Tables II and IV and Tables III and V, respectively). This observation hints at initial overtraining. A larger test database would also allow to draw statistically significant conclusions from the experiments. The current setup, although allowing a good guess about the performance of different approaches, is not suitable for a final judgement. Thus, no technique participating in the comparison should be discarded because it reached a low score on the used test set. Regarding statistical segmentation approaches, the number of training images used to generate the presented results should also be taken into account. As it is difficult to enforce teams to use only a limited training set, we deliberately gave them the freedom to use additional proprietary datasets. However, we are conscious that a larger training base generally leads to better results for these methods. Therefore, a *method* evaluated in the presented study should be understood as ensemble of algorithm and training data. Using this definition, the comparison provides a fair and unbiased evaluation.

The employed evaluation measures deliver mostly consistent results: looking at Tables II and III, it is apparent that a method reaching a high score for one of the evaluation measures generally also reaches high scores for the other measures. One exception with relatively low correlation with the other measures is the relative volume difference. This had to be expected, since two volumes can be identical without any actual overlap between the corresponding segmentations, as explained in Section II-B. Incorporating certain measures into the scoring system (or leaving out others) thus does have some influence on the ranking of the various segmentation methods. However, due to the mostly consistent results of the employed evaluation measures, we do not expect the overall ranking to change drastically.

Another question is how far the overall ranking is influenced by correlation between the included measures: average surface distance and rms surface distance produce similar scores for many methods, which results in a slight bias towards distance-based measures. Still, for the reasons presented in Section II-B, we did not want to drop any of these two metrics. As both measures produce results that correlate very well with the human notion of segmentation accuracy, we feel that a slightly higher weight on them might even be beneficial for the final ranking.

In this study, we have evaluated systems in terms of accuracy of produced results, according to our scoring system. In practice, when choosing a particular algorithm for use in clinical application, other arguments may also be important. For example, applications in image-guided surgery, computation time may be a limiting constraint for the segmentation. Ease of use may be an important factor when designing an interactive system for clinical practice. However, high accuracy of segmentation will always be regarded as extremely valuable, if not absolutely necessary for clinical applications. In this regard, we have made the following observation: although no automatic method reaches the high reliability of the best interactive methods (yet), many yield very satisfying results on the employed test images. The method of Kainmüller *et al.* is already very close to manual performance, but the best automatic results were reached by majority voting of the top five approaches (Table II). In fact, this majority voting reaches the overall best scores regarding overlap error, volume difference, and average surface distance, even in comparison with interactive methods. We can thus indeed deduce that automatic segmentation systems cannot only reach manual performance, but even surpass it. Still, the obvious advantage of most interactive systems participating in our study is the complete user control over the result. This user control is required for clinical applications as long as automatic methods still fail on certain image data.

D. Towards Better Segmentation Methods for the Clinic

The study described in this paper was an evaluation of state-of-the-art methods for segmentation of clinical image data. As such, it provides hints regarding the accuracy and robustness of various techniques, which can be used to direct future research. However, to refine and improve one of the existing methods, one must first be able to reproduce the original method. As mentioned in the introduction, open source toolkits with a large repository of algorithms are one answer to this problem. Currently, the algorithms they provide are very general solutions to allow building a variety of different, more complex applications. For specific applications (as liver segmentation), it would be helpful to establish an open source layer above these basic algorithms. A component of this higher layer would e.g., be `SegmentLiverFromCTWithMultiAtlasMatching()` and encapsulate methods as used in the present study. Starting from this level, progress could be made faster than possible with today's tools. Unfortunately, establishing such a repository of dedicated segmentation methods for particular tasks faces many challenges, including the struggle between advancement of scientific research and commercial interests. A publicly funded initiative could be a first step into this direction.

V. Conclusion

After evaluating 16 automatic and interactive methods for liver segmentation, we have found many approaches to be suitable for application in clinical practice. To further improve current techniques, objective evaluation on a common database is an excellent opportunity to prove the performance of new methods. All training and test data from the presented study is publicly available on the web,[13] and results of new systems can be uploaded for evaluation and ranked with existing systems. We would be happy to see more people using this data, increasing transparency in medical image analysis and potentially establishing a common benchmark for segmentation performance.

Eventually, medical image analysis will benefit from a more data-driven approach to research. In computer vision, there is already a visible trend to evaluate new works by their performance on large databases instead of solely on their novelty. Establishing suitable databases for medical images and strengthening their role in research is one of the grand challenges we see for the years to come.

Acknowledgment

The authors would like to thank CHILI GmbH (Heidelberg, Germany) for the organization of the MICCAI 2007 Grand Challenge workshop, which laid the foundation for this paper.

*T. Heimann is with the Division of Medical and Biological Informatics, German Cancer Research Center, 69121 Heidelberg, Germany (e-mail: t.heimann@dkfz.de).

H.-P. Meinzer, A.-M. Rau, and I. Wolf are with the Division of Medical and Biological Informatics, German Cancer Research Center, 69121 Heidelberg, Germany.

B. van Ginneken, Y. Arzhaeva, and E. M. van Rikxoort are with the Image Sciences Institute, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands (e-mail: bram@isi.uu.nl).

M. A. Styner is with the Department of Psychiatry and Computer Science, University of North Carolina, Chapel Hill, NC 27514 USA (e-mail: martin_styner@ieee.org).

V. Aurich and A. Beck are with the Institute for Computer Science, Heinrich-Heine University Düsseldorf, 40225 Düsseldorf, Germany.

C. Bauer, H. Bischof, and A. Bornik are with the Institute for Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austria.

R. Beichel is with the Department of Electrical and Computer Engineering and the Department of Internal Medicine, The University of Iowa, Iowa City, IA 52242 USA.

E. Sorantin is with the Department of Radiology, Medical University Graz, A-8036 Graz, Austria.

C. Becker is with the Department of Clinical Radiology, University Hospital of Munich, 81377 Munich, Germany.

A. Córdova is with the Department of Oncology, Clínica Alemana de Santiago, Santiago, Chile.

L. Grenacher is with the Department of Diagnostic Radiology, University Hospital of Heidelberg, 69118 Heidelberg, Germany.

G. Bekes, M. Fidrich, G. Németh, and L. Ruskó are with GE Hungary ZRT, Healthcare Division, 6720 Szeged, Hungary.

F. Bello is with the Department of Biosurgery and Surgical Technology, Imperial College London, SW7 2AZ London, U.K.

P. M. M. Cashman, Y. Chi, and R. I. Kitney are with the Department of Bioengineering, Imperial College London, SW7 2AZ London, U.K.

G. Binnig and G. Schmidt are with Definiens AG Research, 80339 Munich, Germany.

B. M. Dawant and R. Li are with the Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37235 USA.

[13]See http://www.sliver07.org.

B. Lennon, S. Li, and J. M. Waite are with Pathfinder Therapeutics, Inc., Nashville, TN 37203 USA.

J. D. Furst, D. S. Raicu, and R. Susomboon are with Intelligent Multimedia Processing Laboratory, School of Computing, College of Computing and Digital Media, DePaul University, Chicago, IL 60604 USA.

D. Furukawa, H. Kobatake, and A. Shimizu are with Tokyo University of Agriculture and Technology, Japan.

J. Hornegger and A. Wimmer are with the Chair of Pattern Recognition, Department of Computer Science, Friedrich-Alexander University Erlangen-Nuremberg, 91058 Erlangen, Germany.

G. Soza is with Computed Tomography, Healthcare Sector, Siemens AG, Forchheim, Germany.

D. Kainmüller and H. Lamecker are with Zuse Institute Berlin, 14195 Berlin, Germany.

T. Lange is with the Department of Surgery and Surgical Oncology, Charité-Universitätsmedizin 10117 Berlin, Germany.

J. Lee is with the Department of Digital Media, The Catholic University of Korea, Korea.

M. Rousson and K. A. Saddi were with the Department of Imaging and Visualization, Siemens Corporate Research, Princeton, NJ 08540 USA.

D. Seghers and P. Slagmolen are with Medical Image Computing (ESAT/PSI), Faculties of Medicine and Engineering, University Hospital Gasthuisberg, 3000 Leuven, Belgium.

REFERENCES

[1] K. Price, "Anything you can do, I can do better (no you can't) . . .," *Comput. Vis. Graph. Image Process.*, vol. 36, no. 2–3, pp. 387–391, 1986.

[2] S. G. Armato, G. McLennan, M. F. McNitt-Gray, C. R. Meyer, D. Yankelevitz, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, A. P. Reeves, B. Y. Croft, and L. P. Clarke, Lung Image Database Consortium Research Group, "Lung image database consortium: Developing a resource for the Med. Imag. research community," *Radiology*, vol. 232, no. 3, pp. 739–748, 2004.

[3] A. Shimizu and S. Nawano, "Preliminary report of competition for liver region extraction algorithms from three-dimensional CT images," in *Proc. CARS*, 2004, p. 1361 [Online]. Available: http://www.tuat.ac.jp/~simizlab/CADM/cadm_index.html

[4] A. Shimizu, T. Kawamura, Y. Mekada, Y. Hayashi, D. Deguchi, and S. Nawano, "Preliminary report of CAD system competition for liver cancer extraction from 3-D CT images and fusion of the CADs," *Int. J.CARS*, vol. 1, pp. 525–526, 2006.

[5] H.-P. Meinzer, M. Thorn, and C. E. Cardenas, "Computerized planning of liver surgery-an overview," *Comput. Graphics*, vol. 26, no. 4, pp. 569–576, 2002.

[6] A. Radtke, S. Nadalin, G. C. Sotiropoulos, E. P. Molmenti, T. Schroeder, C. Valentin-Gamazo, H. Lang, M. Bockhorn, H. O. Peitgen, C. E. Broelsch, and M. Malagó, "Computer-assisted operative planning in adult living donor liver transplantation: A new way to resolve the dilemma of the middle hepatic vein," *World J. Surg.*, vol. 31, no. 1, pp. 175–185, 2007.

[7] E.-L. Chen, P.-C. Chung, C.-L. Chen, H.-M. Tsai, and C.-I. Chang, "An automatic diagnostic system for CT liver image classification," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 6, pp. 783–794, Jun. 1998.

[8] M. Gletsos, S. G. Mougiakakou, G. K. Matsopoulos, K. S. Nikita, A. Nikita, and D. Kelekis, "A computer-aided diagnostic system to characterize CT focal liver lesions: Design and optimization of a neural network classifier," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 3, pp. 153–162, Sep. 2003.

[9] T. Heimann, M. Thorn, T. Kunert, and H.-P. Meinzer, "New methods for leak detection and contour correction in seeded region growing segmentation," in *Proc. 20th ISPRS Congress*, Istanbul, Turkey, 2004, vol. XXXV, Int. Arch. Photogrammetry Remote Sens., pp. 317–322.

[10] L. Soler, H. Delingette, G. Malandain, J. Montagnat, N. Ayache, C. Koehl, O. Dourthe, B. Malassagne, M. Smith, D. Mutter, and J. Marescaux, "Fully automatic anatomical, pathological, and functional segmentation from CT scans for hepatic surgery," *Comput. Aided Surg.*, vol. 6, no. 3, pp. 131–142, 2001.

[11] H. Park, P. Bland, and C. Meyer, "Construction of an abdominal probabilistic atlas and its application in segmentation," *IEEE Trans. Med. Imag.*, vol. 22, no. 4, pp. 483–492, Apr. 2003.

[12] H. Lamecker, T. Lange, and M. Seebaß, Segmentation of the liver using a 3-D statistical shape model Zuse Inst., Berlin, Germany, 2004.

[13] T. Heimann, I. Wolf, and H.-P. Meinzer, "Active shape models for a fully automatic 3-D segmentation of the liver," in *Proc. MICCAI*, 2006, vol. 4191, LNCS, pp. 41–48, Springer.

[14] E. Chaney, G. Ibbott, and W. R. Hendee, "Methods for image segmentation should be standardized and calibrated," *Med. Phys.*, vol. 32, no. 12, pp. 3507–3510, 2005.

[15] M. Styner, H. C. Charles, J. Park, J. Lieberman, and G. Gerig, "Multi-site validation of image analysis methods-assessing intra and inter-site variability," in *Proc. SPIE Med. Imag.*, 2002, vol. 4684, pp. 278–286.

[16] L. P. Clarke, R. P. Velthuizen, M. A. Camacho, J. J. Heine, M. Vaidyanathan, L. O. Hall, R. W. Thatcher, and M. L. Silbiger, "MRI segmentation: Methods and applications," *Magn. Reson. Imag.*, vol. 13, no. 3, pp. 343–368, 1995.

[17] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. J. Gee, and G. Gerig, "User-guided 3-D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *NeuroImage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[18] W. J. Niessen, C. J. Bouma, K. L. Vincken, and M. A. Viergever, "Error metrics for quantitative evaluation of medical image segmentation," in *Performance Characterizat. Comput. Vis.*. Norwell, MA: Kluwer, 2000, pp. 275–284.

[19] G. Gerig, M. Jomier, and M. Chakos, "Valmet: A new validation tool for assessing and improving 3-D object segmentation," in *Proc. MICCAI*. New York: Springer, 2001, 2208, Lecture Notes Comput. Sci., pp. 516–523.

[20] S. Bouix, M. Martin-Fernandez, L. Ungar, M. Nakamura, M.-S. Koo, R. W. McCarley, and M. E. Shenton, "On evaluating brain tissue classifiers without a ground truth," *NeuroImage*, vol. 36, no. 4, pp. 1207–1224, 2007.

[21] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, Jul. 2004.

[22] T. Heimann, M. Styner, and B. van Ginneken, Eds., in *Proc. MICCAI Workshop on 3-D Segmentation in the Clinic: A Grand Challenge*, 2007 [Online]. Available: http://mbi.dkfz-heidelberg.de/grand-challenge2007/sites/proceed.htm

[23] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognit.*, vol. 29, no. 8, pp. 1335–1346, Aug. 1996.

[24] T. T. Tanimoto, An elementary mathematical theory of classification and prediction IBM Res., 1958, Tech. Rep..

[25] P. Jaccard, "Etude comparative de la distribution florale dans une portion des alpes et des jura," *Bull. Soc. Vaudoise Sci. Nat.*, vol. 37, pp. 547–579, 1901.

[26] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Wu, "An optimal algorithm for approximate nearest neighbor searching," *J. ACM*, vol. 45, no. 6, pp. 891–923, 1998.

[27] D. Huttenlocher, D. Klanderman, and A. Rucklige, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, Sep. 1993.

[28] D. Kainmüller, T. Lange, and H. Lamecker, "Shape constrained automatic segmentation of the liver based on a heuristic intensity model," in *Proc. MICCAI Workshop 3-D Segmentation Clinic: A Grand Challenge*, 2007, pp. 109–116.

[29] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam, "The use of active shape models for locating structures in medical images," *Image Vis. Comput.*, vol. 12, no. 6, pp. 355–365, 1994.

[30] M. S. Floater and K. Hormann, "Surface parameterization: A tutorial and survey," in *Advances in Multiresolution for Geometric Modelling*. New York: Springer, 2005, pp. 157–186.

[31] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, 2nd ed. Hoboken, N.J.: Wiley-Interscience, 2008.

[32] J. Weickert, B. M. T. H. Romeny, and M. A. Viergever, "Efficient and reliable schemes for nonlinear diffusion filtering," *IEEE Trans. Image Process.*, vol. 7, no. 3, pp. 398–410, Mar. 1998.

[33] M. Sonka, V. Hlavac, and R. Boyle, "Mathematical morphology," in *Image Processing, Analysis, and Machine Vision*. New York: Thomson, 2007.

[34] T. Heimann, H.-P. Meinzer, and I. Wolf, "A statistical deformable model for the segmentation of liver CT volumes," in *Proc. MICCAI Workshop 3-D Segmentation in the Clinic: A Grand Challenge*, 2007, pp. 161–166.

[35] T. Heimann, I. Wolf, and H.-P. Meinzer, "Automatic generation of 3-D statistical shape models with optimal landmark distributions," *Methods Inf. Med.*, vol. 46, no. 3, pp. 275–281, 2007.

[36] J. Kittler and F. M. Alkoot, "Moderating k-NN classifiers," *Pattern Anal. Appl.*, vol. 5, no. 3, pp. 326–332, 2002.

[37] L. J. Fogel, A. J. Owens, and M. J. Walsh*, Artificial Intelligence Through Simulated Evolution*. New York: Wiley, 1966.

[38] H.-P. Schwefel*, Evolution and Optimum Seeking*. New York: Wiley, 1995.

[39] T. Heimann, S. Münzing, H.-P. Meinzer, and I. Wolf, "A shape-guided deformable model with evolutionary algorithm initialization for 3-D soft tissue segmentation," in *Proc. IPMI*. New York: Springer, 2007, vol. 4584, Lecture Notes Comput. Sci., pp. 1–12.

[40] K. Li, S. Millington, X. Wu, D. Z. Chen, and M. Sonka, "Simultaneous segmentation of multiple closed surfaces using optimal graph searching," in *Proc. IPMI*. New York: Springer, 2005, Lecture Notes Comput. Sci., pp. 406–417.

[41] K. A. Saddi, M. Rousson, C. Chefd'hotel, and F. Cheriet, "Global-to-local shape matching for liver segmentation in CT imaging," in *Proc. MICCAI Workshop 3-D Segmentation Clinic: A Grand Challenge*, 2007, pp. 207–214.

[42] A. Tsai, J. Yezzi, A. , W. Wells, C. Tempany, D. Tucker, A. Fan, W. Grimson, and A. Willsky, "A shape-based approach to the segmentation of medical imagery using level sets," *IEEE Trans. Med. Imag.*, vol. 22, no. 2, pp. 137–154, Feb. 2003.

[43] M. Rousson and D. Cremers, "Efficient kernel density estimation of shape and intensity priors for level set segmentation," in *Proc. MICCAI*. New York: Springer, 2005, vol. 3750, Lecture Notes Comput. Sci., pp. 757–764.

[44] G. E. Christensen, R. D. Rabbitt, and M. I. Miller, "Deformable templates using large deformation kinematics," *IEEE Trans. Image Process.*, vol. 5, no. 10, pp. 1435–1447, Oct. 1996.

[45] G. Schmidt, M. A. Athelogou, R. Schönmeyer, R. Korn, and G. Binnig, "Cognition network technology for a fully automated 3-D segmentation of liver," in *Proc. MICCAI Workshop 3-D Segmentat. Clinic: A Grand Challenge*, 2007, pp. 125–133.

[46] Y. Chi, P. M. M. Cashman, F. Bello, and R. I. Kitney, "A discussion on the evaluation of a new automatic liver volume segmentation method for specified CT image datasets," in *Proc. MICCAI Workshop 3-D Segmentat. Clinic: A Grand Challenge*, 2007, pp. 167–175.

[47] , S. Osher and N. Paragios, Eds.*, Geometric Level Set Methods in Imaging, Vision, and Graphics*. New York: Springer, 2003.

[48] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Stat. Probabil.*, 1967, pp. 281–297.

[49] L. Ruskó, G. Bekes, G. Németh, and M. Fidrich, "Fully automatic liver segmentation for contrast-enhanced CT images," in *Proc. MICCAI Workshop 3-D Segmentat. Clinic: A Grand Challenge*, 2007, pp. 143–150.

[50] R. Pohle and K. D. Toennies, "Segmentation of medical images using adaptive region growing," in *Proc. SPIE Med. Imag.: Image Process.*, 2001, vol. 4322, pp. 1337–1346.

[51] D. Seghers, P. Slagmolen, Y. Lambelin, J. Hermans, D. Loeckx, F. Maes, and P. Suetens, "Landmark based liver segmentation using local shape and local intensity models," in *Proc. MICCAI Workshop 3-D Segmentat. Clinic: A Grand Challenge*, 2007, pp. 135–142.

[52] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.

[53] D. Furukawa, A. Shimizu, and H. Kobatake, "Automatic liver segmentation based on maximum a posterior probability estimation and level set method," in *Proc. MICCAI Workshop 3-D Segmentat. Clinic: A Grand Challenge*, 2007, pp. 117–124.

[54] A. Shimizu, R. Ohno, T. Ikegami, H. Kobatake, S. Nawano, and D. Smutek, "Segmentation of multiple organs in non-contrast 3-D abdominal CT images," *Int. J.CARS*, vol. 2, no. 3–4, pp. 135–142, 2007.

[55] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.

[56] E. van Rikxoort, Y. Arzhaeva, and B. van Ginneken, "Automatic segmentation of the liver in computed tomography scans with voxel classification and atlas matching," in *Proc. MICCAI Workshop 3-D Segmentat. Clinic: A Grand Challenge*, 2007, pp. 101–108.

[57] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[58] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999.

[59] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellen, and W. Eubank, "PET-CT image registration in the chest using free-form deformations," *IEEE Trans. Med. Imag.*, vol. 22, no. 1, pp. 120–128, Jan. 2003.

[60] T. Rohlfing, R. Brandt, R. Menzel, D. B. Russakoff, and C. R. Maurer, "Quo vadis, atlas-based segmentation?," in *The Handbook of Medical Image Analysis-Volume III: Registration Models*. Norwell, MA: Kluwer Academic, 2005, pp. 435–486.

[61] R. Susomboon, D. S. Raicu, and J. Furst, "A hybrid approach for liver segmentation," in *Proc. MICCAI Workshop on 3-D Segmentation in the Clinic: A Grand Challenge*, 2007, pp. 151–160.

[62] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man Cybern.*, vol. 3, no. 6, pp. 610–621, Nov. 1973.

[63] R. O. Duda, P. E. Hart, and D. G. Stork*, Pattern Classification*, 2nd ed. New York: Wiley Interscience, 2000.

[64] R. Beichel, C. Bauer, A. Bornik, E. Sorantin, and H. Bischof, "Liver segmentation in CT data: A segmentation refinement approach," in *Proc. MICCAI Workshop 3-D Segmentat. Clinic: A Grand Challenge*, 2007, pp. 235–245.

[65] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient n-d image segmentation," *Int. J.Comput. Vis.*, vol. 70, no. 2, pp. 109–131, 2006.

[66] A. Bornik, R. Beichel, and D. Schmalstieg, "Interactive editing of segmented volumetric datasets in a hybrid 2-D/3-D virtual environment," in *Proc. ACM Symp. Virtual Reality Software Technol.*, 2006, pp. 197–2006.

[67] A. Bornik, R. Beichel, E. Kruijiff, B. Reitinger, and D. Schmalstieg, "A hybrid user interface for manipulation of volumetric medical data," in *Proc. IEEE Symp 3-D User Interfaces*, Mar. 2006, pp. 29–36.

[68] A. Beck and V. Aurich, "HepaTux-a semiautomatic liver segmentation system," in *Proc. MICCAI Workshop on 3-D Segmentat. Clinic: A Grand Challenge*, 2007, pp. 225–234.

[69] B. M. Dawant, R. Li, B. Lennon, and S. Li, "Semi-automatic segmentation of the liver and its evaluation on the MICCAI 2007 grand challenge data set," in *Proc. MICCAI Workshop on 3-D Segmentat. Clinic: A Grand Challenge*, 2007, pp. 215–221.

[70] S. Pan and B. M. Dawant, "Automatic 3-D segmentation of the liver from abdominal CT images: A level-set approach," in *Proc. SPIE Med. Imag.: Image Process.*, M. Sonka and K. M. Hanson, Eds., 2001, vol. 4322, pp. 128–138.

[71] J. Lee, N. Kim, H. Lee, J. B. Seo, H. J. Won, Y. M. Shin, and Y. G. Shin, "Efficient liver segmentation exploiting level-set speed images with 2.5D shape propagation," in *Proc. MICCAI Workshop 3-D Segmentat. Clinic: A Gand Challenge*, 2007, pp. 189–196.

[72] R. Malladi, J. Sethian, and B. Vemuri, "Shape modeling with front propagation: A level set approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 2, pp. 158–174, Feb. 1995.

[73] R. T. Whitaker and X. Xue, "Variable-conductance, level-set curvature for image denoising," in *Proc. Int. Conf. Image Process.*, 2001, vol. 3, pp. 142–145.

[74] S. G. Armato, M. L. Giger, C. J. Moran, H. MacMahon, and K. Doi, "Automated detection of pulmonary nodules in helical computed tomography images of the thorax," in *Proc. SPIE Med. Imag.: Image Processing*, 1998, vol. 3338, pp. 916–919.

[75] A. Wimmer, G. Soza, and J. Hornegger, "Two-stage semi-automatic organ segmentation framework using radial basis functions and level sets," in *Proc. MICCAI Workshop on 3-D Segmentation in the Clinic: A Grand Challenge*, 2007, pp. 179–188.

[76] J. A. Sethian*, Level Set Methods and Fast Marching Methods*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[77] J. C. Carr, R. K. Beatson, J. B. Cherrie, T. J. Mitchell, W. R. Fright, B. C. McCallum, and T. R. Evans, "Reconstruction and representation of 3-D objects with radial basis functions," in *Proc. SIGGRAPH*, 2001, pp. 67–76.

[78] P. Slagmolen, A. Elen, D. Seghers, D. Loeckx, F. Maes, and K. Haustermans, "Atlas based liver segmentation using nonrigid registration with a B-spline transformation model," in *Proc. MICCAI Workshop 3-D Segmentati. Clinic: A Grand Challenge*, 2007, pp. 197–206.

[79] F. Maes, A. Collignon, D. Vandermeulen, and G. M. P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, Apr. 1997.