Case Study: An Evaluation of User-Assisted Hierarchical Watershed Segmentation

Joshua E. Cates, Ross T. Whitaker, Greg M. Jones

UUCS-04-006

School of Computing
University of Utah
Salt Lake City, UT 84112 USA

February 27, 2004

Abstract

While level sets have demonstrated a great potential for 3D medical image segmentation, their usefulness has been limited by two problems. First, 3D level sets are relatively slow to compute. Second, their formulation usually entails several free parameters which can be very difficult to correctly tune for specific applications. The second problem is compounded by the first. This paper describes a new tool for 3D segmentation that addresses these problems by computing level-set surface models at interactive rates. This tool employs two important, novel technologies. First is the mapping of a 3D level-set solver onto a commodity graphics card (GPU). This mapping relies on a novel mechanism for GPU memory management. The interactive rates level-set PDE solver give the user immediate feedback on the parameter settings, and thus users can tune free parameters and control the shape of the model in real time. The second technology is the use of region-based speed functions, which allow a user to quickly and intuitively specify the behavior of the deformable model. We have found that the combination of these interactive tools enables users to produce good, reliable segmentations. To support this observation, this paper presents qualitative results from several different datasets as well as a quantitative evaluation from a study of brain tumor segmentations.

Case Study: An Evaluation of User-Assisted Hierarchical Watershed Segmentation

Joshua E. Cates, Ross T. Whitaker, Greg M. Jones

Scientific Computing and Imaging Institute
University of Utah
50 S. Central Campus Drive Rm. 3490
Salt Lake City, UT 84112
{cates, whitaker, gjones}@sci.utah.edu

Abstract. This paper evaluates the effectiveness of an interactive, three-dimensional image segmentation technique that relies on watersheds. This paper presents two user-based case studies, which include two different groups of domain experts. Subjects manipulate a graphics-based front end to a hierarchy of segmented regions generated from a watershed segmentation algorithm, which is implemented in the Insight Toolkit. In the first study, medical students segment several different anatomical structures from the Visible Human Female head and neck color cryosection data. In the second study, radiologists use the interactive tool to produce models of brain tumors from MRI data. This paper presents a quantitative and qualitative comparison against hand contouring and another semi-automatic technique based on deformable models. To quantify accuracy, we estimate ground truth from the hand-contouring data using the Simultaneous Truth and Performance Estimation algorithm. We also apply metrics from the literature to estimate precision and efficiency.

The watershed segmentation technique showed improved subject interaction times and increased inter-subject precision over hand contouring, with quality that is visually and statistically comparable. The watershed results also compare favorably to results using the deformable models. The analysis also identifies some failures in the watershed technique, where edges were poorly defined in the data, and noted a trend in the hand-contouring results toward systematically larger segmentations, which raises questions about the wisdom of using expert segmentations to define ground truth.

1 Introduction

Image segmentation is arguably the most ubiquitous and difficult technical problem in medical image processing. The problem of *partitioning an image into meaningful pieces* or, alternatively, *delineating regions of anatomical interest* has proven to be as difficult as a host of other computational problems that attempt to mimic the capabilities of human intelligence or perception. The ongoing difficultly of image segmentation is not from a lack of attention; thousands of papers and theses on image segmentation describe a wide variety of approaches ranging from statistics, differential geometry, and partial differential equations to game theory, discrete geometry, and computational mechanics.

As a result, engineers designing clinical systems that require an image segmentation capability are left with wide range of possible approaches—virtually all of which claim to be effective *to some degree*. To help with this, several researchers have proposed mechanisms for evaluating or validating the effectiveness of various segmentation algorithms. However, quantifying the validity of a segmentation has proven to be almost as difficult as the segmentation itself. The challenge stems from the fact that quantifying differences in shapes is also an important, open problem in computer vision and image processing. In the midst of these difficult challenges and proposed solutions one might ask, "How well are we doing?"

This paper is a case study that looks at the effectiveness of a relatively simple, well-known segmentation paradigm, hierarchical watersheds with user interaction. In our experience, this method is moderately effective on a wide range of segmentation problems. We systematically study its effectiveness on two different types of data using several commonly cited validation metrics. Our study is designed to address the question of whether or not one should use a user-assisted watershed segmentation in lieu of hand contouring. The results, however, also provide some insight into the watershed algorithm itself as well as the methodology of validating segmentation algorithms against a user-define *ground truth*.

The remainder of the paper is organized as follows. Section 2 gives some technical background and related work on watershed segmentation, describes our particular implementation, and presents our validation methodology. Section 3 describes the user studies we have conducted and the method by which we collected our data. Section 4 presents the results of our study and gives qualitative observations about these results. Section 5 summarizes what we have learned and discusses in a broader way the implications of this study.

2 Technical Background and Related Work

2.1 Morphological Watersheds

The subject of image segmentation is too broad for an extensive review, but we give a brief overview of watershed segmentation and how it relates to segmentation methods in general. Most segmentation algorithms fall in one or more of three classes: edge-based approaches, classification-based approaches, and region-based approaches. Edge-based approaches segment images by finding the boundaries between regions, often called edges. Classification-based approaches assign pixels to classes based on a set of measures or features at each pixel. Region-based approaches delineate segments based on the similarity of pixels within the segment. Of course, most approaches use some combination these strategies. For instance, Markov-random fields [1] can be used in such a way that they classify pixels in a

statistical manner, while respecting homogeneous neighborhoods and incorporating an explicit edge model. Likewise region-based approaches often rely on a the concept of an edge. Such is the case with the watershed segmentation algorithm, which grew out of mathematical morphology some 20 years ago [2, 3], and takes its inspiration from hydrology and the study of watersheds. The following paragraphs distill the rather large body of research in morphological watersheds to its essential ideas.

Hydrological watersheds partition geographical landscapes based on ridges, or *watershed lines*, and the valleys between them. Water precipitating onto the landscape naturally collects to form pools in low-lying *catchment basins*, where its flow is blocked by dams or ridges. If we treat the values of an image as a relief map describing height in a geographical terrain then an image can be segmented by partitioning it into areas that correspond to catchment basins in the geographical watershed. The *watershed transformation* of an image is a mapping from the original image to a labeled image such that all points in a given catchment basin have the same unique label. Usually the watershed transformation is applied to a *boundary map*, which is a gray scale function, derived from the input image, that has low values within regions and high values along region boundaries. The gradient magnitude of an intensity-based image, for example, is often used as the boundary map, as well as higher order features such as isophote curvature.

There are many different algorithms for computing the watershed transform, but most of them fall into two basic classifications. The first class of algorithms associate pixels with catchment basins according to their shortest *topological distance* from local minima. Thus, a basin in the watershed transform is a set of all points whose paths of steepest descent terminate at the same local minimum [4]. The second approach floods the image from the bottom up, as if the metaphorical landscape were punctured at its local minima and then immersed in water. This is the strategy of the *immersion algorithm* [5]. The immersion algorithm imposes a discrete set of graylevel values on the image and then expands each catchment basin from its minimum graylevel by iteratively adding the closest connected-component regions of the next highest graylevel. Any pixels that are equidistant from two catchment basin are labeled as watershed lines.

The watershed transformation partitions images into patches that coincide with low values of the boundary measure, but it tends to oversegment the image because it creates one catchment basin for every local minimum. Oversegmentation is especially pronounced in noisy or highly detailed images. Several strategies exist to deal with the oversegmentation problem. One common approach is to grow catchment basins using the immersion algorithm only from specific markers (seed points or seed regions), instead of from all image minima. The resulting segmentation is constrained so that it contains only one region per marker [6, 7]. A second strategy, *hierarchical watersheds* [8], produces a multiscale set of

watershed transforms. Catchment basin regions in the initial, oversegmented transform are progressively merged according to some measure such as their depth, size, or shape. The result is a hierarchy of increasingly coarser segmentations across a range of saliency levels. Oversegmentation can also be controlled by careful smoothing and thresholding of the background values in the original image, which flattens out shallow catchment basins in uninteresting regions. A good review of the many classes and variations of the watershed transform, as well as common solutions to the oversegmentation problem is given in [3] and [9]. The latter includes recent work on parallelization of transforms.

2.2 Algorithm Validation

The role of segmentation validation is to understand the strengths, limitations, and potential applications of a particular segmentation algorithm. There are two strategies for validation. One strategy is to study segmentation performance in the context of a particular clinical or scientific question [10, 11]. For instance, the effectiveness of the algorithm within a study that monitors the volumes or sizes of tumors. The second approach is to study to validate segmentation in the absence of a specific clinical application by quantifying the general behavior of the algorithm relative to an *ideal* solution. This paper takes the second approach, and uses general shape metrics to compare watershed segmentation results with the defacto gold standard for clinical applications, which is hand contouring one slice at a time (which we will also call *manual* segmentation) by expert observers. Thus, this study, narrowly defined, examines the question of whether or not watershed segmentation is an appropriate replacement for hand contouring.

Segmentation validation is difficult because of the lack of standard metrics and the difficulty of establishing ground truth in clinical data. Our validation methodology is derived from ideas developed by [12], and others [13–15], who emphasize the importance of quantitative evaluation and statistical metrics. This study concerns a user-assisted segmentation technique, which requires a user-based validation to capture variations in the individual decision-making process. Experimental trials across a number of users and images[16, 17] can generate data appropriate for statistical analysis that account for user variability.

A combination of a variety of factors determines the effectiveness of a segmentation. For instance Udupa et. al[13] propose a quantification of performance based on validity of the results (accuracy), reproducibility of the results (precision), and efficiency of the segmentation method (time). Other researchers have studied the sensitivity of the technique to various disruptive factors such as data artifacts, pathology, or individual anatomical variation (robustness) [18].

Accuracy metrics typically rely on a *ground truth* segmentation—segmentations that are somehow close to this ground truth are considered better than those that are not. Studies with digital or physical phantoms provide a ready definition of ground truth. However, for biological or clinical data sets, ground truth is usually unknown. In this case, researchers typically rely on experts to delineate the ground truth by hand [18, 19]. Experts seldom all agree, but a statistical combination (averaging) of several expert segmentations can account for expert variability. Averaging of multiple nonparametric shapes, however, is itself a difficult problem. One technique for combining multiple segmentations is *Simultaneous Truth and Performance Level Estimation* (STAPLE), [20]. This treats segmentation as a pixelwise classification, which leads to an averaging scheme that accounts for systematic biases in the behavior of experts in order to generate a fuzzy ground truth (W) and an accuracy estimate for each expert. The ground truth segmentation characterizations W are volumes of values between zero and one that indicate the probability of each pixel being in the object targeted by the segmentation.

The accuracy of an individual experimental segmentation is usually given through some measure of a region's overlap and its distance from the ground truth. Common distance metrics include the Hausdorff distance [21] and the root mean squared distance between selected boundary points [14, 15]. Often overlap is characterized by a *similarity* measure between experimental and ground truth volumes. One common similarity measure s is the cardinality of the intersection (in pixels or voxels) of positive classifications in two volumes volumes over the union of the positive classifications [16, 22]. Another overlap metric is the *total correct fraction c*, which is simply the percentage of correctly classified pixels in the image volume (negative and positive) [23].

Another strategy for evaluating a single-object segmentation is to view each pixel as an instance of a detection task, which gives rise to metrics for sensitivity and specificity. Sensitivity, or p, is the true positive fraction of the segmentation, the percentage of pixels in an image correctly classified as lying inside the object boundary. Specificity, or q, is the true negative fraction, the percentage of pixels in a segmentation correctly classified as lying outside the object boundary. Because there is an explicit tradeoff between sensitivity and specificity, researchers have proposed using receiver operator characterizations (ROC), which monitor the behavior of this tradeoff for different segmentation algorithms or parameter settings [24, 13].

The precision of a segmentation method is an indicator of how repeatable the results are using that technique. Thought of another way, it is an indicator or the degree of randomness inherent to the method. Precision does not rely on knowing ground truth and can be estimated by applying the similarity measure s within a set of experimental segmentations

[13]. The mean *s* value from these comparisons gives a characterization of the precision of the method.

The efficiency of a segmentation technique is a measure of the time involved in achieving a segmentation. This can include user interaction and compute times. These two characteristics are usually considered individually, because each has a separate cost and will affect the practicability of the method in a way that depends on the specific application.

3 Methodology

We present results from two separate user studies of the user-assisted watershed (WS) segmentation technique. Both studies are conducted on 3D datasets (volumes). Our first user study applies the WS segmentation technique to three different anatomical structures from the the head and neck section of the National Library of Medicine's Visible Human Female (VHF) color cryosection data [25]— the right eyeball, the right lateral rectus muscle, and the optic nerves (including the chiasm).

In the second validation study, we apply the WS method to the problem of brain tumor segmentation. This includes four brain tumors (chosen at random) from the Brigham and Women's Hospital (HBW) *Brain Tumor Segmentation Database*: two meningioma (cases 2-3) and two low grade glioma (cases 6 and 8)[23, 26]. The HBW database is a set of ten 3D 1.5T MRI images of brain tumor patients sampled from a larger clinical database. The remainder of this section describes the methods for processing data and conducting the user studies.

3.1 Image Preprocessing and Watershed Segmentation

Before WS segmentation all datasets are processed to reduce noise and smooth homogeneous regions. The smoothing preprocessing step reduces the time necessary to compute the WS transform/hierarchy and decreases the sensitivity to small-scale features, noise, and texture. The filtering consists of an edge-preserving, PDE-based smoothing technique known as anisotropic diffusion [27, 24, 28], which preserves gradient edge features in the image while smoothing more homogeneous regions. For the color data, we used an extension of anisotropic diffusion to multiple channel images that diffuses RGB components separately, but computes edge strength as a function of all components [29], using the L_2 norm of the Jacobian as described in [30]. Table 1 shows the values for the parameters associated with this preprocessing. The parameter Δt is the time step taken for each iteration

and the conductance is measured as a fraction of the RMS edge value in each image at each iteration.

Data set	Δt	iterations	conductance
Tumor	0.120	8	0.60
Cryosection	0.120	10	1.0

Table 1. Parameters for the anisotropic diffusion filtering.

A boundary estimator is applied to the smoothed data to produce the input boundary map to the WS transform. For the tumor data (grayscale) this boundary map is the gradient magnitude, and for the color data the boundary map is the L_2 norm of the Jacobian, as used in the conductance for the anisotropic diffusion. The boundary maps are then lower-thresholded to eliminate potential, uninteresting shallow catchment basins in relatively homogeneous regions without any significant edges. For each dataset, a value l was chosen and all intensity values in the image $f(x) < l$ were set to l. The value used for l is typically very small, on the order of 0.1% of the maximum edge value.

The WS segmentation is computed on the thresholded boundary maps using the top-down approach. As implemented in the *Insight Toolkit*, this algorithm groups pixels with topologically closest minima using a gradient descent strategy [31, 32], and has the advantage that, unlike the immersion algorithm, it does not limit the precision of a segmentation by imposing a set of discrete graylevels on the image. Consider again an image f as a height function, where the graph of $f(x),\ \ x \in U$ describes a surface and m is the set of all minima in f. The ITK watershed transform consists of a set of catchment basins B, with each B_i containing exactly one point m_i and all other points in U whose paths of steepest descent terminate at m_i.

Figure 1 illustrates a typical sequence of segmentation using the ITK watershed transform. The boundary map produced from the original image is used as the input to the watershed transform. Colored regions in the transformed image correspond to uniquely labeled catchment basins.

The gradient descent labeling algorithm can be implemented efficiently in a single pass through the image, using neighborhood connectivity in the cardinal directions (6 neighbors in 3D). A second pass resolves any flat connected-component regions of uniform intensity values, or *plateaus*. Plateaus are labeled according to the closest topologically connected basin to their steepest edge. The downhill, or gradient direction, of a pixel is usually computed by examining the connected neighbors, and the degree of connectivity (e.g. 4 versus

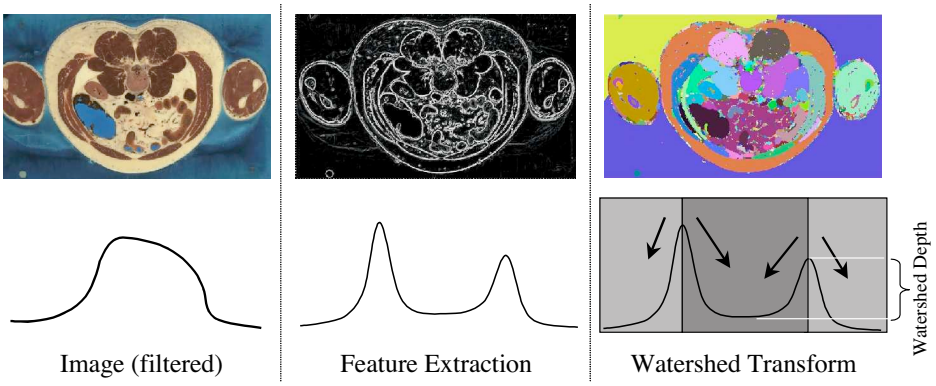


Fig. 1. A typical sequence of segmentation, applied to color cryosection data from the Visible Human Female. An edge map (center) is produced from an image (left) and transformed into a set of segmented regions (right) using gradient descent to local minima.

8 in 2D) is a parameter that depends on the needs of the application. The ITK algorithm uses neighborhoods with connections along cardinal directions (6 neighbors in 3D).

The oversegmentation problem is alleviated using a hierarchical watersheds approach through a sequence of region merges. Merges are determined according to a *saliency measure* $S(B_i)$ for each region or basin B_i. This saliency measure must satisfy two properties. First, it must decrease as regions are merged; that is $S(B_i) \leq S(B_i \cup B_j)$. Second, it must indicate a particular neighboring region with which each region will merge. A succession of merge operations applied in order of increasing saliency produces a hierarchy of watershed transforms in which each transform consists of entirely of regions with saliencies that are greater than some threshold. Alternatively, any choice of saliency threshold (below the maximum saliency of the image) results in a particular segmentation.

The saliency measure for the ITK algorithm is the *watershed depth* of each region, which is defined as the difference in values between the minimum of that region and the lowest saddle point that borders another region. Regions formed by a merge operation have a new depth determined by the minimum of the two merged regions and lowest border value with some other region. The region merging algorithm makes use of a min-heap data structure to store regions according to saliency. The sequence of merges is generated iteratively, by removing the minimum from this heap and creating a new region that combines the minimum region with its neighbor, as indicated by the saliency metric. As the saliency threshold is raised from 0 to the maximum height in the image, region merges can be viewed as nodes in a binary tree, as shown in Fig. 2. A horizontal slice through this tree yields a transform at the associated minimum saliency. To save computation time, the process stops when the number of regions on the list falls below some preset minimum (e.g. several

hundred) that clearly surpasses the needs of the application. For this study we provide users with a user interface into this hierarchy [33], as described in Sect. 3.2.

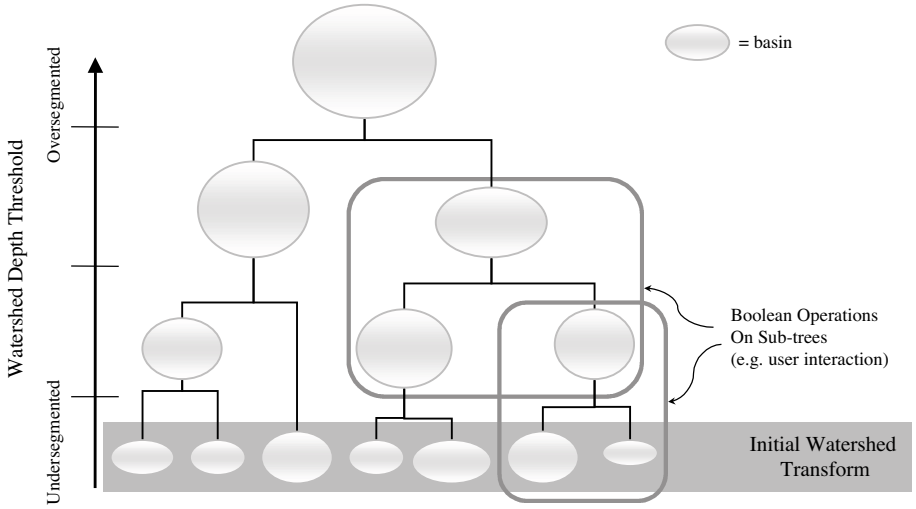


Fig. 2. Enforcing successively higher saliency measures on an initial watershed transform produces a binary tree of region merges and an associated hierarchy of increasingly coarse segmentations.

3.2 User Interface

Using the initial watershed transform and the watershed hierarchy, we can define a semi-automatic approach to segmentation. The *watershed-assisted* methodology allows a user to select any combination of nodes from the watershed hierarchy tree in order to assemble what is similar to a marker-based, region-growing segmentation where markers are selected *after* the watershed transform has been computed [33]. Combining nodes from multiple levels of the tree allows the user to incorporate corresponding multiple levels of detail from the hierarchy. Once the hierarchy is computed, the WS segmentation process is interactive. Effective use of this technique, however, relies on a properly designed user interface.

The interactive graphical user interface (GUI) allows a user to navigate the WS hierarchy from Fig. 2 and select and combine nodes from the tree to produce a segmentation. The segmentation result is stored as a 3D binary mask, with all pixels inside the targeted object boundaries set to 1 and all pixels outside the object boundaries set to 0.

Figure 3 is a screen snapshot of the GUI. The user is presented with slice-by-slice views of the original data (middle window) and an overlay of the segmentation in progress. The segmentation is also shown by itself in a third window, and as a scalable, rotatable 3D isosurface rendering. The left-hand window shows a visualization of the currently selected threshold level in the WS hierarchy, where catchment basins (nodes in the hierarchy) are visualized in contrasting colors. A user selects a catchment basin by clicking on a region in the window and can then add or subtract that region from the current binary mask. A user views different threshold levels in the hierarchy by moving a slider. When the user selects a new depth threshold, the upper right window shows individual segments (using a random coloring scheme) in real-time. Users may also manually correct, select, or deselect voxels in each slice using a two dimensional, circular paint brush.

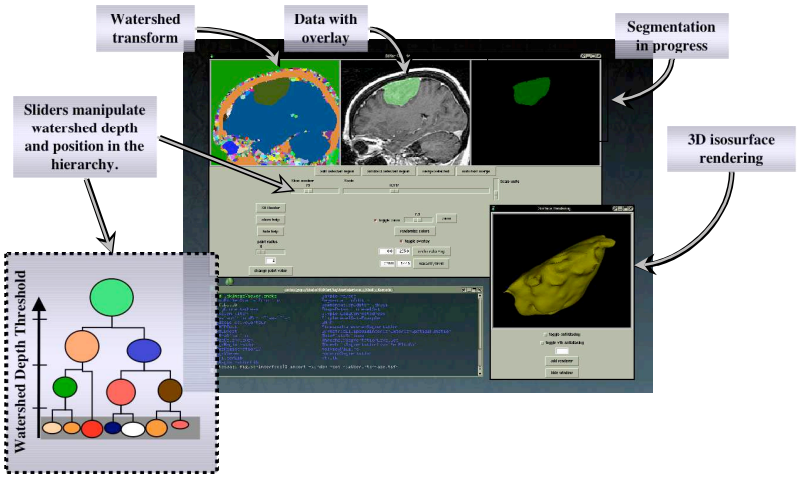


Fig. 3. Screen snapshot of the watershed-assisted GUI. Shown is an interactive segmentation of a brain tumor from MRI data of the head. The GUI has several windows that show, from left to right, the WS transform, the original data, and the segmentation mask. A fourth window displays a 3D rendering of the mask.

In a typical session with the WS GUI, a user selects the coarsest level in the hierarchy that does not appear to erode the boundaries of the target object and selects the appropriate catchment basin(s) to begin the segmentation. The user then selects progressively finer scales in the hierarchy to fill any remaining details. Regions that are too large can be reduced by selecting a basin and subtracting it from the segmentation mask. Once a reasonable segmentation has been generated, the user might touch up areas of the mask manually in several slices.

The WS GUI application is built using the Visualization Toolkit [34] for rendering images and surfaces and the Insight Toolkit for image processing. Filters from these toolkits were

wrapped for the interpreted language, Tcl, to allow the program to be scripted and to interface with Tk for creating widgets and display windows. All components of this software are open source and freely available, including the GUI itself, which can be obtained from [35].

3.3 User Studies

Cryosection Study For establishing the ground truth, we use manual segmentations of the target anatomy from the Surgical Planning Lab (SPL) at the Harvard Brigham and Women's Hospital. The SPL manual segmentations were done by third-year medical students under the supervision of a resident physician. These subjects used the 3D Slicer software from the SPL and the Massachusetts Institute of Technology [36], which provides a slice-by-slice hand-contouring capability. In addition to the SPL segmentations, we gathered additional manual segmentations of eyeball and optic nerve data—conducted by third year medical students at the University of Utah using the same 3D Slicer tool. These additional studies include subject interaction times, which are part of the evaluation of efficiency. In all, a total of 8 medical students contributed hand segmentation data to this study. Each student segmented one or more structures yielding a total of 4 hand segmented eye volumes, 8 lateral rectus volumes, and 3 optic nerve volumes.

For the WS segmentation, the subjects were seven medical students from the University of Utah. All subjects had relevant coursework in the targeted anatomy. The protocol for the segmentation trials was as follows. Each subject was given a brief introduction to the WS GUI and time to practice on data that was similar to that used in the study. They were asked to practice at their own pace until reasonably comfortable with the tool, with an average practice time of about 10 minutes. Following the practice session, subjects were asked to delineate the full, 3D boundaries of the three different anatomical structures in the cryosection data. All experimental trials were timed, but subjects were given no time limit or suggestion of how long to take. Subjects had access to technical help on using the software during the trials. Following the segmentations, subjects were given a brief questionnaire that asked them to rank the difficulty of segmenting each structure and their confidence in the accuracy of their result.

3.4 Tumor Study

Ground truth for the MRI brain tumor study relies on expert hand segmentations available in the *Brain Tumor Segmentation Database* [26]. These consist of four independent *2D*

segmentations of a randomly selected slices for each clinical case. The WS segmentation subjects were three radiologists from the Department of Radiology at the University of Utah Hospital. Each subject was trained to use the software and allowed to practice on several cases from the HBW database that were not included in the study. Total training and practice time combined averaged less than half an hour. Following the practice session, subjects were asked to delineate the full 3D boundaries of the brain tumor in each of the four cases. As with the cryosection study, no time limit was given or suggested and each trial was timed.

3.5 Metrics

This section describes the specific validation metrics and methodology used in this study. We establish ground truth from expert manual segmentations of each data set using the STAPLE algorithm [26]. The STAPLE method combines multiple user segmentations to produce a membership probability function and a sensitivity/specificity for each subject. It employs an iterative EM algorithm that updates pixel-membership probabilities and expert sensitivity/specificity functions asynchronously.

We denote a single subject within a population with the subscript j and the pixels within the image/volume as i. A segmentation for a particular subject consists of an image of binary values D_{ij}. Given sensitivities p_j and specificities q_j for each subject, the degree of confidence that a particular pixel is in the target object is

$$W_i = \frac{g_i \alpha_i}{g_i \alpha_i + (1 - g_i)\beta_i}, \tag{1}$$

where g_i is the prior probability that any pixel would be classified as inside the target object (usually taken to be the fraction of the image that is filled by the object). The values of α and β are

$$\alpha = \left[\Pi_j p_j D_{ij}\right]\left[\Pi_j (1-p_j)(1-D_{ij})\right] \text{ and } \beta = \left[\Pi_j q_j (1-D_{ij})\right]\left[\Pi_j (1-q_j) D_{ij}\right]. \tag{2}$$

Given a probability image W_i, the sensitivity/specificity for each subject can be updated as

$$p_j = \frac{\sum_i W_i D_{ij}}{\sum_i W_i} \text{ and } q_j = \frac{\sum_i (1-W_i)(1-D_{ij})}{\sum_i (1-W_i)}. \tag{3}$$

The full STAPLE algorithm entails iterating on these updates, back and forth between (p,q) and W, until the process converges.

For evaluating accuracy we use the STAPLE algorithm to form, for each segmented object, an aggregate volume that consists of a graded membership function (zero to one). We analyze the accuracy of WS results by evaluating the sensitivity and specificity of each WS subject, using equations 3, relative to these aggregate volumes. We can then make comparisons by computing average sensitivity and specificity for the two groups—subjects using hand contouring and subjects using the WS GUI. Additionally, we can combine values of p_j and q_j to compute a total correct fraction for a subject:

$$c_j = \frac{\sum_i W_i D_{ij} + \sum_i (1 - W_i)(1 - D_{ij})}{\sum_i 1}. \tag{4}$$

Ideally we would analyze the accuracy of manual segmentations using aggregate data from an *independent* group of manual segmenters. A characterization of the accuracy of a small group of manual segmentations using ground truth generated as a complete aggregate of *those same segmentations* contains an obvious bias. To help understand this bias and produce a more complete estimate of manual segmentation accuracy, we make a second, less conservative measurement for comparison using a round-robin *leave-one-out* strategy [37], where p, q, and c values for each D_{ij} are computed using W_k generated by all segmentations $k \neq j$.

We have found that some care must be taken when interpreting accuracy metrics. Where a segmentation technique shows high sensitivity, there is a high confidence level in the results it produces for *negatively* classified pixels. Where a technique shows high specificity, there is a high confidence level for *positively* classified pixels. It is also worthwhile to note that the magnitudes of p and q are incommensurate because they are percentages of different populations of pixels. Total correct fraction is particularly difficult to interpret because it is biased by the ratio of the size of the image volume to the size of the target object. Where this ratio is high, c approaches q. Where the ratio is low, c approaches p. We use total correct fraction here only as a way to compare our results with other published results on the same data. Whenever possible—i.e. when not comparing our results with third-party studies of the same data—we crop experimental and ground truth volumes to a region of interest around the target object before computing accuracy metrics.

For quantifying precision we use the *similarity* s_{jk} of results from subjects j and k,

$$s_{jk} = \frac{2 \sum_i D_{ij} D_{ik}}{\sum_i D_{ij} + D_{ik}}, \tag{5}$$

and average this across all pairs of subjects $j \neq k$. Our accuracy, precision, and efficiency metrics were applied *across* subjects. Given the limited resources for this study and the scarcity of manually segmented data, we were not able to make *intra*-subject comparisons, which require multiple segmentations from the same subject.

4 Results and Discussion

4.1 Cryosection Study

The performance of the subjects in our first study varied case by case, but visual inspection of the experimental VHF segmentations reveal them to be generally of good quality. Figure 4 is a comparison of a typical WS segmentation with a typical manual segmentation for each anatomical structure. Column *a* consists of slices of the original color data with an overlay of the experimental segmentation mask. The surface of the respective experimental and manual masks are rendered in *b* and *c*.

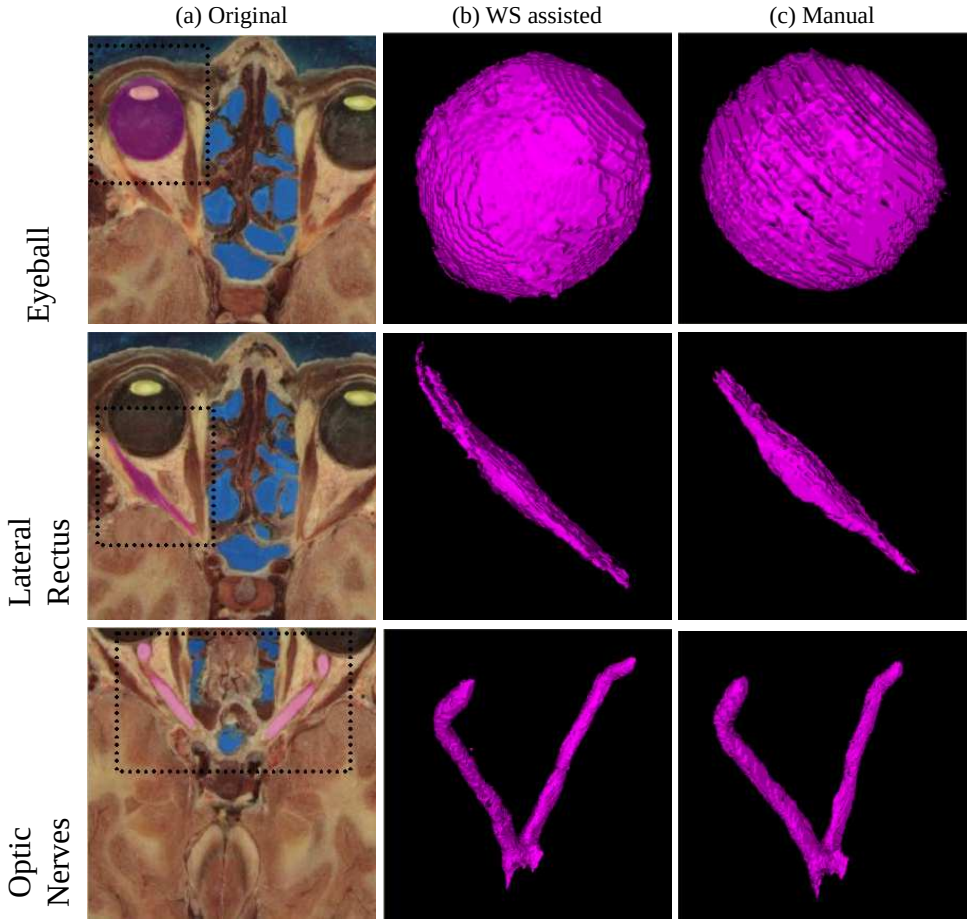


Fig. 4. Visual comparison of typical watershed (b) and manual (c) segmentations of the Visible Human Female color cryosection anatomy. The targeted anatomical structure is highlighted in column (a), which shows the segmentation from (b) superimposed over a transverse slice through the original color data.

Figure 5 shows average questionnaire responses and segmentation times. Interaction times are normalized relative to the highest reported time and normalized on a scale of 0 to 10 (low to high) for comparison with questionnaire responses. Subjects reported high confidence levels in their segmentations across all datasets, with an average rating of $7.4/10.0$. Lower confidence ratings correlate with higher difficulty ratings, which ranged from low to moderate difficulty of segmentation in the lateral rectus and optic nerve, to high difficulty in the eyeball. Subjects reported that our software was easy to use, with the average *ease-of-use* score of $7.5/10.0$.

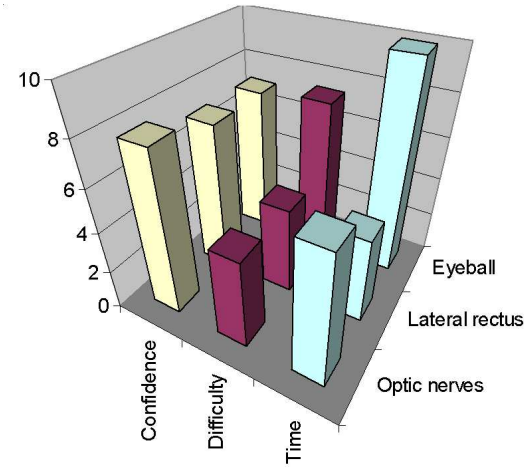


Fig. 5. Summary of questionnaire responses for Visible Human Female segmentations. Values shown are the mean response and normalized to a scale of 0 to 10, low to high. Times are relative to the maximum segmentation time.

The questionnaire results indicate that subjects had the most difficultly segmenting the eyeball. The WS transform failed in several areas of the eyeball where there was not enough color contrast with surrounding tissue for the algorithm to detect an edge. These weak edges were also difficult for manual segmenters to find visually. Figure 6 sheds some light on the problem. The images on the left show slices of the original data. The center and left images show STAPLE aggregate volumes computed for both the manual (W_h) and the WS subjects (W_u) respectively. These images give us an indication of agreement (values close to 1.0) and disagreement (values less than 1.0) among the subjects in each case. The center and right images show the degree of agreement (red where subjects generally agreed and purple where they did not). While WS results were generally in agreement with most manual segmentations, they occasionally leak into the areas of disagreement, which correspond to weak edges in the color data. Furthermore, the manual results exhibit smooth, round segmentations within each slice, which tend to vary slice to slice. This suggests that hand-contouring allowed users to inject some a-priori knowledge about the shape of eyeball. The

WS hierarchy for this particular object, which includes no a-priori shape knowledge, made it difficult for WS users to achieve a round shape for the eyeball.

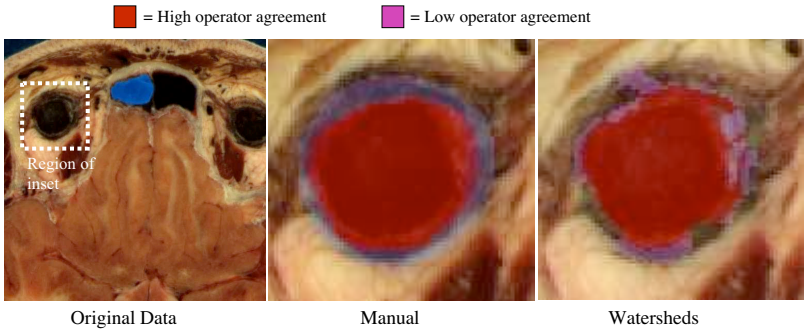


Fig. 6. Pixelwise precision analysis for manual and WS subject segmentations.

Figure 7 shows graphs of average p, q, and c for the subjects in our study versus the manual segmenters. Sensitivity values for WS segmentation are consistently lower, generally falling below the standard deviation of the manual segmenters. Specificity values for the WS method are consistently higher than for manual segmentation. Total correct fraction values for the WS segmentation are generally lower but within the variance of the manual segmentations. WS results compare more favorably with the less biased round-robin (RR) values in all cases.

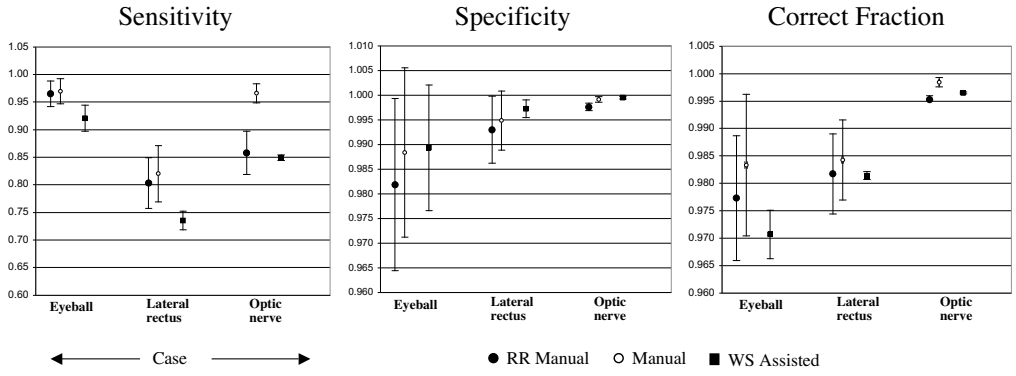
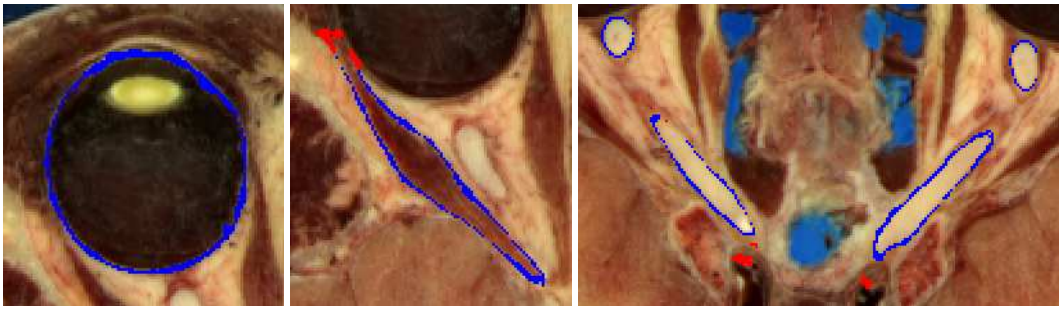


Fig. 7. For the cryosection studies, accuracy measures of sensitivity, specifi city, and correct fraction for WS are generally within the range of results obtained by hand contouring, and compare more favorably to less biased, round-robin (RR) manual values.

To understand why the WS sensitivity values are consistently lower, consider the images in Fig. 8, which shows slices from the difference volumes $W_u - W_h$ in each data set super-

imposed over the original data. Overlay images are transparent where difference values are 0, red where positive (indicating false positive results in the WS segmentations), and blue where negative (indicating false negative results in the WS segmentations). These images are typical of slices throughout the volumes and show a systematic trend in the manual segmentations toward producing larger segmentations. The manual segmentations in this study were on average about 12% larger than WS segmentations, based on total number of positively classified voxels. In many cases the manual segmentations *extend beyond object boundaries indicated by sharp changes in color or contrast*. This trend explains much of the discrepancy between the manual and WS, but it also brings into question the usefulness of these manual segmentations as ground truth.



(a) (b) (c)

Fig. 8. Differences between manual and WS segmentations with read indicated false-positive WS results and blue indicated false-negative WS results: (a) eyeball, (b) lateral rectus, (c) optic nerves.

The WS segmentations in our study show a higher degree of inter-subject precision than the hand segmentations. Mean precision across all WS subjects (and all cases) is $92.67\% \pm 3.75\%$ while the mean precision across manual segmentations is $80.95\% \pm 12.07\%$. This result indicates that independent users are able to produce significantly more consistent segmentations using our tool than when delineating boundaries by hand. For some applications, such as estimating tumor size [19], the consistency of the WS segmentations would be an important advantage.

Average segmentation times of all datasets and subjects using the WS GUI were around 30 minutes, a significant increase in efficiency over the average 2 hour sessions we observed for manual segmentations. Typical preprocessing times, including diffusion and calculation of the WS transform and hierarchy, for VHF color data were 10 to 15 minutes on a Pentium III 1200 Mhz single processor PC.

4.2 Brain Tumor Segmentations

Qualitative analysis of the WS segmentations of the HBW brain tumor volumes show a good correspondence with the visual boundaries of the tumor mass in each case. Slice comparisons with the HBW expert segmentations are also favorable. In general, WS segmentations capture the basic boundaries and size of each tumor. Using the STAPLE technique, we compute fuzzy likelihood volumes W_h and W_u for the expert manual and WS slices, respectively. Figure 9 superimposes W_h and W_u on top of the original data. These overlays are colored red in regions of high subject agreement and purple in areas where subjects disagreed. The difference images $W_u - W_h$ are shown in the last row.

A few distinct misclassifications in the WS results are evident in this figure. We can identify, for example, some leakage at the superior end of the tumor in case 6 and an undersegmented region at the inferior end of the tumor in case 2. The blue, false negative, contour line around each tumor in the difference image again suggests, as in the cryosection study, that manual segmentations tend to produce slightly larger volumes than the semi-automated method. These images also give us an indication of the relative precision of the two techniques; W_h slices clearly show more disagreement than W_u slices over boundary pixels.

Figure 10 shows graphs of average p, q, and c for the users and manual segmenters in this study. Sensitivity, specificity, and total correct volume fraction are typically within standard deviation of manual segmentations. ROC values also compare favorably to those of Lefohn, et al.[17], who use a superset of the same data used in our study and apply a user-assisted level-set based segmentation technique, also with an interactive GUI. WS accuracy values compare more favorably to the RR manual segmentation results, again reflecting the inherent bias in complete-aggregate ground truth volumes. The overall mean precision across users and datasets using our method was $97.73\% \pm 2.57\%$, significantly higher than that of both the level-set users with $92.78\% \pm 3.98$ and the hand segmenters with $84.77\% \pm 5.67\%$.

The accuracy and precision of users in our study also compares well with the automated brain tumor segmentation results of Kaus, et al. [23], who, again, use a superset of the same data used in our study. They report an average correct volume fraction of $99.68 \pm 0.29\%$, while the average total correct volume fraction of our users was 99.76 ± 0.14.

The WS method required an average interaction time of 5-10 minutes, which is similar to times reported by [17] and [23]. The WS method required an additional data preprocessing time of between 5 - 10 minutes, the method in [23] reports processing times of approximately 75 minutes, while that of [17] requires no preprocessing.

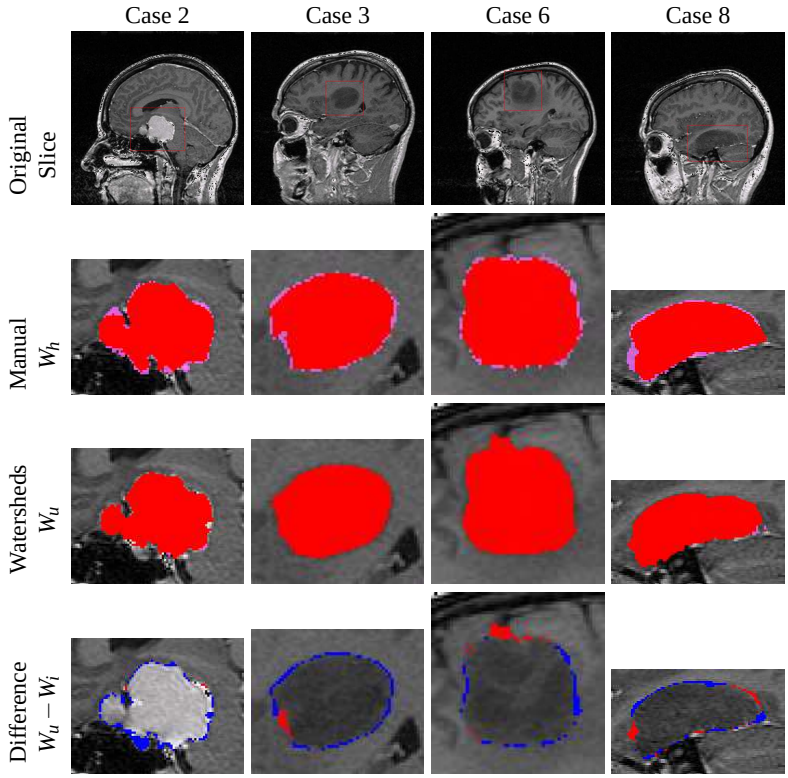


Fig. 9. Comparison of watershed-assisted segmentations with manual segmentations of the HBW brain tumor datasets. The difference images (with red being WS false positives and blue being WS false negatives) show that the WS segmentations were consistently smaller.

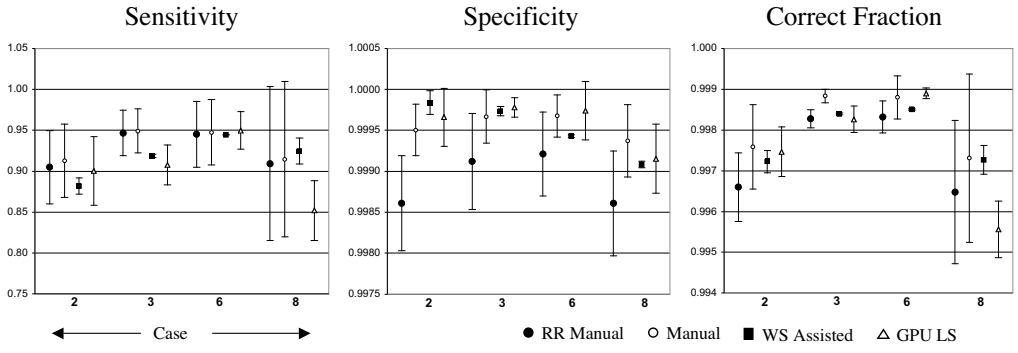


Fig. 10. For the MRI brain tumor study, the accuracy of the WS results are generally within range of segmentations generated by experts using hand contouring, comparing more favorably to the less biased round-robin (RR) manual results.

These quantitative comparisons with experts pertain to only a single two dimensional slice that is extracted from the 3D segmentations. This is a limitation due to the scarcity of expert data. Our experience has been that acquiring good quality expert segmentations is often one of the more difficult and limiting aspects of conducting a user study.

5 Summary and Conclusions

An implementation of a hierarchical morphological WS algorithm offers a powerful tool for interactive 3D segmentation. Users can select and combine catchment basin regions at various levels of detail in the WS hierarchy to produce a model of a targeted object. The quantitative results of using this tool for both color cryosection segmentation and brain tumor segmentation suggest that it compares well with hand contouring and state-of-the-art automated methods. The tool as built and tested is quite general and thus can be used to segment other anatomy. However, we observed that color cryosection data was more difficult for users to segment than the MRI data, suggesting that our tool may be less suitable for delineating complex structures with poorly defined boundaries.

The current limitations are mostly in the preprocessing speed and the interface. Parallel versions of anisotropic diffusion [32] are highly effective in reducing filtering times, and there some promising work on parallelization of WS transforms [9]. An expanded 3D interface that incorporates cutting planes and real-time volume rendering visualization could potentially improve user interaction times and accuracy.

Our experience estimating ground truth from manual segmentations suggests that they may produce a bias toward larger models. Other researchers have reported similar results [16]. Such biases should be taken into consideration when comparing quantitative results from user studies of automated methods.

6 Acknowledgments

Thanks to Peter Ratiu at the HBW Surgical Planning Lab for the hand segmentations of color cryosection data. Thanks to Drs. Steve Stevens, Troy Marlow, and Jay Tsuruda for participating in our study. Also thanks also to Drs. Simon Warfield, Michael Kaus, Ron Kikinis, Peter Black, and Ferenc Jolesz for sharing the online brain tumor database, and to the National Library of Medicine and all other institutes and agencies participating in the Insight project.

References

1. S. Geman, D. Geman, Stochastic relaxation, gibbs distributions and the bayesian restoration of images, IEEE Trans. on Image Processing 6 (6) (1984) 721–741.
2. J. P. Serra, Image Analysis and Mathematical Morphology, Academic Press Inc., 1982.
3. S. Beucher, F. Meyer, The morphological approach to segmentation: the watershed transformation, E.R. Dougherty, 1993, pp. 433–481.
4. F. Meyer, Topographic distance and watershed lines, SP 38 (1) (1994) 113–125.
5. L. Vincent, P. Soille, Watersheds in digital spaces: An efficient algorithm based on immersion simulations, PAMI 13 (6) (1991) 583–598.
6. F. Meyer, S. Beucher, Morphological segmentation, JVCIR 1 (1) (1990) 21–46.
7. R. Lotufo, W. Silva, Minimal set of markers for the watershed transform, in: Proc. ISMM2002.
8. S. Beucher, Watershed, hierarchical segmentation and waterfall algorithm., Kluwer Academic Publishers, 1994, pp. 69–76.
9. J. B. T. M. Roerdink, A. Meijster, The watershed transform: Definitions, algorithms and parallelization strategies, Fundamenta Informaticae 41 (1-2) (2000) 187–228.
   URL citeseer.nj.nec.com/roerdink00watershed.html
10. N. Malpica, C. Solórzano, J. Vaquero, A. Santos, I. Vallcorba, J. García-Sagredo, F. Pozo, Applying watershed algorithms to the segmentation of clustered nuclei, Cytometry 28 (1997) 289–297.
11. J. Sijbers, M. Verhoye, A. Scheunders, A. Van der Linden, D. Van Dyck, E. Raman, Watershed-based segmentation of 3d mr data for volume quantization, Magnetic Resonance Imaging 15 (6) (1997) 679–688.
12. T. S. Yoo, M. J. Ackerman, M. Vannier, Toward a common validation methodology for segmentation and registration algorithms, in: International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI'2000, Lecture Notes in Computer Science, Springer Verlag, Cambridge, England, 2000, pp. 422–431.
13. J. K. Udupa, V. Leblanc, H. Schmidt, C. Imielinska, K. Saha, G. Grevera, Y. Zhuge, P. Molholt, L. Currie, Y. Jin, A Methodology for Evaluating Image Segmentation Algorithms, in: SPIE Medical Imaging, San Diego, 2002.
14. V. Chalana, K. Yongmin, A methodology for evaluation of boundary detection algorithms on medical images, IEEE Trans. Medical Imaging 16 (5) (1997) 642–652.
15. G. Gerig, M. Jomier, M. Chakos, Valmet: A new validation tool for assessing and improving 3d object segmentation., in: MICCAI 2001: Fourth International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer-Verlag, Heidelberg, Germany, 2001, pp. 516–528.
16. A. Zijdenbos, B. Dawant, A. Margolin, Morphometric analysis of white matter lesions in mr images: Method and validation, IEEE Transactions on Medical Imaging 13 (4) (1994) 716–724.
17. A. Lefohn, J. Cates, R. Whitaker, Interactive, gpu-based level sets for 3d brain tumor segmentation, in: MICCAI 2003, 2003, p. To appear.
18. P. Jannin, J. Fitzpatrick, D. Hawkes, X. Pennec, R. Shahidi, M. Vannier, Validation of medical image processing in image-guided therapy, IEEE Trans. on Medical Imaging 21 (12) (2002) 1445–1449.
19. M. Prastawa, E. Bullitt, G. Gerig, Robust estimation for brain tumor segmentation, in: MICCAI 2003, Springer-Verlag, Heidelberg, Germany, 2003.
20. S. Warfield, K. Zou, W. Wells, Validation of image segmentation and expert quality with an expectation-maximization algorithm, in: MICCAI 2002: Fifth International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer-Verlag, Heidelberg, Germany, 2002, pp. 298–306.
21. D. Huttenlocher, G. Klanderman, W. Rucklidge, Comparing images using the hausdorff distance, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (1993) 850–863.
22. K. Leemput, F. Maes, D. Vandermeulen, P. Suetens, Automated model-based tissue classification of mr images of the brain, IEEE Transactions on Medical Imaging 18 (1999) 897–908.
23. M. Kaus, S. K. Warfield, A. Nabavi, P. M. Black, F. A. Jolesz, R. Kikinis, Automated segmentation of mri of brain tumors, Radiology 218 (2001) 586–591.
24. R. T. Whitaker, Geometry-limited diffusion, Ph.D. thesis, The University of North Carolina, Chapel Hill, North Carolina 27599-3175 (1993).
25. M. Ackerman, T. Yoo, D. Jenkins, The visible human project: from data to knowledge, in: Proceedings of CARS2000, Elsevier Press, Amsterdam, 2000, pp. 11–16.

26. S. K. Warfield, A. Nabavi, T. Butz, K. Tuncali, S. G. Silverman, Intraoperative segmentation and nonrigid registration for image guided therapy, in: International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI'2000, Lecture Notes in Computer Science, Springer Verlag, Cambridge, England, 2000, pp. 176–185.
27. P. Perona, J. Malik, Scale-space and edge detection using anisotropic diffusion, IEEE Transactions on Pattern Analysis Machine Intelligence 12 (1990) 629–639.
28. B. M. ter Haar Romeny (Ed.), Geometry-Driven Diffusion in Computer Vision, Kluwer Academic Publishers, 1994.
29. R. T. Whitaker, Geometry-limited diffusion in the characterization of geometric patches in images, Computer Vision, Graphics, and Image Processing: Image Understanding 57 (1) (1993) 111–120.
30. G. Sapiro, D. Ringach, Anisotropic diffusion of multivalued images with applications to color filtering, IEEE Trans. on Image Processing 5 (1996) 1582–1586.
31. A. P. Mangan, R. T. Whitaker, Partitioning 3D surface meshes using watershed segmentation, IEEE Transactions on Visualization and Computer Graphics 5 (4) (1999) 308–321.
32. L. Ibanez, W. Schroeder, L. Ng, J. Cates, The ITK Software Guide, Insight Consortium, http://www.itk.org/ItkSoftwareGuide.pdf (2003).
33. T. Yoo, U. Neumann, H. Fuchs, S. Pizer, T. Cullip, J. Rhoades, R. Whitaker, Direct visualization of volume data, IEEE Computer Graphics and Applications 12 (4) (1992) 63–71.
34. W. Schroeder, K. Martin, B. Lorensen, The Visualization Toolkit, An Object Oriented Approach to 3D Graphics, Prentice-Hall, 1998.
35. I. Consortium, The insight segmentation and registration toolkit.
36. MIT Artificial Intelligence Laboratory and the Surgical Planning Laboratory at Brigham and Women's Hospital, http://www.slicer.org, 3D Slicer Users Guide.
37. J. Tou, R. Gonzalez, Pattern Recognition Principles, Addison-Wesley, 1974.