Evaluation of Interactive Segmentation Algorithms Using Densely Sampled Correct Interactions

S.M. Rafizul Haque, Mark G. Eramian, and Kevin A. Schneider

University of Saskatchewan, Saskatoon, SK, S7N 5C9, Canada
rafizul.cs@usask.ca,
{eramian,kas}@cs.usask.ca

Abstract. The accuracy and reproducibility of semiautomatic interactive segmentation algorithms are typically evaluated using only a small number of human observers which only considers a very small number of the possible *correct interactions* that an observer might provide. A *correct interaction* is one that provides contextual information that would be expected to result in a correct segmentation. In this paper, we demonstrate new evaluation methods for semiautomatic interactive segmentation algorithms that employ simulated observer models constructed from a large number of segmentations computed by uniformly sampling the entire set of possible correct interactions. The advantages of this method are that it is free of observer biases and the large number of segmentations produced for each object of interest to be segmented allow a range of statistical methods to be brought to bear on the analysis of segmentation algorithm performance. The methods are demonstrated using a semi-automated segmentation algorithm for ovarian follicles in ultrasonographic images.

Keywords: interactive segmentation, semiautomatic, correct interaction, reproducibility, performance, evaluation.

1 Introduction

Algorithms for segmentation of semantic objects from images are, ideally, desired to be fully automatic due to the tedious and time-consuming nature of manual segmentation. Some segmentation problems, however, are still very difficult to solve with fully automatic methods such as problems where an arbitrary number of objects of interest need to be both detected and segmented or where the imaging modality results in poorly resolved boundaries between neighbouring objects. This motivates the use of semiautomated segmentation algorithms in which a human operator provides high-level contextual information in an interactive phase, which is followed by an automatic phase where the segmentation is performed under the constraints of the operator-provided guidance.

Evaluation of accuracy and/or reproducibility of semiautomatic segmentation algorithms is typically performed by having a small number of experts in the

A. Petrosino (Ed.): ICIAP 2013, Part I, LNCS 8156, pp. 191–200, 2013.

problem domain who are well-trained in the use of the semiautomatic segmentation system segment a number of cases. Indeed, this has been the case with numerous recent studies that analyze intra- and/or inter-observer variability [5–8, 10, 13, 14]. Of these, only the studies of Stammberger et al. [14] and Claudia et al. [8] used more than 5 observers. Even in the simplest of situations, where the interaction is selecting a seed point somewhere within an object, it is not possible to robustly characterize the inherent variability in segmentation accuracy due to variations in seed point placement using only a small number of example interactions. Recently, some authors [11, 12] have turned to constructing simulated *observer models* to take into account more interactions per case, and to avoid observer bias but the number of seeds and brush strokes used were not sufficient to represent a very diverse set of examples of possible correct interactions.

The main deficiency, therefore, of existing evaluation methods is that an insufficiently diverse sampling of the set of correct interactions for each case are used to draw conclusions about overall segmentation accuracy and reproducibility. One must consider a diverse set of correct interactions in order to compare algorithms fairly and take into account the consequences of poor choices resulting from fatigue or lapses in judgement on the part of the operator. To this end we propose the use of observer models where the user interaction is generated programmatically in a similar way as in [11] and [12]. However, in contrast to these previous methods, our observer models are unbiased in the sense that we systematically and uniformly sample all possible correct interactions for each case to be segmented. Herein we consider a simple interaction mode where a user click is made to supply a seed point for each object to be segmented. For each object, all the segmentations generated from the uniformly sampled set of possible seed points are analyzed statistically to assess the impact of the seed point location on the quality of the resulting segmentation using four models. One of these four models is an overall unbiased one as it includes all the sampled seed points and the other three models are biased as each of these models includes only a specific subset of the sampled seed points based on the locations of these seed points. Using this synthetic interaction model, we have enough data to use statistical methods to test for significant differences in segmentation accuracy between observer models (subsets of correct interactions) and to quantify any differences found. As a vehicle for the demonstration of our methods, we consider an ovarian follicle segmentation algorithm based on binary graph cuts with a shape prior and evaluate the reproducibility of the results, and look for statistically significant differences between the aforementioned four observer models.

2 Methods

2.1 Interactions

We define a *correct interaction* to be the contextual information provided by the operator that would be expected to produce a correct segmentation, e.g. a seed point that is inside the object to be segmented, or a set of brush strokes that correctly indicate areas of foreground and background. Herein we demonstrate some new methods of analyzing the variability of segmentation accuracy over a wide range of uniformly sampled correct interactions using interactive graph cuts.

2.2 Binary Graph Cuts for Follicle Segmentation

Graph cut segmentation [3, 4] utilizes a max-flow/min-cut energy minimization algorithm which eventually generates the optimal segmentation with respect to the weights assigned on the edges of the graph. The minimum cut can be found using efficient solutions to the *maximum flow* problem since the set of saturated edges in a maximum flow solution coincide with the edges in the minimum cut. For details on binary graph cuts, the reader is referred to [4]. Instead of using general graph cut segmentation, a generic shape prior called "star shape" [15] was incorporated to enhance the segmentation accuracy. This shape prior is appropriate due to the roughly elliptical but low-eccentricity shape of ovarian follicles in normal ovaries. Within the Voronoi region of each seed point we added new edges to the graph radiating from each user supplied follicle seed point to encode the shape prior in the graph; see [15] for details.

For our follicle segmentation, asymmetric weights between non-terminal nodes were defined similarly to as in [4]. Terminal weights were derived from intensity distribution models (histograms) built from neighbourhoods defined by user-supplied clicks on example foreground and background regions.

Constraints were added based on the seed points given for each follicle. In our experimental methodology, seed points are selected from a set of correct interactions for each follicle (see Section 2.3).

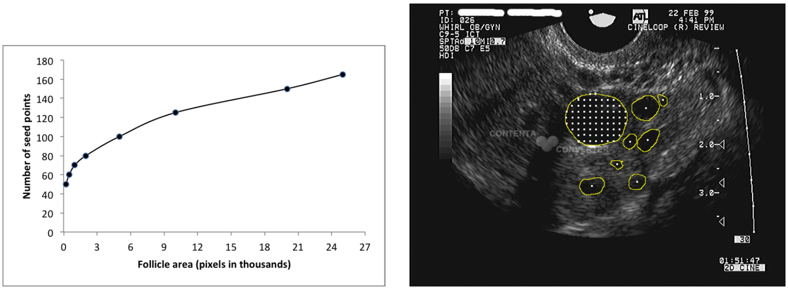
2.3 Experimental Setup

A set of 32 ultrasound *in vivo* human ovarian images, obtained from a previous study [1], were used in this experiment. Size of each image is 640 × 480 pixels and the maximum number of follicles in an image was fourteen. The total number of follicles in all images was 132 among which 53 were very small, having a cross-sectional diameter in the image of less than 2.5mm; this is a significant size threshold because even human observers have difficulty correctly identifying follicles of this size or smaller. These follicles were not considered in our experiment, leaving 79 follicles of diameter greater than 2.5mm. Manually delineated ground truth segmentations of these follicles were provided by a single, highly experienced human operator. For each follicle to be analyzed we constructed a set of correct interactions (seed points). Seed points were sampled on a grid with a spacing of between 2 and 14 pixels depending on follicle size. Seed point locations were sampled more sparsely for larger follicles to maintain computational feasibility. Grid spacing was determined using the following procedure:

1. Determine number of seed points N to be used for the follicle as a function of its area using the piecewise cubic polynomial function in Figure 1(a). Interpolation points were selected empirically based on the data set.

2. Determine the grid spacing as $\lceil\sqrt{A/N}\rceil$ where A is the area (in pixels) of the follicle (determined from the ground truth).

This resulted in sets of correct interactions consisting of between 50 and 375 seed points for each follicle.



(a) (b)

Fig. 1. (a) Piecewise cubic polynomial function used for determining the number of seed points for each follicle. (b)Segmenting a follicle with its set of correct interactions. The large follicle is segmented once with each of the seed points shown while the remaining follicles seed points (at the centroids of their regions) are held constant. This process is repeated for each other follicle in the image.

Each follicle was segmented using each seed point from its set of correct interactions exactly once while the seed points for any other follicles in the image were held constant (Figure 1(b)). These constant seed points were the centroid of the follicle region, determined from the ground truth. Thus each follicle was segmented with each of its correct interactions and the Dice coefficient, HD, and RMSD were determined for each interaction.

For each follicle, sampled seed points were categorized into three groups: central, intermediate and peripheral depending on their locations within the follicle relative to the follicle region's centroid. To determine the category of a seed point, a binary ground truth image was negated and then distance transform of that image was computed. From this transform, distance a of the seed point from the nearest boundary point was determined and distance from the seed point to the centroid b was calculated using the Euclidean distance metric. The seed point category $\mathbf{c}$ was then determined by a double threshold of the quantity $\frac{a}{a+b}$:

$$\mathbf{c} = \begin{cases} \text{peripheral}, & \frac{a}{a+b} \leq 1/3 \\ \text{intermediate}, & 1/3 < \frac{a}{a+b} \geq 2/3 \\ \text{central}, & 2/3 < \frac{a}{a+b} \end{cases} \tag{1}$$

Mean and standard deviation of Dice coefficient, RMSD and HD were computed over each follicle's set of correct interactions and over the central, intermediate and peripheral subsets of interactions for each follicle. Larger values of the Dice coefficient and lower values of RMSD and HD indicate a more accurate segmentation. Coefficients of variation of these segmentation accuracy measures were computed for each follicle and analyzed to evaluate segmentation reproducibility.

3 Analysis of Results

All of the 79 follicles with diameter $>$ 2.5mm, were analyzed. Segmentation accuracy was measured in terms of Dice coefficient [9], root mean squared distance (RMSD) and Hausdorff Distance (HD) [2] of the segmented follicle contour from the manual contour. These measures were determined individually for each follicle.

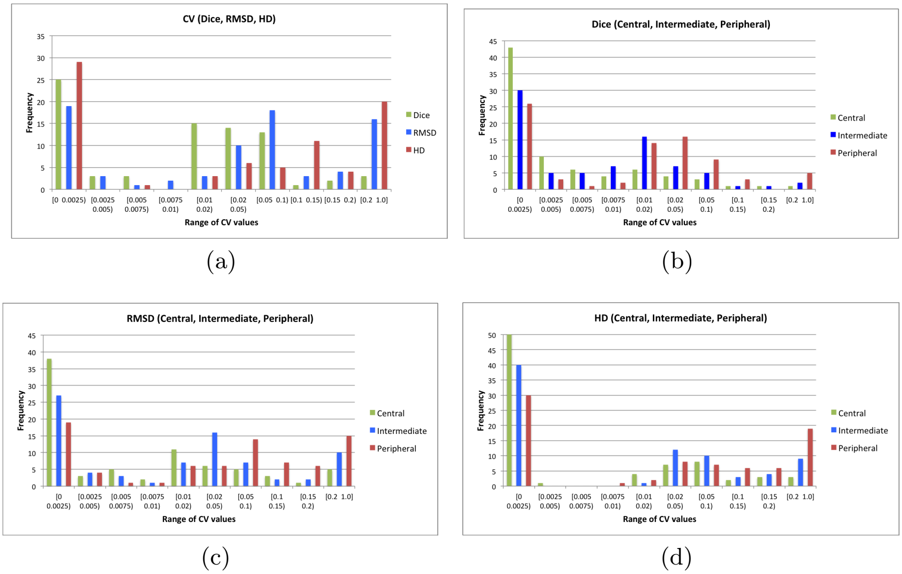
3.1 Coefficient of Variation within Seed Point Categories

Coefficient of variation (CV) is the ratio of standard deviation of a sample to the mean of the sample and indicates the extent of variability in relation to mean of the population. CV of the Dice coefficient, RMSD, and HD were computed over all seed points of each follicle. Histograms of the resulting CV values for the 58 follicles are shown in Figure 2(a). The range of CV values have been divided into ten unequal intervals and have been represented along the horizontal axis; since most of the values are in the range of 0 to 0.2, this interval has been divided into 9 bins illustrate the distribution of CVs within this range. CV values greater than 0.2 have been included in a single bin. The vertical axis represents the number of follicles for which the CV fell into the specified interval. For Dice, RMSD and HD, 58% of the coefficients of variation are less than 5%. Figure 2(b) shows histograms of the Dice coefficient CV values from the 79 follicles calculated for the central, intermediate, and peripheral groups of seed points. The distribution of the these CVs follow the same general trends as the overall distributions; overall 86% of the Dice CV values over all three seed point groupings were less than 5%. The mean Dice CV for the central, intermediate, and peripheral seed point groups, computed over all follicles, were 2.11%, 3.04% and 6.90% respectively, while the overall mean Dice CV across all seed points and follicles was 5.48%. The central and intermediate seed point group's mean Dice CVs were found to be significantly different from the overall mean Dice CV (two-tailed Student's paired two-sample t-test, $p = .00024$ and $p = 0.005$, respectively).

The mean Dice CV for the peripheral region was not significantly different from the overall Mean Dice CV. The mean Dice CV for the central seed point group was significantly less than both the intermediate and the peripheral groups. The mean Dice CV for the intermediate seed point group was significantly less than the mean Dice CV for the peripheral seed point group. This is strong evidence that reproducibility is, on average, higher when the seed points are placed in either the central or intermediate regions of a follicle.

Figures 2(c) and 2(d) show the distributions of CV for the central, intermediate and peripheral seed point groups for RMSD and HD respectively, again showing similar trends. For RMSD, 68% follicles had a CV of less than 5% for all three seed point groups, while for HD, this number was 66%.

The overall mean RMSD CV was 13.5%, and the mean RMSD CV's for the central, intermediate, and peripheral seed point groups were 3.46%, 6.63% and 15.9%; all of these were significantly different from the overall RMSD CV. The mean RMSD CV for the central seed point group was significantly less than both the intermediate and the peripheral groups. The mean RMSD CV for the intermediate seed point group was significantly less than the mean Dice CV for the peripheral seed point group.



(a)

(b)

(c)

(d)

Fig. 2. (a)Coefficient of variation of overall Dice Coefficients, RMSD and HD. (b) Coefficient of variation of Dice Coefficients for three groups of seed points (c)Coefficient of variation of RMSD for three groups of seed points (d)Coefficient of variation of HD for three groups of seed points.

The overall mean HD CV was 14.3%, and the mean HD CV's for the central, intermediate, and peripheral seed point groups were 3.33%, 6.83% and 15.9%; the central and intermediate means were significantly different from the overall HD CV. The mean HD CV for the central seed point group was significantly less than both the intermediate and the peripheral groups. The mean HD CV for the intermediate seed point group was significantly less than the mean Dice CV for the peripheral seed point group.

Again we have very strong evidence that reproducibility is considerably greater, on average, when seed points are confined to the central and intermediate regions.

3.2 Mean Segmentation Accuracy within Seed Point Categories

Figure 3(top row) presents the mean and standard deviations (as error bars) of Dice, RMSD, and HD for each follicle. The follicles are positioned on the horizontal axis in decreasing order of their cross-sectional diameter. A linear regression line fit to this data shows that Dice coefficients generally decrease for smaller follicles; R^2 was significant ($p < 0.05$). There was no evidence of a linear trend for RMSD and HD with decreasing follicle size; the R^2 values for both of these regressions were not significant.

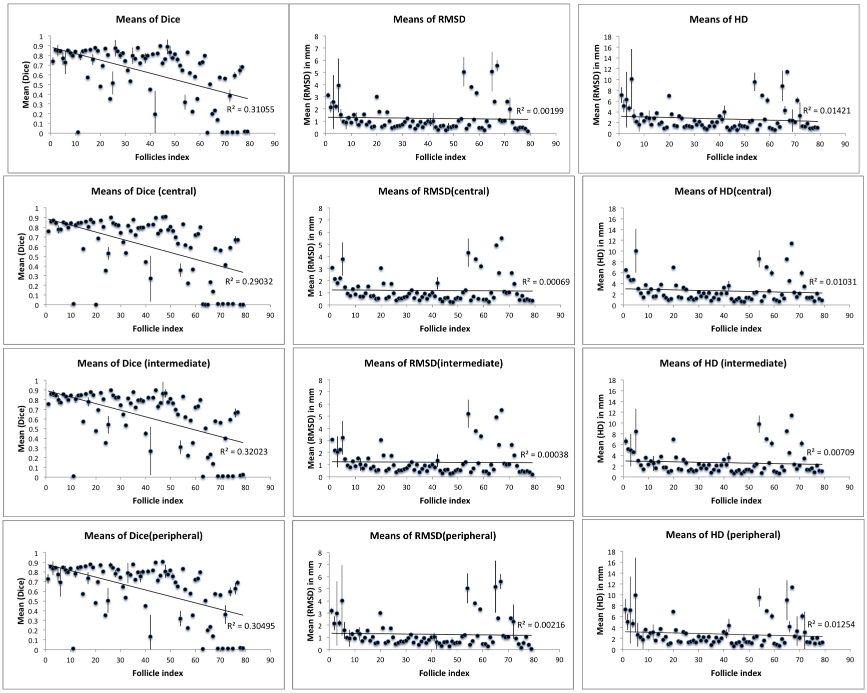

Means of Dice
Means of RMSD
Means of HD
Means of Dice (central)
Means of RMSD(central)
Means of HD(central)
Means of Dice (intermediate)
Means of RMSD(intermediate)
Means of HD (intermediate)
Means of Dice(peripheral)
Means of RMSD(peripheral)
Means of HD (peripheral)

Fig. 3. Mean and Standard Deviation of Dice, RMSD and HD with error bar, 1st row: for overall, 2nd row: for central points, 3rd row: for intermediate points, 4th row: for peripheral points

On the whole, values of RMSD and HD generally remain stable as follicle size decreases, while Dice coefficient worsens. This can be explained by noting that Dice coefficient is a proportional error based on areas where as RMSD and HD are absolute errors based on distances. Consider two follicles region segmentation results with the same Dice coefficient, but where the follicles have vastly different area. Absolute deviations from the true boundary will be smaller for the small follicle than for the larger follicle simply because of the size difference. Thus, it is possible for RMSD and HD to remain stable as Dice coefficient and follicle size decrease. Figures 3 (2nd - 4th row) show the same information but subdivided into subgroups of central, intermediate, and peripheral seed points, respectively. It can be seen that the standard deviations (error bars) are generally smaller for the groups of central seed points and larger for the groups of peripheral seed points. Again, the linear regression line shows, in all cases, a trend of worsening Dice coefficient (R^2 significant, $p < 0.00001$), and no trend in RMSD and HD values (R^2 not significant) as follicle size decreases.

3.3 Results Summary

From our statistical study of groups of correct interactions for the follicle segmentation algorithm we have the following main results.

1. For many, but not all follicles, the overall CV of segmentation accuracy is low to moderate. Overall Dice CV was less than 5% for 76% of follicles; overall CV of RMSD and HD was less than 5% for 48% and 49% of follicles, respectively.
2. The mean Dice CV (respectively HD CV and RMSD CV) for the central seed point groups was significantly smaller than for the intermediate and peripheral groups. The mean Dice CV (respectively HD CV and RMSD CV) of the central and intermediate seed point groups were significantly smaller than the peripheral group, and the magnitude of the difference in means from the peripheral group were quite large.
3. There was a statistically significant trend of decreasing (worsening) mean Dice coefficient with decreasing follicle diameter. However, there was no evidence of a statistically significant trend in the values of mean HD and RMSD with respect to the diameter of the follicles.

3.4 Discussion

In this section we discuss the interpretation and implications of the results in the previous section.

Result #1 tells us, in the broadest of strokes, that the studied segmentation algorithm is not able to generate easily reproducible results on perhaps as many as half of all follicles. is a type of result that could have been obtained from the standard method of examining results from a small number of human observers. However, our methods result in many more samples from which to estimate the mean and variance of the segmentation accuracy measures, possibly resulting in better estimates of the true mean and variance over all correct interactions, and therefore more accurate estimates of coefficient of variation.

Result #2 is very strong evidence that reproducibility is, on average over all follicles, much higher for the central and intermediate seed point groupings than for the peripheral seed point group. This result could not have been obtained using standard evaluation with human observers. Even if there were controls to ensure an equal number of samples in each seed point group, with 10 or fewer observers there would be insufficiently many samples per grouping to obtain any reasonable level of statistical power for our methods. Traditional evaluation would only have provided an estimate of the overall reproducibility and would not have elucidated the magnitude of the degradation of reproducibility with increasing seed point distance from the centroid of the follicle region.

Result #3 indicates that, while the segmentation error with respect to the deviation of the segmented follicle boundary from the ground truth boundary was not related to follicle size, the mismatch between the segmented follicle region and the ground truth follicle region was larger for smaller follicles because said deviations from the true boundary result in a greater proportion of region mismatch for smaller follicles. This result might be obtained using standard evaluation with human observers if reproducibility of the algorithm is already very high, but otherwise, a larger number of samples are needed to get more accurate estimates of mean segmentation accuracy measures on a per-follicle basis.

By judiciously choosing subsets of correct interactions to analyze using our methods, one can potentially obtain evidence that can be used to make recommendations on how operators can best use the system to produce the most accurate and consistent results. In the case of our analysis of the follicle segmentation algorithm, one would likely recommend that observers avoid placing seed points within the follicle's periphery to reduce inter-observer variability.

As our methods sample the set of correct interactions uniformly, our methodology is free from observer bias, observer training effects, and other biases that might result from the instructions/protocols that human observers are instructed to use during data collection. Uniform sampling of correct interactions incorporates into the evaluation those correct interactions that might, under normal circumstances, never be used by human observers because they contradict established usage protocols, but which might nevertheless occur in cases of observer fatigue or a lapse in judgment, and allows us to study the effect and risk of such lapses by studying the appropriate subsets of correct interactions.

4 Conclusion

Instead of characterizing segmentation performance through the actions of a small number of observers, we constructed synthetic observers and characterized their behaviour using a much larger number of segmentations over the set of all correct interactions. Our methods allow for a much richer, statistically backed characterization of interactive segmentation algorithm performance, resulting in new kinds of information that elucidate the best practices for how an interactive algorithm should be used to avoid any inherent sources of error in the algorithm while better understanding those sources of error.

References

1. Baerwald, A.R., Adams, G.P., Pierson, R.A.: Characterization of ovarian follicular wave dynamics in women. Biology of Reproduction 69(3), 1023–1031 (2003)
2. Bowyer, K.W.: Validation of medical image analysis techniques. In: Sonka, M., Fitzpatrick, J.M. (eds.) Handbook of Medical Imaging: Medical Image Processing and Analysis, vol. 2, SPIE press (2000)

3. Boykov, Y., Jolly, M.P.: Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In: Proceedings of the International Conference on Computer Vision (ICCV 2001), pp. 105–112 (2001)
4. Boykov, Y., Lea, G.F.: Graph cuts and efficient n-d image segmentation. International Journal of Computer Vision 70(2), 109–131 (2006)
5. Byrum, C.E., MacFall, J.R., Charles, H.C., Chitilla, V.R., Boyko, O.B., Upchurch, L., Smith, J.S., Rajagopalan, P., Passe, T., Kim, D., Xanthakos, S., Ranga, K., Krishnan, R.: Accuracy and reproducability of brain and tissue volumes using a magnetic resonance segmentation method. Psychiatry Research: Neuroimaging 67, 215–234 (1996)
6. Cates, J.E., Lefohn, A.E., Whitaker, R.T.: Gist: an interactive gpu-based level set segmentation tool for 3d medical images. Medical Image Analysis 8, 217–231 (2004)
7. Coehn, B.A., Barash, I., Kim, D.C., Sanger, M.D., Babb, J.S., Chandarana, H.: Intraobserve and interobserver variability in renal volume measurements in polycystic kidney disease using a semiautomated mr segmentation algorithm. American Journal of Roentgenology 199, 387–393 (2012)
8. Dach, C., Held, C., Wenzel, J., Gerlach, S., Lang, R., Palmisano, R., Wittenberg, T.: Evaluation of an interactive cell segmentation for flourescence microscopy based on the graph cut algorithm. In: Microscopic Image Analysis with Applications in Biology, Heidelberg, Germany (September 2, 2011)
9. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology 26(3), 297–302 (1945)
10. McGuinness, K., O'Connor, N.E.: A comparative evaluation of interactive segmentation algorithms. Pattern Recognition 43(2), 434–444 (2010)
11. Moschidis, E., Graham, J.: A systematic performance evaluation of interactive image segmentation methods based on simulated user interaction. In: IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 928–931 (2010)
12. Nickisch, H., Rother, C., Kohli, P., Rhemann, C.: Learning an interactive segmentation system. In: Proceedings of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2010), pp. 274–281 (2010)
13. Schenk, A., Prause, G.P.M., Peitgen, H.-O.: Efficient semiautomatic segmentation of 3D objects in medical images. In: Delp, S.L., DiGoia, A.M., Jaramaz, B. (eds.) MICCAI 2000. LNCS, vol. 1935, pp. 186–195. Springer, Heidelberg (2000)
14. Stammberger, T., Eckstein, F., Michaelis, M., Englmeier, K.-H., Reiser, M.: Interobserver reproducibility of quantitative cartilage measurements: Comparison of b-spline snakes and manual segmentation. Magnetic Resonance Imaging 17(7), 1033–1042 (1999)
15. Veksler, O.: Star shape prior for graph-cut image segmentation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 454–467. Springer, Heidelberg (2008)