Sampling-Based Ensemble Segmentation against Inter-operator Variability

Jing Huo[1], Kazunori Okada[2], Whitney Pope[1], Matthew Brown[1]
[1] Center for Computer vision and Imaging Biomarkers, Department of Radiological Sciences, David Geffen School of Medicine at UCLA
[2] Computer Science Department, San Francisco State University

ABSTRACT

Inconsistency and a lack of reproducibility are commonly associated with semi-automated segmentation methods. In this study, we developed an ensemble approach to improve reproducibility and applied it to glioblastoma multiforme (GBM) brain tumor segmentation on T1-weigted contrast enhanced MR volumes. The proposed approach combines sampling-based simulations and ensemble segmentation into a single framework; it generates a set of segmentations by perturbing user initialization and user-specified internal parameters, then fuses the set of segmentations into a single consensus result. Three combination algorithms were applied: majority voting, averaging and expectation-maximization (EM). The reproducibility of the proposed framework was evaluated by a controlled experiment on 16 tumor cases from a multi-center drug trial. The ensemble framework had significantly better reproducibility than the individual base Otsu thresholding method ($p<.001$).

Keyword: GBM, ensemble

1. INTRODUCTION

In clinical practice, reproducible and repeatable segmentation is an important prerequisite for longitudinal studies of medical images. Semi-automated segmentation methods are often preferred in common radiographic protocols because they allow expert clinicians to control the segmentation quality which plays a critical role in the final diagnostic decision. However, such semi-automated methods are also inconsistent when input is provided by different readers and/or used with different internal parameter values. This is a trade-off between usability and repeatability, posing a serious technical challenge. Our study focuses on segmenting glioblastoma multiforme (GBM) brain tumors in T1-weighted contrast enhanced MR volumes using the Otsu thresholding method.

GBM tumor segmentation offers an ideal test case for our study because the contrast-enhancement heterogeneity of the tumors makes the current state-of-the-art methods highly irreproducible. The Otsu method is used in this study and has the advantages of efficiency, simplicity and usability, however it also suffers from poor reproducibility due to its need for user input in the setting of inhomogeneous tumors. Different user interaction will lead to different thresholds and thus inconsistent segmentation results. The algorithm parameter setup, in this case the number of thresholding levels, will also result inconsistent segmentation results.

The main purpose of our study is to improve the reproducibility and stability of the semi-automated Otsu thresholding segmentation applied to GBM brain tumors. The final goal is to generate a robust and stable ensemble result given one single manually-drawn user input. The proposed framework combines sampling-based simulations and ensemble segmentation. It generates a set of segmentations by sampling the user interaction space and algorithm internal parameter space, then the set of inconsistent segmentations are ensembled into a single consensus result. The algorithmic sampling of user interaction space is designed to perturb the manual user interaction to simulate the typical inter-operator variability, thus the fusion of all possible segmentations is expected to be stable, reproducible and repeatable.

2. MATERIALS AND METHODS

The brain volumes were pre-processed by skull-stripping using FSL tools[1]. Users define a bounding cube as the volume of interest (VOI), and all the segmentation is done within the VOI. The user interaction is to draw a 2D bounding box that surrounds the target tumor on an arbitrarily chosen 2D slice.

The base segmentation algorithm is Otsu thresholding method. The concept of Otsu's thresholding method is to find the threshold that minimizes the weighted within-class variation: $\sigma_w^2(t) = q_1(t)\sigma_1^2(t) + q_2(t)\sigma_2^2(t)$, with the class probability as $q_1(t) = \sum_{i=1}^{t} P(i)$, class variance as $\sigma_1^2(t) = \sum_{i=1}^{t} [i - \mu_1(t)]^2 \frac{P(i)}{q_1(t)}$ and class mean as $\mu_1(t) = \sum_{i=1}^{t} \frac{iP(i)}{q_1(t)}$. Given an initial $\mu_i(t)$ and $\sigma_i(t)$, the algorithm will exhaustively search by altering the threshold value to find the optimal value. The approach can be generalized to multiple classes.

All the GBM brain tumors were segmented on T1w post contrast images via the base semi-automated algorithm. First, radiologists draw a bounding box around the tumor, intensity values within the box are collected to form a histogram, and Otsu thresholding is applied, with the assumption that the number of classes within the bounding box is 2 or 3. Afterwards, the 3D image volume is thresholded with the highest thresholding value followed by a connected component analysis, morphological opening and closing. The structure element used for morphological operations is circular with radius 1.

The final goal is to generate a robust and stable ensemble result given one 2D manually-drawn user interaction box. There are two steps in the proposed ensemble framework: first, a set of inconsistent segmentations are generated by sampling user interactions and algorithm internal parameters; second, the set of inconsistent segmentations are fused into a single consensus result.

2.1 Sampling user interactions

The purpose of this step is to automatically sample a set of 2D bounding boxes on different slices that simulate typical inter-operator variability given one manual 2D bounding box drawn by the user manually on one single 2D slice. In order to maintain the usability, we wish the user to provide the minimal interaction with only a single 2D bounding box, from which we perform the sampling automatically.

First, given the manually-drawn 2D input box, a 3D segmentation result is generated using the base method. And then, the 3D segmented tumor is uniformly sampled across all slices, yielding N sampled slices. On each of the N slices, one 2D bounding box of the tumor is generated. In the end, the set of N 2D bounding boxes is used as the user interaction variations. The pipeline is shown in Figure 1.

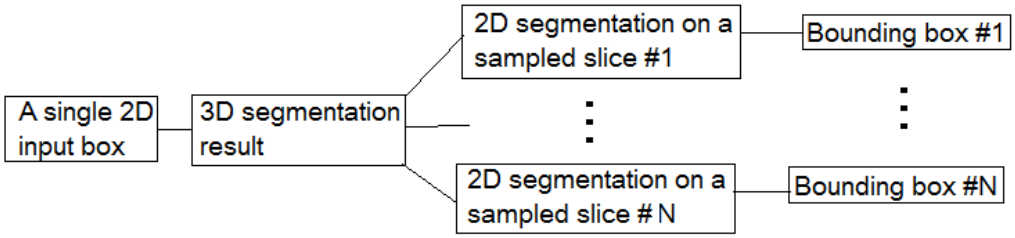

Figure 1 Pipeline of automated sampling of user inputs to simulate inter-operator variability

2.2 Sampling Internal Algorithm Parameters

The internal parameter, L, for the Otsu method is the number of intensity thresholding levels within a user input box. The difficulty of this parameter setup comes from the GBM tumor presentation. GBM brain tumors are composed of tumor cells, edema, and necrotic regions, which present different levels of enhancement, and may or may not be present within a single given tumor. Figure 2 shows one tumor showing different intensity thresholding levels on two difference slices. It is not accurate to set a universal number of levels for all 2D user interaction boxes. Thus, we ran the base method with both L=2 and L=3 in this study.

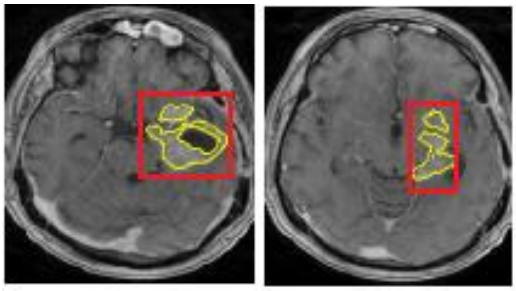

Figure 2. Two slices of one tumor: (a) 3-object problem; (b) 2-object problem

2.3 Ensemble Segmentation

The automatically sampled N user interaction boxes in Section 2.1 are used to perform independent 3D Otsu segmentation, and 2 intensity levels (L=2 and L=3) are applied to each user interaction box. As a result, 2N inconsistent 3D segmentation results are generated accordingly. In this step, the 2N inconsistent segmentation results will be fused into a final ensemble segmentation result.

We compared three ensemble methods: majority voting, averaging and STAPLE-EM algorithm[3], to fuse the 2N inconsistent segmentation results. Majority voting generates a binary segmentation. The averaging result is defined as follows: $PM(i,j,k) = \frac{1}{N}\sum_{n=1}^{N} Seg_n(i,j,k)$. For the details of STAPLE-EM, readers are referred to [3]. Both averaging and EM method generate a probability map (PM), which is thresholded into binary segmentations by 0.3 empirically.

In a summary, the ensemble framework takes a single manual 2D box as the input, and generates one ensemble 3D segmentation as the output. The complete framework is shown in Figure 3. In this study, we set N=8.

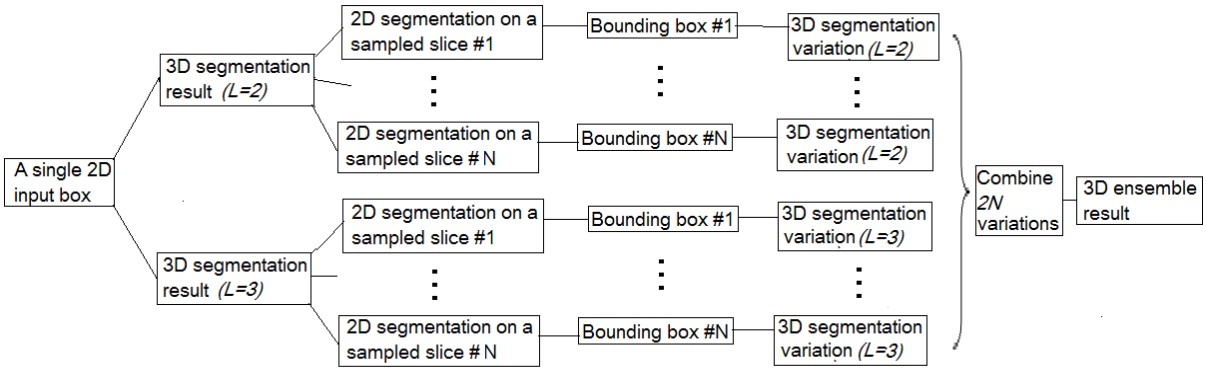

Figure 3 The complete ensemble segmentation pipeline

3. Experiments and Results

We have 16 GBM tumor cases with T1-weighted post-contrast volumetric MR images (voxel size 0.9*0.9*1mm). The ground truth for the segmentation was manually contoured by a neuroradiologist.

The reproducibility of the proposed framework was evaluated by a controlled experiment. For each tumor, the user manually drew eight 2D bounding boxes on 8 uniformly-sampled slices across the whole tumor volume. For each of the manually-drawn boxes, the base method and the ensemble framework were applied respectively for comparison. The overlap ratio of the 8 3D ensemble results and that of 8 base Otsu results were compared (Figure 4). All three ensemble methods showed significant better reproducibility than the base Otsu method with L=3 ($p<.001$) and L=2 ($p<.001$). Even though EM was visually slightly better than averaging and voting (Figure 5), there was no statistically significant difference with this dataset ($p>0.05$).

The accuracy of the proposed framework was evaluated by calculating F-measure between the ground truth and the semi-automated methods. There were in total 128 boxes, with eight manually-drawn boxes for each of the 16 tumor cases; the accuracy of each box was shown in Figure 6. All three ensemble methods showed significant improvement over the Otsu method with L=3 ($p<.001$), but no difference from Otsu method with L=2 ($p>.05$). There was no statistically significant difference between the three ensemble methods in accuracy ($p>.05$) as shown in Figure 7.

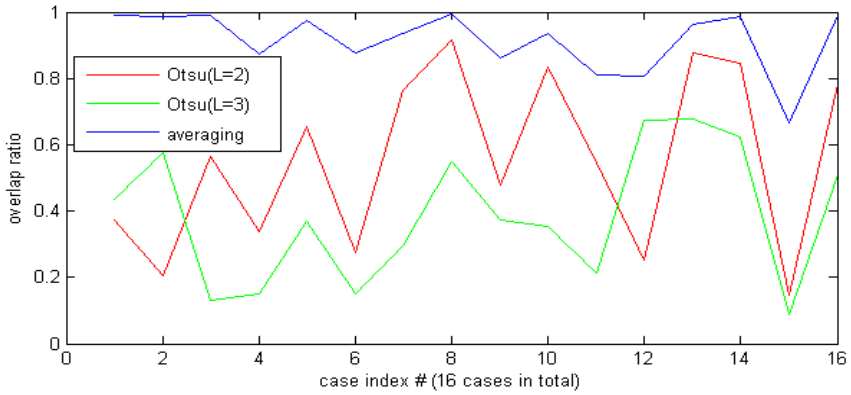

Figure 4. Reproducibility comparison between individual Otsu and averaging ensemble

Figure 8 and Figure 9 showed an example comparing the segmentation results using different manually-drawn user interaction boxes between the base Otsu method and the ensemble framework. Different rows show different manually-drawn user interaction boxes; while columns show different slices of the same tumor. Figure 8 showed that the base Otsu segmentations are inconsistent using different manually-drawn boxes; and Figure 9 showed that the ensemble segmentations are consistent.

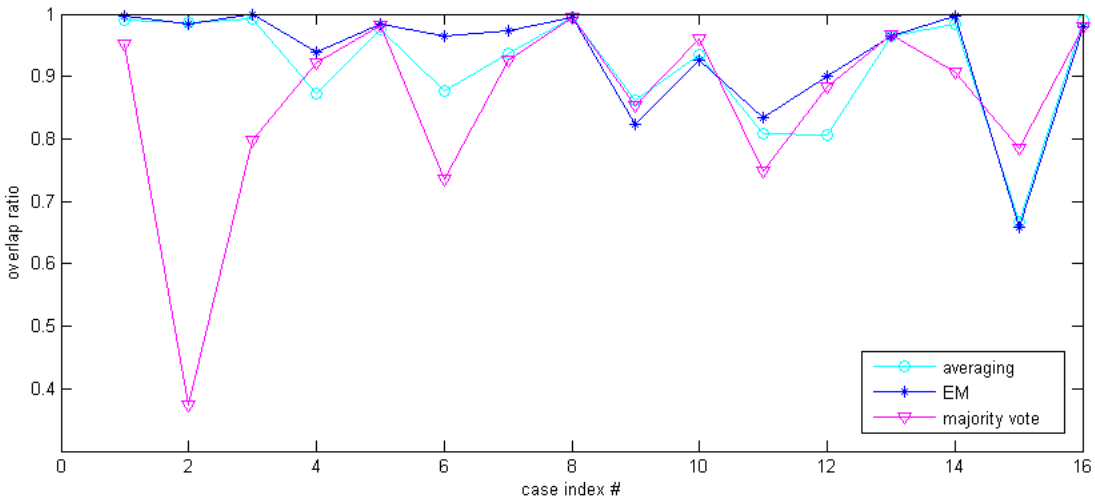

Figure 5 The reproducibility comparison between three ensemble methods: vote, averaging and EM

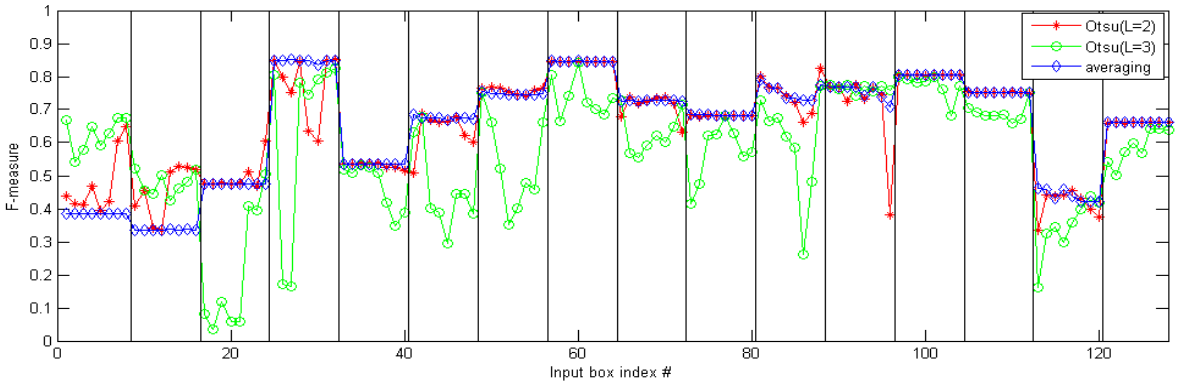

Figure 6 F-measure of the base method and the averaging ensemble.

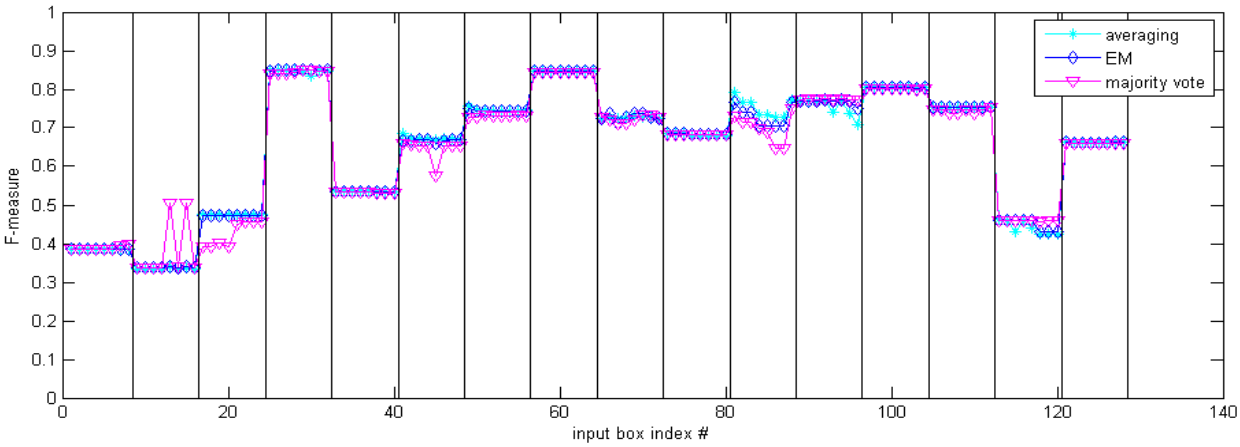

Figure 7 F-measure of the three ensemble methods.

4. DISCUSSION AND CONCLUSION

We developed an ensemble framework aiming to improve the reproducibility of the semi-automated segmentation, and applied the framework to GBM brain tumor segmentation on T1w post contrast images. First, we invented an automated sampling of user interactions that simulates typical inter-operator variability, in order to allow for minimal user interaction, as well as sampling internal algorithm parameters. Then, we generated the inconsistent segmentation set using the sampled user interaction boxes and ensemble them into a final segmentation. We evaluated the performance on a difficult task of the single-channel GBM brain tumor segmentation.

In conclusion, the proposed automated user interaction sampling and ensemble framework tremendously improved the reproducibility compared to the base method on our dataset on GBM brain tumor segmentation. Reproducibility is crucial for semi-automated segmentation methods, the proposed framework shows great potential in improving the consistency and reproducibility.

5. REFERENCES

[1] Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R.K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M. and Matthews, P.M., "Advances in Functional and Structural MR Image Analysis and Implementation as FSL," NeuroImage, 23(S1), 208–219 (2004).

[2] Otsu, N., “A threshold selection method from gray level histograms,” IEEE Trans. Syst. Man. Cybern. 9, 62–66 (1979).

[3] Warfield, S.K., Zou, K.H. and Wells, W.M., “Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation.” IEEE Trans Med Imaging, 23(7), 903-921 (2004).

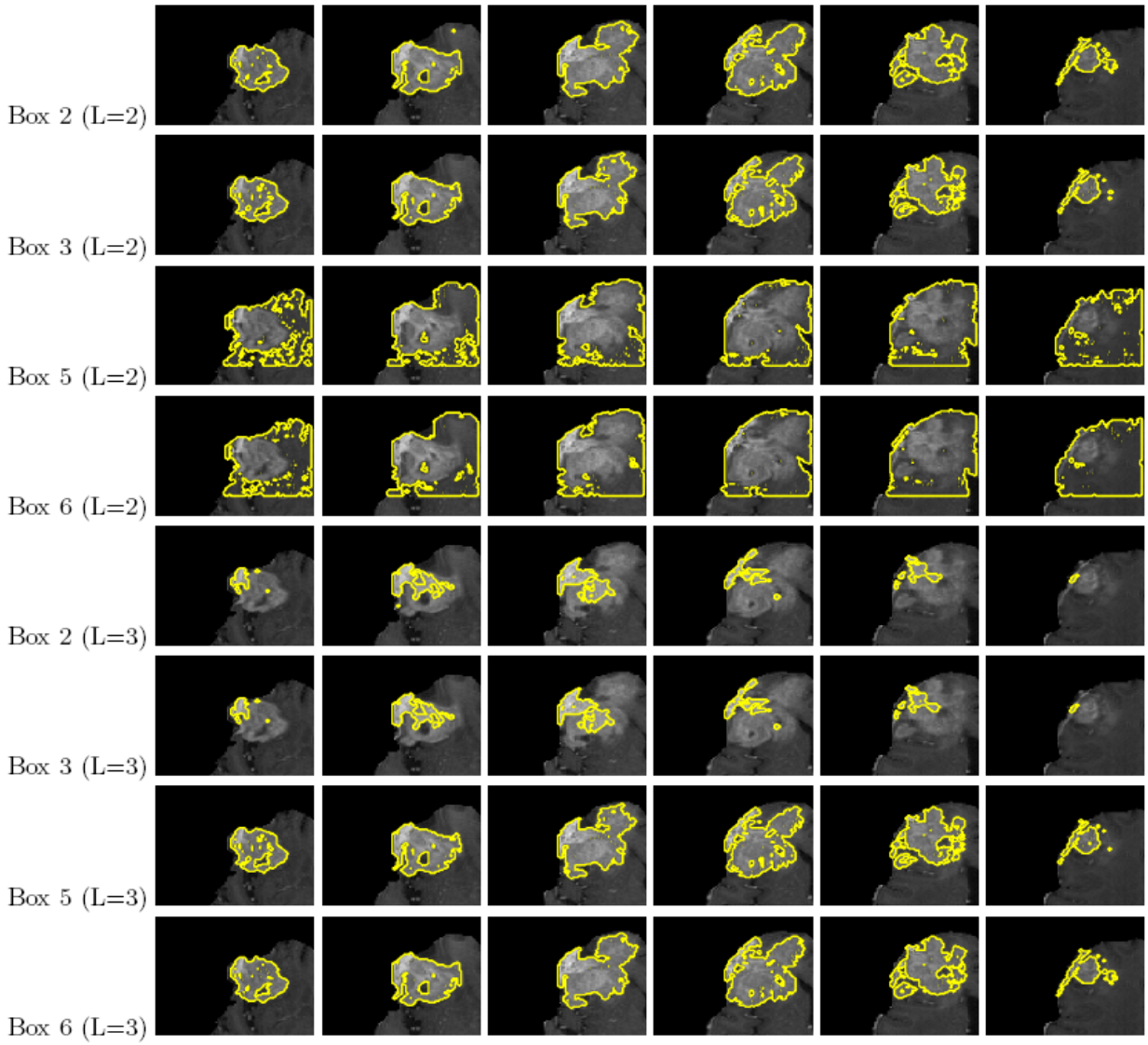

Figure 8 Segmentation results from the base Otsu method using different manually-drawn user interactions: rows are 6 different slices of the same tumor; columns are result from different manually-drawn user interaction boxes.

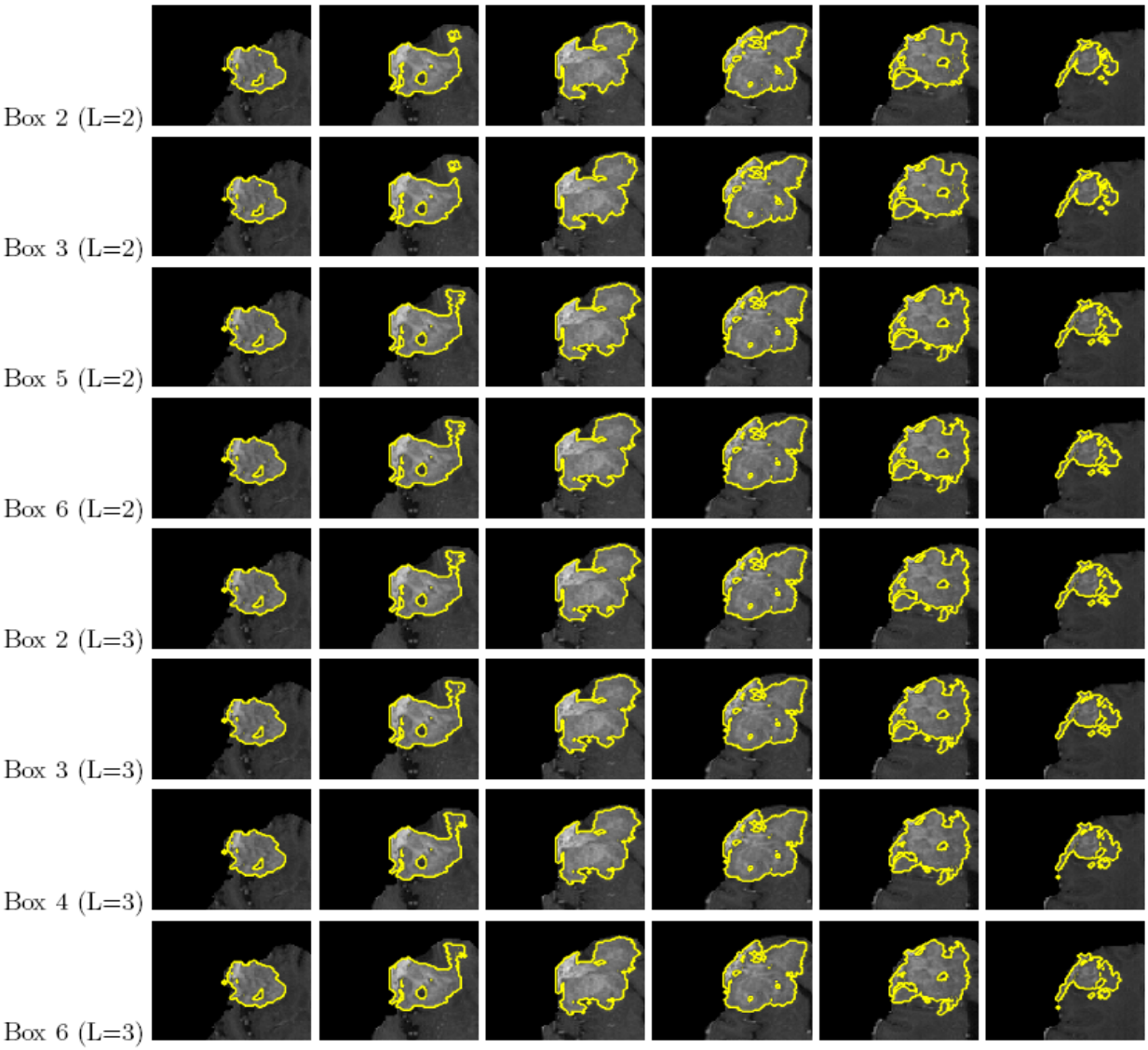

Figure 9 Segmentation results from the ensemble framework using different manually-drawn user interactions: rows are 6 different slices of the same tumor; columns are result from different manually-drawn user interaction boxes.