Lab 3 - Algorithms Evaluation

Songyou Peng
psy920710@gmail.com

I. INTRODUCTION

In this lab, we utilise different measures to evaluate 6 different 2D image segmentation algorithms and then compare the performance of them based on the measures. We also extend our segmentation evaluation to a 3D MRI volume. Throughout the report, some encountered problems will also be mentioned.

II. SEGMENTATION EVALUATION IN 2D

In this part, we first obtain the receiver operating characteristic (ROC) and the corresponding area under the curve (AUC), which are two quite important evaluation measures. And then, two area overlap measures, Jaccard index and Dice index, will also be calculated. Finally, Hausdorff distance will be computed. All these 3 measures are acquired at the best threshold for each algorithm. However, we notice that the best thresholds acquired from ROC are actually not the same as the ones from the area overlap and distance measures. We will discuss this in section II-B.

A. ROC curve & AUC

In order to get the ROC curve, what we need to do is iteratively increasing the threshold for getting binary segmentation images. Each threshold produces an (tp_rate, fp_rate) pair corresponding to a single point in ROC curve.

In order to get tp_rate, we simply element-wise multiply the segmentation image and the ground truth and the result is the intersection of the two binary images. True positive number tp_number is the sum of the pixels inside the intersection and the positive number p_number is the sum of ground truth. Also, the false positive number fp_number is simply the number of non-zero pixels of (segmentation image – intersection). Similarly, the negative number n_number is the sum of (1 - ground truth). The (tp_rate, fp_rate) pair can be written as:

$$tp\_rate = \frac{tp\_number}{p\_number}$$

$$fp\_rate = \frac{fp\_number}{n\_number}$$

After acquiring the pairs of the different thresholds, we can plot the ROC curves, which are shown in Fig. 1.

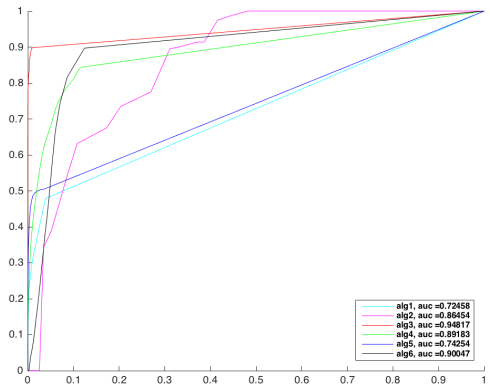
Once ROC has been obtained, one way to evaluate the segmentation quality of an algorithm is to calculate the AUC. Here, we see the zones between two pairs as trapezoids. Then AUC, which is the sum of the area of each zone, is able to be acquired easily. The values of AUC of each algorithm for each image are included in Fig. 1. According to that, we can say that in most cases algorithm 3 outperforms others and algorithm 4 also provides decent results.

It should be noted that in Fig. 1(b), the ROC curves of algorithms 1,3 and 5 are almost diagonal. The reason for that is: the image 2 of algorithm 3 is totally black so only all the pairs are $(0,0)$ except one pair $(1,1)$ acquired when the threshold is smaller than 0. For algorithms 1 and 5, the AUCs are even less than 0.5 because the images contain no true positive information, also contain false positive information. Therefore, the true positive rates are all 0 while false positive rates have some small values.

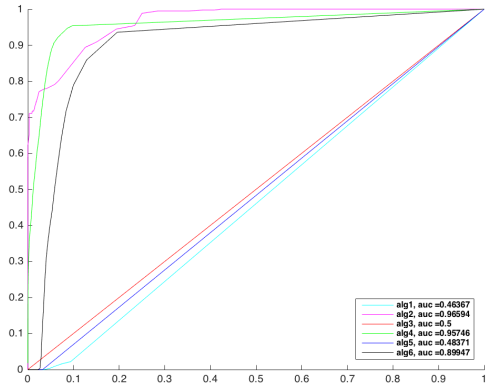
B. Area overlap & distance measures

For the sake of calculating the area overlap and distance measures, we need to get the best threshold at the very beginning based on the ROC. As we know, position $(0,1)$ is the ideal situation in ROC curve, which means the highest true positive rate with the lowest false positive rate. The way we get the best threshold from ROC is to find the threshold which has the smallest Euclidean distance between the corresponding pairs and $(0,1)$ among all 4 images. Once having the "best" threshold for each algorithm, we can calculate the Jaccard Index (Table I), Dice Index (Table III) and Hausdorff distance (Table V, -1 means invalid value).

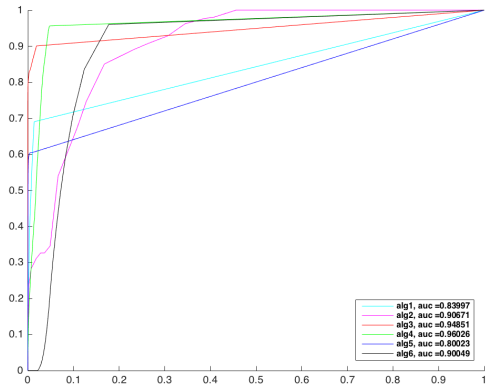
At first, the way of calculating the best threshold for each algorithm is to get the best thresholds separately for 4 images and then get the mean of the 4 thresholds. The problem is sometimes the thresholds diverse with



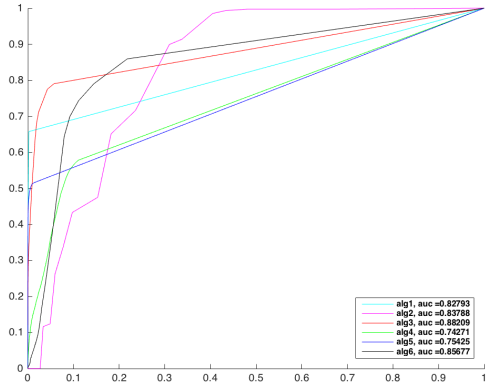
(a) Image 1



(b) Image 2



(c) Image 3



(d) Image 4

Fig. 1. Comparison of the the ROC curves of different algorithms

TABLE I

JACCARD INDEX IN THE "BEST" THRESHOLD FROM ROC

Alg \ Img	1	2	3	4
1	0.0850	0.0010	0.5873	0.3675
2	0	0.0308	0.0535	0.0867
3	0.4447	0	0.2410	0.4101
4	0.0548	0.0423	0.0863	0.2397
5	0.2341	0	0.4166	0.4237
6	0.0549	0.0210	0.0703	0.0793

each other. For example, the thresholds for algorithm 4 are: 3, 3, 23, 3 and the mean is 8, which lead to a bad segmented results for all 4 images. So we choose to get the best threshold out of the 4 best.

Although we can tell the differences among 6 methods from Table I and III, the differences are not quite significant because the "best" thresholds we acquire from ROC are not really the best. The "best" thresholds are the global thresholds for the algorithms. Here we also give Table II and IV, which are the Jaccard and Dice index using the real local best threshold for each image. Compared with the global one, the local thresholds can really help us distinguish the goodness among various algorithms. According to our discussion, we propose a strategy for a segmentation system: first find out which method performs the best based on the different local thresholds, and then choose a trade-off good global threshold for that particular algorithms for the sake of segmenting all the other images.

One problem we faced during the implementation was the calculation of Hausdorff distance. If we considered all the non-zero points of the segmented images and the ground truth, it took years to calculate. As we know, the Hausdorff distance is all about the minimum and maximum distances between one region and the other, so the contour points of regions are enough. Therefore, we applied Canny filter to get the contour of each region in the image and then calculate the distance, which accelerates the calculation process significantly.

C. Comparisons of segmentation results

From the tables, we can find out that algorithm 3 still performs well under these three measures, while the algorithm 4 performs really bad and algorithm 5 has relatively good results. This is quite different from what we get in the last section when AUC was calculated.

TABLE II

JACCARD INDEX IN THE REAL BEST THRESHOLD FROM ROC

Alg \ Img	1	2	3	4
1	0.1729	0.0045	0.5873	0.3675
2	0.0651	0.6067	0.0719	0.2110
3	0.7470	0.0045	0.3487	0.7747
4	0.1955	0.2168	0.1027	0.2677
5	0.2951	0.0045	0.4289	0.5370
6	0.0787	0.0369	0.1261	0.0985

TABLE III

DICE INDEX IN THE "BEST" THRESHOLD FROM ROC

Alg \ Img	1	2	3	4
1	0.1524	0.0020	0.7399	0.5374
2	0	0.0598	0.1015	0.1595
3	0.6156	0	0.3884	0.5816
4	0.1039	0.0811	0.1588	0.3867
5	0.3794	0	0.5881	0.5952
6	0.1041	0.0410	0.1313	0.1470

TABLE IV

DICE INDEX IN THE REAL BEST THRESHOLD FROM ROC

Alg \ Img	1	2	3	4
1	0.2949	0.0089	0.7399	0.5380
2	0.1222	0.7552	0.1342	0.3485
3	0.8552	0.0089	0.5170	0.8730
4	0.3271	0.3563	0.1863	0.4223
5	0.4557	0.0089	0.6003	0.6988
6	0.1460	0.0712	0.2239	0.1793

TABLE V

HAUSDORFF DISTANCE IN THE "BEST" THRESHOLD FROM ROC

Alg \ Img	1	2	3	4
1	767.43	623.71	480.38	594.86
2	541.3917	655.76	532.31	687.01
3	642.23	-1	552.41	238.32
4	757.88	578.88	479.28	587.49
5	835.32	667.28	563.55	675.53
6	846.3297	677.99	572.80	708.35

After analysing the images, we found out that the segmentation images acquired from algorithm 4 have many false positive regions, which are not what we want. In medical case, we expect the false positive rate to be as low as possible. Although the AUCs are high for algorithm 4, it is still a bad segmentation method due to the false positive regions. Based on this, AUC is not really good measures for the medical segmentation case.

As for the comparison between Hausdorff distance and Jaccard/Dice index, it should be noticed that Hausdorff distance is relatively difficult for users to distinguish the differences among various algorithms. One possible reason is: the Hausdorff distances are only non-negative numbers so the distance cannot distinguish whether the segmented lesions are larger or smaller than the ground truth lesions, assuming the gravity center of them are the same. Overall, Jaccard or Dice index are two relatively better measures for the segmentation of medical images.

III. SEGMENTATION EVALUATION IN 3D

Since the given 3D segmented image is already thresholded, we cannot obtain ROC or AUC. Here we calculate Jaccard and Dice Index as well as Hausdorff distance. In order to calculate the Hausdorff distance, we calculate the distance between each non-zero point in one volume and the non-zeros in the other, assuming the spacing of between two slices is 1.

The final results are: 0.5277 for Jaccard Index , 0.6908 for Dice Index and the 26.1916 for Hausdorff distance. Compared with all the previous Tables, all the 3 values almost perform better. We can say that although we don't know what kind of segmentation method it is, this method must be the best if all other 6 algorithms are also applied to 3D images.

IV. CONCLUSIONS

In this lab session, we evaluate various segmentation methods using different measures: ROC, AUC, Jaccard Index, Dice Index and Hausdorff distance. Some comparisons among different methods are made based on the measures. We find out that AUC and Hausdorff cannot clearly show the differences, while the area overlap measures (Jaccard and Dice) are relatively suitable for the medical imaging use.