Medical Image Analysis Lab Report
ALGORITHM EVALUATION TECHNIQUES

Raabid Hussain

April 11, 2016

1 Objective

The objective of the lab-work was to get familiar with some of the algorithm evaluation techniques that are being used used in the biomedical industry. Different area overlap and distance measures were used to evaluate different segmentation algorithms for both 2D and 3D volumes. The algorithms' performances for different cases were compared to decide which one is the best suited for particular problems at the end.

2 Introduction

With the evolution of different algorithm, it became necessary to devise comparison methodologies. Unfortunately, there is no algorithm in medical imaging domain that has been classified as robust and perfect. So, it is necessary to have some sort of criteria to compare the algorithms in use. In order to understand the relative merits of these alternatives, it is necessary to evaluate them, enabling a prompt decision on which algorithm to carry on with.

To evaluate an algorithm, qualitative analysis is not sufficient when dealing with computer related methodologies. As a result, a number of quantitative analysis methodologies have been proposed. Common ones from these algorithms among the medical imaging community are distance measures like Harsdouf distance and area overlap measures like the Jaccard index and Dice

similarity co-efficient.

One of the most common evaluation technique is to compose a receiver operating characteristic (ROC) curve. ROC is a graphical interpretation of the performance of a binary classifier system. Some variable, usually the threshold, is varied to obtain different outputs for the system. A curve is generated using these values which plots true positive rate (TPR), also known as sensitivity, against the false positive rate (FPR). This ROC curve is then used to determine the best threshold to be used by the system. The accuracy of the system is determined by the area under (AUC) the ROC curve. An ideal ROC curve goes from [0,0] to [0,1] to [1,1]. Hence the ideal area under curve is 1 whereas area under curve for a random or useless test is 0.5. Anything below 0.5 is seen to be doing something else rather solving the proposed system.

After choosing the best threshold from the ROC curves, the evaluation techniques are applied at results of that threshold. One common technique is to compute the Harsdouf distance (HD) which measures how far a similar point is in an image compared to the other image. HD works by first computing the boundaries of the output image and the ground truth. Every boundary point in the ground truth image is matched with every point in the output image and the euclidean distance between them is calculated. The minimum distance is kept and the rest are discarded. This is repeated for all the points in the ground truth image. After all the distance have been computed, the maximum distance is saved. The above procedure is repeated for the algorithm output image to compute its maximum distance. Finally, the maximum distance of the two distances is called the harsdouf distance. Ideally the HD should be zero.

Area overlap measures are used to evaluate the performance of the algorithms. These include Jaccard index (JI) which is a statistic used to compare the similarity and diversity of images and the Dice Similarity coefficient (DC or QS) which measures the similarity of two images. Both of these use different forms of areas (true positive - TP, true negative - TN, false positive - FP and false negative - FN) to compute the similarity between the ground truth and the algorithm output.

The formulas for all the above mentioned measures are provided in figure 1.

$$TPR = TP/P = TP/(TP + FN)$$
$$FPR = FP/N = FP/(FP + TN)$$

$$d_H(X, Y) = \inf\{\epsilon \geq 0;\ X \subseteq Y_\epsilon \text{ and } Y \subseteq X_\epsilon\}.$$

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

$$QS = \frac{2|X \cap Y|}{|X| + |Y|}$$

Figure 1: Formulas for the evaluation measures used.

3 2D Volume Segmentation Evaluation

We were provided with 4 different cases from MIAS public database. Segmentation results for 6 different algorithms were provided along with the ground truth for each of the images. So a total of 24 segmentation outputs were used in evaluation. First ROC curves were computed for all the images followed by determining the best thresholds for each algorithm. These best thresholds were used to compute values for the above mentioned evaluation statistics.

3.1 Receiver Operating Characteristic Curves

The first task was to select the optimal thresholds for each algorithm. For this, the outputs of the algorithms were thresholded at all gray level values. Then the thresholded image was compared with the ground truth images to compute area measures like false positives, true positive, false negatives and false positives. These statistics were used to obtain false and true positive rates. TPR and FPR values were obtained for all thresholds for the image. The corresponding points were plotted with each other obtain the ROC curves. This was repeated for all images and algorithms. The results were plotted in two forms. Firstly, ROC for all the images per algorithm were plotted in the same window as shown in figure 2.

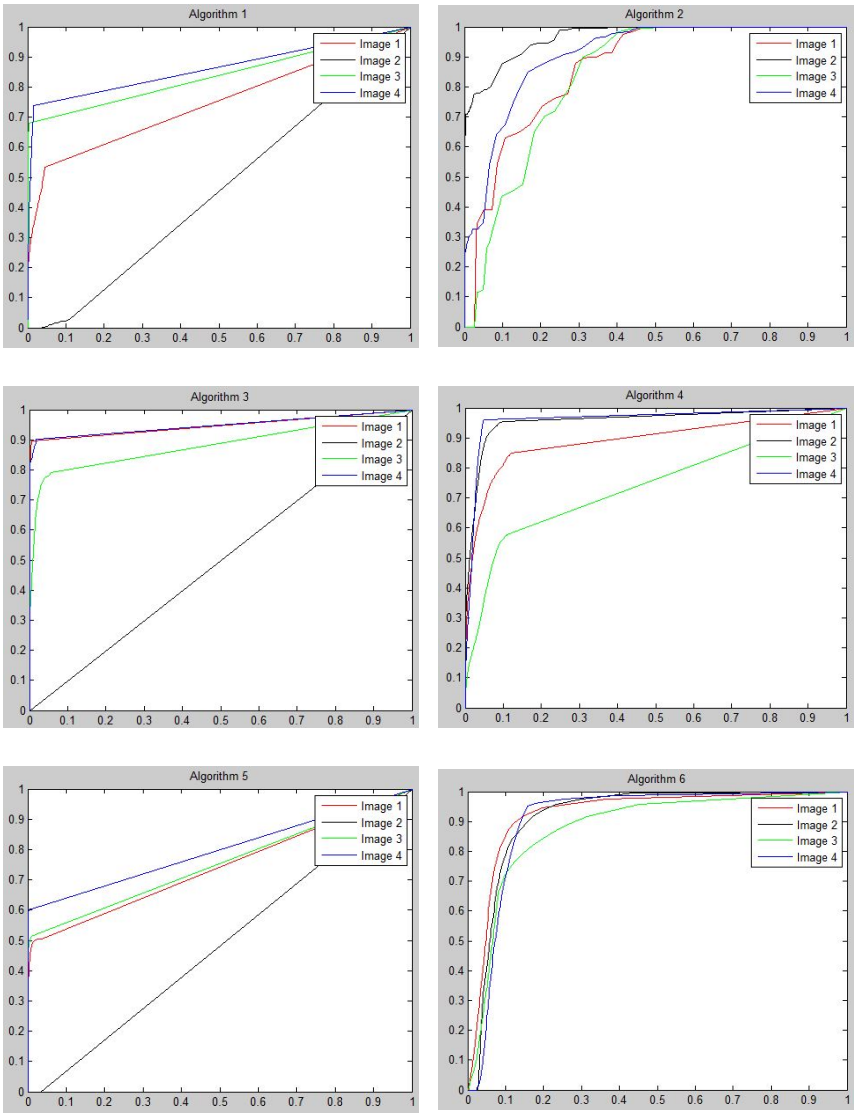

Figure 2: ROC curves (TPR vs FPR) LRUD: Algorithm 1, 2, 3, 4, 5 and 6 for each test image

This was done to compare which algorithms works better for which image. The results are indicated in the table below:-

Alogrithm	Algo 1	Algo 2	Algo 3	Algo 4	Algo 5	Algo 6
Best Image	Img 4	Img 2	Img 1 and 4	Img 2 and 4	Img 4	Img 1, 2 and 4

Table 1: Best image results for each algorithm

Secondly, ROC for all the algorithms per image were plotted in the same window as shown in figure 3.

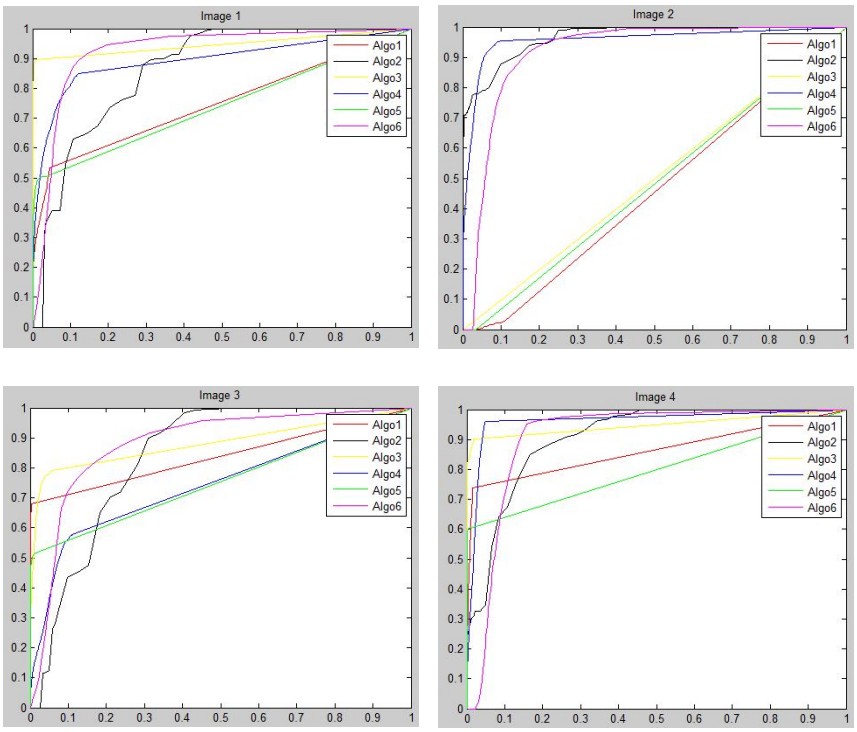

Figure 3: ROC curves (TPR vs FPR) LRUD: Test images 1, 2, 3 and 4 for each algorithm segmentation result

This was done to compare which image works better for which algorithm. The results are indicated in the table below:-

Image	Img 1	Img 2	Img 3	Img 4
Best Algorithm	Algo 3 and 6	Algo 2, 4 and 6	Algo 3 and 6	Algo 4

Table 2: Best algorithm results for each image

From the above ROCs, it can be easily seen that algorithm 6 works better than the others on most occasions while algorithms 1 and 5 work really poorly.

This was followed by the computation of the area under curve for each ROC. The average areas for each algorithm were computed and displayed in tabulated form in the table below.

Alogrithm	Algo 1	Algo 2	Algo 3	Algo 4	Algo 5	Algo 6
Area under Curve	0.7275	0.8940	0.8197	0.8894	0.6592	0.9086

Table 3: Average area under curve for each algorithm

The next was to extract the optimal threshold that can be used as a fixed parameter for practical applications as in practical demonstrations, it is not very convenient to see the image and then select the threshold for that particular case. Different option were available to determine the best threshold. Mean threshold was ultimately chosen. To determine the mean threshold, euclidean distance was calculated for each point in the ROC curve with the ideal point (0,1). The threshold that gave the minimum distance was kept. This was repeated for all the images and an average of the thresholds was characterized as the best threshold for that algorithm. The results are given in tabulated form in the table below.

Alogrithm	Algo 1	Algo 2	Algo 3	Algo 4	Algo 5	Algo 6
Optimal thresholds	2	90	4	7	16	7

Table 4: Average best thresholds for each algorithm

All the tasks in this section are programmed in the 'roc2D.m' script file.

3.2 Evaluation Measures

The results for the best thresholds computed in the previous section were used to determine different evaluation measures for each case. A point to note is that since the mean thresholds for each algorithm are used here, they do not produce the best result for each image. Rather they produce satisfactory results for all images combined. However, in two cases they produced a black image.

For the Jaccard index and the Dice similarity coefficients, area overlap measures were computed using the best thresholds. These results were input to the equations mentioned in figure 1 to compute the measures for each case. The results are shown in the tables below.

Algorithm—Image	Img 1	Img 2	Img 3	Img 4
Algo 1	0.0818	0.0010	0.5920	0.3695
Algo 2	0.0651	0.0180	0.0530	0.0578
Algo 3	0.3788	0	0.1995	0.3865
Algo 4	0.0565	0.0441	0.0875	0.2435
Algo 5	0.0937	0	0.2519	0.3710
Algo 6	0.0700	0.0289	0.0935	0.0956

Table 5: Jaccard index for each case

Algorithm—Image	Img 1	Img 2	Img 3	Img 4
Algo 1	0.1512	0.0020	0.7437	0.5396
Algo 2	0.1222	0.0353	0.1007	0.1092
Algo 3	0.5495	0	0.3327	0.5575
Algo 4	0.1069	0.0844	0.1609	0.3916
Algo 5	0.1714	0	0.4024	0.5412
Algo 6	0.1308	0.0562	.1710	0.1744

Table 6: Dice similarity coeffecients for each case

Apart from area measures, a distance measure known as harsdouf was also computed. An initial loop was run that extracted the boundaries from

the images. Then all the boundary points were compared with each other using the equations in figure 1 to compute the distance measure. However, this was taking too long to compute. So while computing the boundaries, a separate 2D array was created that contained all the coordinates of the boundary points. This reduced redundancy in the system and made the computation much faster.

Algorithm—Image	Img 1	Img 2	Img 3	Img 4
Algo 1	767	615	480	592
Algo 2	747	672	548	698
Algo 3	655	0 (no white part in image)	552	240
Algo 4	757	580	478	584
Algo 5	840	616	562	612
Algo 6	842	520	567	699

Table 7: Harsdouf distances for each case

The bad values in the above table are a reflection of the fact that the algorithms do not give results that are any close to the ideal ones. All the tasks in this section are programmed in the 'ddparams.m' script file.

4 3D Volume Segmentation Evaluation

In this section, the evaluation measures explained in the previous section were computed for the 3D volume. The results of which are displayed in the next table. For 3D case the entire volume was treated as one rather than treating every slice separately. This was achieved by simply adding a third dimension in the previous code. Since we were given an already thresholded image result, so ROC curve could not be built for this case.

Evaluation measure	Jaccard index	Dice similarity coefficient	Harsdouf distance
3D Volume 1	0.5277	0.6908	26.1916

Table 8: Harsdouf distances for each case

The evaluation results were better for 3D volume case since the algorithm output was in high resemblance with the ground truth provided. All the tasks in this section are programmed in the 'ddd.m' script file.

5 Conclusion

In this lab work different algorithm results were evaluated using different evaluation techniques. Both distance methods like harsdouf distance and area measures like jaccard index and dice similarity coefficient were computed. ROC curves were also used to compare the algorithms. These measure were computed for 24 different 2D cases and 1 3D case, the results of which have been displayed in tabular and graphical forms.