Berkeley Segmentation Dataset and Benchmark - Lab 7 - Computer Vision

Yesith Juez
University of the Andes
jy.juez@uniandes.edu.co

Nicolas Florez
University of the Andes
n.florez228@uniandes.edu.co

Abstract

Segmentation of objects present in images is a important computer vision objective as we can use it to identify objects in the images and then interpreting its content. BSDS or Berkeley segmentation datasets is an approach of a database that represents a set of natural images to test methods against human anotations to observe and evaluate the performance of the machines and its implemented algorithms. We used this database to compare the state of the art algorithm UCM against a implementation of kmeans and watersheds with different colorspaces. The best algorithm was watershed with an average performance of 0.38 classification accuracy. Still is a very poor result that is incomparable with the UCM algorithm. We could improve the results by using more information of the images as the texture.

1. Introduction

To obtain a quantitative evaluation method in applications with segmentation, the University of California created a database (BSDS500) that contained images of different natural scenes, where each image had annotations of different segmentations made by humans [1].

The objective of this laboratory is to use the BSDS500 to evaluate the two best segmentation methods performed by us, and compare them with the evaluation results of the UCM method [2]. Different functions in Matlab that approximate the fast version of the original BSDS Benchmark code will be used to obtain the recall-precision evaluation curves.

2. Materials and Methods

2.1. Description of the Database

The database used for the realization of this laboratory is a portion of the Berkeley Segmentation Dataset and Benchmark (BSDS500) is a database made by the Berkeley Computer Vision Group specifically to be used in the computer vision problems of Contour Detection and Image Segmentation.

Its a second version of a previous created one BSDS300 they used the 300 images of that database for training and validation and they added 200 more images with human annotations for testing.

Each image of the database is human annotated by 5 subjects in average.

The images are colored formated as JPEG that could be in landscape or portrait 481 x 381px.

Figure 1. 4 random example images of the database

2.2. Segmentation Methods

The segmentation methods selected were the k-means method (RGB space and Lab-XY space) and the watershed method in the RGB Space. This selection was made because the level of segmentation for a given hyperparameter was much better with respect to the other methods.

For the segmentation with the k-mean, the hyparameter to be established is the number of clusters (k). In the case of the watershed method it is necessary to establish the number of markers.

To choose the number of clusters in the k-mean method and the number of markers in the watershed method, one must perform trial and error tests, taking a specific range

for each hyperperameter, and executing the method for each value of the hyperparameter.

2.3. Test methodologies

The BSDS is a database that contains 500 images of natural scenes with different annotations made by humans. The images in JPG format are divided into three groups: training, evaluation and testing (200 training images, 200 evaluation images and 100 test images). Groundtruth are the respective annotations (segmentations made by humans) that are available for each image in the database.

If the coverage that has the method is plotted with respect to precision for each hyperparameter value, a curve called "precision-recall curve" is obtained, and it allows us to know how the evaluated method behaves in a more generalized manner. This form of evaluation is widely used in a standardized way to make comparisons with other different methods using the benchmark database.

As the name says it compare the precision with the recall. The precision represent the relation between the true positives and the true positives with the false positives. So its an indicator of the performance of the detections, when incrementing the false positives the precision lowers. If there are not false positives the precision must be 1. The recall represent the relation between the true positives and the true positives with the false negatives. Its an indicator of the performance of the detections when the algorithm missclassify data (false negatives). When the algorithm did not detect a positive detection the recall lowers.

The Benchmark code allows to obtain the precision-recall curves, by means of a set of matrices that represent the segmentations of the images for certain values of a hyperparameter.

For the Kmeans method, the number of clusters used was 5 (k = 3, k = 6, k = 9, k = 12 and k = 15). The selected k was used by the database in that we consider the mean amouunt of segemented obejcts in the iamges are 9. So we used k's up and below this number. the ammount of k was selected to just 5 in order to have the best computational time of calculation if we selected more k's the computer most surely will spend a lot of time. For the watershed method, the number of the height used was 5 (h = 90, h = 95, h = 100, h = 105 and h = 110), and was selected following the same method as kmeans finding the h that we thought was the best and selecting 2 heights upper and lower of it.

3. Results

As mentioned before to validate the different methods we calculated recall-precision curves. the general idea is to compare the methods of kmeans RGB, kmeans LAB+XY, watersheds RGB and the state of the art method overall UCM.

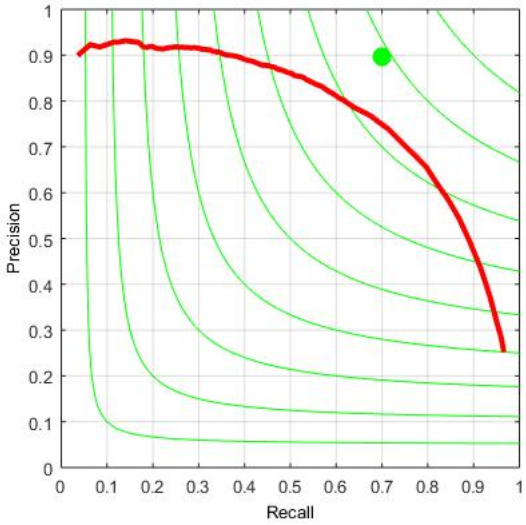

Figure 2. Precision - Recall curve from the UCM algorithm

The recall-precision curves for the UCM method is shown in the Figure 2 we are going to use this as the comparison method to observe the behaviour of our methds. The recall-precision curves of kmeans method for both train an test sets are shown in figures 3,4,5. Finally the curves for watersheds are shown in the figures 6 and 7. To compare all methods we collected every single curve of the test database in a single figure using different colors to diferenitate.

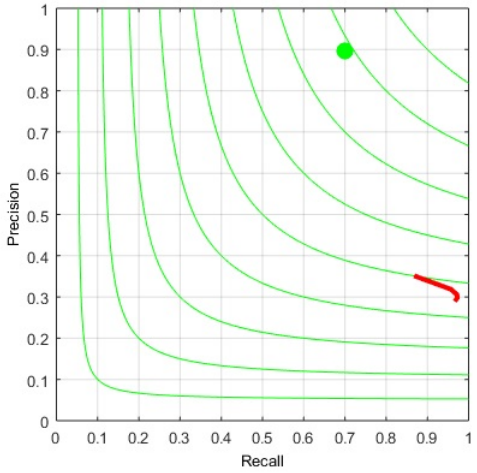

Figure 3. Val set Kmeans-RGB 3,6,9,12,15 Clusters

4. Discussion

The curves were calculated for the set of validation and test images. If we compare the validation set with the test set for each method, it can be deduced that the curves do not have any significant change.

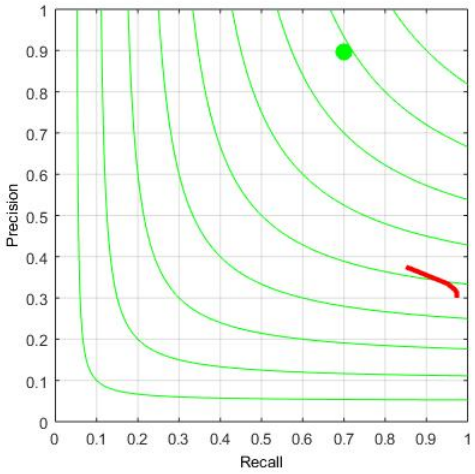

Figure 4. Test set Kmeans-RGB 3,6,9,12,15 Clusters

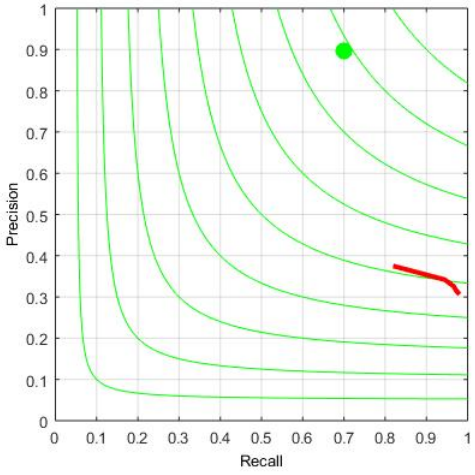

Figure 5. Test set Kmeans-LAB+XY 3,6,9,12,15 Clusters

When comparing those precision recall curves between the two methods(Fig 8), it is observed that the method of watershed is a little more precise than the kmeans, but the kmeans has higher recall. The curves of the methods evaluated are not very generalized; this occurs because of the following: Because of the selected k for the k means method in order for the curve to have a higher range of recall vs precision, the sweep of the clusters must be much greater.For the watershed method the case is similar to of the Kmeans method chaging the number of height of the watershed.

If we compare the segmentation methods performed with respect to the UCM method (see figure 1), we can conclude that the UCM method has a much greater recall. In addition, the performance found is far above that of the kmeans and watershed methods. Figure 7 shows all the segmentation methods in the training stage. In the recall-precision curve for the kmeans method, the change of color space

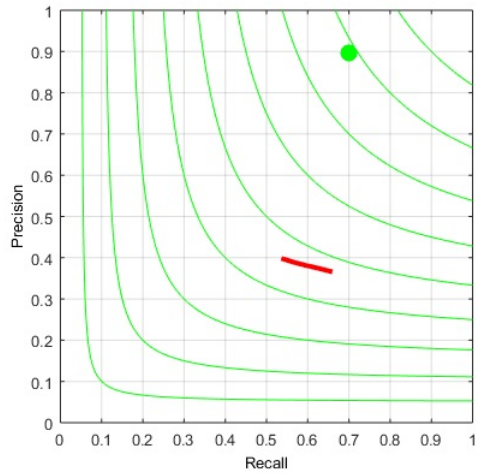

Figure 6. Val set WatershedsRGB 90,95,100,105,110 h

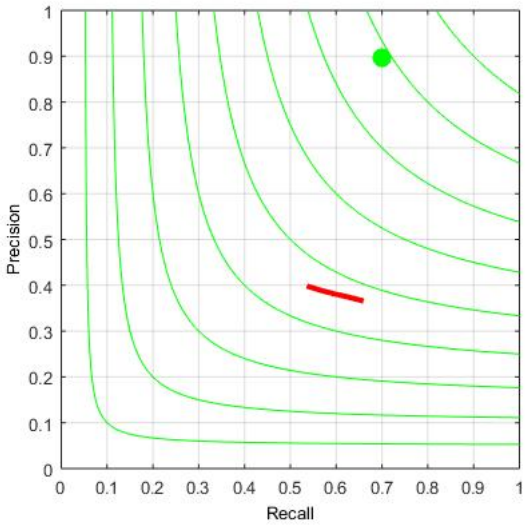

Figure 7. Train set WatershedsRGB 90,95,100,105,110 h

from RGB to LAB-XY does not matter much.

To do a comparison method we obtained the average performance given by each curve and with that we obtained that the methods of watershed and kmeans, have an average performance of watershed = 0.38 and kmeans = 0.35.

With the evaluation of the methods it can not be confirmed that the best segmentation method is still watershed, in the range of selected hyperparameters. This is because the method of kmeans has better recall (of almost 1), although the watershed does have better accuracy.

A possible option to improve the algorithms developed for the segmentation, could be using different types of representations of the image (Not only in spaces of color, but also take into account their shape and texture).Then, apply the clusters or markers to the combination of representa-

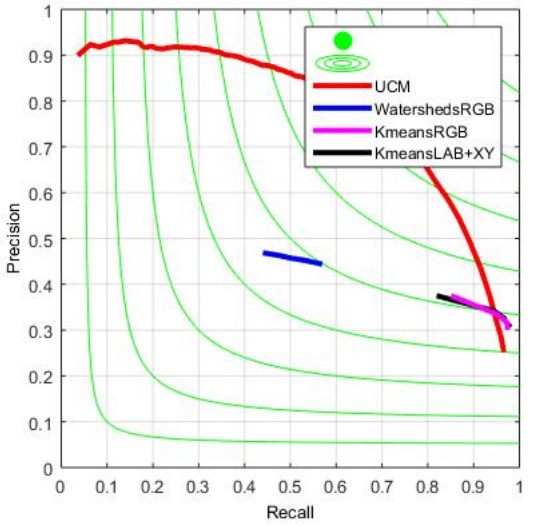

Figure 8. Train set ALL

tions made.

5. Conclusions

The methods of watershed and kmeans, have an average performance (watershed = 0.38 and kmeans = 0.35), so comparing it with the BSDS database are not good methods for image segmentation, taking into account only the range of hyperparameters used.

The BSDS database allows us to evaluate different methods of segmentation in a more generalized way, allowing at the same time to compare them with segmentation methods.

The high performance of the UCM segmentation method is verified, by comparing the best methods developed in the previous laboratory.

There is no significant difference between the recall-precision curves in the evaluation and test stages for each method.

The segmentation performance of the evaluated methods depends strictly on the defined hyperparameters. In order to observe the minimum error it would be necessary to use a very wide range of hyperparameters.

References

[1] David Martin, Charless Fowlkes, Doron Tal, Jitendra Malik, ”A Database of Human Segmented Images and its Applications to Evaluating Segmentation Algorithms and Measuring Ecological Statistics”, University of California, 2001.

[2] Pablo Arbelaez, ”Boundary Extraction in Natural Images Using Ultrametric Contour Maps”,Worshop on Perceptual Organization in Computer Vision, 2006.