Comparison of Segmentation Methods on Berkeley Segmentation Dataset

Natalia Andrea Durán Castro
Universidad de los Andes
na.duran@uniandes.edu.co

Ana M. Velosa Orduz
Universidad de los Andes
am.velosa@uniandes.edu.co

Abstract

This paper presents a comparison between gaussian models with different clusters in the segmentation with regard to the Berkeley data set and algorithms. It demonstrates that a big number of clusters is not accurate for the segmentation of this data base in a gaussian distribution.

1. Introduction

Image segmentation is one of the most important image processing problem because it's was wrongly defined because not even humans could agree on a unique segmentation of a single image. That problem its related with the level of details that a human gives to an image, and to sort that problem the ground truth it's a group of segmentation made by different subjects with different level of details.

To evaluate the performance of different methods of classifiers we used the precision-recall curve. This plot is based on two basic evaluation measures, recall it's the sensitivity of the method and precision is similar with positive predictive value[1]. The goal of this graph it compares the performance of different algorithms of classification observing how much the curve of me method approximate to 100 percent recall and precision.

Also, with the perdition-recall plot we can find the harmonic mean of precision and recall, that tells us the accuracy of the method we are evaluating. This harmonic mean will have different values depending on how many images are you evaluating your algorithm. For example, if you are evaluation you algorithm with just one image, OIS, will be bigger than evaluating the harmonic mean, ODS, with all of the images. This difference anyways should not be big, otherwise the method has problems.

Methods of segmentation

Three different methods of segmentation were evaluated: Watershed, Mixture of Gaussian's and K-means. Then, we realize that the best method was mixture of Gaussian because it presented an histogram with a nearly distribution to the groundTruth data. The way to evaluate this was without having account the correspondence between clusters,because these could be randomly selected due to the order change in every run. In this way, the contrast was developed having account the maximum and minimum cluster of the groundTruth and the segmentation made with gaussians of three images.

To evaluate the effectiveness of this method it was developed an algorithm to save three images segmented with the method of gaussians in a cell array .mat. These three methods presented a different number of clusters, the clusters chased were 60, 90 and 240. This process was developed with the training and test set. The average time to segment each image was three minutes.For this reason, this process was the longest part of the algorithm. After this, the segmented images were evaluated with the benchmarks functions included in the data-set of Berkeley Segmentation. These functions organize the data and compare five groundTruth segmentations with the three methods saved in the cell. With this information, the recall and precision curved is calculated. It is necessary that the cell array is in grayscale and with only positive integers. In other way, the metrics comparisson can not be developed. Finally, the evaluation return the curved of precision and recall with information about this in the command window. In the figure 1 is presented the different segmentation of gaussian with differents clusters.

For the evaluation of the method we choose we used the BSDS500 (The Berkeley Segmentation Dataset and Benchmark). That dataset its an extension of the BSDS300 that has 300 images more, where the original 200 images now are used for training and the 200 new are used for validation.Those were RGB and gray images with a 321x481x3 dimension organized in two separete groups, one for the RGB images and the other for gray images. Also inside this two groups they divided the images into smaller groups of 25 images each one.All of the this images have annotations from 5 diferent human subjects which makes more than 3269 annotations, and a benchmarking code to be able to compared diffent algorithms and get closer to human-level performance that can be tracked over time .That code

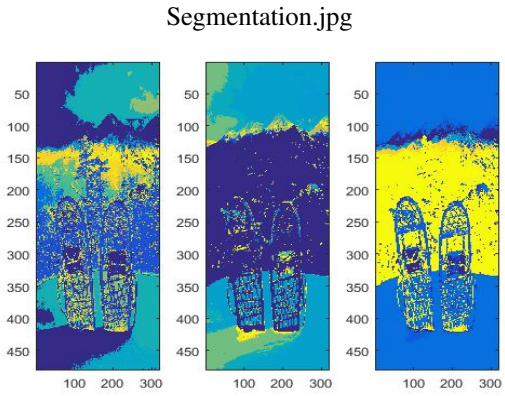

Figure 1. Segmentation with Gaussian Mixture model with k=60,k=90,k=240

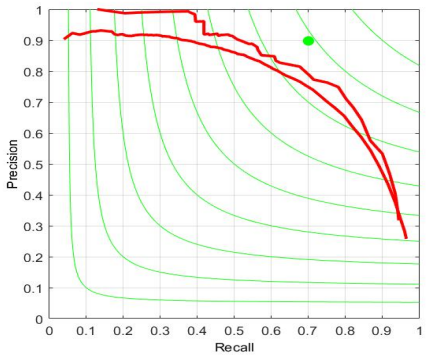

Figure 2. Precision recall of UCM method developed by Pablo Arvelaez. The upper line represents the result with few images of test and the other with all the images of the test set.

has available the UCM2 method made by Pablo Arvelaez to be compare with our methods. In the Figure 2 is presented the results of segmentation by the ultrametric contour maps. It can be apreciate a difference between the two curves, the upper has more pics because less images were used to developed it. On the other side, the other line includes the evaluation of all the test set. Due to the time of all the process, it was developed an algorithm to compare

2. Results

With our method base on Gaussian Mixture model using 60,90 and 240 Gaussian Mixtures we obtain the segmentation we can observe on 1. We can observe that the result seems to be runned with few images; however, it was realized with 132 images of the test set.

3. Evaluation of Results

When we evaluate out method in comparison with UCM we observed that our method we could observed that the performance of our method was worst as we can observed

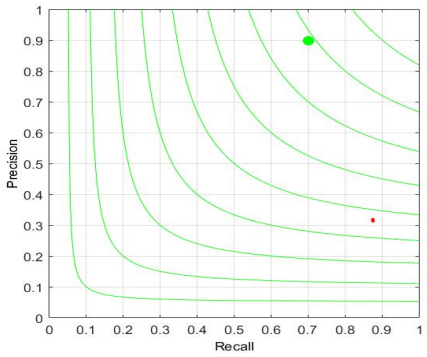

Figure 3. Precision-recall plot of our method and UCM

in 3, because the value we took as a basis for k was well above the value of elements that were in the ground truth of the image. Therefore, it can be said that our method is over-segmented and finds a much larger number of objects than what is found in reality.

Keep in mind that depending on the problem you are attacking one result may be better than another. If in our problem we need a lot of sensitivity and we are willing to sacrifice a little positive predictive value, UCM is much better, on the other hand if we want to have an excellent positive predictive value and sensitivity is not so important, our method is better. Equally, one must be careful about the problem of over-segmentation.

In addition to the graphs of precision and recall we also have the values of the harmonic average of our method . The ODS value of our method tells us that it is correctly segmenting less than half of the entertainment images and that our optimal k value is 60. On the other hand, because it gives a difference of the values between ODS and IDS does not vary much, it reiterates to us that the problem that our method has is related to the parameters that were chosen to make the models and not to the internal functioning of the same since it is behaving in a consistent manner.

4. Conclusions

Our biggest limitation when generating our algorithms was the processing time of each of them by image since it was not easy to interacturally quickly with the results obtained with each method. Considering the above, it was difficult to take an objective deisicon with respect to the relevant hyperparameters of the method and not to make mistakes of over-segmentation. To improve our algorithm it would be interesting before choosing the number of elements of an image to make a decision tree in order to take into account much more information before making decisions about the image. Having the above mentioned in the same way it would be possible to reduce the processing time of each one of the images.

References

[1] RUI. Introduction to the precision-recall plot. último acceso 17 Marzo 2019. 1