Berkeley Segmentation Dataset and Benchmark

Maria Camila Escobar
Universidad de los Andes
Bogotá D.C, Colombia
mc.escobar11@uniandes.edu.co

Laura Gongas
Universidad de los Andes
Bogotá D.C, Colombia
l.gongas10@uniandes.edu.co

***Abstract*- The Berkeley Segmentation Dataset and Benchmarks 500 (BSDS500) is a dataset that transformed the research field of image segmentation and boundary detection. The change in this field was due to a set of images and ground-truths that allowed researchers to expand their analysis and the implementation of a standarized metric, called the precision-recall (PR) curve, so that all methods could be evaluated equally. Here we implement our previously developed segmentation method in the BSDS500 and evaluate the results through a PR curve and three other segmentation metrics.**

1. Introduction

Segmentation can be considered the first step in most of computer vision applications. For example, in order to create a self driving car that detects other vehicles and pedestrians on the streets it is necessary to first develop an algorithm that will be able to segment the pedestrians from the rest of the objects. Segmentation is the process of partitioning an image into different regions based on different features such as space of color, brightness, texture or spatial coordinates [5].

There can be different approaches to segmentation depending on the method that is used for grouping pixels and the features that such grouping will be based upon. We previously tested an algorithm for four different clustering methods in six possible feature spaces and found that the best results were obtained with kmeans in rgb colorspace and gaussian mixture model (gmm) in Lab colorspace with spatial coordinates [3]. The criteria to determine the best configuration were the Best covering criteria given by the evaluation method and the time that the algorithm took running.The results of this methods can be compared with the other possible clustering methods in figure 1, it is possible to qualitatively identify that kmeans and gmm gave the best results. Another parameter that was taken into account when selecting the methods was that they took significantly less time running, since the BSDS500 has around 200 images per category time was considered an important factor.

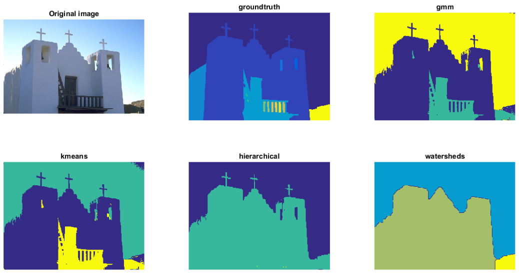

Figure 1. Comparison of segmentation methods with original image and groundtruth

K means is a partitional algorithm that assumes that the number of clusters is known. The main idea of the algorithm is to represent clusters with their centroids and find the centroids and clusters that minimize squared distances between elements and centroids. Lloyd's algorithm is an approximation that consists of: choosing k random centroids, assigning elements to clusters using a Voronoi diagram, computing new centroids and iterating until convergence is reached. The convergence is defined when there's no reassignments of elements to other clusters, no variance of centroids or when there's a minimum decrease in the sum of squared error [1].

Other algorithms of partitional clustering are model based like mixture of gaussians (gmm). The main idea of this method is to represent groups with gaussian distributions. In other words, clusters are formed by describing the probability density function of variables by a mixture of multivariate normal densities. Therefore, the objective is to find the parameters of a mixture of Gaussians that explain best the data [1].

The Berkeley Segmentation Dataset and Benchmarks 500 (BSDS500) is a dataset created by the Computer Vision group at Berkeley university. The goal of the BSDS500 is to provide a basis for research on image segmentation and boundary detection [2]. Before the BSDS was introduced, segmentation and boundary detection were thought to be ill-posed problems, since every human can create a different segmentation of the same image. However, the data in the BSDS includes ground-truth segmentations from various humans and it was found that, even though the ground-truth segmentations were not exactly the same for each case, there were regions that all ground-truths had in common.

Another contribution to research made by the BSDS was the creation of a standardized metric for judging the quality of a segmentation method. Martin et. al formulated boundary detection as a classification problem where the goal is to tell apart boundary pixels from non boundary pixels. A precision-recall curve was then made for varying hyperparameters of the algorithm and the maximum F-measure $(\frac{2\times Precision\times Recall}{Precision\times Recall})$ was calculated with respect to human ground-truth boundaries [4]. The ideal result of a precision-recall curve would be of a 1 in precision and 1 in recall, this would mean that the segmentation predicted every border in the ground-truth. However, given the nature of the segmentation problem, the consistency of the humans segmentation was calculated as 0.8 so this would be the best result that a segmentation algorithm could aim for.

The objective of this paper is to use two segmentation methods: kmeans and gmm with features in different colorspaces: rbg and lab on the BSDS500 and analyze the results through a PR curve and other segmentation metrics.

2. Materials and methods

First, the BSDS500 dataset consists of 300 training and validation images and 200 test images. The images are in jpg format and there are some in horizontal orientation (481 x 321) and others in vertical orientation (321x481). The dataset comes with a ground-truth for each image in .mat format, inside the ground-truth there are segmentations performed by five different subjects. Performance is evaluated by measuring Precision-Recall on detected boundaries, it also uses covering of ground-truth segmentations, Probabilistic Rand Index and Variation of Information benchmarks.

The development of this method is divided in three parts. In the first part, the algorithm made in the previous lab was tested for the train dataset. This was done in order to select the K number of clusters that would yield the best result and give a longer PR curve with better recall and precision. Once the number of clusters was determined, we ran the test dataset with our two best configuration. For each of these we obtained the metrics of the evaluation method and compared them with each other and with the Ultrametric Contour Map segmentation algorithm.

3. Results

First, two range of K clusters were selected. The first was from the range of clusters that was possible to identify analytically, consisting of a range from 2 to 10 clusters per segmentation. Next, we tried a different range of clusters in order to see if we could get a better coverage in the PR curve, for the second range we selected five clusters [2,6,10,20,50]. We ran both K configurations with kmeans and the results can be seen in figure 2. Here we found that the configuration of 2 to 10 clusters worked better than the other range, this because it shows a better precision with the same recall. Additionally, the 5 clusters configuration did not widen the range of the PR curve but gave worst results both in precision and recall for certain cluster numbers.

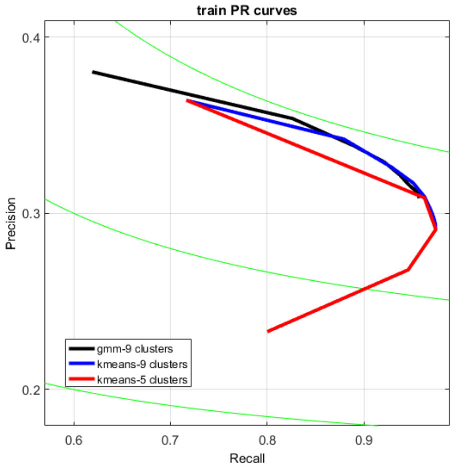

Figure 2. Precision-recall curve for the train dataset with different configurations

This result can be seen qualitatively in figures 3 and 4. In figure 3 it is possible to see the best and the worst segmentation results for the configuration of 9 possible cluster ranges. Here we can identify that the result for 3 clusters works good taking into account the groundtruth. On the other hand, figure 4 shows the best and worst configurations for the second range of possible clusters. Here we see that the selection of 50 clusters for kmeans will not give a result in any way similar to the one in the groundtruth, hence the low recall and precision for some ranges in that curve. Taking this into consideration, we decided that the best cluster range was the one from 2 to 10 and we calculated the results for gmm with this range as well (figure 2).

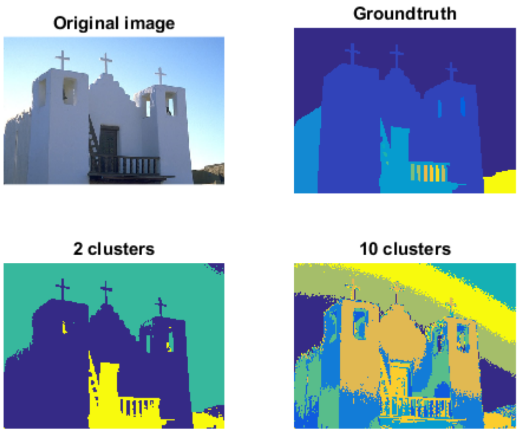

Figure 3. Best and worst segmentations for the chosen cluster range

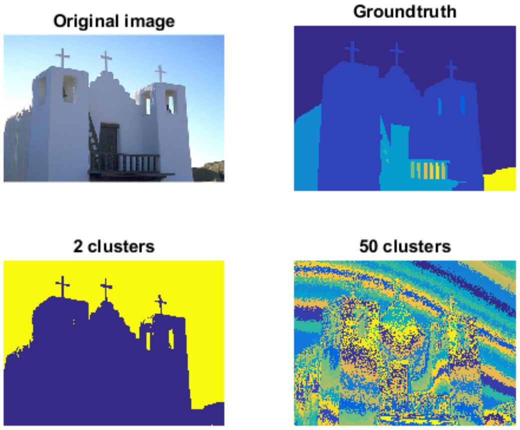

Figure 4. Best and worst segmentations for the cluster range that was not chosen

Once the range of clusters was selected we evaluated our best methods in the BSDS500 test datasets. The results for PR-curves can be seen in figure 5 and table 1 has the F-Measure for the Optimal data scale (ODS), Optimal image scale (OIS) and Area below each curve. Our performance is also compared with the ucm method in order to analyze how are our methods performing in the real segmentation field, the results of PR curve including ucm can be seen in figure 6. Also, since our method consists of segmentations and not border detections we tought necessary to obtain as well the results for the Covering, Probability of rand index (PRI) and variation of information (VI) in order to obtain a better analysis. The results for the segmentation metrics can be found in table 2.

Table 1. F-Measure for test data

	ODS	**OIS**	**AP**
gmm	0.52	0.56	0.14
kmeans	0.52	0.53	0.10
ucm	0.73	0.76	0.73

Table 2. Segmentation metrics for test data

	Covering			**PRI**		**VI**	
	ODS	**OIS**	**Best**	**ODS**	**OIS**	**ODS**	**OIS**
gmm	0.35	0.38	0.40	0.69	0.71	2.64	2.62
kmeans	0.35	0.38	0.40	0.70	0.72	2.67	2.66
ucm	0.59	0.65	0.74	0.83	0.86	1.69	1.48

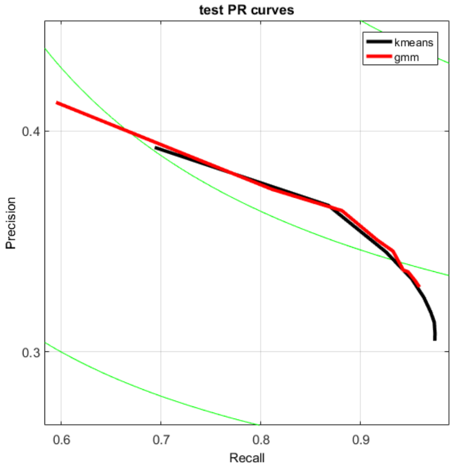

Figure 5. Precision-recall curve for the test dataset with kmeans and gmm

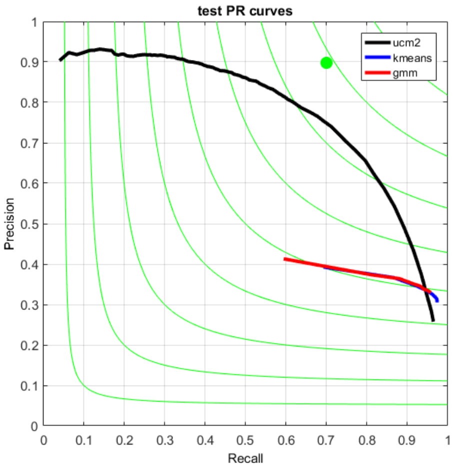

Figure 6. Precision-recall curve for the test dataset with kmeans,gmm and ucm

4. Conclusions

There are various metrics to compare the results we obtained and not all of them are consistent in selecting the best method. For instance, in table 1 gmm has a better F-Measure than kmeans in the optimal image scale. However, since the developed methods produce segmentations and not borders we find it relevant to analyze the results of the segmentation metrics in table 2. Here we find that kmeans has a better result in PRI. A bigger result in VI means that the clusters produced by the method are more distant in entropy to the ones in the groundtruth. Nonetheless VI is not adapted to work with multiple groundtruths like the ones given in this dataset therefore the results of covering and PRI are considered to be more trustworthy.However, since there is not a significant difference between the PRI results of kmeans and gmm it is possible to take only into account the results of the F-Measure and the PR curve. Also, it is valid to analyze the OIS results since it gives the metric for the best cluster configuration in each image and since methods like ucm and others do not have a predefined number of clusters it puts the methods on the same ground.

Overall, the best performance between our two methods corresponds to gmm. This could be attributed to the fact that kmeans uses a hard assignment to each value while gmm takes into account variance and therefore assigns probabilities to belong to a certain cluster.

Finally, when comparing our method with ucm it is possible to see there is still a vast room for improvement in our algorithm. The main limitation with our method is the fact that the clusters have to be selected manually because it can have bias from the person that is selecting the cluster range. Therefore, an improvement for our method could be to choose a number of clusters automatically depending on the image.

References

[1] P. Arbelaez. Lecture 5: Clustering. *Universidad de los Andes*, 2018.

[2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):898–916, May 2011.

[3] M. C. Escobar and L. Gongas. *Segmentation*. Computer Vision. 2018.

[4] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE transactions on pattern analysis and machine intelligence*, 26(5):530–549, 2004.

[5] N. R. Pal and S. K. Pal. A review on image segmentation techniques. *Pattern recognition*, 26(9):1277–1294, 1993.