BSDS

Federico Ulloa
Universidad De Los Andes
Cra. 1 18a-12, Bogotá, Colombia
f.ulloa10@uniandes.edu.co

Esteban Vargas
Universidad De Los Andes
Cra. 1 18a-12, Bogotá, Colombia
e.vargas11@uniandes.edu.co

Abstract

*In this laboratory, our two best segmentation methods were chosen, to run the complete BSDS500 dataset. We chose K-means and GMM in La*b* color space to perform segmentation. The results show that GMM works better, as expected, but both methods have limitations like ignoring different features and having no pre-processing. In conclusion, thse methods give an acceptable segmentation based in the benchmark evaluation of the database but to obtain much better results, other more refined algorithms must be done as UCM, for instance.*

1.. Introduction

One of the main problems in segmentation, is the evaluation of the methods.[1] This is because ground truths vary depending on the person that makes the annotation. Hence, same images can have several different ground truths, which was said to mean that segmentation cannot be successfully solved, as not even humans can obtain the same results.

The Berkeley Segmentation Dataset and Benchmark (BSDS), created a generalized evaluation method to solve this problem. [2] Basically, the collection of the different human segmentations makes up the ground truth of each image and they developed a comparison algorithm that is very general for all applications possible for the dataset. Also, they came up with a metric to evaluate how good humans perform in segmentation, as it was mentioned that different results are obtain among several people, called the Berkeley Benchmark.A problem in evaluation for segmentation existed as several ground truths exist, and it would be wrong in principle to evaluate the performance of an algorithm based in a biased ground truth made by one human.

The Berkeley benchmark consists in fusing the different ground truths into one general ground truth and with this generating a precision-recall curve. Precision being the probability of a boundary pixel generated by the algorithm is a true boundary, and recall is how much ground truth is detected. [2]

This way, different methods for different applications can be evaluated in this same dataset with its benchmark in order to have comparable results between algorithms and so that progress can be tracked with respect to human accuracy. [2]

2.. Materials ans methods

Segmentation can be done through many different methods. In the previous laboratory, a segmentation method was built with different clustering methods in different color spaces. After running experiments with these different variants, we concluded that the method with better results were K-means and GMM clustering in La*b* color space. Hierarchical clustering also gave good results, but it was discarded as it is too expensive computationally. On the other hand, watersheds was also discarded as its performance overall gave more error than K-means and GMM.

The dataset used was the complete BSDS500 dataset, which includes 200 images for training and test set and 100 images for validation set. All the images have a size of 481 x 321 or 321 x 481, both portrait and landscape orientations. The dataset contained images of natural landscapes, different vehicles, animals, varied people making different activities and even buildings. In other words, it is a very varied dataset with typical images taken from regular people. It was organized pretty randomly among these types of images.

As mentioned, the 2 chosen methods were K-means and GMM in La*b* color space. The K-means clustering method consists in starting with k centroids located randomly in the data. Then, each data is assigned in the cluster of the closest centroid. The centroid of this new cluster is recalculated and this process repeats until convergence. The GMM clustering is a more generalized K-means, in which it tries to fit the data into a mixture of gaussians.

To evaluate these two methods, they were run in the BSDS dataset. The training phase was done in the training images and the test phase was done in the test set. First, to check the correct functioning of the algorithm, all this was made in a random sub-sample of the dataset of 80 training images and 80 test images. Then, when results were checked, the same methods were run in the complete dataset. To build the Precision-Recall curves, a range of 2-40 for every second k was tested.

3.. Results

First, after running the methods in the sub-sampled databese, Precision-Recall curves obtained are shown in figures 1 and 2. As shown, GMM gives better results than K-means, but the curves are not well defined, as the dataset was sub-sampled. When thes reults were obtained, we proceeded to run the experiments with the same algorithm among the full dataset.

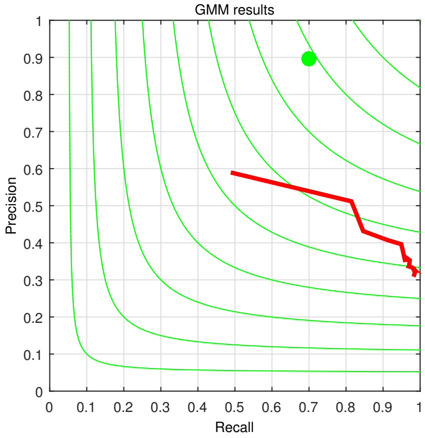

Figura 1: Results of GMM in sub-sampled dataset

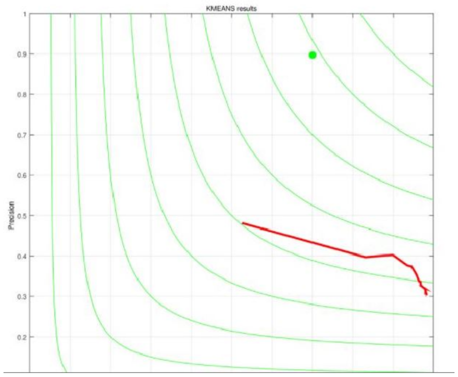

Figura 2: Results of k-means in sub-sampled dataset

Figures 3 and 4 show the Precision-recall curves for the same methods in the complete dataset. Again, as in the sub-sampled dataset and the previous laboratory, GMM gives better results. Note that the curve is lower in the full dataset, which is expectes as more images are present, so more variability is introduced. GMM woks better because, even though it is very similar to K-means, the data fitting into gaussian mixtures makes it more refined, despite it assumes normal distribution.

Lastly, it was found that the best performance of our methods were GMM with k=5 and K-means with k=4. This means that with these k's, the distance frm the curve to the ideal result (the green dot which is human performance) is minimum. Examples for these segmentations are shown in figures 5 and 6.

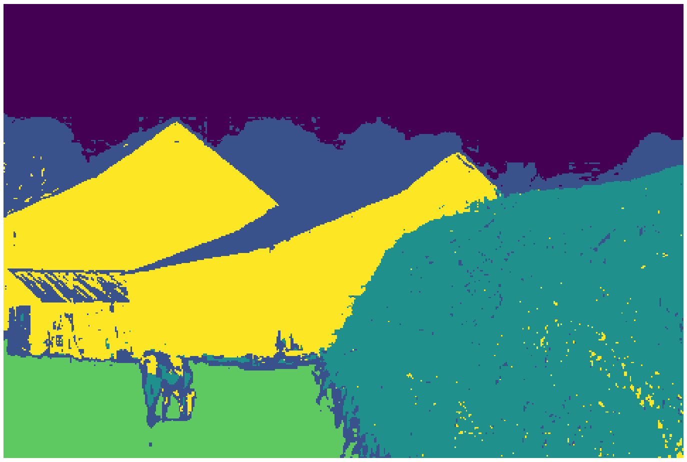

Figura 3: GMM segmentation with k=5

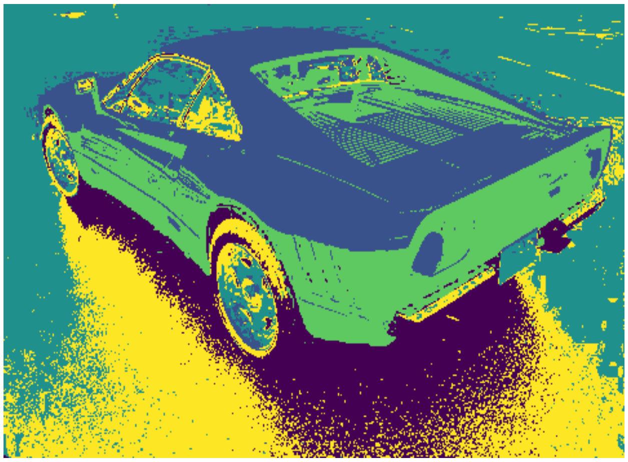

Figura 4: K-means segmentation with k=4

Clearly, this method has several limitations. This is because images have no pre-processing and segmentation is performed with the most basic clustering methods. This segmentation methods are nowhere near to the segmentation method made by Pablo and his team, obviously. Mainly, the most relevant limitation in these clustering methods are mainly color based and other relevant features are not taken into account.

To improve the methods, pre-processing is recommended. Another important note is that images were not re-sized and this also made the algorithm to take a very long time to run (more than 30 hours). Other AI techniques could work a lot better, SVM for instance. With a good re-size and pre-processing of the images, and taking different features to train an SVM, the algorithms could give way better results.

4.. Conclusions

In conclusion, the two chosen methods give decent results for the BSDS500 dataset in segmentation. Specifically, GMM performs better than K-means in La*b* color space, as it is a similar but more refined method.

Though, compared to more elaborate methods like Pablo's UCM, these algorithms show their big limitations. These limitations include the few features taken into account, the lack of pre-processing of the images, and the fact that these clustering methods are the most basic ones and the dataset has annotations. For this reason, other features shoul be taken into account for better results, like contour detection, as it is done in the UCM algorithm. [3]

Referencias

[1] "Performance Evaluation of Image Segmentation", Pdfs.semanticscholar.org. [Online]. Available: https://pdfs.semanticscholar.org/a5e5/d7f10980806f2528417c780665ed76fcc247.pdf.

[2] The Berkeley Segmentation Dataset and Benchmark. (2007). Www2.eecs.berkeley.edu. Retrieved 14 March 2018, from https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/

[3] Contour Detection and Hierarchical Image Segmentation. P. Arbelaez, M. Maire, C. Fowlkes and J. Malik. IEEE TPAMI, Vol. 33, No. 5, pp. 898-916, May 2011.