07: Berkeley Segmentation Dataset and Benchmark

Sergio Steven Leal Cuellar* and Mateo Rueda Molano†,
Department of Biomedical Engineering, University of the Andes, Bogota, Colombia
Email: *ss.leal10@uniandes.edu.co, †ms.rueda10@uniandes.edu.co,

Abstract—**The objective of this work is to compare methods of segmentation by clustering previously found as better for the segmentation of the Berkeley Segmentation Dataseten database [1] with the work and algorithm of Contour Detection and Hierarchical Image Segmentation developed by Arbelaez et al. Berkeley researchers who developed this method to improve the segmentation of the Berkeley BSDS500 database [3]. The evaluation and subsequent comparison takes into account in most part the accuracy-recall curve of our algorithms versus those developed by the Berkeley researchers. Likewise, metrics such as Optimal Dataset Scale (ODS), Optimal Image Scale (OIS) and Average Precision (AP) will be taken into account for the comparison of segmentation performance. Finally, mechanisms will be provided to improve the performance of our algorithm against what was proposed by the Berkeley researchers.**

I. Introduction

Image segmentation is widely used in medical image processing, face recognition, pedestrian detection, iris recognition, video suirvellance, etc. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze [4]. Segmentation based on clustering is an important tool for many researches. There are many different methods but in this article we will emphasize in Gaussian Mixture Models and hierarchical clustering. Gaussian mixture models involve the usage of a Mahalanobis distance to relate the samples with their cluster centers instead of using their nearest neighbors. In addition, a group of samples are trying to be described with Gaussian distribution and so the whole data can be described with a mixture of Gaussian distributions. As a consequence, a soft assignment is done for every cluster and responsibilities are calculated for each group. Finally, the parameters of the Gaussian distribution are being re calculated taking into account the responsibilities previously assigned for each group [2]. Hierarchical clustering implies the generation of very small clusters for similar data at first and then generate bigger clusters conformed by the association of smaller clusters. With that sense, a set of clusters of different sizes made up from other clusters is obtained [6].

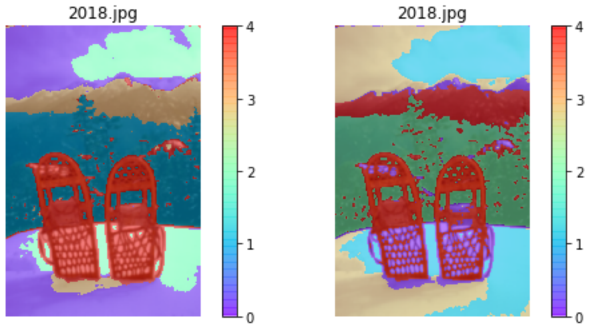
In 2010, a segmentation algorithm developed by Arbeláez, Maire, Fowlkes and Malik emerged which was more robust in BSDS500 The Berkeley Segmentation Dataset than clustering-based segmentation methods. This algorithm was based on a higher performance in contour detection combining both global and local information. In addition,they performed a method to transform the contour signal into hierarchical regions preserving the quality of the contours themselves. Regarding the contour detector, gPb was used to gather multi-scale information of color, brightness and texture to a powerful globalization framework using spectral clustering. The local cues, computed by applying oriented gradient operators at every location in the image, defines an affinity matrix representing the similarity between pixels. To produce the proper segmentation of images, the algorithm combined the contour detector with a generic grouping algorithm. For the grouping, an image transformation called Oriented Watershed Transform was used or constructing a set of initial regions from an oriented contour signal. Second, using an agglomerative clustering procedure, they form these regions into a hierarchy which can be represented by an Ultrametric Contour Map, the actual-valued image obtained by weighting each boundary by its scale of dissemination [3]

In this report we will present the performance differences in regions segmentation of some clustering algorithms found more efficient for the Berkeley BSDS500 database in previous studies [1] specifically Gaussian Mixture and Hierarchical clustering. These algorithms will be compared with the Contour Detection and Hierarchical Image Segmentation algorithm of the previously mentioned authors. The performance comparison will be evaluated from the precision-recall curve.

II. materials and methods

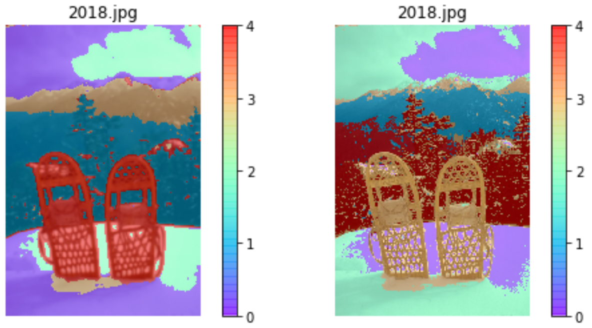
The Berkeley Segmentation Dataset and Benchmark BSDS500 is a large dataset of natural images that have been manually segmented by five different subjects on average. The dataset consists of 500 natural images, ground-truth human annotations and benchmarking code. The data is explicitly separated into disjoint train, validation and test subsets. Train subset has 200 images, validation set has 100 images and test subset has 200 images. Until now, performance has been evaluated by measuring Precision / Recall on detected boundaries and three additional region-based metrics [7].

A. Methods description



(a) Lab color space segmentation with GMM model (b) RGB color space segmentation with GMM model

Figure 1. Comparison of the segmentation between the Lab color and RGB color spaces



(a) Gaussian filtering segmentation with GMM model (b) No Gaussian filtering segmentation with GMM model

Figure 2. Comparison of the segmentation with and without gaussian filtering

The methods we selected in order to develop the evaluation and compare with the Berkeley benchmark were Gaussian Mixtures and Hierarchical clustering both with the Lab color space. The reason to chose Lab as the color space for both clustering methods was because of the chromatic information that it provides due to it's a and b channels. In addition, it also provides relevant information contained in the L channel related with the gray intensity of the objects. The other possible options for the color space were RGB and HSV, RGB was quickly discarded because of it's limitation, it only discriminates objects by their color. Also results of the segmentation had a lower quality than in the lab color space (see figure 1). HSV was the second best option due to the fact that it has gray scale information channel (V), chromatic information (H) and a third one related to the saturation (S) which may provide a scaled distance from white to pure colors. Nevertheless, Lab was chosen over HSV supported with the results previously obtained in [1] which determined that the Lab color space had slightly better results than HSV. The clustering methods selected were Gaussian mixture (GMM) and Hierarchical clustering. Gaussian mixture was selected because clusters are not restricted to have a spherical shape and also because it provides a soft assignment for every cluster. In addition, Agglomerative Clustering was used because it produces a hierarchical structure which may adapt better to the data. Additionally, both clustering methods gave overall better results n the previously done experiments over kmeans and watersheds [1].

The methods found as better in the previous laboratory, a Gaussian filter was added to soften the image. The filter was implemented because we found that with certain numbers of clusters the image was over segmented due to artifacts and texture changes inside the objects without the process of filtering. This can be better visualized in the figure 2 that compares the segmentation using gaussian filtering and without using it.

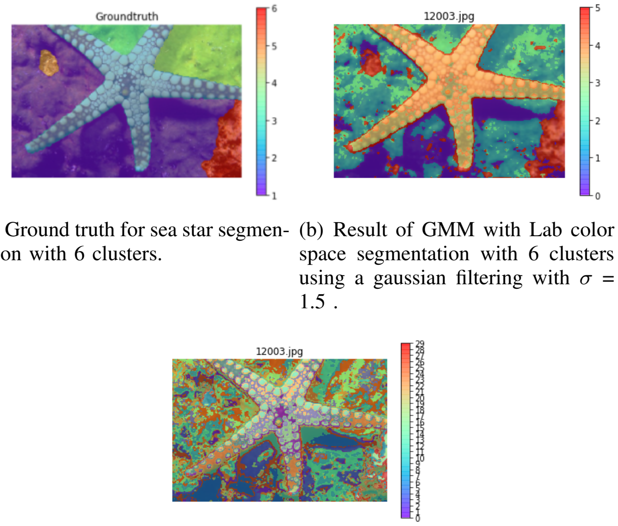
Regarding the Gaussian Mixture model, the most important parameters consist of: the number of neighbors, the type of covariance, the maximum number of iterations and the tolerance [8]. The number of neighbors was iterated when testing the segmentation of the BSDS500 database due to the fact that we wanted to be able to represent the largest possible portion of the precision and recall curve when evaluating the performance of our algorithm. The type of covariance was set to full because each component (K) has a different covariance than another component or cluster. Finally, the number of iterations was set to 100 with a tolerance of 1e-03.

On the other hand, the hierarchical method has as more important parameters the number of clusters, the type of distance between each one of them and the criteria for the definition of each cluster (linkage). As it was mentioned above the number of clusters was iterated as much as possible, given the computational constraints when testing clustering algorithms with a very high number of clusters (K) [9]. The iteration of K was in order to cover a large area in the curve of precision and recall. The type of distance and the linkage were previously tested in [1] and those that presented a better performance in the metric used in this study were used. Thus, the type of distance used was Manhattan and the linkage used was complete, this one separates the clusters by maximum distances between the observations of the two sets. Finally, the number of clusters was iterated in a K vector of [20,15,10,5,3] to know the different behaviors in the precision and recall curve in both methods. Cluster numbers were not expanded due to computational time that generate high K values

Precision-Recall curve is a useful measure of success of prediction when the classes are very imbalanced. In segmentation tasks, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are segmented. The precision-recall curve shows the tradeoff between precision and recall for different threshold (in our algorithms the number of clusters). A system with high recall but low precision returns many segmentation labels, but most of its predicted labels are incorrect when compared to the training labels. A system with high precision but low recall is just the opposite, returning very few segmentation labels, but most of its predicted labels are correct when compared to the training labels [10]. An ideal system with high precision and high recall will return many segmentation labels, with all results labeled correctly. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate segmentations (high precision), as well as returning a majority of all positive segmentation labels in comparison to the groundtruth annotations (high recall).

The methods of evaluation in Arbelaez et al. work were ODS corresponds Optimal Dataset Scale (ODS) or best F-measure on the dataset for a fixed scale, the Optimal Image Scale (OIS) or aggregate F-measure on the dataset for the best scale in each image, and the Average Precision (AP) on the full recall range (equivalently, the area under the precision-recall curve) [3]. Aditionally, F-max corresponds to the minimum distance between a point in the algorithm's curve and the point [1,1] in the precision-recall curve We will discuss this evaluation metrics and compare them between our clustering algorithms and the Contour Detection and Hierarchical Image Segmentation.

III. Results

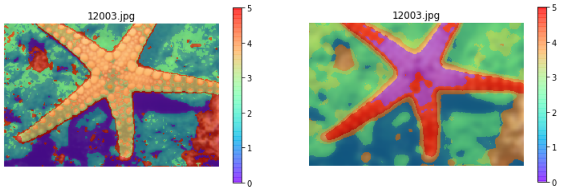


(a) Ground truth for sea star segmentation with 6 clusters.

(b) Result of GMM with Lab color space segmentation with 6 clusters using a gaussian filtering with σ = 1.5 .

(c) Result of GMM with Lab color space segmentation for 30 clusters using a gaussian filtering with σ = 1.5 .

Figure 3. Grountruth and comparison of the segmentation with Gmm clustering in Lab for different number of clusters



(a) Result of GMM with Lab color space segmentation with 6 clusters using a gaussian filtering with σ = 1,5 .

(b) Result of GMM with Lab color space segmentation with 6 clusters using a gaussian filtering with σ = 5 .

Figure 4. Comparison of the segmentation with Gmm clustering in Lab for different σ.

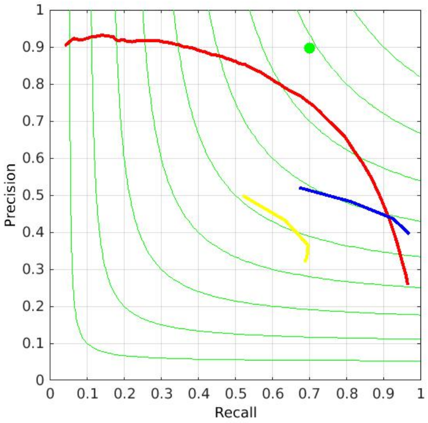


Figure 5. Results of the UCM, GMM and Hierarchical methods presented in red, blue and yellow respectivelly.

Method | Criteria	ODS	OIS	AreaPR	GtCov
UCM	0.73	0.76	0.73	0.74
GMM	0.61	0.64	0.14	0.52
Hierarchical	0.52	0,55	0.08	0.42

Table I
Boundary and region results for UCM, GMM and Hierarchical methods.

IV. Discussion

From the figure 5 can be determined that the method which worked better was the method of Gaussian mixture using the Lab color space, which also happened for the results obtained in the last practice (for more details check [1]). This can be explained because of Gaussian mixture enables to establish non-spherical clusters around the data. Because of that, Gaussian mixture better fits to the data avoiding to restrict the data to be grouped up totally spherical which rarely occurs in real applications. On the other hand, agglomerative clustering presented similar behavior for the precision on the segmentation. Nevertheless, for the k set (3, 5, 10, 15, 20) agglomerative clustering presented important shortcomings related to the recall compared to the Gmm method. However, this diagnosis is based on the behavior of more than the left of the coverage accuracy curve with considerably low Ks. For later work, a test with larger cluster numbers can be performed to see the total behavior of the curve and draw better conclusions.

In contrast, the UCM method presented overall best results both for precision and recall, as a consequence to better F-max. This can be explained because of how every algorithm is designed. First, UCM method is specialized in detecting contours for objects due to the comparisons of local histograms every part in the image and generating saliency for borders with a frequent appearance on the hierarchical clustering. On the other hard, both Gmm and hierarchical are grouping up pixels based on the chromatic or gray scale information totally omitting spacial information. Also, no histogram information is used to compare across local information of pixel related with it's neighbourhood and it's values as it's done in UCM. Likewise, the representation of edges in [3] has greater representation than our algorithm of clustering based on intensities. Researchers consider brightness, color and texture as representation spaces that can represent a greater number of visual characteristics of objects

Returning to the figure 5, the GMM and hierarchical algorithms present considerably low precision because of the method itself. Both GMM and agglomerative clustering have considerable limitations in discriminating pretty similar objects, which is supported with the low effectiveness that presented the method on the last evaluation method. This limitations affects the precision of the method because the algorithm isn't able to effectively classify the pixels that belong to a specific cluster (or object) and causes that most of the time pixels are being miss classified. In fact, for the GMM + Lab method the mean precision is around 0.5, it means that with a probability of 50% a pixel from the recovered ones is going to be incorrectly classified and labeled wrong. And for the Hierarchical method the average precision is near 0.4 (with a even low recall) which means that just 4 of every 10 pixels will be properly classified.

For the method of GMM there also a high tendency of presenting a high recall compared to the relatively low precision as shown in blue in the figure 5. This phenomenon is a consequence of the over-segmentation that the methods tend to generate. In general, both hierarchical and gmm present and high recall for the considerable low precision, as shown in figure 4(a), many artifacts are being segmented. Because of that, a over segmentation is being done on both methods which explains the high value of recall for the methods with a low precision and ,as a consequence, a low F-max. Thus, a selection of higher values of k ($>$20) will just move both, blue and yellow, curves for the right causing the precision to diminish while the recall increases because many other clusters will be used to group up noise or details of objects as can be seen for the background of the sea star in image 3(c).

As can be seen from figure 4(a) and 3(a), even using the same number of clusters that the groundtruth, over-segmentation is done which may explain the low precision that can be seen in the figure 5 for Gmm and hierarchical. This may indicate that one of the most important limitations of the algorithms is their low base precision and the tendency of over-segmenting objects into their small details. Even though a Gaussian filtering was done, many details disappeared but some other remain and increasing the σ for the filtering may reduce the over-segmentation while reducing the size of the segmented objects as shown in the figure 4(b) making the method even worse and directly affecting the performance in the evaluation.

Finally, as can be seen on table I, UCM presents the higher ODS and OIS respectively, followed by GMM and Hierarchical at last. Both criteria, OIS and ODS, are relatively closer values that represent similar performance on individual images and the whole image set for every method. This two criteria along with Gt covering give and idea of the performance of the method, as discussed earlier, GMM seems to be the most competent method above hierarchical and UCM the better of all the methods because it's higher values of ODS, OIS and Gt covering. The area-PR criteria represents the area below the curve (i.e.integral) that is described in the figure 5, both GMM and Hierarchical show low values for Area-PR are explained by the inefficiency of this tow methods in reaching a higher value of precision with a low recall. In addition, due to the problems with low precision from low values of K that were mentioned earlier.

V. Conclusions

Our clustering algorithm has great disadvantages when compared to the segmentation algorithm of the Berkeley researchers. As a first aspect we must consider the contours in order to delimit the proposed regions in a better way and not consider it as an isolated problem of grouping of regions. For this, it can be used as a representation of features such as those of sketch tokens, these include straight lines, t-junctions, and-junctions, corners, curves, parallel lines, etc. [11]. In addition to the above we can use features indexed directly in the channels, understanding that the channels are composed of the color, gradient and oriented gradient of a patch extracted from the color image, typical of the sketch tokens method. That is, contours based on more attributes than just the intensity of the pixel, this with the aim of expanding our representation matrix and considering more aspects of the objects.

References

[1] S. Leal and M. Rueda, "06: Segmentation. Universidad de los Andes. "https://github.com/MateoRuedaMolano/IBIO4490/blob/master/"

[2] R. Szeliski, "Computer Vision: Algorithms and Applications." Springer, Sep-2010.

[3] P. Arbelaez, C. Fowlkes, and D. Martin, " The Berkeley Segmentation Dataset and Benchmark," David A. Patterson EECS at UC Berkeley, 2007. [Online]. Available: https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/. [Accessed: 08-Mar-2019].

[4] N. Dhanachandra, K. Manglem, and Y. J. Chanu, "Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm," Procedia Computer Science, vol. 54, pp. 764–771, 2015.

[5] Y. Song and H. Yan, "Image Segmentation Techniques Overview," 2017 Asia Modelling Symposium (AMS), 2017.

[6] H. Tibshirani , "Gaussian Mixture Models." Stanford University, California, 2008.

[7] P. Arbelaez and J. Malik, "The Berkeley Segmentation Dataset and Benchmark BSDS500." University of Berkeley, California, 2007.

[8] Pedregosa, F. et al. "Scikit-learn: Machine Learning in Python -Gaussian Mixture Model". Journal of Machine Learning Research. vol. 12, pp. 2825–2830, 2011

[9] Pedregosa, F. et al. ”Scikit-learn: Machine Learning in Python - AgglomerativeClustering”. Journal of Machine Learning Research. vol. 12, pp. 2825–2830, 2011

[10] Pedregosa, F. et al. ”Scikit-learn: Machine Learning in Python - Precision-Recall curve”. Journal of Machine Learning Research. vol. 12, pp. 2825–2830, 2011

[11] P. Arbelaez, “Lecture 11: Grouping 04.” Universidad de los Andes, Bogotá D.C.