Image Segmentation with different clustering types

Julio Nicolás Reyes Torres
Universidad de los Andes
jn.reyes10@uniandes.edu.co

Juan David Triana
Universidad de los Andes
jd.triana@uniandes.edu.co

Abstract

This article presents the implementation of 4 clustering methods to segment a set of images. The algorithms implemented are: k-means, gmm, hierarchical and watershed, it is sought to evaluate how the method's precision varies depending on the number of clusters (k) and also how they act in different color spaces such as: rgb, lab, xyz and hsv. The evaluation of the methods is done by comparing the segmented representation of each algorithm with the 5 annotations made by humans, through the DICE score evaluation that overlaps the images to know their mutual relationship.

1. Introduction

Among the different computer vision fields, one of the most relevance and research-driven one is segmentation. It is based on image division into different natural objects or sections [15]. The main purpose of this field is to obtain an object interpretation through a computer in a similar way to what humans interpret.

To approach this problem, a numerous amount of algorithms have been developed. One of these algorithms is the clustering algorithm K-means, which is unsupervised [6]. Other algorithms include: watersheds, hierarchical segmentation, Gaussian Mixture (GMM) segmentation, among others.

The methods known for this problematic evaluation interact with image overlap intersection, such as in the intersection over union metric. Other evaluation methods report accuracy and precision through AP-curves [7]. In this work a various segmentation methods going to be evaluated on the Berkeley database (BSDS). Different parameters are going to be used, depending on the segmentation method, as well as different color spaces. Regarding evaluation metrics, the DICE score is the one for implementation.

2. Methods

2.1. Data set

The data set used for this practice was a small compilation of the Berkeley Segmentation Dataset and Benchmark database (BSDS500) [4]. The train set consists of 60 images annotated with a *.mat* file, which contains boundaries and segmentation from five different humans. The validation set contains 28 images with its respective annotations.

2.2. Clustering Algorithms

K-means

It is one of the simplest unsupervised learning algorithms to solve the problem of clustering. It seeks to group objects in **k** clusters depending on their characteristics. The algorithm solves an optimization problem by minimizing the sum of the quadratic distances each object to the centroid of its cluster [14]. In images, the intensity of the color is the quantization vector that provides the information about the attributes in an image. [3]

gmm

The Gaussian Mixture Method (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities [9], rather than identifying clusters by "nearest" centroids as in K-means, each cluster is represented by its centroid (mean), covariance, and the size of the cluster (Weight). A set of k gaussians are fitted to the data, after is estimated the gaussian distribution parameters such as mean and variance and weight for each cluster. [12, 11]

Watershed

The Watershed method is based on the topographical idea of a surface divided between basins, inside of it, the water falls converges in the same location making up lakes. The Watershed lines separate each catchment basins. Analogously, the gradient magnitude of an image is considered as a topographic surface where the catchment basins should

correspond to objects and the watersheds to their contours [3, 5]. It's important to highlight the watersheds are applied to scalar (gray level) images, in our case the surface is flooded by the markers given for Sobel filter.

Hierarchical

Rather than the other methods, the main idea of hierarchical clustering is to not think of clustering as having groups to begin with [2]. The main goal of Hierarchical clustering's is be depicted as a tree or dendrogram of fusions between clusters [1]. The method implemented was the named **Ward**, the only one among the agglomerative clustering methods based on a classical sum-of-squares criterion, it produce different groups that minimize within-group dispersion at each binary fusion.[8]

2.3. Segmentation parameter tuning and Image preprocessing

The execution time of the algorithm depends on some factors such as the implemented method and the number of clusters, it oscillates approximately between 6 - 10 seconds, we consider that it is not an excessive processing and therefore the images were not scaled.

The number of clusters (k) in the 4 algorithms implemented was clearly the most sensitive parameter to the changes reflected in the final segmentation, a number of low groupings, does not represent all the linked objects in the images, but a very high value generates a over-grouping that does not allow to differentiate objects either. Therefore, choosing a suitable value is of vital importance for a good segmentation, it could be observed that on average for all the images, they give favorable results in: kmeans (7-10), watershed (30-50), gmm (7 -12), hierarchical (15-20).

2.4. Evaluation

The current evaluation methods include square distance measures, couples counting, statistical approach. Some of these are very well known, such as Jaccard index (couples counting) or segmentation difference (distance measure) [13].

To evaluate the implemented algorithms it was necessary to use the entire ground truth set. This means that all of the 5 human annotations per image were used to increase fidelity. More specifically, the DICE score evaluation metric was used. This is an arithmetic mean metric of all segments of evaluation. It relates two area segments with their mutual overlap as follows [10]:

$$dice(E_1, E_2) = \frac{2|E_1 \cap E_1|}{|E_1| + |E_2|} \quad (1)$$

Since every image is annotated five times by different people, five different scores were obtained per image during evaluation. Afterwards, a total score was obtained using the mean of all those values. This was proven to be a significant metric in Santner and collaborators work [10].

3. Results

Note: The results are presented in two stages, at first, an image example to represent the different space colors and its corresponding segmentation, and finally the scores that show the similarity of the segmented image by the clustering methods and the segmentations made by people.

3.1. Color Space and Segmentations

Space color differences

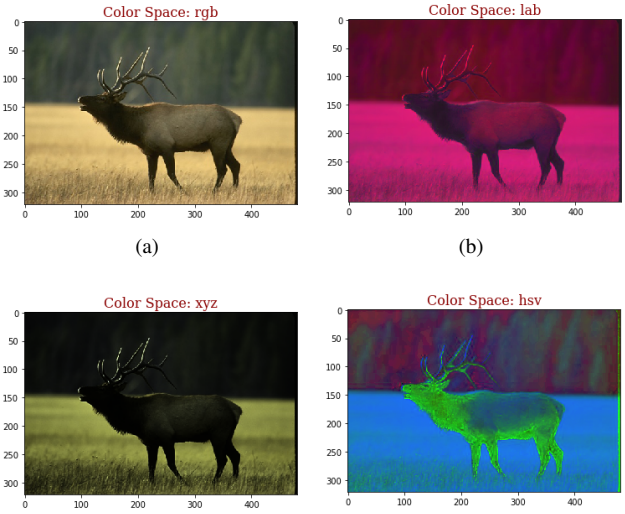
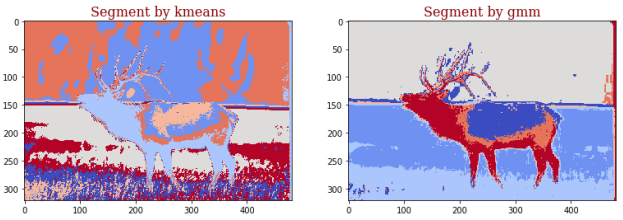


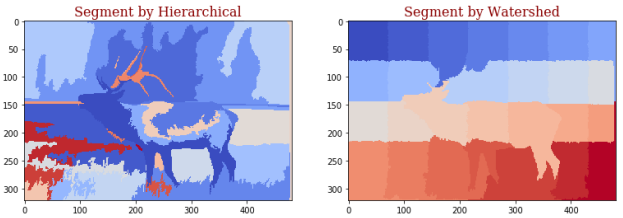
Figure 1: Normalized Spaces Colors: a) rgb, b) lab, c) xyz, d) hsv.

Clustering Methods

1. **Good Segmentation**



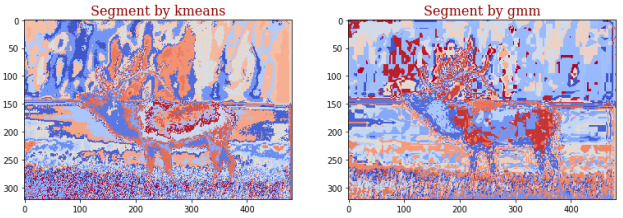
(a) Kmeans, $k = 7$ (b) gmm, $k = 7$



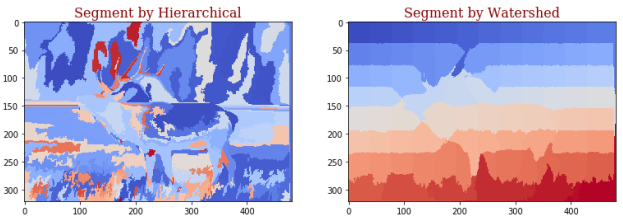
(c) hierarchical, $k = 30$ (d) Watershed, $k = 40$

Figure 2: Clustering Methods. (The images are in rgb)

2. **Over-Segmentation**



(a) Kmeans, $k = 40$ (b) gmm, $k = 40$



(c) hierarchical, $k = 200$ (d) Watershed, $k = 100$

Figure 3: Over-Segmentation. (The images are in rgb)

3.2. DICE score

As it was mentioned before, the evaluation metric used for this segmentation problem was the mean value of multiple DICE arithmetic means. After varying different parameters of the function implemented (cluster number, color space, and clustering method), the following tables were obtained regarding the DICE score obtained on each case:

Mean Segmentation Score(k=5)		Color Space					
		RGB	HSV	Lab	RGB+xy	HSV+xy	Lab+xy
Clustering Method	K-Means	0,512	0,591	0,719	0,64	0,66	0,661
	GMM	0,624	0,619	0,665	0,645	0,648	0,652
	Hierarchical	0,597	0,739	0,391	0,618	0,573	0,588
	Watersheds	0,8	0,8	0,8	0,8	0,8	0,8

Figure 4: Mean DICE scores of different segmentation methods using different parameters with a cluster number of 5.

Mean Segmentation Score(k=10)		Color Space					
		RGB	HSV	Lab	RGB+xy	HSV+xy	Lab+xy
Clustering Method	K-Means	0,728	0,675	0,676	0,661	0,713	0,705
	GMM	0,749	0,617	0,738	0,745	0,727	0,733
	Hierarchical	0,736	0,742	0,694	0,718	0,676	0,68
	Watersheds	0,8	0,8	0,8	0,8	0,8	0,8

Figure 5: Mean DICE scores of different segmentation methods using different parameters with a cluster number of 10.

4. Discussion

In the tables of figure (4,5)

Observing figures 4 and 5, it is easy to see that the number of clusters changes the score results drastically. It seems that cluster number needs to be greater for the clustering methods. Although this apparent increase is very satisfying, there is a threshold where the number of clusters start to decrease the evaluation score. This is because the methods are over-segmenting the image and thus, identifying different artifacts that don't belong to the image's objects and contours. Moreover, after a few runs on the algorithm it was clear to see that the results varied for the different methods excepting the hierarchical method, which sustained a constant value. This means that K-Means and GMM create randomized clustering segmentation due to arbitrary arrangement of the different "classes".

As it was seen on figures 4 and 5, the watersheds algorithm had no change at all on each of the samples taken. This is not a normal behaviour, since watersheds should vary its results depending on color spaces and number of clusters. This is because, the Voronoi diagram obtained through the contour of the different structures changes and thus, the watersheds segmentation. In result, our segmentation technique for watersheds was a failure and data regarding this method should not be taken into account.

Regarding the evaluation method performed in this practice, it is a very precise metric to take into account, since it introduces the significance of each human annotation. It is clear that every single person has a different perspective on how to segment an image on their own, meaning that even

these types of ground truth will have a certain deviation on their borders. Basically, the arithmetic mean is a good indicator for a small amount of multiple ground truth annotations. If there was a case where each image had more than 30 annotations, there could be a reason to use a probability function and describe a normal distribution descriptor.

The best method used was the hierarchical clustering method. It is evident that there were higher scores on other methods, but there was no strong deviation regarding the hierarchical method. This means that there is more accuracy regarding the fidelity of the algorithm, and thus, it is more reliable than any of the other algorithms.

5. Conclusions

- The number of clusters (k) was the most sensitive parameter, a low number of clusters does not represent all the linked objects in the images, and in the other hand, a very high value generates a over-grouping that does not allow to differentiate objects either. So, is very important to choose a suitable value for a good segmentation. It could be observed in the figure (2) a good segmentation, but in the figure (3) the case of over-segmentation, it can be appreciated the objects almost can not be recognized.

- Image segmentation varies depending on the different hyper-parameters used on the clustering methods. Different scores and segmentation areas are obtained on each run of an algorithm due to the random aspect of cluster assignment on k-means and GMM.

- As it can be appreciated in the figure (1) visually there is a noticeable difference between the color spaces, however, as it is observed in the different scores, and in the figure (2), the difference of the color space for the segmentation is not as relevant, in fact as it is observed in the figure (5), in general the best color space is the rgb.

References

[1] Hierarchical Clustering Big Ideas. Technical report.

[2] Distances between Clustering, Hierarchical Clustering. Technical report, 2009.

[3] P. Arbelaez. Computer Vision Perceptual Grouping - Segmentation. Technical report, Universidad de os Andes.

[4] BSDS. The Berkeley Segmentation Dataset and Benchmark.

[5] U. de Toronto. Topic 6: Hierarchical image representations. Technical report, Universidad de Toronto.

[6] e. a. Dhanachandra, N. Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54(1):764–771, 2015.

[7] J. Jordan. Evaluating image segmentation models.

[8] F. Murtagh and P. Legendre. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, 31:274–295, 2014.

[9] D. Reynolds. Gaussian Mixture Models *. Technical report.

[10] P. T. B. H. Santner, J. Interactive multi-label segmentation, s.f.

[11] Scikit-learn. Gaussian mixture models.

[12] A. Soni. Clustering with Gaussian Mixture Model.

[13] S. Srubar. Comparison of segmentation evaluation methods. *WSCG 2013*, 1(1):0, 2013.

[14] Universidad de Oviedo. El algoritmo k-means aplicado a clasificación y procesamiento de imágenes.

[15] H. Y. Yuheng, S. Image segmentation algorithms overview, s.f.