BSDS500 Segmentation by clustering

Saul Gómez
Universidad de Los Andes
sc.gomez11@uniandes.edu.co

Diego Valderrama
Universidad de Los Andes
df.valderrama@uniandes.edu.co

Abstract

Image segmentation is an important task in computer vision that even nowadays has not been solve at all. BSDS500 (Berkeley Segmentation Dataset) is a tool that has been widely used for the development of segmentation methods. It is composed by 500 natural images distributed in three sets: train, validation and test. This dataset was used to evaluate the performance of an implemented segmentation by clustering methods and compare against the one developed by Berkeley Computer Vision group. The best algorithm implemented is based on filtering and k-means clustering in the RGB space color. The results shown a poor performance compared to gPb-OWT-UCM method because of the limitations presented, obtaining a maximum F-measure equals to 0.52 while gPb-OWT-UCM presents an F-measure of 0.73. Finally, the limitations of the methods implemented are discussed and new considerations are established to improve the performance in the implementation of future methods.

1. Introduction

Understanding the image and extracting information from the image to accomplish some works is an important area of application in computer vision [7]. Often it is not necessary to focus on the image as a whole, but only for some certain areas which provides us information of interest. For this reason, image segmentation has become a crucial task in image processing and pattern recognition. It is a process based on certain criteria to divide an input image into different regions in order to extract the area we interested in [7], [4].

A broad family of approaches to segmentation involve integrating features such as brightness, color, or texture over local image patches and then clustering those features based on, such as, fitting mixture models mode-finding, or graph partitioning [1]. Based on this, there are many commonly used algorithms for image segmentation, among the main we can find the threshold segmentation, which is based on the determination of an optimal threshold to differentiate the structures of the background, the regional growth segmentation, whose main idea is to group pixels that in similar characteristics in regions, or the segmentation by clustering algorithm, based on the similarity between things as the criterion of class division, that is, it is divided into several subclasses according to the internal structure of the sample set, so that the same kind of samples are as similar as possible, and the different are not as similar as possible, and many others more [7].

2. Materials & Methods

2.1. Dataset

The database used is BSDS500 and was created by Berkeley University of California. This database consists in 500 natural images with two different sizes (381x421 and 421x381) and is divided in train, validation and test sets with 200, 100 and 200 images, respectively [1]. The data also have benchmarks and labels in order to train algorithms and evaluate the performance of the methods. The labels are different number of manual segmentations for each image (groundTruths) with an average of 5 groundTruths per image [1]. Figure 1 show some examples of the images included in the dataset and the regions segemented by humans.

2.2. Segmentation methods

Based on the results found in [5], the methods that showed a better performance were selected in order to improve them and get better results when evaluating them on the complete database.

2.2.1 Watershed

This technique segments an image into several catchment basins, which are the regions of an image, interpreted as a height field, where rain would flow into the same lake. An efficient way to compute such regions is to start flooding the landscape at all of the local minima and to label ridges wherever differently evolving components meet. Unfortunately, watershed segmentation associates a unique re-

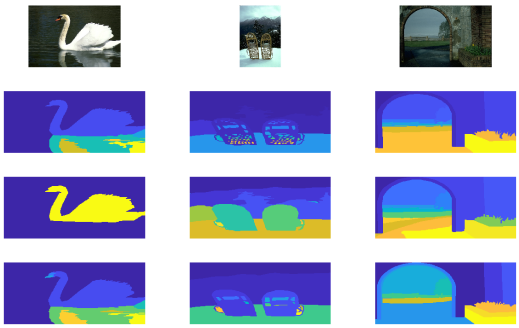

Figure 1. Berkeley Segmentation Dataset [1]. *From top to bottom:* Image and groundtruth made by three different human subjects. The BSDS500 consist of 200, 100 and 200 images for train, validation and test, respectively.

gion with each local minimum, which can lead to over-segmentation. Therefore, it is often necessary to first marks seed locations that correspond to the centers of different desired components, process commonly named minimum imposition [6].

Previously, we found that 1 peaks per label and 48 clusters are the hyerparameters that maximizes the performance using watershed as segmentation method [5]. We used this method in two ways. The first one without image preprocessing and the second with a gaussian filter as image preprocessing

2.2.2 K-means

The K-Means algorithm clusters data by trying to separate samples in a certain number of groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. The algorithm has three steps: The first step chooses the initial centroids randomly. Then, assigns each sample to its nearest centroid. Further, creates new centroids by taking the mean value of all of the samples assigned to each previous centroid. The difference between the old and the new centroids are computed and the algorithm repeats these last two steps until this value is less than a threshold, that is, until the centroids have a significant displacement [2].

In our first approach, we found that 2 number of clusters and 162 number of iterations before coverging to a value, are the better hyperparameters for this method [5]. Furthermore, as same as with watershed method we used this method in two ways. The first one without image preprocessing and the second with a gaussian filter as image preprocessing

2.3. Evaluation

To evaluate and compare the segmentation methods proposed a Precision-Recall curve was implemented. The ideal result of this metric is a point in the coordinates (1,1), which means a 100% of precision and a 100% of recall is achieved. However, since annotations were made by humans it is clear that a perfect performance would not be achieve. Even more, humans F-measure performance is about 80%, which suggests that reasonable results should approximate below that value. F-measure corresponds to the measurement of accuracy that a test has. Used in the determination of a single weighted value of precision and recall [3]. Additionally, it was employed a fixed threshold for all images in the dataset, calibrated to provide optimal performance on the training set named optimal dataset scale (ODS) and we also evaluate performance when the optimal threshold is selected by an oracle on a per-image basis, called optimal image scale (OIS), which may provided even better segmentations, both provided by BDSD benchmark [1].

On the othe hand, segmentation covering is another region-based metrics often employed to evaluate the performance of segementation methods. Its basis lies in overlapping between two regions. The covering of the segmentation is define by

$$C(S' \rightarrow S) = \frac{1}{N} \sum_{R \epsilon S} |R| \cdot \max_{R' \epsilon S'} \frac{|R \cap R'|}{|R \cup R'|} \qquad (1)$$

In this case, to achieve perfect covering the machine segmentation must explain all of the human data.

3. Results

3.1. Train and Validation

Training and validation sets were used to train the segmentation methods implemented. In order to improve the results obtained previously, it was decided to pre-process the images by applying a Gaussian filter to remove small structures and artifacts of the images that could increase the error in the segmentation and thus process images only with low frequencies corresponding to large edges. In table 1, the results obtained in the performance of the methods used and their variation with the pre-processing of the images can be observed.

Table 1. ODS/OIS and Area resulting table from ground-truths in train and validation set

Set	Method	ODS	OIS	AP
Train	Watershed	0.29	0.30	0.08
Train	K-Means	0.49	0.52	0.09
Validation	Watershed	0.30	0.31	0.08
Validation	Watersed with filtering	0.22	0.24	0.05
Validation	K-Means	0.48	0.51	0.08
Validation	K-Means with filtering	0.52	0.53	0.17

These results (Table 1) denote a higher performance for the K-Means method with filtering compared to the unprocessed image method, indicating that image filtering represents an improvement for this algorithm, from 0.49 to 0.52 in ODS F-measure and increasing the AP about 8% . However, in the case of the watershed segmentation method, performance decreased when the image was filtered, decreasing from 0.29 to 0.22 in ODS F-measure. Overall, k-means with Gaussian filter had a superior performance, reaching a ODS equal to 0.52 and an AP of 17%. In this order of ideas, the methods of k-Means with Gaussian filter and Watershed were selected to evaluate their performance in the evaluation set. This also can be notice in figures 3 and 4, where watershed method shows better precision but poor recall, contrary to k-means methods, where the recall was higher than watershed with a similar precision.

3.2. Test

The Table 2 shows the quantitative results in test set using our methods and the method used in [1]. As same as in validation set, k-means with Gaussian filter had superior performance than watershed with 0.52 and 0.31 in ODS F-measure, respectively. Nevertheless, our better method is far below the method proposed by Arbelaez and co-workers as their method has 0.73 in ODS F-measure. These results, are consistent with the precision-recall curve because as shown in Fig 5 our methods do not have an accuracy greater than 0.43. This Figure also shows that with our methods the recall is dependent on the method used because k-means with Gaussian filter has better recall than watershed. Additionally, in the Figure 5 is evident that gPb-OWT-UCM (UCM) method is better than our methods in most of the recall range and we only be better than this method when we want a very high recall.

Table 2. ODS/OIS and Area resulting table from ground-truths in test set

Method	ODS	OIS	AP
Watershed	0.31	0.32	0.09 %
K-Means with filtering	0.52	0.54	0.16
UCM2	0.73	0.76	0.73

The metric used also allows us to know the number of clusters with which our algorithms obtain a better performance in the test set. For the method performed in [1] this metric determines the threshold that maximizes the performance. We found that 6 and 55 are the number of cluster with which we obtain better results using k-means with gaussian filter and watershed, respectively. On the other hand, according with Arbelaez *et al*. the threshold associated with the best performance for their method is 0.13 [1]. Figure 2 shows the segmentations of three images using our methods and ucm method. In this Figure you can see that the segmentation performed using watershed are not similar with the ground-truths (see Fig. 1) Likewise, when you compare the ground-truths with the segmentation made by k-means with gaussian filter, it is observed that these results have a higher level of similarity than the segmentations obtained with watershed but it is lower than the results obtained by arbelaez et al, which is consistent with the quantitative results.

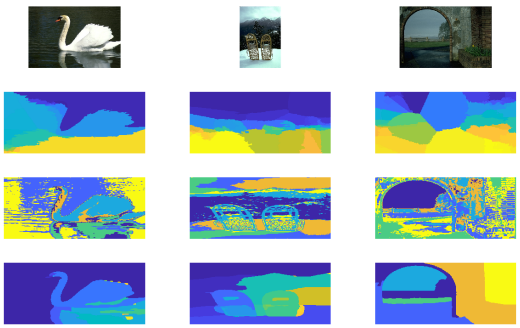

Figure 2. *From top to bottom and left to right:* Watershed segmentations, K-means with filter image pre-processing segmenations and UCM segmentations. Watershed used 55 clusters, K-means 6 clusters and UCM used a threshold of 0.13

4. Discussion

4.1. K-means with filtering and Watershed methods comparison

As denoted in the table 2, among the methods implemented, k-Means with Gaussian filter presented a higher performance compared to the Watershed method. Similarly, the figure shows the results obtained for the precision-recall curve of both methods implemented, where a superior performance can be seen in both, precision and recall, by the k-Means method. This is due to a bias on the Watershed method, because of a certain number minimums imposed tends to segment the image into regions of similar size, as can be seen in the figure 2. This presents an optimal response with large objects, which is reflected in broad precision results (See figures 2, 5). However, it omits fine features of the images or groups them in the same region, so it does not manage to correctly segment all regions of interest in the image, seriously compromising the recall of the method. On the other hand, filtering reduces the presence of fine structures which contributes to the accuracy of k-means method. However, although k-means outperforms watershed method, it is based on the assignment of each pixel of the image to a certain group depending on the euclidean distance, but this can lead to an incorrect segmentation as large uniform regions. Figure 2 shows an example

for which background results in an incorrect partition due to a large distance between pixels, likewise, it can be notice that, for example, the swan is divided into different regions even though pixels are very similar between them and correspond to the same object.

However, the results obtained in the current study show a contradiction with those observed in [5]. Particularly, l-means improved its performance by being evaluated using the complete database. This can be due to the evaluation metric employed previously. Since this metric only takes into account the match of the indices between segmentation and ground truth, the image may have been partitioned into similar regions to Ground Truth but when labeled with different indices or even the fact that the segmentation was performed with a number of regions, K, different from the annotations, decreases the method's success rate with respect to the evaluation metric [5]. For this reason, it was possible to generate an error in the choice of the method with better performance. Also, the image pre-processing developed contributed significantly to improve the performance of k-Means, unlike Watershed, where the performance was diminished.

4.2. Evaluation comparison with gPb-OWT-UCM

Now, comparing the results obtained with the gPb-OWT-UCM method developed by Arbelaez *et al*., there is a clear difference in the performance of the methods implemented, these being much lower as shown in the figure 5. This is because Arbelaez and cols. use a contour detection method based on globalized probability of boundary (gPb). This is the sum of local and spectral signals given by multiscale probability of boundary (mPb), which provides information of all the edges, and the spectral probability of boundary (sPb), which extracts only the most salient curves in the image. From the output of the contour detector, the Oriented Watershed Transform (OWT) is used to recover the initial regions from the contours found. Finally, they build an Ultrametric Contour Map, which allows a resulting hierarchical segmentation retain real-valued weights indicating their likelihood of being a true boundary. For a given threshold, the output is a set of closed contours that can be treated as either a segmentation or as a boundary detector for the purposes of benchmarking [1]. For this reason, this method provides a great versatility in the segmentation of the image, whose accuracy will depend on the threshold used (ODS-OIS), which determines the weight of the contours that will partition the image into regions. This can be evidenced in the higher performance (see figure 5 and table 2) and the well defined regions such as in figure 2.

4.3. Limitations and errors

The most important limitation of our methods is related with the algorithms that we selected since k-means and watershed are segmentation methods used in problems in which one does not know what are looking for, so the algorithm is not properly trained because do not use the ground-truths of train set in order to learn to recognize patterns in objects as they possess very different visual characteristics. This limitation also contribute to errors when the algorithm assign labels for each pixel which leads to a considerable reduction in the performance.

Other errors are more specific of each method, e.g., watershed try to divided the image in regions with same number of pixels so this method makes mistakes assigning the labels for each pixel between regions that are near or similar to each other. On the other hand, k-means has a problem assigning labels when the regions are so big because this method makes a grouping of pixels with similar characteristics in the original image in different regions.

5. Conclusions

As noted, the problem of image segmentation is a difficult task to address because of the lack of information available in the algorithm in its training. In order to improve the performance of our algorithms is necessary use the ground-truths of the train set since human segmentations allow the algorithm to have a better training. It is a good idea to perform an algorithm that takes into account the form of the objects as well as the texture and color distribution of the image, where each of these patterns have an specific weight when the algorithm make grouping the pixels in different regions.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, May 2011.

[2] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics*, 2007.

[3] S. M. Beitzel. *On Understanding and Classifying Web Queries*. PhD thesis, Chicago, IL, USA, 2006. AAI3220872.

[4] H. Cheng, X. Jiang, Y. Sun, and J. Wang. Color image segmentation: advances and prospects. *Pattern Recognition*, 34(12):2259 – 2281, 2001.

[5] S. Gomez and D. Valderrama. Clustering segmentation using different methods and color spaces. Online. Availible at: https://github.com/scgomez17/IBIO4490/blob/master/06-Segmentation/Clustering$_segmentation.pdf$, 2019.

[6] R. Szeliski. *Computer vision*. Springer, 2011.

[7] S. Yuheng and Y. Hao. Image segmentation algorithms overview. *CoRR*, abs/1707.02051, 2017.

6. Images

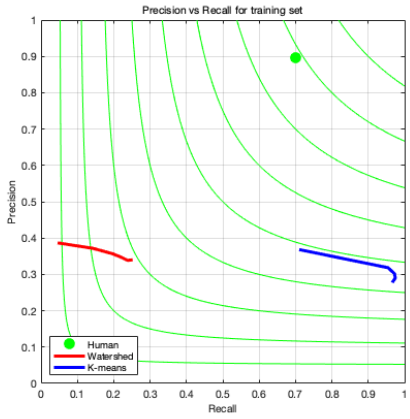

Figure 3. Precision-Recall curve for training set. Watershed method segment fewer true regions than K-means. Nevertheless, watershed is better assigning the labels for each segmented region

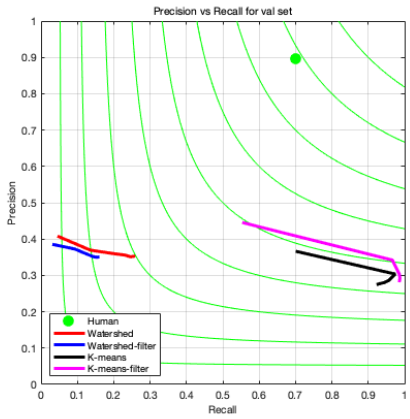

Figure 4. Precision-Recall curve for validation set. Applying a gaussian filter previously of segmentation method improve the results for K-means method. Nevertheless, the filter is not a good idea when we use watershed as segmentation method

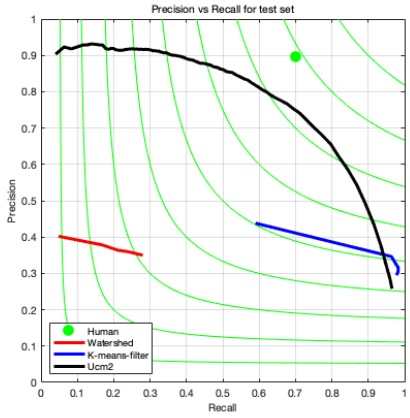

Figure 5. Precision-Recall curve for test set.