BSDS

Mauricio Neira
Universidad de los Andes
Cra 1 N^o 18A - 12, Bogotá - Colombia
m.neira10@uniandes.edu.co

Daniel Rodriguez
Universidad de los Andes
Cra 1 N^o 18A - 12, Bogotá - Colombia
da.rodriguez1253@uniandes.edu.co

Abstract

Image segmentation has been one of the most important ongoing problems in computer vision. In this work, we perform image segmentations using a total of 5 algorithms: k-measn, mixture of gaussians, watersheds, hierarchical clustering and gPb-UCM on the LAB color space and evaluate them using precision-recall curves and the f-measure score on the BSDS-500 dataset. We conclude that the best performing algorithm is gPb-UCM with an f-measure of 0.72 followed by the hierarchical clustering algorithm with an f-measure of 0.53. Additionaly, k-means and gmm underperform significantly since they donnot take spatial information into account. The poor performance of watersheds can be attributed to a poor initialization of markers. Finally, we propose a series of improvements on the algorithms and new methods that could even outperform the gPb-UCM algorithm.

1. Introduction

Computer vision seeks to automate tasks that the human visual system can do. Three main tasks in the discipline are object segmentation, object detection and image classification. In its beginnings, the field strived to developed countour detection algorithms to extract features that distinguish objects in a picture. However, given the small datasets available, this approach was very rudimentary and did not allow for an objective comparison of the methods. With the advent of the internet, larger datasets became available and it became necessary to introduce a universal metric. In 200X, Pablo et al. suggested the Precision-Recall curve, originally from the information retrieval field, as a proper measure for the tasks in computer vision.

In this paper, we evaluate the PR curve on 5 different segmentation methods applied to 200 images on the BSDS dataset. Previously, we had evaluated 4 of these methods on the same databaset with a different metric that was biased to background pixels. Here, we address their difference in performance based on their procedures. Finally we discuss changes required to improve the accuracy of their classifiers and the scenarios where they may be better suited.

2. Database

The images were taken from a subset of the Berkeley segmentation data set (BSDS) [2] which can be obtained in the following link: `http://157.253.196.67/BSDS_small.zip`. Each image in the data set is of variable length and width and has 5 related annotations. The data set is partitioned into training, test and validation. The test set is comprised of a total of 200 images. The classification algorithms were run on those images for evaluation.

3. Segmentation Methods

The nonsupervised methods implemented were: Kmeans, GMM, Hierarchical, and Watersheds. These were later compared to the Ultrametic Contour Map method on the PR framework. We did not make any modifications to the segmentation methods in [Lab 06] because we wanted to compare their normal output against the benchmark. The main issue in [Lab 06] was the metric used that did not properly reflect the performance of the algorithms. Since we did not have a quantitative way of evaluating performance, we chose to run every algorithm using the established benchmark to see how the algorithms performed.

Using the ill-defined metric in [Lab 06], we observed that the color space had no significant effect on the performance of the segmentation algorithms. However, this result is not reliable as the metric did not reflect the correct performance of the algorithms. Thus, ideally, every segmentation algorithm should be tested with every color space. However, due to a lack of computational resources, all algorithms were run on the $L, a*, b*$ color space as it is the color space that provides a correct notion of distance between colors.

Next is a brief discussion on the clustering algorithms evaluated:

3.1. K-means

Clusters the data points into k spherical concentrations. This is done by calculating the center of mass of the points that are nearest to each centroid.

3.2. GMM

This clustering algorithm is analogous to k-means except that it adjusts k n-dimensional gaussians to the datapoints.

3.3. Hierarchical

The hierarchical method implemented is known in the literature as agglomerative clustering. It is a clustering algorithm that takes a bottom up approach, combining every pair of closest pixels in a color space in each step until the entire image forms a single group. This construction produces a dendogram that can be cut at any stage of the process to obtain k partitions.

3.4. Watershed

This algorithm starts with k markers corresponding to the k most significant local minima of the image. The algorithm then simulates a flood starting from these k markers and determines the point at which the water from the different pools of water will meet. The formed pools correspond to the image segments.

3.5. gPb with Ultrametic Contour Maps

This method is a multiscale improvement on the Pb method that assigns a posterior probablity $Pb(x, y, \theta)$ of a boundary with position x, y and orientation θ. The probability is based on a gradient oriented computation from an intensity image. Briefly, features are extracted from the channels of the color space in addition to textons. Then, selecting a circular region around each pixel, histograms are computed in each channel and compared with the ξ^2 distance.

Ultrametric contour maps were applied in the context of image segmentation in [1]. At its baseline, the method uses agglomerative clustering (here hierarchical) to cluster features. Then, based on the distance and similarity between pixels it constructs a dendogram that represents their overall hierarchy of the regions in the image (the root is the whole image). The regions are then compared with an ultra-metric (hence the name).

4. Experiment methodology

The precision-recall curve is a performance measure on information retrieval. Let a set of data be assigned in two categories, relevant and non-relevant. If we then classify the data, the precision term answers the question how many selected items are relevant, and the recall term answers how many relevant items are selected? Another way to think about this measure is to introduce two types of error of classification, the false positives and false negatives. False positives concern with the data that was marked relevant without being so, and false negatives are the data that were marked non relevant while being so. The following graph summarizes the previous discussion.

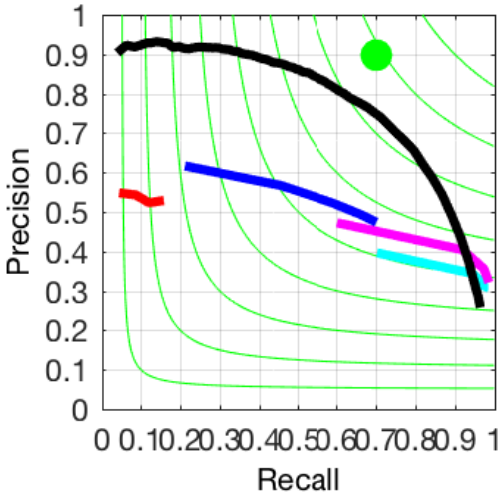

Figure 1. Precision-recall curve for the methods described in section 3. The black line represents Pablo Arbelaez's method - probability of boundary and ultrametric contour maps. The red, blue, magenta and cyan line correspond to the wathersheds, hierarchical, GMM and K-Means method, respectively.

As all performance measures, the precision-recall curve is biased and needs to be applied with consideration. In particular, the PR curve ignores the true negatives, that is the data that were correctly marked as non relevant. This is the reason why the PR curve should be employed in settings with a large amount of non relevant data in contrast to few relevant data.

Another popular performance measure that encompass the true positives, false positives and false negatives is the F measure, defined as the harmonic average between the Precision Recall. Once again, there are biases involved with these measures, so that, according to the application, some coefficients can be introduced to properly balance the results.

5. Results

Results for the BSDS benchmark (full test set - 200 images) for each of the methods described in section 3.

It is clear from the precision-recall curves that the best performing algorithm is the data set's benchmark - probability of boundary with ultrametric contour maps. The next best performing algorithm uses the heriarchical

Figure 2. Original image used as an example to show the performance of the various algorithms.

method, followed by GMM. K-means and watershed are the worst performing algorithms but they cannot be directly compared as they take located at two different extremes of the precision-recall curves. The reasons for their relative order of performance will be explored in the following section.

From figure 3, apart from the ground truth, it is evident that the best performing algorithm is the heirarchical clustering with $k = 5$. We find that for all algorithms but watershed, $k = 20$ is heavily oversegmented and that $k = 2$ produces images with significant undersegmentation. We also find that the watersheds clustering method is the worst performing out of all the proposed methods.

Additionally, GMM and k-means produce segmentations that are very granular and produce regions which are disjoint.

6. Discussion and Conclusions

The best classification algorithm was obtained using the method known as gPb with an F-measure of 0.72. Apart from gPb, the algorithm thar works best is the hierarchical an F measure of 0.58. Even though GMM and K-Means have similar F measures note that the are very skewed in the recall regime. This shows the limitations of the F measure as these two methods would only be useful in a limited number of applications (when the cost of a false negative greatly exceeds that of a false positive, e.g fog estimation in climate research).

In our previous comparison we selected the watershed segmentation as the best classification algorithm. However, as stated there, the metric employed had several flaws. In particular, it was biased against true negatives (see above). Consequently, it translated to the worst performing algorithm using the data set's metric.

Method	**F Measure**
gPb-UCM	0.72
Hierachical	0.58
Watersheds	0.24
K-means	0.53
GMM	0.57

Table 1. F-measure for all of the tested algorithms.

The watershed algorithm was on the precision regime and it has low recall power because properly classifying small categories had little value in the metric. This can be explained by the poor choice of initial markers. The watershed algorithm depends heavily on the initial marker setup. The algorithm proposed set the markers in the minima of the the image. As a result, it is no surprise that the segmented regions are found in the darkest area of the image.

On the contrary, kmeans and GMM worked best on the recall regime. Both of these algorithms tend to oversegment images as their clustering does not take into account the spatial information. They produce a lot more false positives than usual and trigger a higher recall response.

Unlike k-means and GMM, hierarchical clustering does not oversegment the image until a high order k. The main reason behind this is that the implemented algorithm clumps neighbouring pixels based on their distance in the LAB space. In other words, its segmentation takes into account the spatial dimensions. This is probably the reason why it was the best performing after gPb.

Even though we use the same hyper parameters on each method to construct the PR curve, its length varies dramatically along methods. This implies that the methods have different sensibility to the number of clusters. In principle, methods that traverse more F measure level sets should be preferred above since an increase in the number of clusters generally leads to higher quality segmentation. For comparison, the UMC method ranges from a 0.1 to 0.72 F measure level set. This attribute is desirable since a more dynamic range allows for more applications.

The groundtruth data or human clustering in Figure 2 shows a consistent number of regions 5 across subjects. We required to employ $k = 2, 5, 10, 15, 20$ number of clusters because we the images had no constant number of clusters. From Figure 3 it seems that $k = 5$ is a good approximation to the clusters identified by humans. Also, K Means and GMM have a bad scalation property where increasing the number of clusters, selects for clusters with no spatial cohesion. This is no the case for Hierarchical clustering, even for values of $k = 20$ the representation appears like an overdetailed *human* segmentation. We believe this is the main reason why the Hierarchical method crosses 4 F measure level sets from $k = 5$ to $k = 20$.

In conclusion, the best performing algorithm that is not

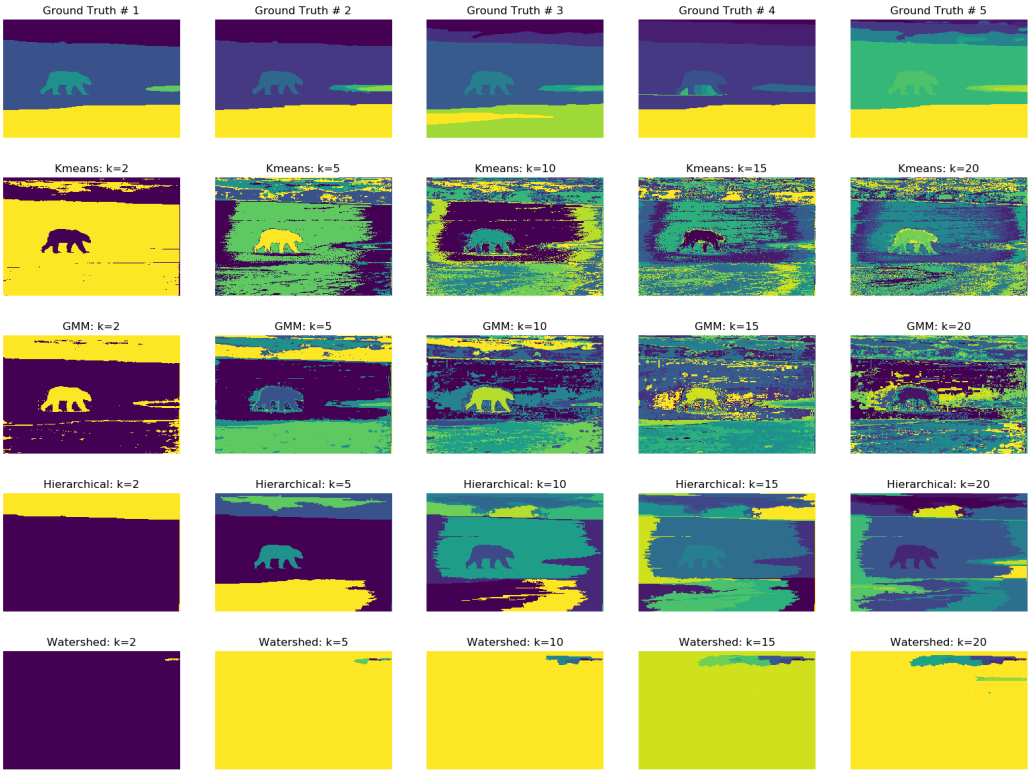

Figure 3. Segmentations results for the image in figure 5. The first row corresponds to the ground truth segmentations.

gPb-UCM that works out of the box without major tweaks is heirarchical clustering [1].

7. Improvements

The simplest and probably the most significant improvement of all the algorithms is to modify the marker selection of watersheds clustering. Because the markers were selected to choose the local minima, most of the segments will come from dark sections of the image which will most commonly not correspond to the groundtruth segmentations as shown in Figure 3. An alternative is to randomly select the starting markers. This should be an improvement if the random points land on different segmentations. If not, the image is likely to be again poorly segmented. A better approach for the selection of markers could be to use the centroids resulting from an initial k-means or GMM classification. A marker at the pixel closest to the centroid is likely to be at a wanted segment. Yet another improvement is the use of hierarchical watersheds where only minima of a relative depth k are selected.

A simple change that could improve the performance of all algorithms is to run a Gaussian kernel to smooth out the images. Since most of the algorithms generate oversegmented images, smoothing the images will make the algorithms segment images on rougher changes in color and generate less oversegmentation. This could decrease the noise in K Means and GMM methods segmentation with a high number of clusters $k > 10$.

Unrelated to the algorithms presented in the paper above, the state of the art segmentation methods are all based on convolutional neural networks (CNNs). It would be worth training a coder-decoder network that produces a segmented image. This approach is likely to produce an f-measure much closer to human classification as it is learning from the segmentation of humans.

[1] Watersheds could outperform it given appropriate initialization markers

References

[1] P. Arbelaez. Boundary extraction in natural images using ultrametric contour maps. In *null*, page 182. IEEE, 2006.

[2] 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, 2011.