# Evaluation of unsupervised clustering segmentation in BSDS500

Stephannie Jimenez Gacha
Universidad de los Andes
Cra 1 Este No 19A - 40 Bogotá, Colombia
s.jimenez16@uniandes.edu.co

Sergio Galindo León
Universidad de los Andes
Cra 1 Este No 19A - 40 Bogotá, Colombia
sa.galindo10@uniandes.edu.co

## Abstract

*The Berkeley segmentation dataset and benchmark (BSDS) represents a great advance in the segmentation problem development as it corresponds to the fist publicly available large collection of natural images with multiple human segmentations and specially, because of the introduction of a common evaluation framework for segmentation. For this, we evaluate a segmentation algorithm based on GMM and K-means clustering using color and spacial features in the lab+xy and rgb+xy space over the BSDS500 train, validation and test sets using the benchmark provided. The best results were obtained using the lab+xy feature space with a maximal F measure of 0.59 for areas and an average precision of 0.16 for boundaries using the GMM. Despite being acceptable for clustering segmentation, they remain distant from the actual state of art, Convolutional Oriented Boundaries. For this, the inclusion of local information in the descriptor as texture and magnitude if gradient is suggested to improve the clustering performance in segmentation tasks.*

## 1. Introduction

Segmentation, as a computer vision task, was considered an ill-posed problem during various decades of the ending of the XX century. The main reasons were the lack of a consistent evaluation metrics, the lack of a standardized dataset to compare the performance of the different algorithms and the absence of a formulation of the segmentation problems for humans. In addition, the segmentation was considered an extension of the boundary detection problems which was thought were solved where the canny edge detector was published.

Despite this, in 2001, the Berkeley segmentation Dataset and benchmark was released, containing a realistic dataset of natural images,an appropriate formulation of the problem from multiple annotations and a quantitative and comparable framework for segmentation evaluation[2]. In this dataset, the segmentation problem is defined as a well posed problem using the fact that the segmentations produce by multiple human are consistent between them, leading to the possibility of evaluating segmentations as it, not exclusively in the context of object recognition. Also, all the necessary tools to perform and evaluate segmentation are proposed and given in this dataset[2].

The images in the BSDS correspond to 1000 representative RGB images of the Corel database with a size of 481x321 pixels. All images belong to natural scenes with at least one object. Respecting the annotations, they were made using a Java application with high flexibility and quality which made possible to posses more than one segmentation per image[2].

For evaluation, the benchmark included one function to evaluate border and area correspondence which returned precision and recall curves for a given set of segmentations of the dataset, and the maximum F measure. However, the most important part of the BSDS is it's public availability, which offered a way to evaluate all the segmentation algorithms and promoted a quick improvement of the algorithms as well as a standardized dataset.

Due to the importance of the BSDS in segmentation, we will evaluate two clustering segmentation algorithms that use the lab+xy and the rgb+xy feature space with a Gaussian mixture and k means in this dataset. After obtaining the segmentation by this method the performance will be calculated using the benchmark of the BSDS and the results will be compared with other methods.

## 2. Materials and Methods

In order to use the BSDS to evaluate the performance of a clustering algorithm for segmentation, the first step is to compute the segmentations for the train, validation and test images. After that, the benchmark was used to obtain the corresponding Precision-Recall curves and the quantitative performance parameters. The BSDS500 is composed of 500 .jpg images of 481x321 pixels both landscape and portrait orientation. The dataset is split into 200 train images, 100 validation images and 200 images for test for a total of 500. In addition, the groundtruth is made up of 500

.mat files, one per image containing 3 human made segmentations.

### 2.1. Segmentation and parameter tuning

The segmentation of the images was performed using the *SegmentByClustering* function in python 3 developed previously. This function uses the color and spacial information to cluster the pixels of an image into k clusters, using K-means, Gaussian mixture, watersheds or hierarchical clustering. For the BSDS500, the segmentation was performed using the *lab+xy* and *rgb+xy* feature space with a Gaussian mixture model and K-means, as they present the best results according to our evaluation quantitative and qualitative metrics.

For each image in the train, validation and test subset, a segmentation using 10 different equally spaced values of k between 3 and 30 was performed and the segmentations were stored into a cell array containing in each position the segmentation for a predetermined k. This cell arrays were stored as .mat files whose names correspond to the original image number. We selected this particular range of clusters as it spans the average number of objects in each image starting from a rough segmentation to a high detail segmentation, also, because it is approximately half of the maximum number of regions found in the train set. For each image, 20 segmentations were obtained for each method, 10 using the *lab+xy* feature space and 10 using the *rbg+xy* feature space.

### 2.2. Evaluation

The evaluation of the segmentation was performed using the allBenchfast function in Matlab, provided in the dataset. This function evaluates the border correspondence (True positive border) with a confidence threshold that is variated to obtain the precision in all the 0-1 recall interval, returning the corresponding curve and the calculated maximum F-measure. Also, the function evaluates the area of the segmented regions by means of the Jaccard index between them and the annotations, returning the quantitative results for both border and area evaluation.

The function handles the multiple annotations by adding them into a single image which is coherent enough to evaluate the segmentation. The evaluation for the train, validation and test subsets was performed independently and the plots for each set were obtained using the ploteval function, given in the benchmark too. All the results were obtained using the recall given by k=[3:3:30]=10 (Recall in [0.55-0.95] for GMM and [0.2-0.83] for kmeans). No other parameter was varied.

To avoid training over the test set, this images were removed from the dataset during the train an validation and were incorporated only during testing. However, as the segmentation methodology does not uses a supervised approach, the parameter tuning using the test set, despite being totally unacceptable, does not represent considerable aid.

## 3. Results

The images of the train, validation and test set were segmented using both rgb+xy and lab+xy feature space with a Gaussian mixture model and K means. After representing the images as a collection of pixels with the rgb or lab color space and the pixel position and normalizing the channels, a quantization was performed by clustering. The GMM clustering algorithm assumes that the data is distributed parametrically in k normal distribution in the feature space. Then the algorithms uses k-means to initialize the weights and means of the distributions and iterates the parameters to reduce the covariance and increase the means among the distributions until convergence which is called the expectation maximization process.

On the other hand, the k means algorithm sets k random centroids in the feature space, calculates the normal distribution parameters, and assigns the labels based on the euclidean distance to the centroid. The algorithm calculates the centroids and labels recursively based previous assignation until convergence. The segmentations were stored in different folders for further evaluation. Different segmentation for the image 55075 with k=5 using both feature spaces and clustering methods are shown below.

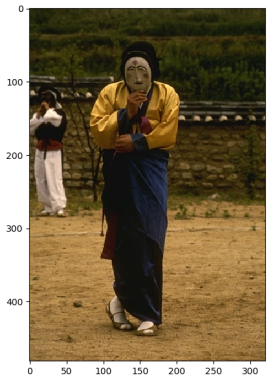


Figure 1. Original 55075 image.

As the addition of the spacial information in the feature Space improves the segmentation, the lab+xy and rgb+xy feature spaces were selected. Also, the usage of GMM clustering corresponds to the advantages of this method respecting the soft assignment and the variability in the parametric distributions that allows a better label prediction. The clustering was performed using the expectation maximization algorithm with a *"full"* covariance which maximizes the mean of the distributions while reducing variance (if possible), however, these hyper-parameters can be changed to other types like spherical for covariance or variational inference for prediction. The segmentation methodology was not

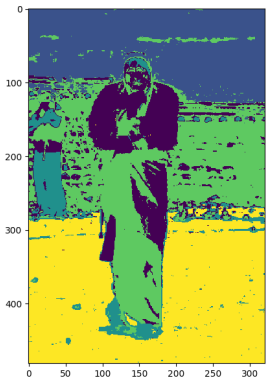

Figure 2. Segmentation of '55075' image using RGB+xy feature space and GMM clustering with k=5 clusters.

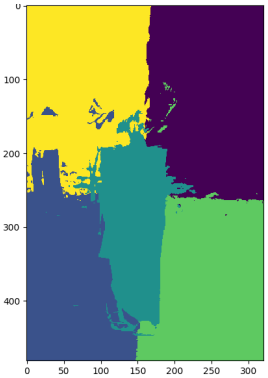

Figure 3. Segmentation of '55075' image using lab+xy feature space and K means clustering with k=5 clusters.

changed respecting the previous implementation by clustering. Respecting K means, this method was used as it corresponds to a robust and fast clustering algorithm with similar and homogeneous cluster sizes and some rigidity that might reduce segmentation noise while still working fine for non-normal distributed data.

After running the *allbenchfast* function, the precision and recall curves for the training, validation and test set were obtained. This curve represents the precision of the algorithm, the true positives among all detections, when the recall is varied (the number of total detections) by thresholding a confidence metric of the detections over the interval 0-1. The ideal precision and recall curves should be a straight horizontal line (P=1), corresponding to an algorithm that does not present errors to any confidence level. Also, the PR curve gives the maximum F measure (ideally 1), a metric of the algorithm performance and the PR area (ideally 1), which corresponds to the average precision.

From the precision and recall curves it can be seen that the performance of the algorithm in the train, validation and test was similar for each clustering method. The k (=10) used to generate different segmentations spanned a recall between 0.55 and 0.95 in all sets when using GMM and

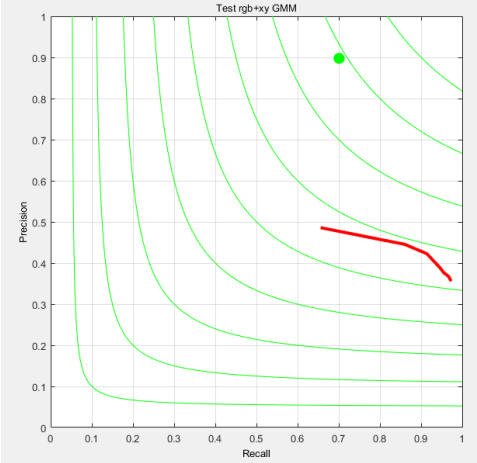


Figure 4. PR curve for the test set using the rgb+xy feature space and GMM clustering.

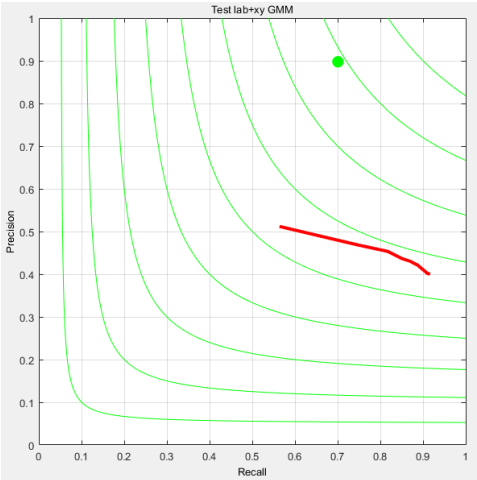


Figure 5. PR curve for the test set using the lab+xy feature space and GMM clustering.

a recall between 0.2 and 0.83 when using Kmeans.This is due to the fact that the same algorithm was used for all sets without changes in the parameters. In addition, the maximal F measure remain between the iso-F 0.55 and 0.65 both clustering methods being higher for the GMM model.

Comparing the different results, the best performance was obtained with the *lab+xy* feature space using the GMM clustering method. with this combination, a maximum F measure of 0.63 and an average precision of 0.16. was achieved. The kmeans method offered a maximum F measure of 0.61 and an average precision of 0.31. The difference between using this feature space and the combinations that utilizes the *rgb+xy* space (Fmax=0.62, AP=0.14 for GMM and Fmax=0.57, AP=0.6 for Kmeans ) is moderate, however, it can be said that moth feature spaces behave similar.

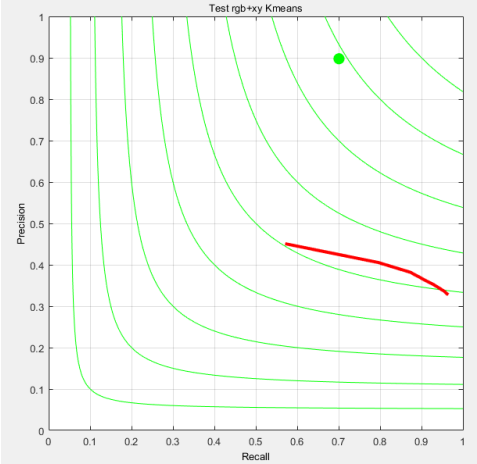


Figure 6. PR curve for the test set using the rgb+xy feature space and Kmeans clustering.

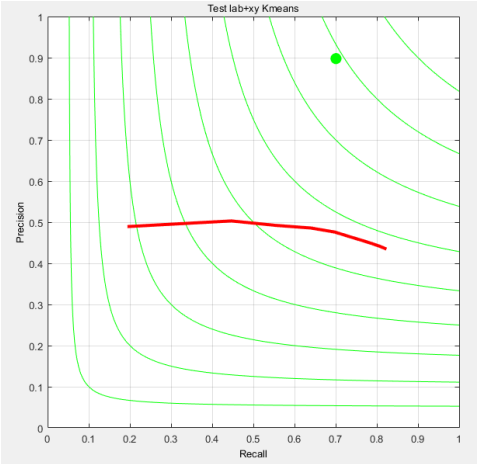


Figure 7. PR curve for the test set using the lab+xy feature space and Kmeans clustering.

Table 1. Evaluation of the boundary(B) and regions(R) for the segmentations using the allBenchFast function in the lab+xy (C) and the rgb (D) feature space for the GMM clustering method

| Set | | Train | | Validation | | Test | |
|---|---|---|---|---|---|---|---|
| Metric | | C | D | C | D | C | D |
| B | Area P-R | 0.15 | 0.13 | 0.15 | 0.13 | 0.16 | 0.14 |
| | F max | 0.63 | 0.61 | 0.61 | 0.59 | 0.63 | 0.62 |
| R | GT ODS | 0.39 | 0.40 | 0.38 | 0.38 | 0.39 | 0.38 |
| | GT OIS | 0.46 | 0.45 | 0.44 | 0.43 | 0.45 | 0.43 |
| | Best | 0.60 | 0.55 | 0.58 | 0.52 | 0.59 | 0.53 |

Table 2. Evaluation of the boundary(B) and regions(R) for the segmentations using the allBenchFast function in the lab+xy (C) and the rgb (D) feature space for the Kmeans clustering method

| Set | | Train | | Validation | | Test | |
|---|---|---|---|---|---|---|---|
| Metric | | C | D | C | D | C | D |
| B | Area P-R | 0.29 | 0.16 | 0.30 | 0.15 | 0.31 | 0.16 |
| | F max | 0.60 | 0.55 | 0.59 | 0.55 | 0.61 | 0.57 |
| R | GT ODS | 0.34 | 0.33 | 0.34 | 0.33 | 0.34 | 0.34 |
| | GT OIS | 0.39 | 0.37 | 0.38 | 0.37 | 0.38 | 0.37 |
| | Best | 0.50 | 0.45 | 0.49 | 0.45 | 0.50 | 0.45 |

## 4. Discussion

Clustering methods are fairly appropriate to segment images based on color and spacial information. The selection of an adequate feature space is one of the most important characteristics of the methodology and should be handled with care, for example, using the lab color space seems more discriminative than using rgb and, when using the spacial information, this should not have more weight than color to avoid equally spaced clusters. Lab color space offers a better segmentation because of the contrast difference it offers respecting rgb, also, because the luminosity information ,that is one of the most important for the human visual system, is enclosed in an exclusive channel, the L. For this, the segmentations obtained using the GMM method with the *lab+xy* feature space corresponds to the best results.

Also, the selection of the clustering method is a important methodology parameter. The GMM model seems more flexible and appropriate to segment images when represented with spacial and color information than Kmeans. This last method tends to return similar clusters which, for this feature space is easily achieved by giving more importance to the spacial information, returning equally spaced clusters with little semantic meaning. For Kmeans, the spacial information seems more discriminative due to the algorithm characteristics. The flexibility of the GMM is desired when segmented based on color and spacial information as the regions might be parametrically distributed with more variate parameters than assuming normal distribution (that cannot be assured actually). In addition, the GMM model tends to produce continuous and well defined region borders while Kmeans tends to produce noisy borders.

Respecting the precision and recall curves, the limited recall interval was due to a small number of segmentations performed for image. To obtain the whole recall range the number of k must be considerable high. For this, the results of the average precision could be below expected. The

PR area between both methods are not comparable as the obtained recall for the same number of segmentations per image is different. Respecting the area evaluation we obtained adequate results for a simple methodology. Despite a great number of segmentation errors, due to pixels with similar colors in different spacial locations grouped in the same cluster, the majority of the objects possessing the same color have their pixels correctly labeled.

The actual evaluation of the BSDS dataset offered better results than our previous evaluation metrics for the algorithm. Furthermore, they are still consistent. For both of them the *lab+xy* feature Space using a GMM as clustering methodology offers the best results for segmentation. Our previous metrics in which an average Jaccard index for the whole dataset was obtained is not comparable with the actual BSDS benchmark. Despite this, the both evaluations take into account the Jaccard index in area evaluation, which supports the consistency between them.

Finally, our results are not comparable with the ones obtained by the Convolutional Oriented Boundaries method, the actual state of art for the BSDS, which posses an maximal F measure of 0.793[1]. The clustering method using color and spacial information has multiple limitations such as contrast dependency, spacial artifacts such as equally distributed clusters, channel normalization dependency the possibility to segment different color objects only.Also the impossibility to construct a trained model for the problem as it corresponds to an unsupervised task.

For this, in order to achieve better performance of the segmentation by clustering we suggest to add local information into the representation space such as texture and magnitude of gradient. A better descriptor for the images could be composed of a texton histogram and the magnitude of the color gradient, which are supposed to remain constant in the same objects added to a local color and spacial information from a local patch in a small window. The main cost of this changes might be computing time at least.

## 5. Conclusions

The BSDS represents the first open access large dataset and benchmark for the segmentation problem. Between its' main attributes are that the images correspond to natural scenes and that, having multiple consistent human segmentations for each image, the segmentation problem itself was defined as a well posed problem. Moreover, its'main contribution is a common benchmark to evaluate the segmentation and boundary detection problems that rapidly propelled the algorithmic improvement in this area.

One of the unsupervised segmentation approaches corresponds to clustering, in which, starting from a feature space, the quantization of the images in k clusters might segment the objects present on them. For segmentation purposes, the *lab+xy* feature space combined with a Gaussian mixture offer appropriate and better results respecting the Kmeans method for both *lab+xy* and *rgb+xy* space. This methodology was evaluated in the BSDS benchmark obtaining a maximal F measure of 0.59 for the areas and an average precision of 0.16 for the boundaries.

The Precision and recall curves were obtained for a recall interval of 0.55-0.95 and 0.2-0.83 which might reduce the actual AP. This was due to the limited number of segmentations produced using different number of clusters (n=10). To conclude, our results are not comparable to the actual state of art of the dataset, however, they are quite acceptable. Some possible modifications to improve the performance are including local information as texture or the magnitude of gradient and exploring other clustering methods.

## References


[1] K.-K. Maninis, J. Pont-Tuset, P. Arbelaez, and L. Van Gool. Convolutional Oriented Boundaries: From Image Segmentation to High-Level Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):819–833, 4 2018.

[2] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE Comput. Soc.


## 6. Code and Images

The code is available at the team's repository at *https://github.com/steff456/IBIO4680/tree/master/07-BSDS/Answers*

## 7. Appends

### 7.1. PR curves for training and validation sets for GMM

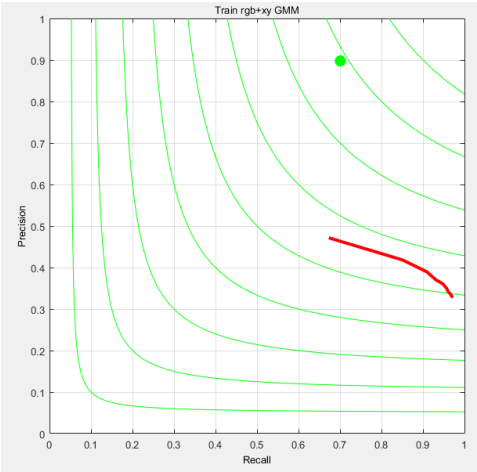


Figure 8. PR curve for the training set using the rgb+xy feature space and GMM clustering.

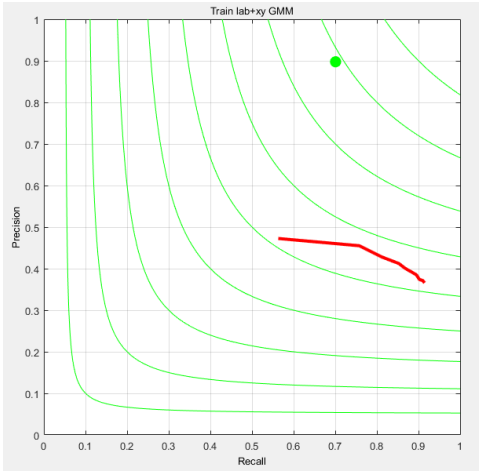


Figure 9. PR curve for the training set using the lab+xy feature space and GMM clustering.

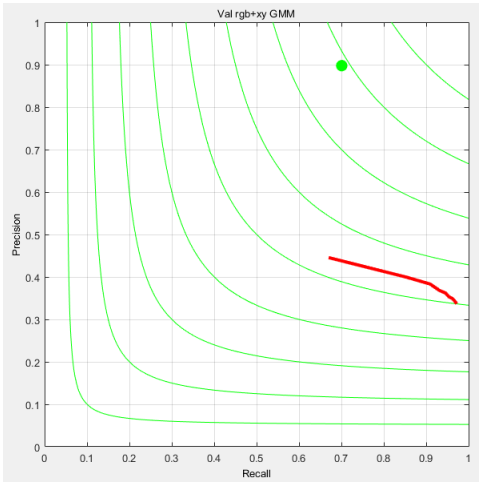


Figure 10. PR curve for the validation set using the rgb+xy feature space and GMM clustering.

## 7.2. PR curves for training and validation sets for Kmeans

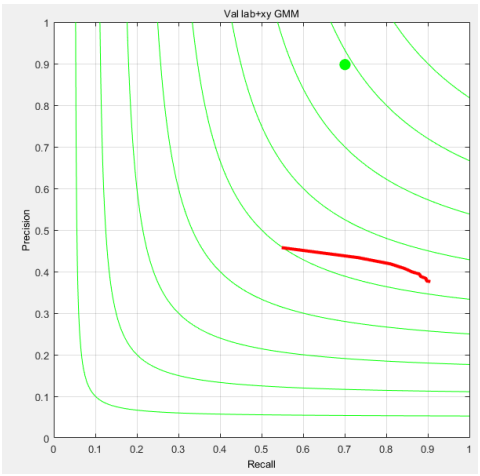


Figure 11. PR curve for the validation set using the lab+xy feature space and GMM clustering.

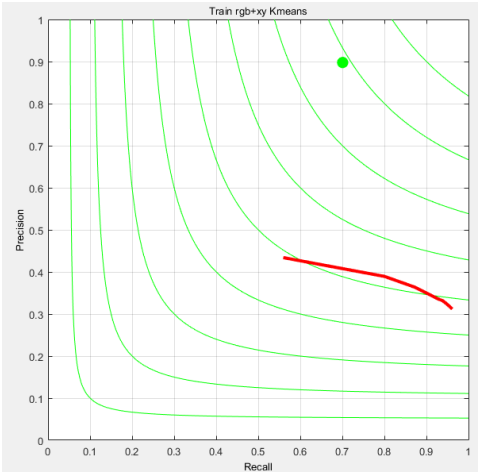


Figure 12. PR curve for the training set using the rgb+xy feature space and Kmeans clustering.

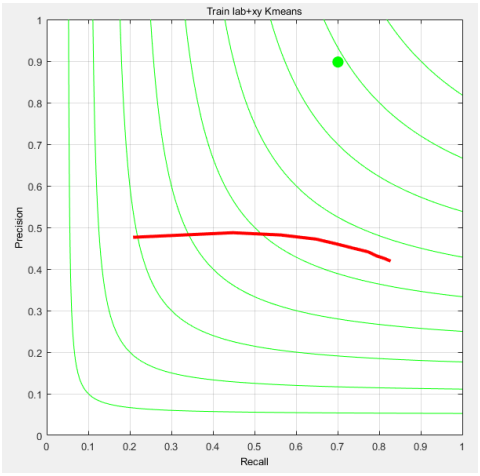


Figure 13. PR curve for the training set using the lab+xy feature space and Kmeans clustering.

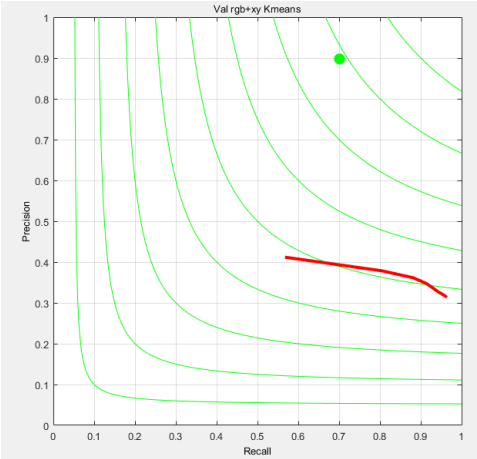


Figure 14. PR curve for the validation set using the rgb+xy feature space and Kmeans clustering.

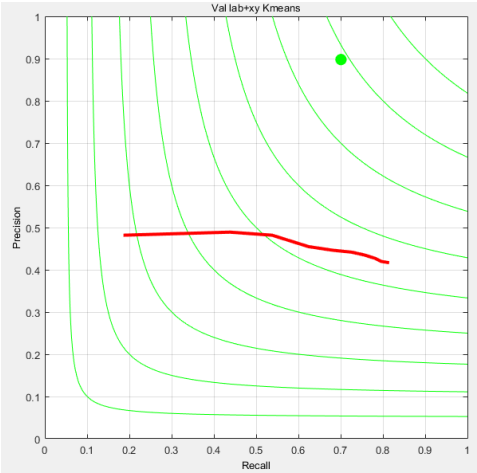


Figure 15. PR curve for the validation set using the lab+xy feature space and Kmeans clustering.