Segmentation using Clustering Methods

Laura Munar Acosta
Universidad de los Andes
l.munar10@uniandes.edu.co

Maria Ana Ortiz
Universidad de los Andes
ma.ortiz1@uniandes.edu.co

1. Introduction

Image segmentation is a vital part of image analysis processing. It differentiates between the objects or instances and their background. Be it a tumor mass in an X-ray image, or a tooth root canal, or a component of a printed circuit board, the image analysis algorithm should find their borders properly and separate the regions without adding or subtracting any extra information [8]. Two variations on the problem exist: supervised and unsupervised segmentation. In the former, models of the texture associated with different entities in the scene are assumed known, and are then applied to the image in the hope of segmenting it into regions corresponding to those entities. This requires a training stage in which human beings group texture exemplars into classes, corresponding to the entities involved, from which the corresponding model parameters are then learnt. In the unsupervised case, no models are known a priori. Instead, the aim is to discover similarities in the data that betray the existence of one or more distinct classes into which the data can be divided [8].
Clustering is one of the simplest methods for unsupervised segmentation and has been widely used in segmentation of grey level images. In the present study, four algorithms are going to be evaluated (k-means, Gaussian Mixture Models, Hierarchical, Watersheds) as a potential use for the segmentation of a small part of the BSDS database with the use as features different space colors such as rgb, lab, hsv and spatial x and y components.
One of the most used clustering algorithm is k-means clustering. It is simple and computationally faster than the hierarchical clustering. This algorithm aims to minimize the sum of squared distances between all points and the cluster centre. The main disadvantage of k-means is that produces different cluster results for each number of clusters; also, different value of initial centroid would result in a different cluster. So selection of proper initial centroid is an important task [3].
On the other hand, a Gaussian Mixture Model (GMMs) is a parametric probability density function represented as a weighted sum od Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continous measurements or features in a biometric system. Its parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm[7].
Also, Hierarchical clustering is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters where each cluster is distict from each other cluster and the objects within eache cluster are broadly similar to each other. Hierarchical clustering starts by treating each observation as a separate cluster. Then, it repeatedly executes the following two steps: (1) identify the two clusters that are closest together, and (2) merge the two most similar clusters. This continues until all the clusters are merged together. The choice of distance metric should be made based on theoretical concerns from the domain of study. That is, a distance metric needs to define similarity in a way that is sensible for the field of study [4].
Finally, the watershed segmentation technique has been widely used in medical image segmentation. The algorithm originated from mathematical morphology that deals with the topographic representation of an image.The set of pixels with the lowest regional elevation corresponds to the regional minimum. The minima of an image are the groups of connected pixels with their grey level strictly lower than their local neighboring pixels. Advantages of the watershed transform are that it is a fast, simple and intuitive method. More importantly, it is able to produce a complete division of the image in separated regions even if the contrast is poor [6].

2. Materials and Methods

2.1. Number of clusters determination

In order to determine the number of clusters, the Elbow Method was performed. The main idea of this method is to run k-means clustering on the dataset of a range of values and for each value of k, calculate the sum of squared errors (SSE). The number of cluster determination is given by the line chart that resembles as an arm, then the elbow (the point of inflection on the curve) is a good indication that the underlying model fits best at that point. The figure 1 presents the results of the elbow method for the image 12003.jpg from the database. It is important to notice that the method used works for one image. So, the approximation consists in seeing the results for different images and to choose a common k value.

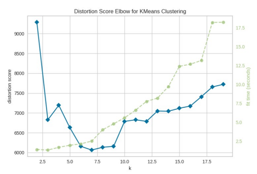

Figure 1. Computational time and Disortion score for Elbow Method

2.2. Segmentation Function

The segmentation function development had multiple options:

- *Number of desired clusters*: Number fo clusters or objects the image will be segmentated
- *Method*: method for clustering or segmentation as watersheds, k-means, hierarchical and aussian mixture model
- *Color space*: RGB, L$*$a$*$b, hsv and versions of the previously mentioned using spatial information as RGB+xy, L$*$a$*$b+xy and hsv+xy. It is important to notice that in order to streamline the hierarchical method, the image was resized. As in every resized image, some information is been lose but it is a trade between the computational time and the results.

Additionally to the function the image was a mandatory parameter. All color spaces were normalized. In general, the distribution of color values in an image depends on the illumination, which may vary depending on lighting conditions, cameras, and other factors. Color normalization allows for object recognition techniques based on colour to compensate for these variations. *Watersheds*: Methodology follow for this method was different considering that watersheds its not a clustering method, and it segmentates according to local minimums. Following this path of ideas there was established a relationship between number of clusters and components obtained by imposing H minimums in the image, taking into account that lower H lower clusters found but not necessarily exactly to the H minima imposed.

2.3. Evaluation Method

First of all, because of the different types of images in the dataset, the evaluation of the method was made with only one image. In order to ensure that the difference in outcome of the method is due to the parameters of feature space and clustering method and not because of the differences between images. Additionally it is important to take into account the varibility between the groundtruths, meaning the variability on human segmentation, mainly for the evaluation methodology propose in these report only first segmentation was used, ignoring variability between humans. The proposed evaluation method consisted in calculating jaccard index per object segmented:

$$JaccardIndex = \frac{Groundtruth \cap Segmentation}{Groundtruht \cup Segmentation} \quad (1)$$

Once every jaccard index is computed a matrix of jaccard index per instance its obtained of size $SegmentationO * NgroundObjects$. These matrix will have higher values of jaccard index per instance segemented when it contains more pixels of the groundtruth. After obtaining the matrix the general segmentation metric for the method will be the average of the best of each column, refering two finding and average value of segmentation of each instance.

$$FinalMetric = average(MAX(JaccardMatrix(:,i))) \quad (2)$$

3. Results

The results of each of the methods segmentation with its corresponding metric are presented below,

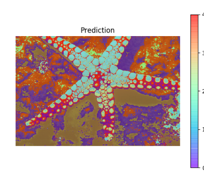

Figure 2. Prediction for the GMMs algorithm with number of clusters=5 and rgb color space

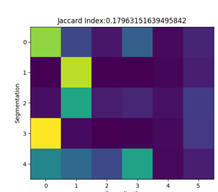

Figure 3. Proposed metric results for the GMMs algorithm with number of clusters=5 and rgb color space

4. Discussion

According to what was mentioned before, the number of clusters determination is one of the main tasks when

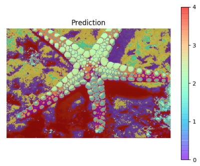

Figure 4. Prediction for the k-means algorithm with number of clusters=5 and rgb color space

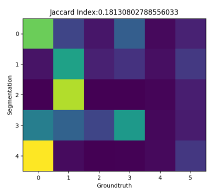

Figure 5. Proposed metric results for the k-means algorithm with number of clusters=5 and rgb color space

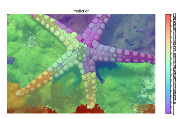

Figure 6. Prediction for the Watersheds algorithm with number of clusters=5 and rgb color space

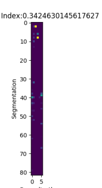

Figure 7. Proposed metric results for the Watersheds algorithm with number of clusters=5 and rgb color space

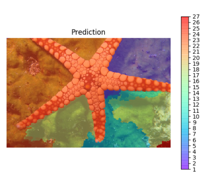

Figure 8. Prediction for the Watersheds algorithm with number of clusters=5 and lab color space

an unsupervised segmentation is been developed. Unfortunately, there is no definitive answer to this task. The optimal number of clusters is somehow subjective and depends on the method used for measuring similarities and the parameters used for partitioning. One popular solution consists of inspecting the dendrogram produced using hierarchical clustering to see if it suggest a particular number of clusters. Other method commonly used is gap statistic. This compares the total within intra-cluster

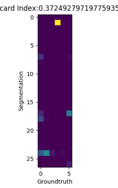

Figure 9. Proposed metric results for the Watersheds algorithm with number of clusters=5 and lab color space

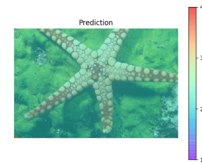

Figure 10. Prediction for the Watersheds algorithm with number of clusters=5 and hsv + xy color space

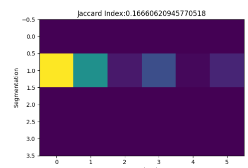

Figure 11. Proposed metric results for the Watersheds algorithm with number of clusters=5 and hsv + xy color space

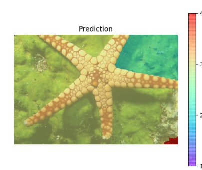

Figure 12. Prediction for the Watersheds algorithm with number of clusters=5 and lab + xy color space

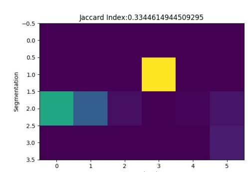

Figure 13. Proposed metric results for the Watersheds algorithm with number of clusters=5 and lab + xy color space

variation for different values of k with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be value that maximize the gap statistic (i.e, that yields the largest gap statistic). This means that the clustering structure is far away from the random uniform distribution of points. In the present work,

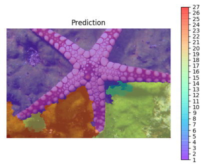

Figure 14. Prediction for the Watersheds algorithm with number of clusters=5 and hsv color space

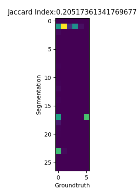

Figure 15. Proposed metric results for the Watersheds algorithm with number of clusters=5 and hsv color space

the Elbow Method was performed. As it is seen in Figure 1 the computational time increases as k increases and it can be said that the elbow is found in between 2.5-5 number of clusters but the answer can be subjective.

It is known, that color images can provide more information than gray level images. Although there is not a common opinion about which color space is the best choice, some papers identify the best color space for a specific task. For example, HSV color space is highly recommended for skin colour detection. Considering that RGB color space can be understood as all possible colors that can be made from three colourants for red, green and blue; it is not the most adequate for color image analysis because it can not represent all the colors perceptually visible for humans. Other color spaces can provide more information in each of its channels and are perceptually uniform [5].

As it can be seen in the Figures 3, 5 and 7, the best performance according to our proposed metric is the Watershed algorithm. This could be explained because of the nature of the image was only described by this method. For example, one of the main disadvantages of k-means is its sensitivity to outliers, besides the number of clusters determination. K-means assumes the clusters as spherical, so does not work efficiently with complex geometrical shaped data(Mostly Non-Linear). So, this specific image could have outliers that only allow to have a performance index of 0.1813. Considering that, the main disadvantage of the Watershed Transform is that for most natural images it produces excessive oversegmentation, so finding the adequate number of imposed minima is very important to this algorithm [2]. It could be said that the number of cluster equals to five was a good approximation to the imposed minima of this image.

According to the Figures 7,9, 9, 11, 13 and 11 it is shown that the space color that yield the best result is lab. This could be explained because the L channel represents the lightness and can give a lot of information of the image. Also, the image that is been tested has colors between blue and yellow and green and red, so this channels can also give relevant information for the algorithm. As each channel gives specific information of the image, there are channels that are more discriminative than others. But this is according to each problem, because the nature of the problem comes with inherited and specific color information.

In order to improve the best method, it would be relevant to consider other descriptors besides color spaces. For example, textures are good descriptors than can provide more information with color spaces. Also, an edge detection descriptor would work with this task because the aim is to segmentate instances. Also, as mentioned before a correct tuning of the hyperparameters can improve the performance of the method.

The main limitation of the methods proved in this study is the selection of the number of clusters. This hyperparameter determines the quality of the segmentation. Although, there are approximations for its selection, the results are subjective. Other limitations inherent of the nature of the problem are tuning the hyperparameters. For example, GMMs have many parameters that influence on the results and the wrong selection according to the problem would give different results.

4.1. Evaluation Method

Evaluation method proposed in this report even though it represents a valid value for analyzing a general evaluation needs to included all possible groundtruths, a mean of the metric choose is valid if the problem is considered in general. The metric that may be used its the one proposed by berkeley, using boundaries precision recall[1]. The evaluation method proposed have shortcuts in the way that it does not take into account all proposed groundtruths available in the database, and if many clusters are used the metric tends two be lower due to variability in clusters size, making little accuracy per class generating lower jaccard indexes.

5. Conclusions

The relevance of the hyperparameter number of cluster has been proved in the present study. It was shown that this parameter is inherent of each dataset and; although there are some approximations. This value has to be iterated in order to obtain a better performance.

On the other hand, for the 12003.jpg image on the BDSD dataset the best clustering method was watersheds. It can be said, that all the algorithms studied have benefits and disadvantages. However, the selection of which one to use is going to be determined with the dataset. As a recommendation for future works, initiate with the simplest method (k-means) and then, start to tune it or change the clustering algorithm. Evaluation method for this type or problems are significantly important mainly because they have to potrait development of the algorithm, and as mention the methodology proposed is sensible to larger number of clusters, generating the need to better fitted problem evaluation methods. Finally, as future work is highly recommended to use more descriptors to segmentate in a proper way. The use of textures can be useful in the image used in this study because of the patterns inside the sea star.

References

[1] The Berkeley Segmentation Dataset and Benchmark.

[2] N. Amoda and R. K. Kulkarni. Image segmentation and detection using watershed transform and region based image retrieval. *Int. J. Emerg. Trends & Techno. Comp. Sci*, 2:89–94, 2013.

[3] N. Dhanachandra, K. Manglem, and Y. J. Chanu. Image segmentation using k-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54:764–771, 2015.

[4] DisplayR. What is hierarchical clustering? Available: https://www.displayr.com/what-is-hierarchical-clustering/.

[5] A. Jurio, M. Pagola, M. Galar, C. Lopez-Molina, and D. Paternain. A comparison study of different color spaces in clustering based image segmentation. volume 81, pages 532–541, 06 2010.

[6] H. Ng, S. Ong, K. Foong, P. Goh, and W. Nowinski. Medical image segmentation using k-means clustering and improved watershed algorithm. In *2006 IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 61–65. IEEE, 2006.

[7] H. Permuter, J. Francos, and I. Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition*, 39(4):695–706, 2006.

[8] R. Ray, S. Turi. Determination of number of clusters in k-means clustering application in colour image segmentation. *Monash Univerisy*. Available: http://users.monash.edu/ roset/papers/cal99.pdf.