AN IMPROVED K-MEANS CLUSTERING ALGORITHM FOR IMAGE SEGMENTATION

WILLIAM ROBSON SCHWARTZ
University of Maryland, Department of Computer Science
College Park, MD, USA, 20742-327, schwartz@cs.umd.edu

RICARDO DUTRA DA SILVA
Department of Computer Science, Federal University of Paraná
Curitiba-PR, Brazil, 81531-990, rdsilva@inf.ufpr.br

RODRIGO MINETTO
Institute of Computing, State University of Campinas
Campinas-SP, Brazil, 13084-971, rdminetto@yahoo.com

HÉLIO PEDRINI
Department of Computer Science, Federal University of Paraná
Curitiba-PR, Brazil, 81531-990, helio@inf.ufpr.br

Abstract. Image segmentation is a primary step in many computer vision applications, whose purpose is to extract information from the images to allow the discrimination among different objects of interest. This task usually involves the partitioning of the image into a number of clusters, such that the data in each cluster share similar features. This work describes a new clustering algorithm for providing a more suitable coarse segmentation, used for parameter estimation. Experimental results and comparisons to other techniques are presented and discussed to demonstrate the effectiveness of the proposed method.

1 Introduction

Image segmentation is of great interest in a variety of scientific and industrial fields, with applications in medicine, microscopy, remote sensing, control of quality, retrieval of information in graphic databases, among others. The segmentation process is usually based on gray level intensity, color, shape or texture.

Several image segmentation approaches have been proposed in literature [1, 2]. Methods based on different stages generally present more reliable results, since they first obtain a coarse segmentation and afterwards refine it. Furthermore, they allow a hierarchical segmentation, improving the quality of the results and speeding up the overall segmentation process.

A common multi-stage approach performs a coarse segmentation to estimate a set of parameters, which are used to obtain a fine segmentation [3, 4]. The parameters are usually estimated by either using a priori information, such as a training set describing each of the different regions in the image (known as supervised segmentation), or extracting information directly from the images, since it is not available in most cases (known as unsupervised segmentation).

In the unsupervised approach, a clustering algorithm can be applied to produce a coarse segmentation. After generating the clusters, a parameter estimation is performed. Clustering is a common technique for statistical data analysis, which consists of partitioning the data set into clusters (subsets), so that the data in each cluster share certain common characteristics. The similarity between the clusters is defined according to a distance measure.

The main advantages in considering clustering algorithms are that the only extra information needed is the number of classes present in the image (cluster validation criteria could also be used to detect the number of classes [5]) and the output of the clustering algorithm could be directly used to estimate the parameters once each sample corresponds to a region located in the image. However, the clustering algorithms found in the literature for obtaining the coarse segmentation have some disadvantages. First, certain instability may occur due to initialization. Second, the definition of thresholds to compute similarity matrices is not trivial. Finally, the existing methods assign all samples to at least one cluster as result,

which is not adequate to parameter estimation, since mislabeled samples would be considered. Therefore, a clustering algorithm that reduces such problems is more suitable to generate a coarse segmentation.

This work presents a new clustering algorithm for improving the resulting coarse segmentation and, consequently, the quality of the final segmentation. The algorithm is based on multiple executions of the K-means algorithm [6] for computing the similarity matrix used to define which samples belong to a same cluster and those that should be removed before performing the parameter estimation.

The paper is organized as follows. Section 2 describes some relevant related work. The proposed method is presented in Section 3. Experimental results and a comparison to other methods are shown in Section 4. Finally, the conclusions are presented in Section 5.

2 Related Work

In unsupervised image analysis, clustering algorithms partition data into a number of categories or subsets, known as *clusters*. Ideally, the data in each cluster should share common features. The homogeneity between clusters is defined according to distance measures. Typical measures include Euclidean, Manhattan and Mahalanobis distance.

Several clustering algorithms have been proposed in the literature, applied in a wide variety of fields [7]. A possible taxonomy categorizes the clustering algorithms as *hierarchical* or *partitional* [7]. Hierarchical algorithms determine successive clusters using previously established clusters, whereas partitional algorithms directly divide data regions into some predefined number of clusters. Hierarchical algorithms can be classified as *divisive (top-down)* or *agglomerative (bottom-up)*. Divisive algorithms start with the entire set and proceed to subdivide it into successively smaller clusters. Agglomerative algorithms start with each element as a separate cluster and merge them in successively larger clusters.

In hierarchical clustering algorithms, data are organized into a hierarchical structure according to the similarity matrix, whose results are commonly represented by a dendogram or a binary tree [8]. Several agglomerative clustering algorithms have been proposed, they usually differ in the distance function used to determined the merge sequence. Some known algorithms include single link, complete link, average link, centroid link, and Ward's method [8].

In constrast to hierarchical clustering, partitional clustering assigns a set of objects into k clusters without using a hierarchical structure. Some important partitional clustering algorithms are QT (Quality Threshold) clustering algorithm [9], fuzzy c-means clustering [7], K-means algorithm [6], and Gaussian Mixtures [7]. In the K-means algorithm, initially, k clusters are randomly or directly generated, considered as their centers or centroids. Then, each sample point is assigned to the nearest centroid. After that, the centers are recalculated and these two last steps are repeated until a convergence criterion is met. The result of K-means is not guaranteed to be a global optimum. The quality depends on the chosen centroids, such that a different initialization can generate a different resulting clustering.

3 Modified Clustering Algorithm

The proposed clustering algorithm is based on the relationship between pairs of samples by assuming that if two samples are clustered together after several executions of K-means, both should belong to a same cluster in the final clustering. Considering this scheme for defining the clustering algorithm, the result is not very affected by the initialization of K-means. This occurs because local relationships are being considered during many executions of K-means, which reduce considerably the instability found in a single execution.

Based on the previous assumption, the clustering algorithm described in the next paragraphs is applied to compute a coarse segmentation (the steps performed by the proposed algorithm are summarized in Figure 1). However, before its execution, the input image is divided into a number of square windows, such that overlapping between distinct regions is allowed; measures are extracted from each window to compose a feature vector. Afterwards, these resulting feature vectors are used to decide which regions of the image are clustered together.

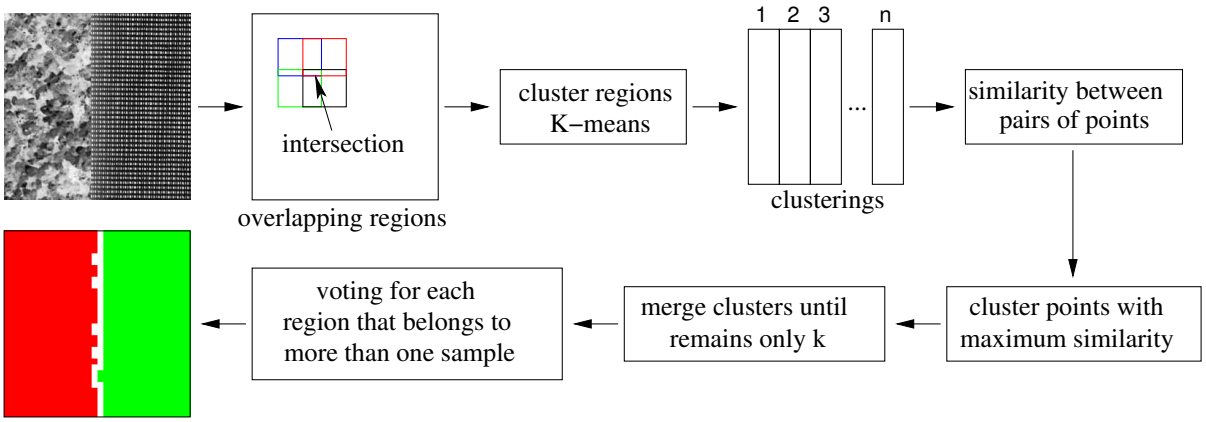


Figure 1: Steps performed in the proposed clustering algorithm. After several executions of K-means, the sample points are clustered according to their similarity. Finally, the number of possible misclassified samples is reduced during the voting step, by considering the labeling of the overlapping windows.

Six measures are extracted using a Daubechies wavelet transform [10], which decomposes the image into subimages. After a two-level Daubechies wavelet decomposition, the resulting six subbands with higher frequencies are used to compute the energy of each block. Once relevant texture information has been removed due to lowpass filtering, the energy of the low frequency subband is not considered as a texture measure [10]. Furthermore, three additional measures are extracted using the texture unit method [11], which is responsible for disclosing the global aspect of a texture. Such measures compose a 9-dimensional feature vector.

Once the image is sampled and the measures are computed, the clustering algorithm may be applied to cluster those feature vectors presenting similar features. A graph is used to encode the relationship between pairs of samples. Let $G = \{V, E\}$ be a complete graph composed of n nodes, where n is the number of sample points considered in the clustering. For each execution of K-means, if two sample points v_i and v_j are set to the same cluster, the weight of the edge e_{ij} is increased by one. Since the edges have weights according to the cluster in which two nodes are assigned, this graph represents a similarity measure among the samples, whose values are computed based on multiple executions of K-means algorithm, instead of a single distance calculation followed by a threshold, as usually done.

Although clustering methods usually split the graph into a desired number of subgraphs, the proposed clustering algorithm creates a new graph based on the described similarity measure. An initially unconnected graph is merged until it reaches the desired number of clusters. The initial clustering contains edges with the maximum possible weight in the graph G; in other words, edges between those pairs of sample points which are assigned to a same cluster by K-means during all its executions. After this first step, the new graph has a forest with some subgraphs and some isolated nodes. One important point is that no subgraph of this graph can be further split because its nodes have the maximum similarity, based on the results of the K-means; therefore, the number of clusters is not greater than the number of isolated components present in this graph.

After having the initial forest, isolated components are linked until their number reaches the desired number of clusters. To perform this task, the average similarity between every two pairs of isolated components is computed and that pair which presents the highest average is merged into a single component.

Since it is allowed to have overlapping of distinct regions before applying the clustering algorithm, a region in the image might belong to more than one sample in the clustering. Knowing this information, a region in the image is labeled only if all samples corresponding to that region are clustered in the same class. This last step, referred to as voting, is performed to minimize the number of mislabeled regions in the coarse segmentation, which is obtained according to the labels assigned to each region in the image.

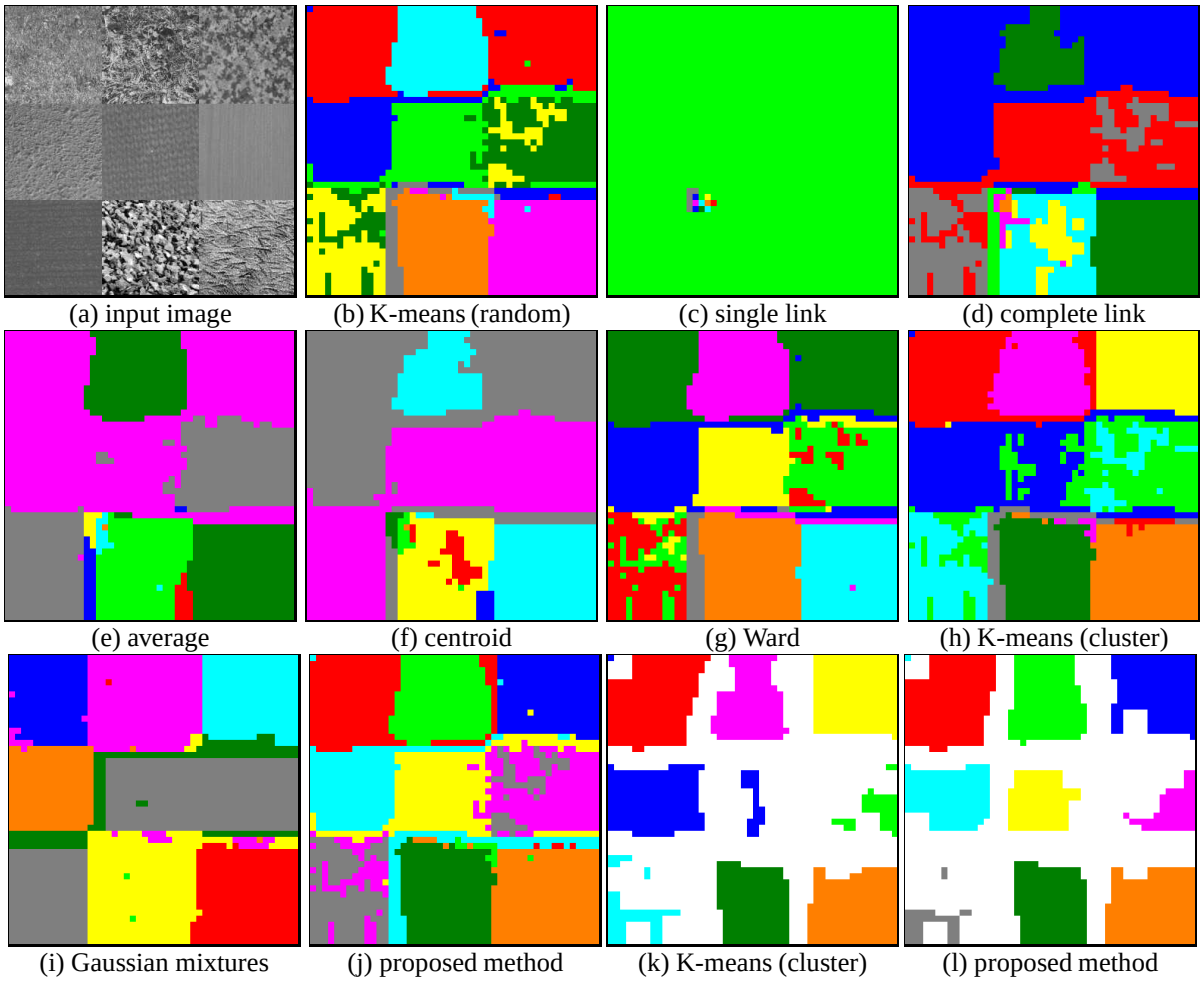


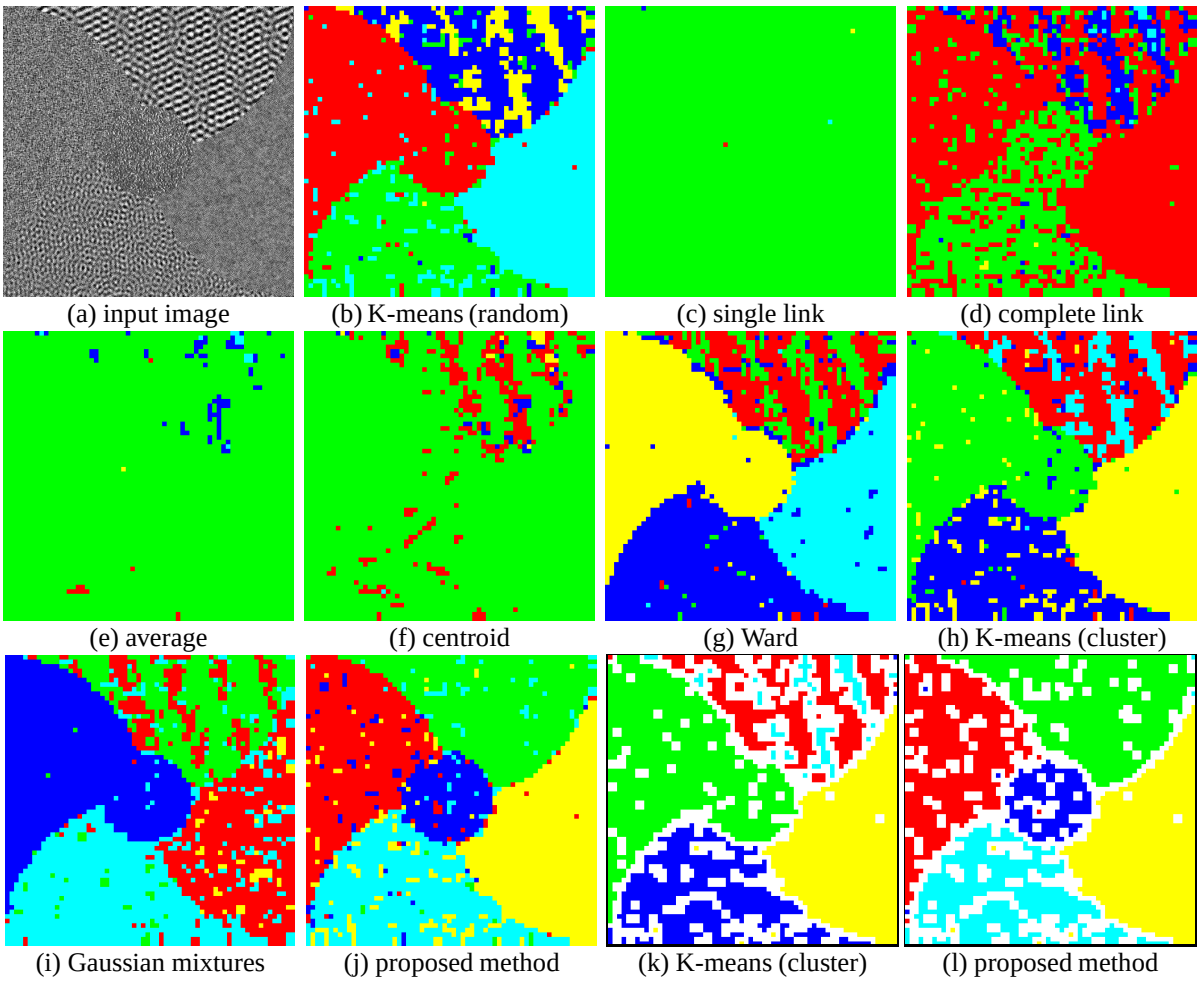
Figure 2: Results obtained by several clustering algorithms in the texture mosaic in (a), composed of 9 classes. (b)-(j) clustering without voting; (k)-(l) results after voting (coarse segmentation) for the method which performed the second best result and for the proposed method (please refer to the color version to a better view of the results).

Therefore, applying the voting step the parameter estimation based on the coarse segmentation can be obtained from regions which are more likely to belong to correct classes.

4 Experimental Results

Results obtained by applying the clustering algorithm described in Section 3 are shown and compared to several other algorithms. The accuracy of the coarse segmentation is also shown, when applied to texture mosaics. In order to have a quantitative evaluation, synthetic texture mosaics from University of Oulu dataset [12] were chosen, since a ground truth map is directly available. These two mosaics (Figures 2(a) and 3(a)) have particular interest for tests since the first presents an elevate number of classes for the segmentation task, and the second has no angular edges between classes and mainly because it has a class with different number of samples (the class in the center has a small number of samples for the clustering).

The algorithms used in our experiments are: K-means, with random initialization and a preliminary clustering phase on random 10% subsample, both executed 50 times and the best clustering was returned; hierarchical algorithms with single linkage, complete linkage, average, centroid, and Ward methods; and mixture of Gaussians with random point initialization. The experiments aim at showing the clustering obtained with the original clustering algorithms, that is, without applying the voting step (Figures 2(b-j)



(a) input image (b) K-means (random) (c) single link (d) complete link

(e) average (f) centroid (g) Ward (h) K-means (cluster)

(i) Gaussian mixtures (j) proposed method (k) K-means (cluster) (l) proposed method

Figure 3: Results obtained by several clustering algorithms in the texture mosaic in (a), composed of 5 classes. (b)-(j) clustering without voting; (k)-(l) results after voting (coarse segmentation) for the method which performed the second best result and for the proposed method.

and 3(b-j)), and after voting for the described method and that one which has the second best accuracy (Figures 2(k-l) and Figures 3(k-l)). These last results correspond to the coarse segmentation.

The accuracy (percentage of samples classified correctly according to the ground truth map) for these two experiments is shown in Tables 1 and 2, where the second column presents the accuracy for the original method (without voting), and the third column shows the accuracy obtained within those regions that were classified after voting (non-white regions in the figures).

Method	Original	Voting
Hier. single link	0.114	0.114
Hier. complete link	0.442	0.545
Hier. average	0.414	0.455
Hier. centroid	0.398	0.456
Hier. Ward	0.711	0.839
K-means (cluster)	0.716	0.978
Gaussian mixtures	0.576	0.698
K-means (random)	0.678	0.856
Proposed method	0.775	0.999

Table 1: Accuracy obtained from Figure 2.

Method	Original	Voting
Hier. single link	0.118	0.115
Hier. complete link	0.451	0.550
Hier. average	0.431	0.436
Hier. centroid	0.370	0.388
Hier. Ward	0.688	0.883
K-means (cluster)	0.786	0.885
Gaussian mixtures	0.767	0.864
K-means (random)	0.762	0.868
Proposed method	0.895	0.992

Table 2: Accuracy obtained from Figure 3.

Although the merge step of the described clustering algorithm is similar to the hierarchical algorithms, as can be seen, the results outperform them. This occurs because the similarity between the samples is computed using not a function of the distance computed only once, but a function of the relation between pairs of samples computed many times to avoid possible unstable behavior once the algorithm used to find the similarity is the K-means. According to the results, mainly those after the voting step, it can be seen that the error in the classification is low (computed only in non-white regions) for the method described in Section 3. Therefore, the sampling for the parameter estimation, obtained from the coarse segmentation, might be done from regions which present low error, avoiding possible mischaracterization of a class. The problem of mischaracterization can be seen in Figure 2(k) and 3(k), where both clusterings present misclassification between two classes. The features were not enough to separate the two classes in the first case, whereas the problem in the second case is due to the reduced number of samples in the class located in the mosaic center.

Finally, once the clustering is only used to find a coarse segmentation from where the parameters are estimated after sampling, this clustering algorithm is suitable for image segmentation because it removes the outliers (misclassified samples), usually located in the frontiers between two classes.

5 Conclusions

The clustering algorithm described allows a more accurate parameter estimation, once the voting step removes some regions from the image in order to decrease the probability of misclassification. Furthermore, without considering the voting step, either this algorithm might be used as a regular clustering algorithm or the graph G described in Section 3 can be used by other methods as a similarity graph.

References

[1] N. R. Pal and S. K. Pal, "A Review on Image Segmentation Techniques," *Pattern Recognition*, vol. 26, no. 9, pp. 1277–1294, 1994.

[2] Y. J. Zhang, "A Survey on Evaluation Methods for Image Segmentation," *Pattern Recognition*, vol. 29, no. 8, pp. 1335–1346, Aug. 1996.

[3] H. Derin and H. Elliott, "Modeling and Segmentation of Noisy and Textured Images Using Gibbs Random Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 1, pp. 39–55, Jan. 1987.

[4] S. Liapis, E. Sifakis, and G. Tziritas, "Colour and Texture Segmentation Using Wavelet Frame Analysis, Deterministic Relaxation, and Fast Marching Algorithms," *Journal of Visual Communication and Image Representation*, vol. 15, no. 1, pp. 1–26, Mar. 2004.

[5] U. Maulik and S. Bandyopadhyay, "Performance Evaluation of Some Clustering Algorithms and Validity Indices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1560–1654, Dec. 2002.

[6] J. B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in *5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California, Berkeley, 1967, vol. 1, pp. 281–297.

[7] R. Xu and D. Wunsch, "Survey of Clustering Algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, May 2005.

[8] A. Jain and R. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1988.

[9] Laurie J. Heyer, Semyon Kruglyak, and Shibu Yooseph, "Exploring Expression Data: Identification and Analysis of Coexpressed Genes," *Genome Research*, vol. 9, no. 11, pp. 1106–1115, Nov. 1999.

[10] G. Van de Wouwer, P. Scheunders, and D. Van Dyck, "Statistical Texture Characterization from Discrete Wavelet Representations," *IEEE Transactions on Image Processing*, vol. 8, no. 4, pp. 592–598, Apr. 1999.

[11] D. C. He and Li Wang, "Texture Unit, Texture Spectrum, and Texture Analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 28, no. 4, pp. 509–512, July 1990.

[12] P. P. Ohanian and R. C. Dubes, "Unsupervised Texture Segmentation Dataset," `http://www.ee.oulu.fi/research/imag/texture/texture.php`.