The Role of Quantitative Metrics in Enhancing Spatial Information Retrieval via Fuzzy K-Means Clustering

Zhengmao Ye, Habib Mohamadian
Southern University, Baton Rouge, Louisiana, USA

Abstract: Remote sensing information is critical to preserve environments in tropical and mountain regions (e.g., costal hurricane mitigation or mountain forest management). In order to partition graphical information into meaningful regions and extract salient objects, segmentation is used to identify visual signatures based on similarity criteria. Due to atmospheric dispersing, nonlinear denoising is needed to capture intrinsic information. To identify illustrious objects and regions, image segmentation using K-Means clustering is generally applied to partition information into diverse clusters. Since each pixel might have certain degree of belongings to multiple clusters, fuzzy K-Means clustering is introduced to depict belongings using fuzzy membership functions, which classifies pixels into two or more clusters. Without general rules to determine an optimal number, the key problem is to specify the desired number of clusters. Quantitative measures are also proposed to identify the actual number of clusters to enhance decision support accuracy and optimize K-Means clustering.

KEY WORDS: Digital, Vision, Engineering, Environment

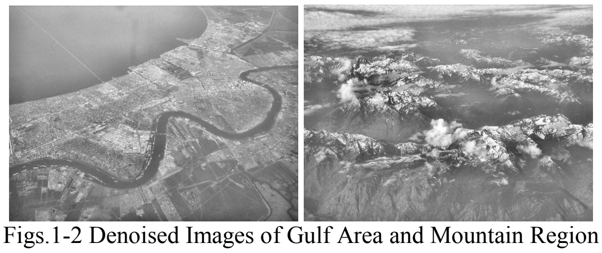
1. INTRODUCTION

Spatial image segmentation is one of leading approaches for automated object recognition in remote sensing. It has various applications on weather forecasting and target detection [1-4]. Since remote sensing data are affected by atmospheric dispersions, denoising techniques should be used to eliminate noise and preserve true information [5-6]. Diverse segmentation methodologies have been proposed. In some cases, artificial intelligence is introduced. Clustering using swarm algorithms is an alternative to hierarchical and K-Means clustering. Particle Swarm Optimization (PSO) clustering performs a global search for the corrupt images due to occlusions. It generates compact clustering results [7]. PSO can also be used to implement mixel decomposition. It combines with linear mixels decomposition model. Mixel decomposition of the remote sensing images is to improve qualities of feature extraction like distortion in linear mixel decomposition. It presents robustness to environment [8]. Ant colony optimization (ACO) based clustering provides improvement in clustering the 3D data with the higher density and complexity. By hybridization of its foraging behavior and K-Means, the quality and processing time become much better than other techniques [9]. ACO can also be used to solve complex remote sensing classification. It takes into account data correlation between the attribute variables. Discretization technique is incorporated so that classification rules can be induced from large data sets of remote-sensing images. It yields better accuracy than decision tree methods [10]. A supervised classification technique for hyperspectral imagery is utilized as a feature extractor to generate a particular feature eigenspace for each class presented in hyperspectral data. It consists of both the greedy modular eigenspace and positive Boolean function. Compared to principal components analysis (PCA), it significantly increases the accuracy and dramatically improves the decomposition computational complexity [11].

Taking into account of the computational complexity and uncertain nature of remote sensing, as an unsupervised clustering technique, K-Means clustering is still among the most applicable approaches. It is to classify an image into parts that have a strong correlation with objects to reflect the actual information collected in real world. In practice, there is no distinct boundary between clusters, thus fuzzy K-Means clustering can be further proposed for data analysis. Fuzzy membership functions are used to describe the degrees of belonging to clusters. The fuzzy membership function is dependent on the distance between the image pixel and those independent cluster centroids. Fuzzy partitioning is carried out through iterative optimizing a well defined cost function. The iteration continues until the cost function converges to the minima. K-Means clustering requires that certain number of clusters for partitioning be specified and its distance metrics be proposed to quantify relative orientation of objects. The optimal algorithms can be categorized as threshold based, region based, edge based or surface based. How to choose a number of clusters will directly affect the overall outcome of segmentation [12-16]. This article is dealt with the impact of "number of clusters" on remote sensing data processing. Besides the decision making via visual appealing, quantitative metrics are also introduced to evaluate the outcomes of spatial image segmentation using fuzzy K-Means clustering [4, 12, 17].

2. DISCRETE WAVELET TRANSFORM DENOISING

Discrete wavelet transform is introduced for noise filtering of spatial digital images. It is a multiple level transformation. The decomposition outputs at each level include: approximation, horizontal detail, vertical detail and diagonal detail. The detail components will be retained and approximation component can be further decomposed into multiple levels. For denoising using DWT, wavelet coefficients of detail components are subject to thresholding, while the approximation component at each level would be retained for image reconstruction. Soft thresholding is selected instead which shrinks nonzero wavelet coefficients towards zero. In general a small threshold produces a fine but noisy estimation while a large threshold produces a smooth but blurring one. The median value at each decomposition level is selected. Using DWT, two resulting denoising images taken from the tropical area and mountain region are shown in Fig. 1 and Fig, 2, respectively.



Figs.1-2 Denoised Images of Gulf Area and Mountain Region

3. NONLINEAR FUZZY K-MEANS SEGMENTATION

K-Means clustering classifies data sets through K numbers of clusters, where each data point inside will be assigned to certain location. In fuzzy K-Means clustering, instead of belonging to one cluster exclusively, an individual point can belong to more than one cluster with certain probability (degree of belonging). Points along borders between the clusters would have lesser degree of belonging than those points around centroids. Each point x has a degree of belonging to a cluster. The sum of the degrees of belonging for any individual point is defined to be 1.

$$\sum_{i=1}^{K}\mu_i(x)=1 \quad (1)$$

In fuzzy K-Means clustering, the centroid of a cluster is the mean of all points, weighted by the fuzzy membership. The degree of belonging is related to the inverse of the distance metric to the cluster centroid. It is then normalized and fuzzified with the specified parameter m > 1.

$$c_i=\sum_{j=1}^{K}\mu_{ij}^{m}x_j \,/\, \sum_{j=1}^{K}\mu_{ij}^{m} \quad (2)$$

$$\mu_{ij}=1/\sum_{j=1}^{K}\frac{||x_i^j - c_j||}{||x_i^j - c_j||^{2/(m-1)}} \quad (3)$$

Fuzzy K-Means clustering optimizes the cluster centroid by iteratively adjusting positions and evaluating a defined cost function. The cost function can be formulated as:

$$F=\sum_{j=1}^{K}\sum_{i=1}^{N}\mu_{ij}^{m}||x_i^j - c_j||^2 \quad (4)$$

where μ_{ij} is the fuzzy membership of a point xi in a cluster j, xi is the ith element of N-dimensional data set, cj is the j-dimension centroid of the clusters, and m is a constant that defines the fuzziness of segmentation outcomes. F can reach the global minimum when points around centroids are assigned bigger membership values, while those points far away from centroids are assigned with smaller membership values. This procedure is continued until optimal object assignment for all clusters is reached. With fuzzy K-Means, any cluster centroid is the statistical mean of all points, weighted by the degree of belonging to clusters (fuzzy memberships). However, the ultimately generated cluster assignment highly depends on the initial number of clusters and initial cluster assignments. Especially, results are very sensitive to the actual choice of the cluster number. In this case, further analysis needs to be conducted to determine the desired number of clusters to optimize the segmentation process. There is no distinct rule to follow how to come up with the best number of clusters. In this case, this article intends to study the impact of the number of clusters from both subjective and objective points of view.

4. Fuzzy Segmentation with K Centers

Segmentation of remote sensing data requires the number of clusters should be specified for partitioning, while the centroid of each cluster must also be defined initially, representing mean values of all data points with fuzzy membership functions for that cluster. In this section, two denoised spatial images are both selected for segmentation. Fuzzy K-Means clustering is implemented and the number of clusters is assigned to be from 2 to 16 to make suitable comparisons for each selection, respectively. The clustering outcomes are shown in Figs 3-11, respectively, where spatial images of Mississippi-Gulf area are placed to the left and spatial images of US-CA mountain region are placed to the right. From cases with 2 clusters to 6 or 7 clusters, visual appealing does improve dramatically. Beyond 7 clusters until 16 clusters, visual appealing does not illustrate significant improvement, though.

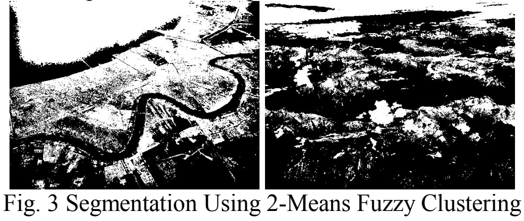


Fig. 3 Segmentation Using 2-Means Fuzzy Clustering

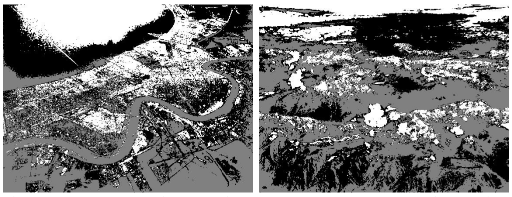


Fig. 4 Segmentation Using 3-Means Fuzzy Clustering

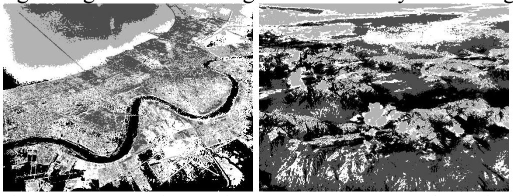


Fig. 5 Segmentation Using 4-Means Fuzzy Clustering

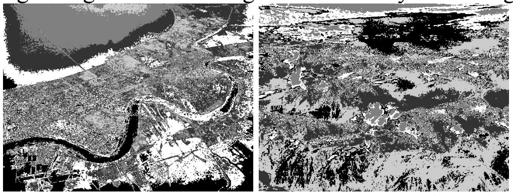


Fig. 6 Segmentation Using 5-Means Fuzzy Clustering

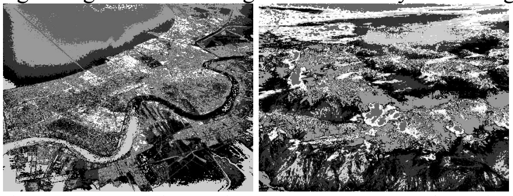


Fig. 7 Segmentation Using 6-Means Fuzzy Clustering

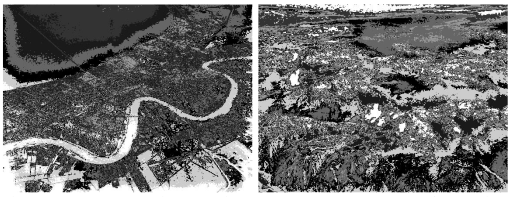


Fig. 8 Segmentation Using 7-Means Fuzzy Clustering

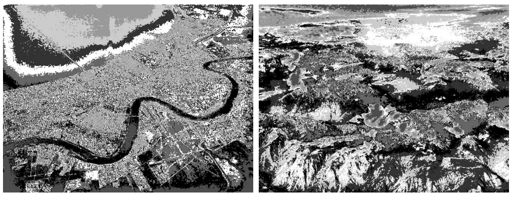


Fig. 9 Segmentation Using 8-Means Fuzzy Clustering

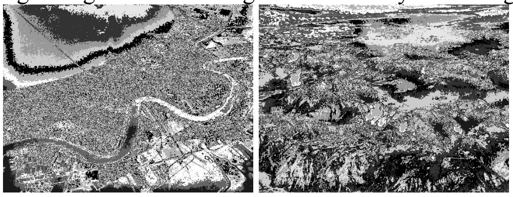


Fig. 10 Segmentation Using 10-Means Fuzzy Clustering

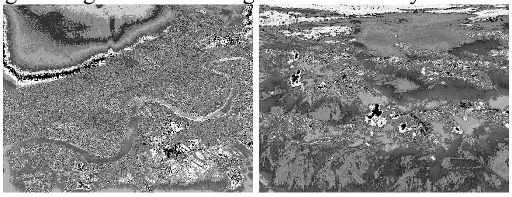


Fig. 11 Segmentation Using 16-Means Fuzzy Clustering

5. Quantitative Analysis

Subjective evaluation is conducted. In this section, objective evaluation will also be made. All digital images with M × N pixels have been considered. Occurrence of the gray level is described as co-occurrence matrix of relative frequencies. The occurrence probability is then estimated from its histogram. All these quantitative measures will be applied for fuzzy K-Means clustering evaluation.

5.1 Discrete Entropy

The discrete entropy is a measure of information content, which can be interpreted as the average uncertainty of the information source. The discrete entropy is the summation of products of the probability of the outcome multiplied by the logarithm of the inverse of probability of the outcome, taking into considerations of all possible outcomes {1, 2, …, n} as the gray level in the event $\{x_1, x_2, \ldots, x_n\}$, where p(i) is the probability at the level i, which contains all the histogram counts (5).

$$H(x)=\sum_{i=1}^{k} p(i)\log_2\frac{1}{p(i)} = -\sum_{i=1}^{k} p(i)\log_2 p(i) \tag{5}$$

5.2 Discrete Energy

The discrete energy measure indicates how the gray level elements are distributed. Its formulation is shown in (6), where E(x) represents the discrete energy with 256 bins and p(i) refers to the probability distribution functions at different gray levels, which contains the histogram counts. For a constant value of the gray level, the energy measure reaches its maximum value of one. The larger energy is corresponding to a lower gray level number and the smaller one is corresponding to a higher gray level number.

$$E(x)=\sum_{i=1}^{k} p(i)^2 \tag{6}$$

5.3 Correlation

Correlation is a standard measure of image contrast to analyze linear dependency on the gray levels of neighboring pixels. It indicates the amount of local variations across the gray level image. The higher the contrast is, the sharper the structural variation is. This measure is formulated as (7):

$$COR=\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\frac{(i-\mu_i)(j-\mu_j)}{\sigma_i\sigma_j}g(i,j) \tag{7}$$

where i and j are coordinates of the co-occurrence matrix; M and N represent total numbers of pixels in row and column of the digital image; g(i, j) is the element in the co-occurrence matrix at the coordinates i and j. $\mu_{i,j}$ and $\sigma_{i,j}$ are the horizontal means and variances, respectively.

5.4 Dissimilarity

The dissimilarity between two gray level images is regarded as the distance between two sets of co-occurrence matrix representations. It is based on the local distance representation, which is formulated as (8):

$$DisSim=\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} g(i,j)\,|i-j| \tag{8}$$

where g(i, j) is an element in the co-occurrence matrix at the coordinates i and j; M and N represent total numbers of pixels in the row and column of the digital image.

5.5 Homogeneity

This measure is a direct measure of the local homogeneity of a gray level image, which relates inversely to the image contrast. Higher values of homogeneity measures indicate less structural variations and lower values indicate more structural variations. Larger values are corresponding to higher homogeneity and smaller values are corresponding to lower homogeneity. It is formulated as (9):

$$Homogeneity=\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\frac{1}{1+(i-j)^2}g(i,j) \tag{9}$$

5.6 Mutual Information

Another metric of the mutual information I(X; Y) can also be applied, which is employed to describe how much information one variable tells about the other variable. The relationship is formulated as (10).

$$I(X;Y)=\sum_{X,Y} p_{XY}(X,Y)\log_2\frac{p_{XY}(X,Y)}{p_X(X)p_Y(Y)} \tag{10}$$

where H(X) and H(X|Y) are values of the entropy and conditional entropy; p_{XY} is the joint probability density function; p_X and p_Y are marginal probability density functions. It can be explained as information that Y can tell about X is the reduction in uncertainty of X due to the existence of Y.

6. Numerical Simulations

Using a set of well defined information metrics, all computation results of two segmented spatial images across diverse numbers of clusters are documented, which are shown in Table 1 and Table 2, respectively. Two sets of simulation results are also plotted in Figs. 12-13. It is indicated again (Tables 1-2 and Figs. 12-13) that the cases with 6 or 7 clusters are the most desirable choice for fuzzy K-Means clustering. Additional number of clusters will increase computational cost, however, values of quantitative metrics vary very little.

Table 1 Quantitative Metrics of Lake View

Clustering # Metrics	**K=2**	**K=3**	**K=4**	**K=5**	**K=6**
Discrete Entropy	2.7530	4.9148	5.8619	6.3566	6.7467
Discrete Energy	0.3535	0.0944	0.0450	0.0352	0.0299
Correlation	0.7624	0.6347	0.6700	0.5446	0.6119
Dissimilarity	0.8534	1.1375	0.9862	1.3426	1.0331
Homogeneity	0.8699	0.7576	0.7512	0.6566	0.6749
Mutual Information	4.4353	2.2736	1.3264	0.8317	0.4417
Clustering # Metrics	**K=7**	**K=8**	**K=9**	**K=10**	**K=16**
Discrete Entropy	7.0030	7.1477	7.3633	7.4560	7.6877
Discrete Energy	0.0238	0.0172	0.0156	0.0138	0.0058
Correlation	0.6481	0.5133	0.5860	0.4135	0.2422
Dissimilarity	1.0721	1.4108	1.3043	1.6335	1.5946
Homogeneity	0.6660	0.6135	0.5943	0.5706	0.5286
Mutual Information	0.1853	0.0406	0.0375	0.0267	0.0149

Table 2 Quantitative Metrics of Mountain View

Clustering # Metrics	**K=2**	**K=3**	**K=4**	**K=5**	**K=6**
Discrete Entropy	3.1045	4.6158	5.6773	6.1570	6.7621
Discrete Energy	0.2553	0.1141	0.0584	0.0421	0.0239
Correlation	0.7852	0.6881	0.7545	0.5392	0.6908
Dissimilarity	0.6157	0.9665	0.7962	1.4507	1.1150
Homogeneity	0.9043	0.7546	0.7810	0.6473	0.6689
Mutual Information	3.8486	2.3373	1.2758	0.7960	0.1909
Clustering # Metrics	**K=7**	**K=8**	**K=9**	**K=10**	**K=16**

Discrete Entropy	6.8877	7.0718	7.3280	7.3533	7.6193
Discrete Energy	0.0204	0.0193	0.0139	0.0141	0.0078
Correlation	0.5033	0.7062	0.4950	0.5496	0.4740
Dissimilarity	1.3846	1.0979	1.3962	1.4183	1.0121
Homogeneity	0.6257	0.6452	0.5743	0.5872	0.6430
Mutual Information	0.0653	0.0626	0.0400	0.0375	0.0118

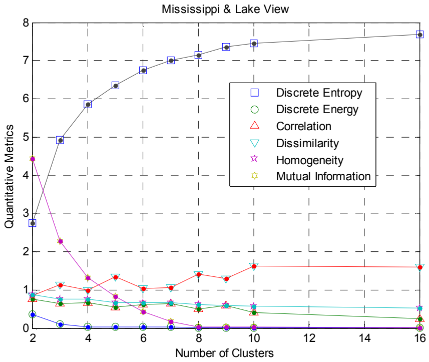


Fig. 12 Metrics of Lake View Fuzzy Clustering (K=2-16)

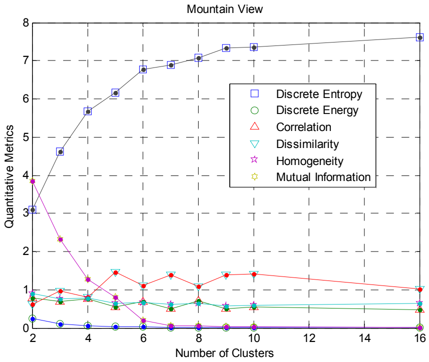


Fig. 13 Metrics of Mountain View Fuzzy Clustering (K=2-16)

7. Conclusions

Nonlinear fuzzy K-Means segmentation has been presented to enhance spatial information retrieval. To reduce blurring effects stem from the atmospheric media, 2D wavelet denoising is employed without distorting key features. As an enhanced unsupervised learning approach, fuzzy K-Means clustering has potential to simplify computation complexity and accelerate the convergence. Fuzzy K-Means clustering also serves as a statistical algorithm, where soft clusters with degrees of belong (fuzzy membership functions) are applied rather than hard clusters with a distinct set of points in conventional K-Means clustering. Determination of actual number of clusters is a trial and error procedure where no rule can be followed. It turns out to be the shortcoming of K-Means algorithms. This research is to seek for the most desirable number of clusters on the segmentation outcomes using fuzzy K-Means clustering. Both qualitative observation and quantitative evaluation are conducted, where a convincing set of quantitative information metrics (discrete entropy, energy, correlation, dissimilarity, homogeneity and mutual information) are applied to depict the impact of the actual cluster number on segmentation results. Based on outcomes, the ideal number of clusters is indicated.

References

[1] R. Gonzalez, R. Woods, “Digital Image Processing,” 3rd Edition, Prentice-Hall, 2007

[2] R. Duda, P. Hart, D. Stork, “Pattern Classification,” 2nd Edition, John Wiley and Sons, 2000

[3] Simon Haykin, “Neural Networks – A Comprehensive Foundation”, 2nd Edition, Prentice Hall, 1999

[4] David MacKay, "Information Theory, Inference and Learning Algorithms", Cambridge Univ. Press, 2003

[5] Z. Ye, H. Mohamadian and Y. Ye, "Information Measures for Biometric Identification via 2D Discrete Wavelet Transform", Proceedings of the 2007 IEEE International Conference on Automation Science and Engineering, pp. 835-840, Sept. 22-25, 2007, Scottsdale, Arizona, USA

[6] M. Ghazel, G. Freeman, and E. Vrscay, “Fractal-Wavelet Image Denoising Revisited”, IEEE Transactions on Image Processing, Vol. 15, No. 9, September, 2006

[7] S. Bedawi and M. Kamel, “Segmentation of Very High Resolution Remote Sensing Imagery of Urban Areas Using Particle Swarm Optimization Algorithm”, Image Analysis and Recognition, Lecture Notes in Computer Science, 2010, Volume 6111, pp. 81-88, 2010

[8] D. Wang, X. Wu and D. Lin, "Particle Swarm Mixel Decomposition for Remote Sensing", IEEE International Conference on Automation and Logistics, 5-7 Aug. 2009, pp. 212 - 216, Shenyang, China

[9] F. Samadzadegan, S. Saeedi, H. Hasani, "Evaluating the Potential of Ant Colony Optimization in Clustering of LIDAR Data", GIS Ostrava 2010, Jan 24-27, 2010

[10] X. Liu, X. Li, L. Liu, J. He, B. Ai, "An Innovative Method to Classify Remote-Sensing Images Using Ant Colony Optimization", IEEE Transactions on Geoscience and Remote Sensing, Vol. 46, No. 12, pp. 4198-4208, 2008

[11] Y. L. Chang, "Greedy Modular Eigenspaces and Positive Boolean Function for Supervised Hyperspectral Image Classification", Optical Engineering 42(09), 2003

[12] Z. Ye, H. Cao, S. Iyengar and H. Mohamadian, "Medical & Biometric System Identification for Pattern Recognition and Data Fusion with Quantitative Measuring", Systems Engineering Approach to Medical Automation, Chapter Six, pp. 91-112, Artech House Publishers, Oct. 2008

[13] Z. Ye, Y. Ye, H. Mohamadian, et. al., "Fuzzy Filtering and Fuzzy K-Means Clustering on Biomedical Sample Characterization", 2005 IEEE International Conference on Control Applications, pp. 90-95, 2005, Toronto, Canada

[14] Z. Ye, J. Luo, P. Bhattacharya and Y. Ye, "Segmentation of Aerial Images and Satellite Images Using Unsupervised Nonlinear Approach", WSEAS Transactions on Systems, pp. 333-339, Issue 2, Volume 5, February, 2006

[15] Z. Ye, “Artificial Intelligence Approach for Biomedical Sample Characterization Using Raman Spectroscopy”, IEEE Transactions on Automation Science and Engineering, Vol. 2, Issue 1, pp. 67-73, January, 2005

[16] M. Jaffar, N. Naveed, B. Ahmed, A. Hussain, A. Mirza, "Fuzzy K-Means clustering with spatial information for color image segmentation", International Conference on Electrical Engineering, 6 pp., April 2009, Lahore, Pakistan

[17] Z. Ye, H. Mohamadian, Y. Ye, "Discrete Entropy and Relative Entropy Study on Nonlinear Clustering of Underwater and Arial Images", 2007 IEEE International Conference on Control Applications, pp. 318-323, Oct. 2007, Singapore