Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 170 (2020) 129–136

Procedia Computer Science

www.elsevier.com/locate/procedia

The 11th International Conference on Ambient Systems, Networks and Technologies (ANT)
April 6-9, 2020, Warsaw, Poland

Discovering similarities in Landsat satellite images using the K-means method

Ariza-Colpas, Paola Patricia[a,b]*, Oviedo-Carrascal, Ana Isabel[b], De-la-Hoz-Franco, Emiro[a]

[a]*Universidad de la Costa, CUC, Barranquilla, Colombia. Street. 58 # 55 - 66 Barranquilla – Colombia*
[b]*Universidad Pontificia Bolivariana, Medellin Colombia. Street. 1 #70-01, Medellín, Colombiac*

Abstract

This article different ways for the treatment and identification of similarities in satellite images. By means of the systematic review of the literature it is possible to know the different existing forms for the treatment of this type of objects and by means of the implementation that is described, the operation of the K-means algorithm is shown to help the segmentation and analysis of characteristics associated to the color. In this type of objects, a descriptive analysis of the results thrown by the method is finally carried out.

Keywords: Clustering, Multiclustering, Multimedia Multidimensional Georreferenced Objects, Satellite Images.

* Corresponding author. Tel.: +57 - 3002287498;
E-mail address: pariza1@cuc.edu.co

10.1016/j.procs.2020.03.017

1. Introduction

A satellite image represents in a visual form the information captured through sensors from an artificial satellite. These sensors collect the information reflected from the earth's surface, which is processed in such a way that it allows the analysis of the identified area. The images obtained by satellite have managed to put in context different fundamental aspects of the Earth and its resources. Currently, making use of remote sensing, has achieved the achievement of profitable information for the achievement of multiple applications where aspects such as orientation, shape and size of objects in a satellite image are identified. Some of the most common applications are: 1) Identification of elements built by man such as roads, airports, vehicles, buildings, etc [1]; 2) Update existing maps [2]; 3) Delimitation between land and water [3]; 4) Quantification of human growth and development [4]; 5) Definition of landuse [5]; 6) Generation of digital elevation models [6]; 7) Distinction of rocks and soil [7]; 8) Delimitation of lands with marshes [8]; 9) Delimitation of the depth of the water [9]; 10) Cataloging of land cover; among other applications [10].

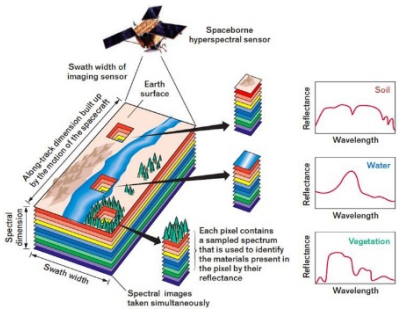

Fig. 1. Graphic scheme of a satellite image

Satellite images are composed of spectral bands, which allow measurements of reflectance at different wavelengths to identify terrain features. Depending on the number of spectral bands, satellite images can be classified as panchromatic, multispectral, super-spectral, hyperspectral and ultra-spectral. The panchromatic images have a single band. Multispectral images have between 2 and 10 bands. Super-spectral images can have between 11 and 100 bands. Hyperspectral images between 101 and 1000 bands. Finally, the ultra-spectral images present more than 1000 bands. The type of satellite image depends on the capture sensor, the most common being multispectral images with bands in visible blue, visible green, visible red, near infrared, medium infrared, among others. As a result of the exploration of the state of the art, the histogram analysis of the images has been constituted as one of the accepted solutions to analyze the color of the different images. In the framework of this experimentation, color analysis will be performed on LandSat satellite images using color histograms and clustering algorithms

2. Brief review of literature

The mining of images has taken a great relevance in research in recent years mainly due to the potential existing in the discovery of patterns among a conglomerate of images. Most of the works that have been developed are focused on the process of identifying the size and position of different types of objects. Kendal, performed this type of analysis applied to the thyroid nucleus, in order to reduce the number of false positives in the detection of this type of

pathologies, using dataset ultrasound images, for which proposed a algorithms called SNDRLS (Spatial Neutrosophic Distance Regularized Level Set) - this algorithm takes as input an approximate region where the nodules are to be found. This algorithm is compared with different existing solutions such as NLM clustering, Active Contour Without Edges (ACWE), Fuzzy C-Means (FCM). To validate the SNDRLS method, the manual demarcations of three expert radiologists are used as a reality of the terrain. The SNDRLS produces the limits closest to the reality of the terrain compared to other methods as revealed by six evaluation measures (true positive rate is 95.45 ± 3.5%, the false positive rate is 7.32 ± 5.3% and the superposition is 93.15 ± 5. 2%, with an absolute mean the distance is 1.8 ± 1.4 pixels, Hausdorff distance is 0.7 ± 0.4 pixels and given metric is 94.25 ± 4 , 6%). The experimental results show that the SNDRLS can delineate multiple nodules in ultrasound images of the thyroid with precision and efficacy. The proposed method reaches the limit of automated nodules even for low contrast, blurred, and noisy ultrasound thyroid images without any human intervention. The software used for the development of this algorithm is Matlab 7.9. [11]

To perform SAR type image analysis, Alias proposes a novel method to reveal differences between this type of images. In order to identify the regions that change and that do not change in the images, the FUZZY K-means algorithm is used, based on the process of treating the mottle noise found in them. In order to be able to identify the existing changes in this type of images, first, the Wavelet Bayesian noise elimination technique is applied to reduce the noise existing in the image. Subsequently, the image fusion technique is introduced to generate an image that contains the differences between the initial image and the one that has subsequently been processed. Later, with the help of FCM, the variation in SAR images is detected. The main advantage of the proposed method is its mastery in the reduction of speckle noise and its computational simplicity. [12]

One of the most relevant tasks in the process of image analysis through data mining techniques is the segmentation process of these images, in different significant classes, which is an important task for the automatic image analysis technique. The finite gaussian model is one of the most popular models for parameterization of images based on segmentation model. In this sense, Banerjee presents a new grouping algorithm, called probabilistic grouping, judiciously integrating the merits of the approximate sets and a new probability distribution, called "Stomped Normal Distribution" (SN). The intensity distribution of any image is modelled as a mixture of finite number of SN distributions. The expectation of maximization of the algorithm is used to estimate the parameters of each class. The incorporation of a random Markov field frame hidden in the probabilistic grouping is proposed as a new method for accurate and robust image segmentation. The performance of the proposed segmentation approach, together with a comparison with related methods, is demonstrated in a set of images of HEp-2 cells, and synthetic and real brain MR images for different polarization fields and noise levels. For the development of this experimentation, the following software was used: MRI: statistical parameter mapping (SPM) software version 8 [13] and FMRIB Software Library (FSL) version 5.0. [14]

Although there are a lot of grouping algorithms proposed in the literature, the results of the grouping of the existing algorithms usually depend in large part on the parameters specified by the user, and it is usually difficult to determine optimal parameters in this way. Hou proposes the use of a data similarity matrix as the input, the dominant sets of the grouping are shown to be a grouping of data to achieve an effective image segmentation approach, in part because of their ability to discover the data structure underlying and determine the number of groups automatically. In order to eliminate the dependence of the parameter specified by the user, it is studied how dominant sets influence the measures of similarity. As a result, it is proposed to transform similarity matrices and the equalization of the histogram before grouping. Although this transformation is shown to eliminate sensitivity to similarity measures effectively, it also results in excessive segmentation. Therefore, in the next stage, a cluster extension method is presented to overcome the over-segmentation effect and generate more reasonable grouping results. [15]

The techniques defined above tend to achieve a good performance of the preservation of important details of the image, while removing the noise in the segmentation process. In this context, Zhang presents a multiobjective fuzzy clustering evolutionary algorithm (MOEFC) to convert diffuse clustering problems by segmenting images into multi-objective problems. The multi-objective problems are optimized by means of an evolutionary algorithm of multiobjective decomposition. The decomposition strategy is adopted to project the multiobjective problem into several sub-problems. Each sub-problem represents a fuzzy grouping problem with local information for image segmentation. Opposition-based learning is used to improve the search capability of the proposed algorithm. The results of this experimentation conclude that the synthetic and real images that illustrate the proposed algorithm can achieve a balance between the preservation of details and the elimination of noise for image segmentation. [16]

In the same way, many investigations have been made regarding the number of optimal groups that must be obtained in the process of analysing the images and the distance between them. To solve this problem, Chang proposes a niching grouping algorithm based on individual connectivity (CIRD) for unsupervised classification without prior knowledge. The goal of this algorithm is to automatically evolve the optimal number of groups, as well as the centres of the conglomerates of the data set based on the k-distance compact environment algorithm. It also provides an application of the color CIRD clustering algorithm of the segmentation image. The experimental results show that the CIRD clustering algorithm has high performance and flexibility.

The representation of the image is done through vectors of representation of the image. Reboul also makes contributions regarding the identification of the number of subsets required for the processing of images through a stochastic algorithm for the identification of homogeneous subgroups of images. The goal of the method is to generate improved 2D class averages that can be used to produce a reliable 3D match model in a quick and unbiased manner. We demonstrate that our method overcomes the limitations of the widely used cluster and proceed to test the approach in six publicly available cryo-EM experimental datasets. The representation of the data is done through a matrix of color intensity levels that is entered into the algorithm.[17]

In this order of ideas Dubey proposes an algorithm based on the traditional algorithm fuzzy c-means (RIFCM) grouping algorithm for the segmentation of brain magnetic resonance imaging (MRI). In the first place, researchers propose a new automated method to determine the initial values of the centroid group using the diffuse roughness measure, which is obtained by analysing the upper approximation of the set based on the histogram. A new diffuse complement function of image representation is proposed to consider the lack of homogeneity of intensity and noise in brain MR images [18]. The results of the segmentation of the proposed algorithm are compared with the fuzzy grouping algorithms that were taken as the basis for the construction of this algorithm. The experimental results show the superiority of the proposed algorithm. The representation of the data is done through the matrix of analysis of color levels. (Dubey, 2016).

Jian says that the Fuzzy C-means clustering algorithm (FCM) considering spatial constraints (FCM_S) is considered as an effective solution to support in image segmentation processes. However, FCM_S has a high computational complexity and still lacks enough robustness to noise and outliers, which will limit its usefulness [19]. To overcome these difficulties, he proposed a new algorithm (LCFCM_S) and its simplified model (LCFCM_S1). The grouping algorithm can efficiently highlight the weights of the samples that are near their centres of the corresponding conglomerates. The experimental results on the synthetic and real images show the superiority of the method in terms of precision and robustness for the segmentation of images with the lack of homogeneity of intensity and noise, when compared with several approaches of the state of the art [20].

3. Materials and Methods

3.1. Description of the experimentation

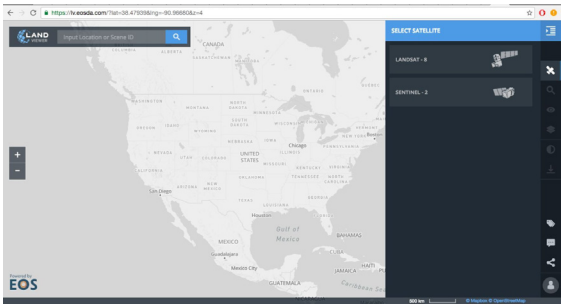

Fig. 2. Software for images download.

For the development of this experimentation the following resources and tools were used: 300 multispectral satellite images taken from the LandSat satellite, through the Land Viewer online tool. Available at https://lv.eosda.com/, Matlab Software and Weka Software. They were first selected from the geographic locations of the Study, using the online tool: https://lv.eosda.com/. In the figure 2, it is shown the software for download the images.

Then, using the tool mentioned above, each of the bands of the images of the selected places were downloaded.

Fig. 3. Image with band combination

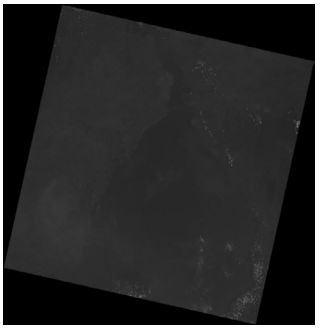

Fig. 4. Red Band

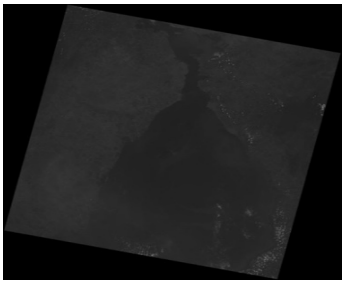

Fig. 5. Green Band

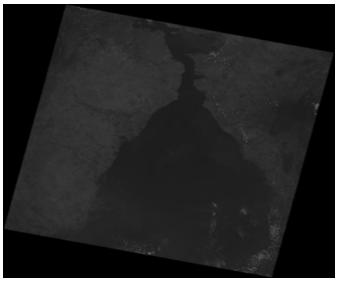

Fig. 6. Blue Band

Using the imhist instruction you can extract the color histogram of each of the images which are exported to an Excel file.

```
I299 : imread("Washington_B2.tiff");
A299=imhist(I299);
B299=A299;

I300 : imread("Washington_B3.tiff");
A300=imhist(I300);
B300=A300;

I301 : imread("Washington_B4.tiff");
A301=imhist(I301);
B301=A301;
HorArray1=(B299,B300,B301)
```

Fig. 7. Matlab script for extract the color histogram.

Considering the information gathered from these images, the following arff file can be built with the following structure:

Table 1. Dataset Structure.

Attribute	Description
Satellite	Satellite used to capture the image
Brightness	The amount of light available in the image
Angle	Angle from which the image is taken
Description	Information on the geographical location of the image.
Hb1p1 ... .. hb1p256	256 columns that refer to the histogram of band 1
Hb2p1 ... .. hb2p256	256 columns that refer to the histogram of band 2
Hb3p1 ... .. hb3p256	256 columns that refer to the histogram of band 3

4. Experimentation

Considering the specific characteristics of each of the arrangement of the images, we proceed to the preparation of the dataset, for the implementation of the K-means method to the data. For the application of the method weka software was used, in order to obtain similarities in the images considering the components of each of their bands. The results of the application of the algorithm were the following:

Table 2. Data Aggregation Results with the Chosen Clustering Method

Cluster	Number of Record
Satellite	Satellite used to capture the image
Brightness	The amount of light available in the image
Hb3p1 ... .. hb3p256	256 columns that refer to the histogram of band 3

5. Results

As a result of the implementation process, the following clusters that were formed can be evidenced. In cluster 0, where 55% of the records are located. Group R, G, B bands between pixels 112 to 116, higher concentration of R. In the case of Cluster 1, where 15% of the records are represented, Group R, G, B bands between pixels 32 to 36, higher concentration of G, and the cluster 2, where 21% of the record are located. Group R, G, B bands between pixels 33 to 42, higher concentration of G.

6. Conclusions

From the approximations found in the literature, it is possible to identify the high dimensionality of each of the components of georeferenced satellite images. Therefore, challenges arise such as: The analysis of the characteristics that can be extracted from this type of objects present challenges related to the integration of different intrinsic aspects of these and the accuracy in processing. The analysis of the completeness of the intrinsic variables that can be associated to an image such as the incorporation of audio, video, metadata, text, geolocation.

References

[1] Ariza-Colpas, P., Morales-Ortega, R., Piñeres-Melo, M. A., Melendez-Pertuz, F., Serrano-Torné, G., Hernandez-Sanchez, G., ... & Collazos-Morales, C. (2019, October). Teleagro: Software Architecture of Georeferencing and Detection of Heat of Cattle. In Workshop on Engineering Applications (pp. 159-166). Springer, Cham.

[2] Ariza, P., Pineres, M., Santiago, L., Mercado, N., & De la Hoz, A. (2014, November). Implementation of moprosoft level I and II in software development companies in the colombian caribbean, a commitment to the software product quality region. In 2014 IEEE Central America and Panama Convention (CONCAPAN XXXIV) (pp. 1-5). IEEE.

[3] Calabria-Sarmiento, J. C., Ariza-Colpas, P., Pineres-Melo, M., Ayala-Mantilla, C., Urina-Triana, M., Morales-Ortega, R., ... & Echeverri-Ocampo, I. (2018). Software applications to health sector: A systematic review of literature.

[4] Echeverri-Ocampo, I., Urina-Triana, M., Patricia Ariza, P., & Mantilla, M. (2018). El trabajo colaborativo entre ingenieros y personal de la salud para el desarrollo de proyectos en salud digital: una visión al futuro para lograr tener éxito.

[5] Jimeno Gonzalez, K. J., Ariza Colpas, P. P., & Piñeres Melo, M. (2017). Gobierno de TI en pymes colombianas.¿ mito o realidad?.

[6] Ariza-Colpas, P., Morales-Ortega, R., Piñeres-Melo, M., De la Hoz-Franco, E., Echeverri-Ocampo, I., & Salas-Navarro, K. (2019, July). Parkinson Disease Analysis Using Supervised and Unsupervised Techniques. In International Conference on Swarm Intelligence (pp. 191-199). Springer, Cham.

[7] Ariza-Colpas, P., Piñeres-Melo, M., Barceló-Martinez, E., De la Hoz-Franco, E., Benitez-Agudelo, J., Gelves-Ospina, M., ... & Leon-Jacobus, A. (2019, July). Enkephalon-technological platform to support the diagnosis of alzheimer's disease through the analysis of resonance images using data mining techniques. In International Conference on Swarm Intelligence (pp. 211-220). Springer, Cham.

[8] Ariza-Colpas, P. P., Piñeres-Melo, M. A., Nieto-Bernal, W., & Morales-Ortega, R. (2019, July). WSIA: Web Ontological Search Engine Based on Smart Agents Applied to Scientific Articles. In International Conference on Swarm Intelligence (pp. 338-347). Springer, Cham.

[9] Piñeres-Melo, M. A., Ariza-Colpas, P. P., Nieto-Bernal, W., & Morales-Ortega, R. (2019, July). SSwWS: Structural Model of Information Architecture. In International Conference on Swarm Intelligence (pp. 400-410). Springer, Cham.

[10] Ariza-Colpas, P., Oviedo-Carrascal, A. I., & De-la-hoz-Franco, E. (2019, July). Using K-Means Algorithm for Description Analysis of Text in RSS News Format. In International Conference on Data Mining and Big Data (pp. 162-169). Springer, Singapore.

[11] Koundal, D., Gupta, S., & Singh, S. (2016). Automated delineation of thyroid nodules in ultrasound images using spatial neutrosophic clustering and level set. Applied Soft Computing, 40, 86-97

[12] Alias, H. M., Rekha, K. S., & Anitha, R. (2016). Reveal Difference in Synthetic Aperture Radar Images Implementing Fuzzy Clustering Along With Improved MRF Energy Function and Wavelet Denoising Technique. Procedia Technology, 24, 1325-1332

[13] Ariza-Colpas, P., Morales-Ortega, R., Piñeres-Melo, M. A., Melendez-Pertuz, F., Serrano-Torné, G., Hernandez-Sanchez, G., & Martínez-Osorio, H. (2019, September). Teleagro: iot applications for the georeferencing and detection of zeal in cattle. In IFIP International Conference on Computer Information Systems and Industrial Management (pp. 232-239). Springer, Cham.

[14] Banerjee, A., & Maji, P. (2016). Rough-probabilistic clustering and hidden Markov random field model for segmentation of HEp-2 cell and brain MR images. Applied Soft Computing

[15] Hou, J., Liu, W., Xu, E., & Cui, H. (2016). Towards parameter-independent data clustering and image segmentation. Pattern Recognition, 60, 25-36

[16] Zhang, H., & Dai, G. (2016). Improvement of distributed clustering algorithm based on min-cluster. Optik-International Journal for Light and Electron Optics, 127(8), 3878-3881.

[17] Reboul, C. F., Bonnet, F., Elmlund, D., & Elmlund, H. (2016). A Stochastic Hill Climbing Approach for Simultaneous 2D Alignment and Clustering of Cryogenic Electron Microscopy Images. Structure, 24(6), 988-996.

[18] Jin, X., & Kim, J. (2016). Video fragment format classification using optimized discriminative subspace clustering. Signal Processing: Image Communication, 40, 26-35.

[19] Viloria, A., & Lezama, O. B. P. (2019). An intelligent approach for the design and development of a personalized system of knowledge representation. Procedia Comput. Sci, 151, 1225-1230.

[20] Pineda Lezama, O. B., & Reniz, J. (2019). Recommendation of collaborative filtering for a technological surveillance model using Multi-Dimension Tensor Factorization.