SATCLUS: An Effective Clustering Technique for Remotely Sensed Images

Sauravjyoti Sarmah and Dhruba K. Bhattacharyya

Dept. of CS & Engg., Tezpur University, India
{sjs,dkb}@tezu.ernet.in

Abstract. This paper presents a grid density based clustering technique (SATCLUS) to identify clusters present in a multi spectral satellite image. Experimental results are reported to establish that SATCLUS can identify clusters of any shape in any satellite data effectively and dynamically.

1 Introduction

A multi-spectral satellite image is a remotely sensed image of the earth's surface which is a collection of huge amount of information in terms of number of pixel data. In such an image, each pixel represents an area on the earth's surface. Segmentation or clustering a multi-spectral satellite image is a process of discovering finite number of non-overlapping as well as overlapping regions or clusters in an image data space which has been a complex problem for a long time. Region-growing [1] is a local optimization procedure, while pixel clustering is a global analysis of color space. The segmentation technique based on pixel clustering is an important approach. Segmentation using clustering involves the search for points that are similar enough to be grouped together in the color space. Many clustering algorithms have been developed in the past few years to detect clusters from satellite images. K-means [2] and ISODATA [3] are two popular algorithms widely used for detecting clusters in satellite images. Other approaches to segmentation of remotely sensed satellite images include fuzzy thresholding techniques reported in [4], combining SOM and FCM as in [5], using mean-shift algorithms as in [6], Genetic algorithm as a classifier ([7]), combination of region-growing algorithm with mean shift clustering technique [8]. High resolution multi-spectral satellite images cause problems for clustering methods due to clusters of different sizes, shapes and densities as they contain huge amount of data. Due to this reason, most algorithms for clustering satellite data sacrifice the correctness of their results for fast processing time. The processing time may be greatly influenced by the use of grids. In this paper, we propose a grid density based clustering technique, SATCLUS, for multi-spectral satellite images. We have used a combination of density based and grid based clustering due to the fact that density based clustering gives clusters of good quality of different shapes and sizes ([2], [9], [10]); and grid based clustering has the added advantage of fast processing time([2]). SATCLUS can handle the

S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 110–115, 2011.

detection of irregular shaped clusters by pixel level processing of the cluster borders. Most region growing algorithms are executed in a left to right, top-down manner, whereas SATCLUS initiates the cluster expansion process starting with the maximum hue value. Due to the use of grid-based technique, SATCLUS is scalable even for large datasets. SATCLUS does not require the initial of cluster centers; neither does the number of clusters play any role in the clustering process. SATCLUS was tested on a large number of multi-spectral satellite imagery and the cluster results are found very satisfactory.

2 The Proposed Technique

The aim of our clustering algorithm is to discover clusters over multi spectral high resolution satellite images. Following definitions [11] provide the basis of SATCLUS.

Definition 1. *Difference value of a pixel w.r.t. the seed pixel is the distance (dist) between the HSI values of that pixel and the seed. If dist $\leq \theta$, then the difference value of that pixel is considered as 1 otherwise 0, where dist, is any proximity measure (here we have used Mahalanobis distance).*

Definition 2. *Population-object ratio is the ratio of the population count (number of ones in each grid cell) and cell density (number of pixels within a particular grid cell) of a grid cell, i.e., population_count/ cell_density*

Definition 3. *If the difference of the population-object ratio of the current cell and one of its neighbors is greater than or equal to some threshold α then α is the confidence between them. For two cells p and q to be merged into the same cluster the following condition should be satisfied:*

$\alpha \leq\mid P_o(p)-P_o(q) \mid$, *and* $\alpha = \theta \times P_o(seed)$ *where* P_o *represents the population-object ratio of that particular cell and seed is the cell which initiates the expansion of a cluster.*

Definition 4. *Cell reachability: A cell p is reachable from a cell q if p is a neighbor cell of q and cell p satisfies the confidence condition w.r.t. cell q.*

Definition 5. *Rough cluster: A rough cluster is defined to be the set of points belonging to the set of reachable cells. A rough cluster C w.r.t. α is a non-empty subset satisfying the following condition,*
$\forall p, q$: *if* $p \in C$ *and q is reachable from p w.r.t.* α, *then* $q \in C$, *where p and q are cells.*

Definition 6. *Border cell: A cell p is a border cell if it is part of a rough cluster C_i and at least one of its neighbors is part of another rough cluster C_j.*

Definition 7. *Noise: Noise is simply the set of points belonging to the cells not belonging to any of its clusters. Let $C_1, C_2, \cdots C_k$ be the clusters w.r.t. α, then* $Noise = \{no\_p \mid p \in n \times n, \forall i : no\_p \notin C_i\}$, *where no_p is the set of points in cell p and* C_i $(i = 1, \cdots, k)$.

2.1 Procedure of SATCLUS

The algorithm starts with dividing the image space into $n \times n$ non-overlapping square grid cells, where n is a user input, and maps the image pixels to each cell. It then calculates the density of each cell. Next, it converts the RGB values of each pixel to its corresponding HSI values. The algorithm uses the cell information (density) of the grid structure and clusters the data points according to their surrounding cells. The clustering process is divided into two steps. In the first step a rough clustering of the image space is obtained and the second step deals with cluster smoothening for quality cluster identification. The execution of the rough clustering algorithm (Step I) includes the following steps:

Input: The Image dataset

1. Create the grid structure.
2. Compute the density of each cell.
3. Convert the RGB values of each pixel into their HSI values.
4. Identify the maximum hue value as the seed.
5. Calculate each pixels difference value w.r.t. the seed.
6. The population count of each grid cell is computed and the corresponding population-object ratio calculated.
7. Traverse the neighbor cells starting from the grid-cell having the highest population-object ratio value.
8. Merge the cells and assign cluster_id.
9. Repeat steps 5 through 9 till all cells are classified.

The value of θ mostly depends on the values of the h, s, i (i.e., hue, saturation and intensity) and the resolution of the image. However, based on our exhaustive experimentation over various multi-spectral and pan-chromatic satellite images, it has been observed that an effective range of θ for multi-spectral images is $15 \leq \theta \leq 30$ and for pan-chromatic images, the range is found to be $0.5 \leq \theta \leq 4.0$. The rough clusters obtained in Step I are grainy in nature which is a drawback of a grid based algorithm. To obtain clusters with smooth and accurate borders, the border cells are detected and re-clustered using a partitioning based approach as given in Step II. The algorithm for the cluster smoothening is given below:
Input: q border cells; k seeds corresponding to the k rough clusters obtained from Step I

1. Start with an arbitrary border pixel x
2. Find the distance of x to each of the k seeds
3. Assign x to the cluster to which it has minimum distance w.r.t. the seed
4. Repeat steps 1 to 3 till all border pixels have been reassigned

The partitioning of the dataset into $n \times n$ non-overlapping cells results in a complexity of $O(n \times n)$. The complexity of rough clustering in step I is $O(k \times p)$, where p is the number of cells in a cluster so formed and $p \ll n \times n$ in the average case and k is the number of clusters obtained. Similarly in step II, for identification of the q border cells require $O(q)$ times where $q \ll n \times n$ and assignment of r pixels to k clusters requires $O(k \times r)$ times, where r is the total number of pixels in q cells. However, $O(k \times r)$ dominates the overall time complexity since n, p, $q \ll r$.

3 Performance Evaluation

To evaluate SATCLUS in terms of quality of clustering, we used several synthetic and real datasets. SATCLUS was implemented using Java in Windows environment with Pentium-IV processor with 1 GHz speed and 4 GB RAM. We have tested SATCLUS on several synthetic and multi-spectral satellite image datasets and seven of them are reported in this paper. SATCLUS has been found capable of detecting arbitrary shaped clusters as can be seen from Figure 1 (a) and (b). SATCLUS was compared with several relevant algorithms using the image shown in Figure 1 (c) and the results have been found satisfactory, as can be seen from Figure 2. The clusters obtained from Landsat MSS image (4 spectral bands, 79m resolution) of Figure 3 (a) is shown in Figure 3(b). Figure 3(c) shows the plain built up area of Sonari in Sibsagar district of Assam (Cartosat-I image, 4 spectral bands, 2.5m resolution). SATCLUS automatically detects 5 clusters (Figure 3(d)) corresponding to Brahmaputra river, road, agricultural land, water bodies and human settlements. SATCLUS automatically detects four clusters for the IRS LISS II image of Kolkata, West Bengal (Figure 4(a)) as observed in Figure 4(b). From our ground knowledge, we can infer that these four clusters correspond to the classes: Water Bodies (black color), Habitation and City area (deep grey color), Open space (light grey color) and Vegetation (white color). Figure 4(c) and Figure 4(d) shows the IRS Kolkata image partitioned using the GA algorithm of [7] and FCM algorithm respectively. From the figure, it can be noted that the river Hoogly and the city area has not been correctly classified by FCM. In fact, those have been classified as belonging to the same

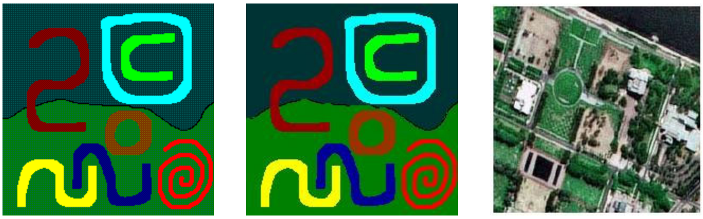

Fig. 1. (a) Example dataset (b) Output of SATCLUS (c) SPOT image of Canberra Australia

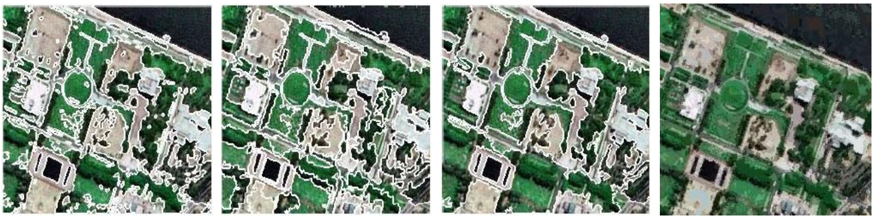

Fig. 2. Outputs of SPOT image of Canberra Australia using (a) Region-growing (b) Mean shift (c) Segmentation method of [8] (d) SATCLUS

Table 1. Homogeneity values for SATCLUS over some satellite images

Dataset	Homogeneity measure
Dataset I	0.9873
Dataset II	0.9915
Dataset III	0.9908
Dataset IV	0.9933
Dataset V	0.9887
Dataset VI	0.9917

Table 2. Comparison in terms of β value and CPU time for different clustering algorithms

Method	beta	CPU time (in hrs)
k-means	5.30	0.11
Astrahan's	7.02	0.71
Mitra's	9.88	0.75
SATCLUS	17.82	0.08

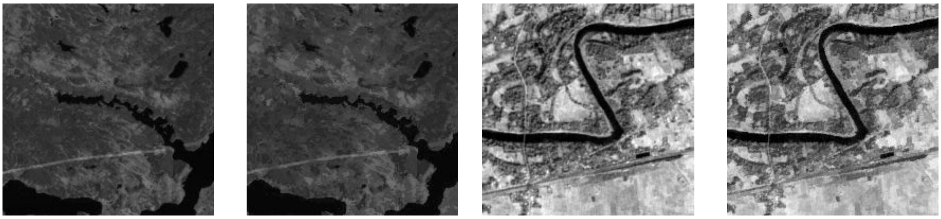

Fig. 3. (a) Landsat-MSS (b) Output of SATCLUS on Landsat-MSS (c) Cartosat-1 image of Sonari, Assam (d) Output of SATCLUS on Cartosat-1 image of Sonari

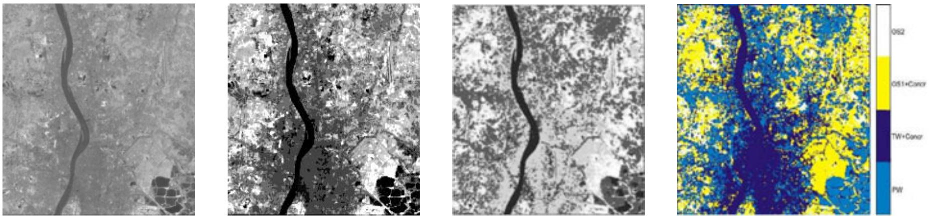

Fig. 4. (a) IRS Kolkata (4 spectral bands, 36.25m resolution) (b) Output of SATCLUS (c) Output of NSGA-II-based clustering technique [7] (d) Output of FCM clustering

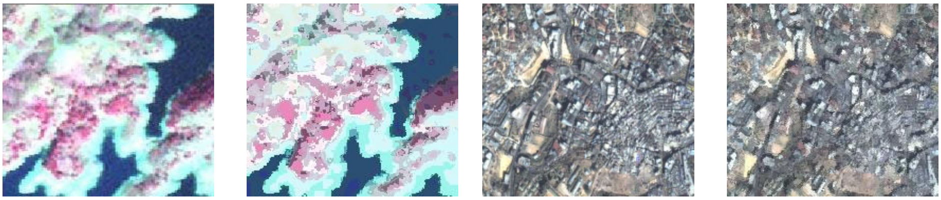

Fig. 5. (a) IRS image of Borapani (b) Output of SATCLUS on image of Borapani (c) Ikonos image of Shillong city (d) Output of SATCLUS on the image of Shillong city

class. Another misclassification is that the whole Salt Lake city has been put into one class. Although some portions have been correctly identified such as canals, the Dumdum airport runway, fisheries, etc. still there is a significant amount of confusion in the FCM clustering result. The characteristic regions of Dataset IRS P6 LISS IV image of Borapani, Meghalaya (Figure 5(a), 4 spectral bands, 5.8m resolution) are the Deep water (Deep Blue color), Wetlands (light blue color), Vegetation (Red and Pink colors) and Open spaces (White color). Executing SATCLUS on this image resulted in the detection of the above four classes as shown in Figure 5(b). The clustered image output obtained by SATCLUS on the IKONOS image of Shillong, Meghalaya (Figure 5(c), 4 spectral bands, resolution of 4m for multispectral and 1m for panchromatic imagery) is shown in Figure 5(d) which relate well with the ground information known to us (concrete structures, roads, open spaces, etc). The performance of SATCLUS in terms of homogeneity measure [2] for some satellite images is shown in Table 1. SATCLUS has also been evaluated quantitatively using an index β [4] and the results for Dataset II can be seen from Table 2. SATCLUS has the highest β value in comparison to the comparable algorithms.

4 Conclusions and Future Work

The proposed SATCLUS was experimentally tested and found capable in detecting the clusters qualitatively and efficiently in light of several real-life satellite images. Work is going on to extend the SATCLUS algorithm for handling hyperion (hyper spectral) data and also to extend the work using a fuzzy-rough set based approach. As a future direction of our work, we a plan to experiment the technique on microwave remote sensing data.

References

1. Baatz, M., Schape, A.: Multiresolution segmentation: an optimization approach for high quality multi-scale image segmentation. Journal of Photogrammetry and Remote Sensing 58(3-4), 12–23 (2000)
2. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Fransisco (2004)
3. Ball, G.H., Hall, D.J.: A clustering technique for summarizing multivariate data. Behavioural Science 12, 153–155 (1967)
4. Pal, S.K., Ghosh, A., Shankar, B.U.: Segmentation with remotely sensed images with fuzzy thresholding and quantitative evaluation. IJRS 21(11), 2269–2300 (2000)
5. Awad, M.M., Nasri, A.: Satellite image segmentation using self- organizing maps and fuzzy c-means. In: IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 398–402 (2009)
6. Comaniciu, D., Meer, P.: Robust analysis of feature spaces: Color image segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 750–755 (1997)
7. Bandyopadhyay, S., Maulik, U., Mukhopadhyay, A.: Multiobjective genetic clustering for pixel classification in remote sensing imagery. TGRS 45(2), 1506–1511 (2007)
8. Bo, S., Ding, L., Jing, Y.: On combining region-growing with non-parametric clustering for color image segmentation. In: CISP, pp. 715–719 (2008)
9. Astrahan, M.M.: Speech analysis by clustering, or the hyper-phoneme method. Stanford A. I. Project Memo (1970)
10. Mitra, P., Murthy, C.A., Pal, S.K.: Density-based multiscale data condensation. IEEE TPAMI 24(6) (June 2002)
11. Sarmah, S., Das, R., Bhattacharyya, D.K.: A distributed algorithm for intrinsic cluster detection over large spatial data. IJCS 3(4), 246–256 (2008)