

Image mining and Automatic Feature extraction from Remotely Sensed Image (RSI) using Cubical Distance Methods

¹S.Sasikala, ²Dr.N.Radhakrishnan

¹Research Scholar, Mother Theresa Women's University, Kodaikanal.

²Managing Director, Geosensing Information Pvt., Ltd., Chennai

¹sasikalarams@gmail.com, ²radhakrishnan.nr@gmail.com

Abstract

Information processing and decision support system using image mining techniques is in advance drive with huge availability of remote sensing image (RSI). RSI describes inherent properties of objects by recording their natural reflectance in the electro-magnetic spectral (*ems*) region. Information on such objects could be gathered by their color properties or their spectral values in various *ems* range in the form of pixels. Present paper explains a method of such information extraction using cubical distance method and subsequent results. This method is one among the simpler in its approach and considers grouping of pixels on the basis of equal distance from a specified point in the image or selected pixel having definite attribute values (DN) in different spectral layers of the RSI. The color distance and the occurrence pixel distance play a vital role in determining similar objects as clusters aid in extracting *features* in the RSI domain.

Key words: Image mining, Feature extraction, Cubical distance, Frequent Item set, Remote sensing Image

1,0 Introduction

In the present paper, digital numbers (DN) values of the selected remote sensing image (RSI) is studied to understand the image mining technique to extract information. This was carried out by grouping of pixels to determine features and predicates the possible integration of applications using clustering techniques. In this context, an algorithm using cubical distance calculation method to cluster pixels is used to extract features. The classification process is implemented with the DN values obtained through pre-processing the RSI. [1]

In cubical distance method, the pixels are grouped or clustered based on the distance of their respective DN values in each layer (B, G, R). The estimated distance helps to group pixels having similar distance value as group or clusters [2]. The data input in the form of macro array, position of each pixel is represented as x, y followed by the respective DN values in the three layers (B,G,R). In this method, distance of pixels to respective cluster interval is segregated as groups, which may be labeled as specific feature may show similar distance criterion [3]. To appreciate the inherent influence of such distance measures, three methods are identified under cubical distance method such as "*from the origin*", "*from the first position of the selected ROI*" and "*frequently occurred pixel*". Algorithms for these three methods are designed and implemented to cluster pixels and in turn, extract features.

2.0 Methodology

Methodology adopted in the study is as follows:

- i. Select appropriate Remote sensing Image (RSI) to identify landuse features
- ii. Preprocess the RSI for geometrical error so that the image represents the ground condition.
- iii. Convert the multi layer image into two dimensional Digital number (DN) values as macro-array data.
- iv. Determine the distance of the pixel using cubical distance determination methods
 - a. From the origin (0,0,0)
 - b. From the first point of the Region of Interest
 - c. From the frequent occurrence point in the ROI
- v. Implementation and clustering of pixels according to their distance for the significance of results using such methods.

3.0 Preprocessing of RSI

Selected RSI is preprocessed for geometrical correction so that the image represents the ground condition in real time [4]. Ground control points (GCPs) are used to carry out geometrical correction with the help of GIS (Geographic Information System) as available in the image processing software, ERDAS Imagine. The geometrically corrected image was later converted into a procesable macro array data format.

3.1 Macro Array Data format

The pixels in the RSI image have specific digital numbers (DN) as their attributes, which vary with the spectral range (Blue, Green, Red and Infrared). Such spectral variations are exploited in image mining so that specific features are clustered individually. In the present study, any three layers mostly RGB are selected from the RSI and converted into macro-array format (X,Y, layer1 value , layer2 value, layer3 value), where X and Y stands for the position of the pixel and layer1, layer2 and layer 3 are the attribute DN values of the pixel in different spectral layers. A sample of extracted DN values in specific data format is shown below.

Layer 1

125 134 143 146
 144 145 148 147
 134 126 126 125
 127 139 145 144
 144 137 146 140
 127 128 131 122

Layer 2

134 141 147 128
 141 139 146 138
 141 135 126 111
 140 139 135 126
 138 139 148 145
 143 139 116 106

Layer 3

137 132 139 139
 141 141 153 151
 145 129 120 130
 123 125 139 137
 131 148 153 143
 139 133 134 132

Constructed macro array

[1,1, 125 ,134,137;
 1,2,134,141,132;
 1,3, 143,147,139;
 1,4, 146,128,139;....]

4.0 Procedure for Cubical Distance Method

Procedure for classification using cubical method

- i. Select the RSI for preprocessing
- ii. Preprocess the captured RSI for analysis
- iii. Select the ROI
- iv. Convert the ROI into cubical digital array using macro array process
- v. Calculate the distance for the chosen three methods:
 - a. Assign the P as (0,0,0) origin value and calculate the distance of each existing pixel from the origin using the formula $OP = \sqrt{((0-a)^2 + (0-b)^2 + (0-c)^2)}$.
 $OP = \sqrt{a^2 + b^2 + c^2}$
 - b. Assign P1 as first pixel of the ROI (x_1, y_1, z_1) and distance for another point (P2) as (x_2, y_2, z_2) value and calculate the distance of each existing pixel from the first pixel of the ROI using the formula $P1Pi = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2]}$
 - c. Assign the highest Frequency as a first point PF (x_1, y_1, z_1) and distance for another point as (x_2, y_2, z_2) . Calculate the distance of the each existing pixel from the highest Frequency pixel (PF) using the formula
 $PFPi = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2]}$.

Calculate the minimum and maximum distance. Assign the required number of clusters as input from the user

- vi. Calculate the range of distance as $maximum - minimum / number\ of\ clusters$
- vii. Verify the distance and compare with the range belongings
- viii. Assign the clustering index value, representing features, for each pixel as per the distance and range belongings
- ix. Group all the similar classification index values as clusters and resultant image is generated.

5.0 Implementation and Result

The selected ROI of 400 x 400 array is manipulated based on the DN's of each layer together. The coding for the ROI to determine the range values of pixels based on equal interval was carried out and executed using Matlab 7.1.

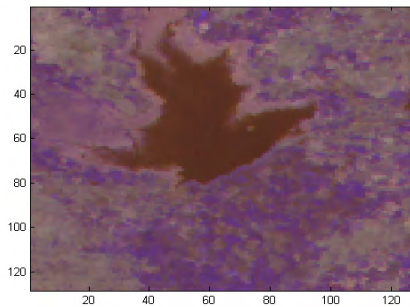


Figure 1. ROI image taken for analysis





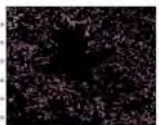
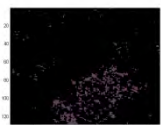



The above-discussed algorithms are applied on the selected ROI and resultant clustering and *feature* extraction are discussed in the following and illustrated in Table 1.

Table 1. Clustering and DN Range values by Cubical Distance Method

A. From the Origin (0,0,0)								
Cluster	No. of Pixel	%of Pixel	Blue		Green		Red	
			Start	End	Start	End	Start	End
1	4290	26.18	84	169	45	126	34	133
2	-	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-
4	129	0.79	84	92	45	47	34	33
5	641	3.91	84	114	49	60	41	49
6	4023	24.55	92	134	53	81	102	69
7	6196	37.82	94	144	55	92	139	96
8	1082	6.60	105	160	63	95	178	129
B. From the First Point of ROI								
Cluster	No. of Pixel	%of Pixel	Blue		Green		Red	
			Start	End	Start	End	Start	End
1	4904	29.93	84	169	45	126	34	133
2	3853	23.52	109	143	95	104	106	108
3	3001	18.32	101	153	82	105	107	115
4	2401	14.65	95	160	74	95	111	129
5	1388	8.47	92	163	57	121	96	133
6	213	1.30	92	161	53	134	102	133
7	218	1.33	91	114	72	60	40	49
8	301	1.84	87	105	49	63	37	178

C. From the first point of Frequent occurrence pixel								
Cluster	No. of Pixel	%of Pixel	Blue		Green		Red	
			Start	End	Start	End	Start	End
1	4290	26.18	84	169	45	126	34	133
2	-	-	-	-	-	-	-	-
3	896	5.47	92	107	53	67	102	85
4	4560	27.83	84	122	45	70	34	79
5	3725	22.74	89	136	46	81	33	90
6	2239	13.67	103	145	63	91	154	113
7	584	3.56	105	159	63	90	178	119
8	82	0.50	143	161	125	118	120	129

All the above outputs in the form of statistical output for all the three methods are also generated as graphical map output, which is illustrated below in Figure 2. The table1 depicts the statistical discussion above in a more lucid manner. Waterbody, obviously, present in the center part of image is clearly brought out in the first cluster image of all the three methods – from the origin, from the first pixel and from the frequent occurrence pixel of the selected ROI, irrespective of various methods of distance estimation for clustering. In the first cluster of the image obtained from the first point, agricultural pattern is seem to be more clearly brought out compared to the other two methods. A dense patch of pixels seen in northeastern and southwestern parts of the image may indicate presence of crop land, which are seen sporadically in the other two methods. Since there is no DN range value of pixel in second and third clusters of *origin* method and second cluster in *frequent occurrence* method, output image are not found. This implies the absence of pixel range values in that particular cluster while distance is calculated. Most of the vegetation such as crop land, scrub and plantation is visible in the second and third clusters of the *first point* method and fourth and fifth clusters of the *FIS* method whereas vegetation mostly crop land is seen in the sixth and seventh clusters of the *origin* method. Except the first cluster, waterbody is conspicuously absent in all the other cluster in all the three methods, though its shape is retained due to the presence of soil moisture and scrub around the waterbody. Presence of scrub and barren soil is well brought out in the seventh and eighth clusters of the *first point* method. Similarly, existence of barren land around the waterbody is seen in the fifth cluster of the first method and fourth cluster in the third method.

Clusters	From the Origin	From the First Point	From FIS
1			
2	No pixels / image		No pixels / image
3	No pixels / image		
4			

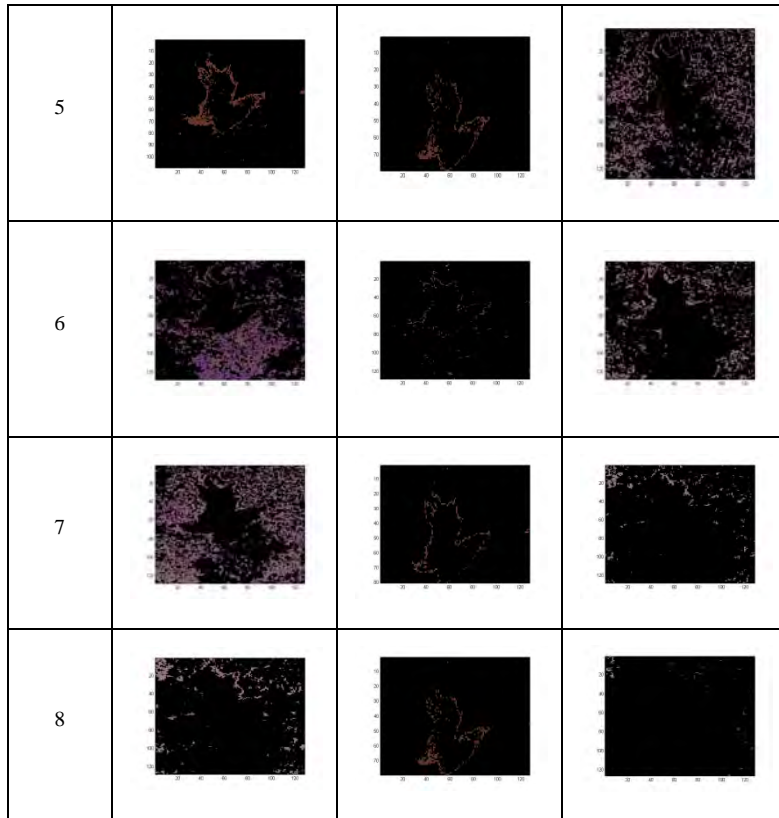


Figure 2 Clustered Image output of ROI by Cubical Distance from Origin

The individual clusters of each method have one or more features that are categorized with the help of difference in colors of pixels in the output image. To refine further and to extract features a second level iteration is carried out for individual clusters and the respective values are tabulated below (Table 2).

Table 2. Clustering and DN Range values by Second level iteration by Cubical Distance

A. Cubical Distance Method Origin(0,0,0)								
Cluster	Feature	Pixel %	Blue		Green		Red	
			Start	End	Start	End	Start	End
1	5	8.73	86	93	47	51	33	36
	1	8.73	98	107	55	65	104	115
	4	8.73	121	131	96	106	100	111
4	5	0.79	87	91	45	48	33	36
5	3	3.91	93	107	53	60	42	51
6	1	12.28	98	108	56	63	93	126
	3	12.28	109	126	71	80	73	86
7	6	18.91	117	130	88	103	93	108
	1	18.91	100	108	58	68	121	154
8	2	6.60	135	150	106	125	108	123
B. Cubical Distance Method from ROI								
Cluster	Feature	Pixel %	Blue		Green		Red	
			Start	End	Start	End	Start	End
1	5	9.98	85	93	46	53	34	36
	1	9.98	98	109	54	66	101	122

	6	9.98	123	131	96	106	100	104
2	3	11.76	124	139	89	95	107	110
	2	11.76	126	137	107	113	104	115
3	7	9.16	113	124	73	83	101	114
	2	9.16	130	142	115	118	111	116
4	1	7.33	106	114	61	77	93	119
	2	7.33	113	158	68	84	65	81
5	1	4.24	97	103	53	59	96	116
	7	4.24	107	110	65	73	62	73
6	1	0.65	94	107	52	59	110	146
	7	0.65	99	115	59	70	54	66
7	7	1.33	97	104	57	61	40	53
8	5	1.84	90	98	49	55	37	42
A. Cubical Distance Method FIS								
Cluster	Feature	Pixel %	Blue		Green		Red	
			Start	End	Start	End	Start	End
1	5	8.73	85	91	46	51	34	37
	1	8.73	98	117	55	80	102	115
	2	8.73	121	138	101	111	100	109
3	1	2.74	94	103	52	62	94	110
	7	2.74	101	106	55	71	65	86
4	5	9.28	87	110	47	68	37	57
	1	9.28	94	102	53	62	108	125
	3	9.28	110	117	69	82	70	89
5	4	11.37	112	122	70	83	104	114
	6	11.37	118	125	87	96	81	102
6	3	6.84	135	139	91	96	102	116
	6	6.84	122	133	100	109	104	114
7	2	3.56	138	145	99	115	113	120
8	2	0.50	150	159	118	128	115	127

Thus, cubical distance estimation among pixels for feature extraction using various methods - from the origin, from the first pixel and frequent item pixel – enumerate the efficiency of segregating pixels as groups, which may inferred with presence of certain predominance of certain feature such as crop, scrub, waterbody, fallow, plantation and so on.

From the above analysis, it is observed that crop feature has a range value of 123, 112, 78 in BGR respectively. Similarly, barren land shows 125,95,100, fallow land as 120,80,83, scrub shows 56, 68, 120, water body shows 85,45,35, soil moisture as 118,85,90 and plantation shows 106,70,85 in that order as shown in the above table that are identified and selected with the help of domain landuse expertise.

6.0 Conclusion

Feature extraction using “*distance*” as the criterion is significantly efficient in clustering the pixels of similar character and thus enables a better clustering and grouping of features. Cubical distance estimation among pixels for feature extraction using various methods - from the origin, from the first pixel and frequent item pixel – enumerate the efficiency of segregating pixels as groups, which may inferred with presence of certain predominance of features such as crop, scrub, fallow, plantation and so on. Such feature extraction since based on distance alone as parameter without any other conditionalities may show certain “bias” or “mixing up of pixels” either overlapping or completely strange pixels. Even though such limitations exist they do not warrant any less in significance of feature extraction, since the method attempts to extract features by determining the DN range values.

References

- [1] Sasikala S and Radhakrishnan N, 2010, Remotely Sensed Image (RSI) analysis using Equal Interval Classification, IJ-ETA-ETS, ISSN: 0974-3588, Page no. 726-732
- [2] Knorr, E. and Ng, R. 1998. Algorithms for mining distance-based outliers in large datasets. In Proceedings of the 24th Conference on VLDB, 392-403, New York, NY.
- [3] Ramaswamy, S., Rastogi, R., and Shim, K. 2000. Efficient algorithms for mining outliers from large data sets, *Sigmoid Record*, 29, 2, 427-438.
- [4] Hyypä, H. (et al.), 2004, Algorithms and Methods of Airborne Laser-Scanning for Forest Measurements. *International Archives of Photogrammetry and Remote Sensing*, Vol XXXVI, 8/W2, Freiburg, Germany.
- [5] Sheikholeslami, G., Chatterjee, S., and Zhang, A. 1998. WaveCluster: A multiresolution clustering approach for very large spatial databases. In Proceedings of the 24th Conference on VLDB, 428-439, New York, NY.
- [6] Kettig, R. L., and Landgrebe, D. A. 1999, Classification of multispectral image data by extraction and classification of homogeneous objects. *IEEE Transactions on Geoscience Electronics*, GE-14(1):19–26,
- [7] Miller Han, J., Kamber, M., and Tung, A. K. H. 2001. *Spatial clustering methods in data mining: A survey*. Taylor and Francis
- [8] Miller Han, J., 2001 (Eds.) *Geographic Data Mining and Knowledge Discovery*, Taylor and Francis.
- [9] Massart, D. and Kaufman, L. 1983. *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*. John Wiley & Sons, New York, NY.
- [10] Quinlan, J. R.: Learning with continuous classes. In Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, pages 343–348. World Scientific, Singapore, (1992).
- [11] Raymond, M. (et al.): *Measures. Laser remote sensing: fundamentals and applications*. Malabar, Fla., Krieger Pub. Co. 510 p. G70.6.M4 (1992)
- [12] Clementking, A. Angel Latha Mary, S. Comparing and identifying common factors in frequent item set algorithms in association rule 18-20 Dec. 2008 ISBN: 978-1-4244-3594-4- 10.1109/ICCCNET.2008.4787769. 2009-02-24.
- [13] JANSSEN, L. (1993): Methodology for updating terrain object data from remote sensing data. The application of Landsat TM data with respect to agricultural fields. Doctoral Thesis, Wageningen Agricultural University, Wageningen