Similarity measure of spectral vectors based on set theory and its application in hyperspectral RS image retrieval

Peijun Du (杜培军)[1,2], Tao Fang (方 涛)[1], Hong Tang (唐 宏)[1], and Pengfei Shi (施鹏飞)[1]

[1] *Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai 200030*

[2] *Department of Spatial Informatics, China University of Mining and Technology, Xuzhou 221008*

Received August 11, 2003

In this paper, two new similarity measure methods based on set theory were proposed. Firstly, similarity measure of two sets based on set theory and set operation was discussed. This principle was used to spectral vectors, and two approaches were proposed. The first method was to create a spectral polygon corresponding to spectral curve, and similarity of two spectral vectors can be replaced by that of two polygons. Area of spectral polygon was used as quantification function and some effective indexes for similarity and dissimilarity were computed. The second method was to transform the original spectral vector to encoding vector according to absorption or reflectance feature bands, and similarity measure was conducted to encoding vectors. It proved that the spectral polygon-based approach was effective and can be used to hyperspectral RS image retrieval.

OCIS codes: 280.0280, 300.6320, 100.5010.

In hyperspectral remote sensing (RS) image, each pixel can be represented by a spectral vector that is composed of the attributes (albedo) of the ground entity corresponding to the pixel on every band in turn. This spectral vector is the basis of further classification and information extraction[1]. After vast hyperspectral RS images were acquired, they would be managed in image library. In order to improve management efficiency, those images were segmented and then some features will be extracted, therefore content-based indexing can be done and the link between image library and feature library can be established. The features that can be used for retrieval include two types: one is original spectral vector and the other is some specific feature extracted from original data, such as normalized difference of vegetation index (NDVI), fractal and so on. Because the latter should be supported by pre-known knowledge and some algorithms available, it is difficult to extract those features. The spectral vector is directly derived from original data, it is easy to be captured and processed, and it is convenient and effective to hyperspectral RS image retrieval. When content-based hyperspectral RS image retrieval is conducted, the retrieval mask is the spectra of a given region or class and the mean vector of this mask can be viewed as a reference spectrum, and the retrieval task is to compare all feature vectors in feature library with the reference spectrum and measure their similarity according to a certain criterion, finally those feature vectors with high similarity will be selected and the image blocks connected with them will be selected by indexing mechanism[2].

After spectral vector is determined as retrieval feature, the other key issue is similarity measure between reference spectrum of retrieval mask and those in feature library. Nowadays, some similarity measure methods including Euclidian distance, spectral angle, correlation coefficient, spectral information divergence (SID), encoding and matching and others are used in common, and each has its advantages and disadvantages[1,3]. In order to improve the precision and efficiency of similarity measure, it is necessary to research and put forward some new methods. In this paper, two new methods based on set theory and set operation will be proposed.

It is known that similarity measure is the operation to compare the similarity of two sets and express it with a quantitative index in essence. Similarity of two sets can be analyzed by some set relationships and set operations. Suppose that A and B are two sets that will be compared, some new sets such as $A \cup B$, $A \cap B$, $\overline{A} \cap B$, $A \cap \overline{B}$, $\overline{A} \cap \overline{B}$ can be generated by specific operation. For every set X, a function $M(X)$ used to describe properties of X quantitatively can be defined, and so, a set can be expressed by some quantitative indexes, and relationship between different sets can be analyzed by those indexes. Here, function $M(X)$ is very important to the results and efficiency.

For two definite sets A and B, $\overline{A} \cap \overline{B}$ is an unlimited set generally, so $M(\overline{A} \cap \overline{B})$ is useless in practice and it is not adopted in further analysis. In addition, it can be known according to set theory that $A \cup B = (A \cap B) \cup (A \cap \overline{B}) \cup (\overline{A} \cap B)$ and $A \cap B$, $\overline{A} \cap B$ and $A \cap \overline{B}$ are independent each other, so it can be drawn that $M(A \cup B) = M(A \cap B) + M(\overline{A} \cap B) + M(A \cap \overline{B})$.

Therefore, seven useful indexes can be used in further analysis[4]. Those are

$$M_1 = M(A \cap B), \tag{1}$$

$$M_2 = M(A \cap \overline{B}), \tag{2}$$

$$M_3 = M(\overline{A} \cap B), \tag{3}$$

$$M_4 = \mathrm{Min} = \mathrm{Min}[M(A), M(B)], \tag{4}$$

$$M_5 = \mathrm{Max} = \mathrm{Max}[M(A), M(B)], \tag{5}$$

$$M_6 = \mathrm{Sum} = M(A) + M(B), \tag{6}$$

$$M_7 = M(A \cup B) = M_1 + M_2 + M_3. \tag{7}$$

According to Stephan Winter's studies on region-based similarity and our analysis and experiments, the follow-

ing indexes are effective to measure similarity between two sets[4].

$\mu_1 = M_1/M_7$. Its domain is [0,1]. When A is same to B, the result equals 1.

$\mu_2 = M_1/M_4$. Its domain is [0,1]. When one polygon is within the other, the result is 1.

$\mu_3 = M_1/M_5$. Its domain is [0,1]. When A is same to B, the result equals 1.

$\mu_4 = M_1/M_6$. Its domain is [0,0.5]. When A is same to B, the result equals 0.5.

And the following are useful to measure dissimilarity

$$d_1 = (M_2 + M_3)/M_7 = 1 - \mu_1,$$

$$d_2 = (M_2 + M_3)/M_5,$$

$$d_3 = (M_2 + M_3)/M_6.$$

It can be seen that the key factor to similarity is M_1, or quantitative value of intersected set, and the key indices to dissimilarity are M_2 and M_3, so those indices can be combined to define a new similarity measure index as follows

$$s_1 = \frac{\mu_1}{d_1} = \frac{M_1}{M_2 + M_3}. \tag{8}$$

The bigger the index is, the more similar the two sets are. For two identical sets, this index is infinity because the denominator is zero. In the following discussions this principle will be used to similarity measure of spectral vectors, and two approaches are proposed. One is similarity measure to original vectors, and the other is to encoding vectors based on absorption or reflectance features.

When the pixel is expressed by original spectral vector, this vector can be demonstrated by a spectral curve in the coordinate system with wavelength as horizontal axis and albedo as vertical axis. Figure 1 is the demonstration of a spectral curve.

From Fig. 1, we can know that a polygon can be formed by spectral curve, horizontal axis and two lines parallel vertical axis and crossing the first and last wavelength points. Here, this polygon is named as spectral polygon. Different ground objects or pixels in image have distinct spectral curves, and their spectral polygons are distinct, so some quantitative properties of spectral polygon can be used to describe the spectral properties of ground objects. Therefore, similarity measure of two spectral vectors is transformed into similarity measure of two spectral polygons, and area is selected to quantify properties of spectral polygon. That means area computation is selected as quantification function $M()$, and $M_1, M_2, M_3, M_4, M_5, M_6$ and M_7 can be calculated. According to former discussions, M_1, M_2 and M_3 are enough here.

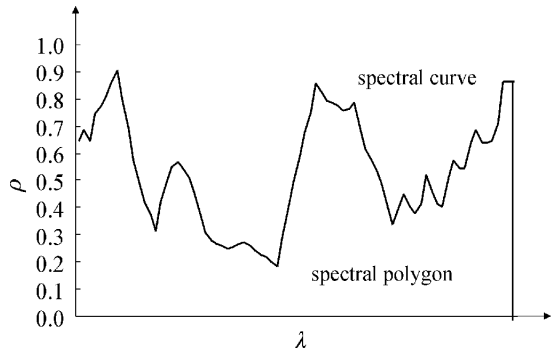

Fig. 1. Spectral curve and spectral polygon.

Although spectral polygon is irregular in shape, it is composed by $N-1$ trapezoids (N is the band number of hyperspectral RS image) and each trapezoid can be viewed as a sub spectral polygon. Each trapezoid is composed of the wavelength of two adjacent bands and their albedo, and its area is easy to compute by $S_i = (\rho_i + \rho_{i+1}) * (\lambda_{i+1} - \lambda_i)/2 (i = 0, 1, \cdots, N-2)$, and area of the spectral polygon can be calculated by $S = \sum S_i$. So the key is to compute area of every trapezoid. For two adjacent bands, the spatial relationships between two different spectral curves include four cases demonstrated in Fig. 2. In the following, $S(A)$ is the area of the trapezoid composed of the ith and $(i+1)$th band in spectral vector A, and $S(B)$ is the area of corresponding polygon in spectral vector B.

Case (I) is characterized by $A_i < B_i$ and $A_{i+1} < B_{i+1}$, so

$$M_1 = S(A);$$

$$M_2 = O;$$

$$M_3 = S(B) - S(A).$$

Case (II) is characterized by $A_i < B_i$ and $A_{i+1} > B_{i+1}$. In order to compute M_1, M_2 and M_3, the intersected point of two lines should be computed at first. Suppose that the intersected point is (λ_0, ρ_0), so

$$M_1 = (\rho_0 + A_i) * (\lambda_0 - \lambda_i)/2 + (\rho_0 + B_{i+1}) * (\lambda_{i+1} - \lambda_0)/2;$$

$$M_2 = (A_{i+1} - B_{i+1}) * (\lambda_{i+1} - \lambda_0)/2;$$

$$M_3 = (B_i - A_i) * (\lambda_0 - \lambda_i)/2.$$

Case (III) is characterized by $A_i > B_i$ and $A_{i+1} > B_{i+1}$, so

$$M_1 = S(B);$$

$$M_2 = S(A) - S(B);$$

$$M_3 = O.$$

Case (IV) is characterized by $A_i > B_i$ and $A_{i+1} < B_{i+1}$. Firstly the intersected point (λ_0, ρ_0) is determined, and then it can be computed, so

$$M_1 = (\rho_0 + B_i) * (\lambda_0 - \lambda_i)/2 + (\rho_0 + A_{i+1}) * (\lambda_{i+1} - \lambda_0)/2;$$

$$M_2 = (A_i - B_i) * (\lambda_0 - \lambda_i)/2;$$

$$M_3 = (B_{i+1} - A_{i+1}) * (\lambda_{i+1} - \lambda_0)/2.$$

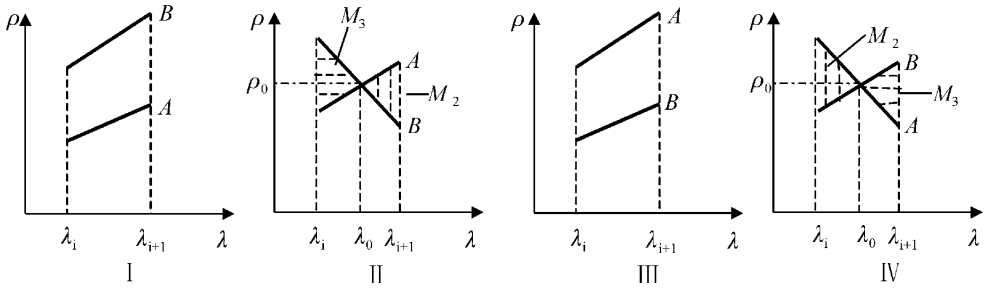

Fig. 2. Four cases of the spatial relationship between two adjacent spectral curves.

After all areas are computed, those indexes used to measure similarity and dissimilarity can be calculated further, so similarity measure of two spectral polygons can be realized and described quantitatively. This is spectral polygon-based approach.

It is well known that noises influence the accuracy to a great extent, and an effective approach should reduce or overcome the impacts of noises. If two spectral vectors are expressed by spectral polygon and their similarity is measured based on set operation and set theory, it can be seen that some minor noises could not cause obvious impacts to intersected and mutual area so this approach is not sensitive to noises including displacement, deformation and local change. But those noises will cause errors in some traditional approaches such as Euclidean distance, correlation coefficient and spectral angle.

In addition to above approach, spectral vector can also be expressed by encoding vector based on feature bands, so similarity measure can be done by those vectors too. For a spectral vector $\overrightarrow{A} = (A_1, A_2, \cdots, A_N)$, it can be encoded according to reflectance feature bands. If albedo on a band is local maximum, it will be encoded as "1", else it will be "0", therefore an encoding vector is got.

After encoding ends, quantitative indexes to encoding vector can be computed. Suppose that encoding vectors of two spectral vectors A and B are x and y, the basic indexes can be computed by the following equations

$$f(x) = \sum x_i, \quad f(y) = \sum y_i;$$
$$M_1 = f(x \cap y) = \sum x_i y_i;$$
$$M_2 = f(x \cap \overline{y}) = \sum x_i(1 - y_i);$$
$$M_3 = f(\overline{x} \cap y) = \sum (1 - x_i) y_i;$$
$$M_7 = f(x \cup y) = M_1 + M_2 + M_3.$$

Based on those values, those parameters can be computed so as to the similarity between two spectral curves can be determined. This is feature-based similarity measure approach.

In order to compare the efficiency and performance of above approaches, we used them to measure the similarity of similar and different spectral vectors. Figure 3 shows six spectral curves used, where A_1 and A_2, B_1 and B_2, C_1 and C_2 belong to a same class respectively. The results of polygon-based approach are listed in Tables 1 – 3, and result of feature-based method is listed in Table 4.

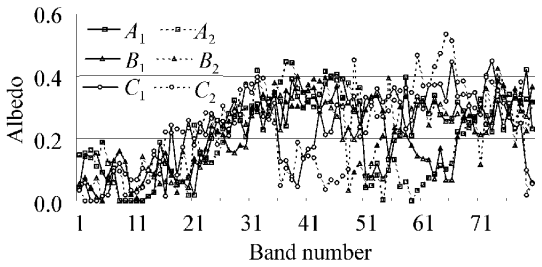

Fig. 3. Experimental data of six spectral vectors.

Table 1. μ_1 Computed by Polygon-based Approach

	μ_1					
	A_1	A_2	B_1	B_2	C_1	C_2
A_1	1	0.77	0.74	0.76	0.65	0.61
A_2	0.77	1	0.73	0.70	0.57	0.52
B_1	0.74	0.73	1	0.76	0.65	0.59
B_2	0.76	0.70	0.76	1	0.68	0.63
C_1	0.65	0.57	0.65	0.68	1	0.81
C_2	0.61	0.52	0.59	0.63	0.81	1

Table 2. μ_2 Computed by Polygon-based Approach

	μ_2					
	A_1	A_2	B_1	B_2	C_1	C_2
A_1	1	0.90	0.87	0.88	0.81	0.77
A_2	0.90	1	0.85	0.83	0.78	0.72
B_1	0.87	0.85	1	0.87	0.85	0.77
B_2	0.88	0.83	0.87	1	0.85	0.80
C_1	0.81	0.78	0.83	0.85	1	0.91
C_2	0.77	0.72	0.77	0.80	0.91	1

Table 3. s Computed by Polygon-based Approach

	$s = \mu_1/d_1$					
	A_1	A_2	B_1	B_2	C_1	C_2
A_1	∞	3.32	2.86	3.12	1.87	1.57
A_2	3.32	∞	2.66	2.31	1.33	1.10
B_1	2.86	2.66	∞	3.13	1.88	1.46
B_2	3.12	2.31	3.13	∞	2.16	1.68
C_1	1.87	1.33	1.88	2.16	∞	4.27
C_2	1.57	1.10	1.46	1.68	4.27	∞

Table 4. μ_1 Computed by Feature-based Approach

	μ_1					
	A_1	A_2	B_1	B_2	C_1	C_2
A_1	1	0.39	0.24	0.23	0.30	0.09
A_2	0.39	1	0.11	0.17	0.21	0.11
B_1	0.24	0.11	1	0.25	0.10	0.16
B_2	0.23	0.17	0.25	1	0.24	0.32
C_1	0.30	0.21	0.10	0.24	1	0.27
C_2	0.09	0.11	0.16	0.32	0.27	1

Table 5. Retrieval Results of Two Approaches

	Polygon-Based Approach				
	I	II	III	IV	V
Index Value	∞	4.27	3.18	2.63	1.94
No.	c	d	q	x	y
	Feature-Based Approach				
	I	II	III	IV	V
Index Value	∞	0.46	0.37	0.34	0.28
No.	c	f	d	o	r

From above tables, it can be seen that the spectral polygon-based approach is effective to similarity measure between spectral vectors, and the feature-based approach is not effective since it only considers feature band locations and neglects other factors, but the location of feature bands maybe displace or change because of noise and other reasons.

In order to compare the performance of the two methods further, a retrieval example is given. Suppose a series of spectral vectors representing different ground objects are stored in a file, one spectral vector (No c) of those is selected as retrieval mask, and it is already known that d is similar to c and other vectors are different from it. The five vectors with maximum matching ratio are got and listed in Table 5.

It can be seen that the spectral polygon-based approach can achieve satisfied effects with the retrieved vector itself as the first matching element and the similar vector as the second. In feature-based method, the first matching element is the retrieved vector, but the second is not the desired one, and it ranks the third, so this method is not effective to retrieval. We also experiment it by other examples and the conclusions are similar.

In this paper the spectral vector similarity measure approaches based on set theory and set operation are proposed. The similarity measure of two spectral vectors (or curves) in hyperspectral RS image is transformed into similarity analysis and operation of two sets. According to the properties of hyperspectral RS image, two measure approaches are put forward and tested. It is shown that the similarity measure method based on spectral polygon is effective to spectral vectors and can be used to content-based image retrieval, but the approach based on spectral feature bands is sensitive to noise, and its performance in similarity measure and retrieval is not very good.

In the future, we would like to give more studies on this issue, especially on its comparison and integration with other metrics and their applications in content-based hyperspectral RS image retrieval and its applications to multispectral RS image retrieval.

This research was supported by the National 863 Program of China (No. 2001AA135091), National Natural Science Foundation of China (No. 60275021) and China Postdoctoral Science Foundation (No. 2002032152). P. Du's e-mail address is dupj@vip.163.com.

References

1. R. L. Pu and P. Gong, *Hyperspectral Remote Sensing and Its Applications* (in Chinese) (Higher Education Press, Beijing, 2000).
2. Y. J. Zhang, *Content Based Visual Information Retrieval* (in Chinese) (Science in China Press, Beijing, 2003).
3. J. X. Sun, *Modern Pattern Recognition* (in Chinese) (National University of Defense Technology Press, Changsha, 2001).
4. S. Winter, ISPRS J. Photogrammetry & Remote Sensing **55,** 189 (2000).