METHOD OF IDENTIFYING ENDMEMBER SPECTRAL VALUES FROM HYPERSPECTRAL IMAGE DATA

FIELD OF THE INVENTION

[0001] The present invention relates to a method of identifying endmember spectral values from multispectral or hyperspectral image data, which is particularly useful in identifying different materials from multispectral or hyperspectral images.

BACKGROUND TO THE INVENTION

[0002] It is known to collect remote sensing data to provide images of scenes to aid in broad scale discrimination of various features of land scanned including identifying mineral deposits and vegetation. Two examples of hyperspectral scanners are NASA's 224 band AVIRIS, which has bands spaced about every 10 nanometers in a range from 400 to 2500 nanometers, and the 128 band Australian commercial scanner, HyMap, which covers a similar wavelength range with about 16 nanometer resolution.

[0003] A goal is therefore to identify the components of each pixel in the hyperspectral image. This can be done by comparison with a library of spectra of "pure" materials. "Pure" materials in a hyperspectral image are often termed endmembers.

[0004] Depending on the resolution of the image obtained from the spectral scanner, an individual pixel may represent an area ranging in size from 5 to 10 meters across in images from an aircraft scan or 10 to 30 meters across from a satellite scan. Each pixel therefore will relate to a portion of a scene which will usually include a mixture of material components. It is not uncommon to find that not all of the pure spectral representations of endmembers are present in a scene.

[0005] Images are also subject to distortion due to noise from various sources including instruments, atmospheric interference, viewing geometry and topography of the area scanned. Corrections for these distortions are still not sufficiently accurate to allow for reliable comparisons to reference libraries. Also, many remotely sensed scenes contain materials not in libraries. Therefore, there are problems with matching spectra with ground-based libraries. There is consequently interest in identifying the component materials represented in a scanned scene, without reference to a library.

[0006] Similar problems occur in other fields where it is desired to determine

endmembers from multispectral, hyperspectral or other data where a signal is detected on a number of channels or bands. For example a similar problem occurs in the analysis of proteomics and genomics array data where the signal represents cell or organism response across a range of proteins, cDNAs or oligonucleotides. In this context, each protein, cDNA or oligonucleotide is regarded as being equivalent to a wavelength or band in the hyperspectral or multispectral context. Similar problems also occur in fluorescence imaging such as fluorescence microscopy.

[0007] In the art the terms multispectral and hyperspectral, multidimensional and hyperdimensional etc. are used, with "hyper" generally meaning more than "multi". This distinction is not relevant for the purposes of this invention. For convenience, throughout the rest of the specification the term "multispectral" will be used to refer to both multispectral and hyperspectral data. The term "multidimensional" and other "multi" terms will likewise be used to mean more than one dimension.

[0008] Current solutions of finding endmembers often involve "whitening" or "sphering" the data and then fitting to the data a multidimensional simplex having a number of vertices equal to the number of endmembers.

[0009] The bands of a multispectral image are usually highly correlated. "Whitening" involves transforming the data to be uncorrelated with a constant variance and preferably an approximately Normal distribution of errors. It is also desirable to compress the dimensionality of the data to reduce calculation time.

[0010] A widely used algorithm to "whiten" the data is to compress the information into a smaller number of bands by use of the Minimum Noise Fraction (MNF) transform. This is disclosed in Green, A., Berman, M., Switzer, P., and Craig, M. (1988). *A transformation for ordering multispectral data in terms of image quality with implications for noise removal. IEEE Transactions on Geoscience and Remote Sensing*, 26:65-74.

[0011] Simplex fitting using the pixel purity index (PPI) method is disclosed in Boardman, J. Kruse, F., and Green, R. (1995) Mapping target signatures via partial unmixing of AVIRIS data. In Green, R. (editor), *Summaries of the Fifth Annual JPL Airborne Earth Science Workshop*, volume 1, AVIRIS Workshop, pp 23-26. JPL Publ. 95-1, NASA, Pasadena, CA.

[0012] One of the main disadvantages of Boardman's method is that it requires considerable manual intervention in processing.

[0013] An alternative to Boardman's method is the N-FINDR algorithm by Winter, M. (1999). Fast autonomous spectral endmember determination in hyperspectral data. In *Proceedings of the 13th International Committee on Applied Geologic Remote Sensing*, Vancouver, vol. 2, pp 337-334. This process is fully automated. After transformation to (M-1) dimensional subspace, this algorithm finds the M-dimensional simplex of maximum volume constrained to lie within the data cloud. Another alternative is to construct the minimum volume simplex enclosing the data cloud, which is provided by Craig, M. (1994). Minimum-volume transforms for remotely sensed data. *IEEE Transactions on Geoscience and Remote Sensing*, 32:542-552.

[0014] These solutions cannot satisfactorily deal with the common situation where pure or almost pure endmembers are absent from the scene. Furthermore, they do not deal well with noise in the data.

SUMMARY OF THE PRESENT INVENTION

[0015] It is an object of the present invention to provide an improved method of identifying endmembers spectral values from multispectral data.

[0016] According to the present invention there is provided a method of identifying endmember spectral values from multispectral image data, where each multispectral data value is equal to a sum of mixing proportions of each endmember spectrum, said method including the steps of:

processing the data to obtain a multidimensional simplex having a number of vertices equal to the number of endmembers, the position of each vertex representing a spectrum of one of the endmembers,

wherein processing the data includes:

providing starting estimates of each endmember spectrum for each image data value;

estimating the mixing proportions for each data value from estimates of the spectra of all the endmembers;

estimating the spectrum of each endmember from estimates of the mixing proportions of the spectra of all the endmembers for each image data value;

repeating estimation steps until a relative change in the regularised residual sum of squares is sufficiently small, the regularised residual sum of squares including a term which is a measure of the size of the simplex.

[0017] Preferably the term used in the regularised residual sum of squares is the sum of the squared distances between all of the simplex vertices.

[0018] Preferably the step of providing the starting estimates includes choosing starting points with a high pixel purity index score. More preferably the starting estimates are well separated.

[0019] Preferably the relative change in the regularised residual sum of the squares is regarded as sufficiently small when the ratio of successive values of regularised residual sum of squares is less than a tolerance. Preferably the tolerance is 0.99999.

[0020] Preferably processing the data includes whitening the data. Preferably whitening the data includes conducting a transform of the data into data that is not band correlated. Preferably processing the data includes removing bands that do not have a high signal to noise ratio.

[0021] Preferably the step of estimating the spectrum of each endmember is conducted using a linear estimation technique.

[0022] Preferably the step of estimating the mixing proportions is conducted using a quadratic programming minimisation technique.

BRIEF DESCRIPTION OF THE DRAWINGS

[0023] In order to provide a better understanding a preferred embodiment of the present invention will now be described in detail, by way of example only, in relation to Figure 1 which is a diagrammatic representation of a simple example of the use of the method of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0024] Multispectral image data is obtained from a multispectral scanner, such as the AVIRIS airborne scanner. A typical scan for mineral applications includes a short wave infra red scan with wavelengths in the region of 2,000 to 2,500 nanometers. This

spectral range is useful for exhibiting distinctive shapes for important clay minerals. Typically this will provide 10's of thousands to millions of pixels or even many more.

[0025] A MNF transform is performed on the relevant bands of data to produce variables which are uncorrelated and approximately Normally distributed with an estimated error variance of 1. It is usual to retain MNF bands with the highest signal to noise ratios.

[0026] We let d be the number of MNF bands retained, N is the number of pixels, and M is the number of endmembers (assumed to be less than or equal to d+1).

[0027] It is convenient to think of the MNF data as an N x d matrix, whose ith row is written as X_i, and whose jth column is written as x_j. Similarly, it will also be convenient to think of the unknown endmembers as an M x d matrix, whose kth row is written as E_k and whose jth column is written as e_j.

[0028] The MNF transformed data can be represented by the following formula:

$$\mathbf{X}_i = \sum_{k=1}^{M} p_{ik}\mathbf{E}_k + \varepsilon_i, i=1,...,N. \qquad (1)$$

[0029] Here ε_i is an error vector, and the p_{ik} are mixing proportions that satisfy the constraints of:

$$p_{ik} \geq 0, k=1,...,M, \qquad \sum_{k=1}^{M} p_{ik} = 1, i=1,...,N. \qquad (2)$$

[0030] If the error term is ignored then (1) and (2) tell us that the data lie inside a simplex in (M-1) dimensional space, and the MNF representations of the M endmembers are at the vertices of the simplex.

[0031] A least squares minimisation of equation (1) is conducted subject to the constraints (2) and a term that constrains the size of the simplex, while being faithful to the model. It can be shown that without the constraints the solution converges to a simplex that is too large. This problem may be solved iteratively: given estimates of the

endmember spectra, the proportions for each pixel are estimated, which is a quadratic programming problem; and given estimates of the proportions, the endmember MNF spectra are estimated, which is a linear estimation problem.

[0032] The constraint is the addition of a term to the residual sum of squares which is a measure of the size of the simplex. A convenient term to add is the sum of squared distances between all of the simplex vertices. It can easily be shown that this is proportional to the sum of the variances of the simplex vertices over the *d* dimensions which is a quadratic function of the vertices and therefore computationally convenient. The regularised solution minimises:

$$R=\sum_{j=1}^{d} \{(x_j - Pe_j)^T(x_j - Pe_j) + \lambda e_j^T D e_j\}, \quad (3)$$

where λ is small, and where $D=I_M - 11^T/M$.

[0033] R is the regularised residual sum of squares; P is a $N \times M$ matrix of proportions of M endmembers for all N pixels; I_M is the $M \times M$ identity matrix and 1 is the vector of length M, where every entry equals 1.

[0034] Formula (3) is minimised iteratively.

[0035] In what follows, P_l will denote the estimated value of P after the lth iteration, and either $e_{j,l}$, $j=1,\ldots,d$ or $E_{k,l}$, $k=1,\ldots,M$ will denote the estimated endmembers after the lth iteration.

1. Let $e_{j,o}$, $j=1,\ldots,d$ denote the starting values for the algorithm and let $l=1$.
2. Let P_l denote the value of P minimising

$$R_{l,1}=\sum_{j=1}^{d} \{(x_j - Pe_{j,l-1})^T(x_j - Pe_{j,l-1}) + \lambda e_{j,l-1}^T D e_{j,l-1}\} \quad (4)$$

subject to (2). This is done using a quadratic programming algorithm.

There are two things to note here. First, the second term in (4) is independent of P, and so only the first term needs to be minimised in this step. Second, we can

separate the minimisation into N separate quadratic programming minimisations for the data at each of the N pixels. Specifically, for i=1,...,N, we find p_{ik},k=1,...,M, which minimise

$$\left(\mathbf{X}_i - \sum_{k=1}^{M} p_{ik}\,\mathbf{E}_{k,l-1}\right)^T \left(\mathbf{X}_i - \sum_{k=1}^{M} p_{ik}\,\mathbf{E}_{k,l-1}\right) \tag{5}$$

subject to (2).

Let $R_{l,1,min}$ denote the minimum value of $R_{l,1}$ achieved.

3. Let $e_{j,l}$,j=1,...,d denote the value of e_j minimising

$$R_{l,2} = \sum_{j=1}^{d} \{(x_j - P_l e_j)^T (x_j - P_l e_j) + \lambda e_j^T D e_j\} \tag{6}$$

This minimisation can be separated into d separate minimisations, and straightforward matrix algebra can be used to show that

$$e_{j,l} = (P_l^T P_l + \lambda D)^{-1} P_l^T x_j,\ j=1,\ldots,d. \tag{7}$$

Let $R_{l,2,min}$ denote the minimum value of $R_{l,2}$ achieved. It can also be shown that

$$R_{l,2,min} = \sum_{j=1}^{d} x_j^T \{I_N - P_l (P_l^T P_l + \lambda D)^{-1} P_l^T\} x_j, \tag{8}$$

where I_N is the N x N identity matrix.

4. Let

$$r_l = R_{l,2,min}/R_{l,1,min} \tag{9}$$

Because each step in the iteration reduces R, we must have $R_{l,2,min} \leq R_{l,1,min}$, or $r_l \leq 1$. When r_l gets very close to 1, the algorithm stops. More specifically:
If $r_l < tol$, let l=l+1 and go to 2. Otherwise, stop and let $e_{j,l}$,j=1,...,d, or equivalently $E_{k,l}$, k=1,...,M, be the final estimates of the endmembers, and let

the *i*th row of P_I give the final estimates of the proportions of each of the estimated endmembers present in the *i*th pixel.

[0036] The process is terminated when a ratio of successive values of the regularised residual sum of squares is less than a tolerance (*tol*). The default tolerance value is 0.99999. Using this value in typical examples, 20 to 100 iterations are required until the process stops.

[0037] Using this method the projections of all the data onto this hyperplane need not lie inside the simplex.

[0038] Figure 1 shows a simulated toy example with a true simplex as a solid line, an (unregularised) least squares solution as a dotted line and a regularised least squares solution as a broken line. The regularised least squares solution provides much better estimates of the true endmembers.

[0039] Most of the information about the simplex is contained in data on or near boundaries of the data cloud. So if only data nearest the convex hull of the data cloud is used computation becomes quicker. In high dimensional problems, points at or near the ends of random one dimensional projections of the data can be used. Alternatively points only on or near two dimensional convex hulls of all $d(d\text{-}1)/2$ MNF band pairs are used.

[0040] The starting points for the iterative process can determine the outcome of the iterative process. Good starting points should be well separated in MNF space. Points with high PPI scores can be useful. The PPI scores are the number of times the data at each pixel are at or near one of the ends of these projections.

[0041] Apart from the estimates of the endmembers an intermediate product of the algorithm is the endmember proportions in each pixel. The proportions give a meaningful idea of how much each endmember is represented in each pixel (assuming that mixing is proportional to area). This can be represented as images/maps. A particularly useful diagnostic is the maximum proportion of each estimated endmember in the scene. The lower the maximum proportions for each endmember spectrum the further the estimated endmember is from the data cloud and the confidence in the estimate will be correspondingly smaller. For endmember estimates having a maximum

proportion less than 0.5 it becomes difficult to estimate the true endmember spectrum.

[0042] Another useful by-product of the algorithm is an image showing the contribution of each pixel to the regularised residual sum of squares. If there are some large residuals and especially if they are spatially clustered it is an indication that the model is not fitting the data adequately, either because the chosen value of M is too small or if only the data on or near the boundaries of the data cloud are used then important observations may have been omitted from the algorithm. Additional observations can be added to the data used and the algorithm re-run to see whether fitting can be improved.

[0043] The present invention may be readily adapted to identify endmember spectral values from multispectral data from a variety of sources, such as remote sensing data, array data, etc.

[0044] Obvious adaptions of the preset invention to make the hereindescribed method suitable to identify endmembers from a nominated source are intended to fall within the scope of the preset invention.

CLAIMS DEFINING THE INVENTION ARE AS FOLLOWS:-

1. A method of identifying endmember spectra values from multispectral image data, where each multispectral data value is equal to a sum of mixing proportions of each endmember spectrum, said method including the steps of:

processing the data to obtain a multidimensional simplex having a number of vertices equal to the number of endmembers, the position of each vertex representing a spectrum of one of the endmembers,

wherein processing the data includes:

providing starting estimates of each endmember spectrum for each image data value;

estimating the mixing proportions for each data value from estimates of the spectra of all the endmembers;

estimating the spectrum of each endmember from estimates of the mixing proportions of the spectra of all the endmembers for each image data value;

repeating estimation steps until a relative change in the regularised residual sum of squares is sufficiently small, the regularised residual sum of squares including a term which is a measure of the size of the simplex.

2. A method according to claim 1, wherein the term used in the regularised residual sum of squares is the sum of the squared distances between all of the simplex vertices.

3. A method according to claim 1, wherein the step of providing the starting estimates includes choosing starting points with a high pixel purity index score.

4. A method according to claim 1, wherein more preferably the starting estimates are well separated.

5. A method according to claim 1, wherein the relative change in the regularised residual sum of the squares is regarded as sufficiently small when the ratio of successive values of regularised residual sum of squares is less than a tolerance.

6. A method according to claim 1, wherein the tolerance is 0.99999.

7. A method according to claim 1, wherein processing the data includes whitening the data.

8. A method according to claim 7, wherein whitening the data includes conducting a transform of the data into data that is not band correlated.

9. A method according to claim 1, wherein processing the data includes removing bands that do not have a high signal to noise ratio.

10. A method according to claim 1, wherein the step of estimating the spectrum of each endmember is conducted using a linear estimation technique.

11. A method according to claim 1, wherein the step of estimating the mixing proportions is conducted using a quadratic programming minimisation technique.