Automatic identification of hyperspectral data: construction of unique signatures

A. Averbuch[1] V. Zheludev[1] N. Ben-Gigi[2] A. Schclar[1] O. Braun[3] Y. Zur[3]

[1] School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

[2]Department of Applied Mathematics, School of Mathematical Sciences

Tel Aviv University, Tel Aviv 69978, Israel

[3] Bar-Kal Systems Engineering, 43 Hamelacha St. P.O.Box 8661

Netanya 42504, Israel

December 21, 2007

Abstract

The paper describes an efficient identification scheme for a specific spectrum against a given database that can run in real time. The scheme finds all the spectra in the database that are similar to the presented spectrum. The similarity measure in our scheme is defined as a number of coinciding geometric characteristics from the spectral curve (up to a tolerance interval). The current geometric characteristics utilized by the scheme are the locations of minima, flat intervals and inflexion points. These characteristics have apparent physical meaning. However, any other set of geometric features can be used by the algorithm. There are no limitations on the number of spectra in the database that the classifier can handle. It can help in performing unmixing where the measurements of the endmembers are not pure but corrupted with noise. A robust features extraction method is presented. The algorithm can be applied to a general curve fitting manipulation.

1 Introduction

The goal of this paper is to design an efficient methodology for automatic identification of pure materials using their spectra in the multitude of wavelengths. For this purpose, characteristic features of each spectrum are determined and extracted. The characteristic features of the spectra from a given database are used for the construction of a reference table. At the identification phase, the characteristic features are extracted from the presented spectrum. Then, a fast algorithm, which uses the reference table, finds the spectra in the database that have common features with the presented spectrum. The size of of the reference table is relatively small in comparison to the volume (size) needed to store the full database. The identification procedure is implemented in real time. Thus, the size of the reference database can be unlimited.

The set of the extracted features that we use is physically justified. Therefore, the coincidence of all (or many) features of two spectra reflects the identity (or the existence of similarity relation) between the materials to which these spectra belong to. The features are:

- Absorption bands that are described by the locations of deep or shallow minima.
- Locations of inflection points.
- Locations of flat intervals in the spectral curve.

These features, which describe the spectral signature, are based on the absorption features and on the backscattering parameters of the material. For instance, the clay minerals Montmorillonite and Kaolinite have an absorption band at 2205 nm (deep minimum), whereas only Kaolinite has an additional small absorption band at 2165 nm (shallow minimum). The proposed method enables to distinguish between these two clay minerals and points to their different features. Both the inflection point location of an absorption band and flat intervals in a spectral signature, are features that can help to distinguish between spectra.

An hyperspectral data will be used to demonstrate the applicabilities of the proposed algorithm.

1.1 Related work and comparison to our scheme

Spectral Angle Mapper (SAM) ([1, 2, 3]) . The reflectance of an hyper spectral pixel can be seen as an N-dimensional vector where N is the number of bands. The length of the vector is its brightness. The direction of this vector can be seen as a spectral feature that is contained in the pixel. The variousness of the illumination affects only the length of the vector but not its angle. The classification of a target is done by calculating the angle between the vector of the analyzed pixel and the vectors (representing materials) from the spectral libraries/database. Each pixel classifies the material with the lowest spectral angel value.

The weakness of the SAM method is that it does not distinguish between positive and negative correlations because only the absolute value is taken into account. The SCM algorithm overcomes this limitation.

Spectral Correlation Mapper (SCM) ([4]) . The difference between SAM and SCM is in the vectors of the reflected spectrum. They are standardized before calculating the angle. Namely, if in SAM, the angle is calculated then in SCM the Pearsonian Correlation Coefficient (PCC) is used. The computation is similar. The difference is that the PCC standardizes the data.

Spectral Identification Method(SIM) ([5, 6]) uses the statistical procedure ANOVA ([7, 8, 9, 10]) that is based on linear regression. It insensitive to the presence of negative correlation. By combining it with the SCM coefficients, it uses the negative/positive information. This technique exhibits three estimates according to the levels of the significance of the materials.

Optimum Index Factor (OIF) ([13, 14]) is used to select the optimal combination of any number of bands in a satellite image to create a color composite. The most significant bands, which contain the highest information, are chosen. The OIF is based on the total variance within bands and on the correlation coefficients between bands. The combination of bands with the smallest correlation coefficient between bands and with the highest total variance within bands will have the maximum OIF. This will be selected.

This approach is different from previous approaches. While others try, in general, to deal with the whole spectrum, this approach seeks a solution for the high-dimensionality of this problem (too much data with high inter-band correlation).

Comparison to our scheme: Our approach in this paper follows, to some extent, the logic of OIF (specific band selection for classification) while achieving an essential dimensionality reduction unlike all the above methods. The proposed method differs from other methods that classify and process spectral signatures in two major aspects . First, in the proposed algorithm, the comparison between spectral signatures is based on the locations of the spectral features in the spectral dimension. It means that the reflectance value of a spectral feature at this location does not affect the comparison. Second, the analysis is done on the reflectance itself and no preprocessing such as normalization is needed as it is done for example in the spectral feature fitting (ENVI software). Other methods such as SAM have to possess the whole vector in order to distinguish between two spectral features. As it was mentioned above, the memory consumption, even for huge databases, is very limited and the identification procedure allows real-time implementation.

1.2 Test data used for the validation of the proposed algorithm

The database that is being used to test and validate the proposed algorithms consists of 1000 spectral in reflectance signatures of natural targets that were measured by a field spectro

radiometer (FR) manufactured by ASD (Boulder, Colorado). The measurements range is between 400 to 2400 nm with a spectral sampling resolution of 1nm. All the spectral signatures are of natural targets with reoccurrence of a few targets measured in different spectral conditions and instruments. Some of the measurements were taken in sun light conditions and others were taken with a contact probe. No noise was added to the measured data. The analysis of the method, which are described in this work, was done on authentic reflectance measurements of real field measurements.

The paper has the following structure: Section 2 describes the extraction procedure of characteristic features from the spectral curves and the construction of the reference table from the characteristic features of spectra in a given database. Section 3 describes the identification procedure of a presented spectrum in the reference database. In section 4, we provide a few examples of the application of the algorithm to the identification of real spectra. The obtained results are discussed in section 5. The appendix (section 6) describes in details the scheme for features extraction and illustrate by a running example the construction of the reference table and the identification procedure. In section 6.2 we present the results from the experiments with natural targets.

2 Construction of the reference table

2.1 Extraction of characteristic features

The spectra, which we use to demonstrate the performance of the algorithm, represent absorbtion values in reflectance. The algorithm can operate in the same way if the spectra came from radiation or any other source of data. The description of the general algorithm is accompanied by a detailed running example of a real database that contains 173 entries (spectra). Each spectrum (vector) in the database is of predefined length which is the number of wavelengths (in our case 2001). We construct a robust classifier that identifies each entry uniquely by the

extracted features. A significant reduction of the dimensionality of the database is achieved by extraction of characteristic features from the vectors in the database.

We use four types of geometrical characteristics, which represent the physical properties of the spectra:

1. Location of deep minima.

2. Location of shallow and one-side minima.

3. Location of near-horizontal flat intervals.

4. Location of inflexion points.

These four features enable to speedup the performance of the algorithm. We can add more features or remove some. For example, the area below two consecutive minima can be another feature. There is no limitations on the number of features and their meaning. Any applicable feature that is chosen can be added.

The algorithm uses some parameters. All the "hardwired" parameters (numbers) that appear throughout the algorithm are parameters that were chosen arbitrarily but they can be modified and changed according to specifics of the problem.

The problem is to correctly locate essential physical events even when strong noise is present. In addition, they have to identify uniquely the spectral entries in the database.

Let $\mathbf{y} = \{y(k)\}$ be an original raw spectrum of length N (number of wavelengths, in our case $N = 2001$). Let

$$Dy(k) = \frac{y(k+1) - y(k-1)}{2}, \quad D^2y(k) = \frac{Dy(k+1) - Dy(k-1)}{2}, \quad k = 2, \ldots, N$$

be the first and second centered differences of the array $\mathbf{y}$, respectively. In real data, all three arrays $\{Y(k)\}$, $\{DY(k)\}$ and $\{D^2Y(k)\}$ are corrupted by noise. Therefore, we filter these arrays by the application of a low-pass shape preserving B-spline filter (see in the Appendix

section 6.3) to the raw spectra. As a result, we get that the arrays $\{Y(k)\}$, $\{DY(k)\}$ and $\{D^2Y(k)\}$ are smoother, while their geometric characteristics (features) are preserved.

Details how to find the characteristic features are presented in section 6.1 of the Appendix. We assume that each spectrum has at most 10 locations for each selected feature.

Example: The construction is illustrated by a running example. We choose entry number 50 (denoted by $\mathbf{X}_{50}$ and displayed in Fig. 1) from the database that was described in section 1.2. The spectrum represents the absorbtion values of a material with 2001 wavelengths.

$$\mathbf{X}_{50} = \begin{array}{l} \text{Deep min} \\ \text{Shal. min.} \\ \text{Flat int.} \\ \text{Inf. points} \end{array} \begin{pmatrix} 1124 & 2091 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 935 & 1467 & 1735 & 2162 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1092 & 1187 & 1287 & 1440 & 1570 & 1642 & 2014 & 2311 & 2345 & 0 \\ 426 & 617 & 811 & 895 & 1034 & 1228 & 1604 & 1701 & 2042 & 2119 \end{pmatrix}. \tag{1}$$

The structure of the matrix in Eq. (1) is: The first row contains deep minima points. There are only three such points. The vacant places in the row are filled with zeros. Second row contains shallow minima points. Third row contains indicators of flat intervals and the last row contains the first ten inflexion points. The spectrum and its features are displayed in Figure 1.

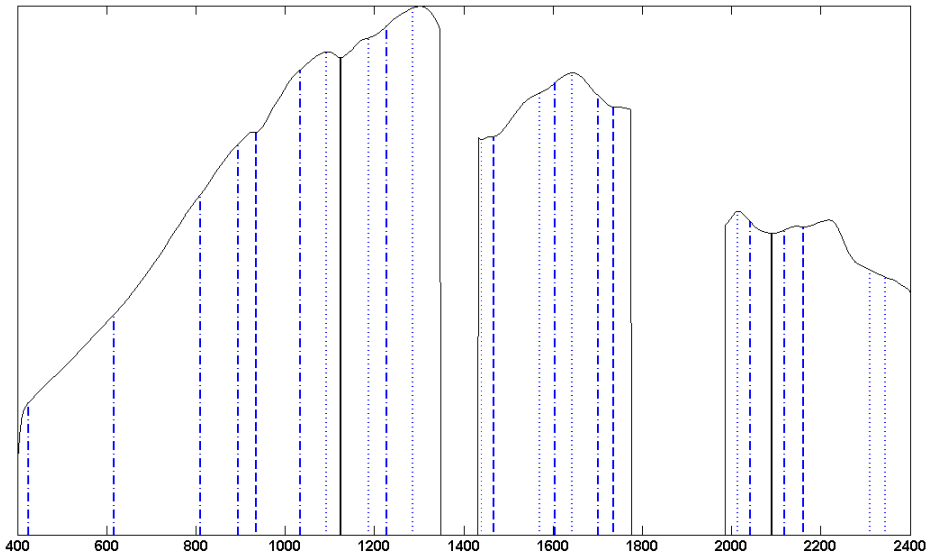

Figure 1: Spectrum # 50 from the database described in section 1.2. Smoothed data is displayed by the curve, solid verticals describe the location of deep minima, dashed verticals describe the shallow minima, dot verticals describe the lat intervals and dashdot verticals describe the inflexion points.

2.2 Construction of the table from the reference database

As a result of above procedures, we obtain a set of the extracted selected characteristic features from all the N spectra in the reference database. For each spectrum, we have at most 40 features, which describe the locations of deep and shallow minima, flat intervals and inflexion points. Assume that the locations of all the features do not exceed R (for our data $R = 3000$). To formalize the distinction between different types of features, we increase each location of the shallow minimum by R, each location of the flat interval by $2R$ and each location of the inflexion point by $3R$. Then, we construct the table $\mathbf{T}$ that describes the distribution of the selected characteristic features among the spectra in the reference database.

We set a tolerance interval whose length I is determined according to the conditions of the experiment. We do not distinguish between two features f_1 and f_2 if their difference is less than I. Namely, if $0 < f_2 - f_1 < I$ then we assume that $f_2 = f_1$. Let K be the number of different (up to the tolerance interval I) features in all N spectra in the database. Table $\mathbf{T} = \{t(i,j)\}, i = 1, ldots, K, j = 1, \ldots, N+1$, is organized as follows. Its first column comprises all K features in ascending order. The remaining columns in $\mathbf{T}$ consist of zeros and ones. Locations of ones in a row k indicate that the spectra have the feature $t(k,1)$. For example, if $t(k,1) = \tau_0 < R$ and $t(k,m) = 1$ then it means that spectrum $\#m-1$ has a feature at the location τ_0 (up to a tolerance interval I). Since $\tau_0 < R$, this feature is a deep minimum. If $t(w,1) = \tau_1,\ R < \tau_1 < 2R$ and $t(w,v) = 1$ then it means that spectrum $\#v-1$ has a shallow minimum at location $\tau_1 - R$ (up to a tolerance interval I).

Example for the construction of a reference table: Table 1 displays the locations of deep and shallow minima in spectra # 36–41 from the reference database.

36	669	1157	4454	5008	5143	5166	5196	5299	5322	0	0	0	0	0
37	1162	1446	5155	0	0	0	0	0	0	0	0	0	0	0
38	673	743	1008	1185	2011	2236	2337	3616	5046	5067	5088	5117	5188	0
39	670	1009	1171	1444	3970	4119	5008	5043	5171	5197	5232	5261	5290	0
40	432	672	973	1167	2001	2148	2343	3619	5039	5062	5088	5117	5181	5215
41	671	1168	1445	3420	3617	5056	5169	5258	5324	0	0	0	0	0

Table 1: Deep and shallow minima locations in ascending order in spectra # 36–41.

The feature Table 1 is transformed into the reference table $\mathbf{T}$ whose portions are displayed in Table 2. For example, the table indicates that only spectrun # 40 has deep minimum at wavelength of 973nm, whereas the deep minimum at wavelength of 1157nm is peculiar to the spectra # 36, 37 and 40. Spectra # 38, 40 and 41 have the shallow minimum at wavelength of 616nm.

Features	36	37	38	39	40	41
973	0	0	0	0	1	0
1008	0	0	1	1	0	0
1157	1	1	0	0	1	0
1168	0	0	0	1	0	1
1185	0	0	1	0	0	0
1444	0	1	0	1	0	1
2001	0	0	1	0	1	0

Features	36	37	38	39	40	41
3616	0	0	1	0	1	1
3970	0	0	0	1	0	0
4119	0	0	0	1	0	0
4623	0	0	0	0	1	1
5008	1	0	0	1	0	0
5039	0	0	1	1	1	0
5056	0	0	0	1	0	0

Table 2: Portions of the reference table **T**. They represent several deep minima of spectra # 36–41 (left table) and shallow minima (right table).

Reference table **T** contains all the information needed for the identification of a test spectrum which does not belong to the reference database. Thus, there is no need to keep in memory the whole database. Obviously, the memory's size needed for storing table **T** is insignificant even for huge databases.

3 Fast identification of a spectrum

Input: A raw *test* spectrum s_n and the constructed reference table **T** (as was described in Section 2.2) are loaded. The test spectrum may or may not belong to the reference database.

Extraction of features: The features of the spectrum s_n are extracted according to the procedure described in Section 6.1 in the Appendix. The features are gathered into a row vector $\mathbf{X}_n = \{x_n(k)\}_{k=1}^{M}$ (M is our case can be at most 40).

Reduction of the feature distribution in table T: Recall that for $i = 1, \ldots, K$, the first column $t(i, 1)$ in table **T** contains all the features from the reference database. Then,

table $\mathbf{T}$ of size $K \times N+1$ is reduced to table $\mathbf{T}_n$ of size $M \times N+1$ in such a way that only the rows $t(i,:)$ are retained. Its first terms $t(i,1)$ coincide (up to the tolerance interval $\mathbf{I}$) with one of the features $x_n(k)$ in the tested spectrum

Detecting spectra that have common features with the presented spectrum: For this purpose, we calculate the sums in columns $t(i,j),\ j=2,\ldots,N+1$, in the reduced table $\mathbf{T}_n$ that correspond to the spectra in the reference database $C_{n,j}=\sum_i t(i,j),\ j=2,\ldots,N+1$.

$C_{n,j}$ determines the number of features in spectrum $\#j$ in the database that coincide (up to the tolerance interval $\mathbf{I}$) with the features of the presented spectrum $\#n$.

Note that we can use all the available sets of features (deep and shallow minima, flat intervals and inflexion points) or any combination of these sets.

Histograms: We present the result from the search by an histogram that displays graphically the total number of spectra that have different numbers of common features with the test spectrum.

All these operations are conducted in real time.

The output from the identification process: Once the histogram related to the presented spectrum # n is displayed, the user indicates the number ν of common features to be used. Then, the algorithm produces a list of spectra that have $\geq \nu$ common features (up to the size of the tolerance interval) with the presented spectrum. Optionally, the plots of these spectra are displayed versus the plot of the presented spectrum. Common features are also displayed in these images.

We exemplify the process of the identification of a spectrum in Section 6.2 in the Appendix.

4 Identification results

In this section, we present a few examples for the identification of natural materials by the above algorithm. We experiment with a database that was described in section 1.2. Some groups of spectra belong to the same material but were measured under different conditions.

Initially, we extracted the characteristic features from the available 173 spectra and constructed the reference table. It appeared that there are 414 different (up to a tolerance interval $I = 10$) features. Consequently, the size of the reference table becomes 414×174, where the first column contains all the features in ascending order. Then, we tested a number of spectra from that database in order to find similar spectra. The results support the validity of the proposed scheme. The spectra that have a number of common features with the tested spectrum belonged either to the same material or to the same group of materials that this spectrum belongs to.

Identification of the spectrum # 25: Figures 2 and 3 illustrate the results from the identification of the spectrum # 25 against the full database using only the deep minima features. This spectrum was emitted from a green vegetable material. Figure 2 displays the distribution of spectra in the database with respect to the number of deep minima common with the deep minima of spectrum # 25. From the histogram we see that two spectra # 26 and 34 have six common deep minima and three spectra # 27, 30 and 31 have five deep minima common to spectrum # 25. We display the spectrum # 25 and five similar spectra in Fig. 3. In addition, the locations of the features are marked in the plots. All these spectra were emitted from green vegetable materials. The spectra # 25 and 26 are emitted from the same material.

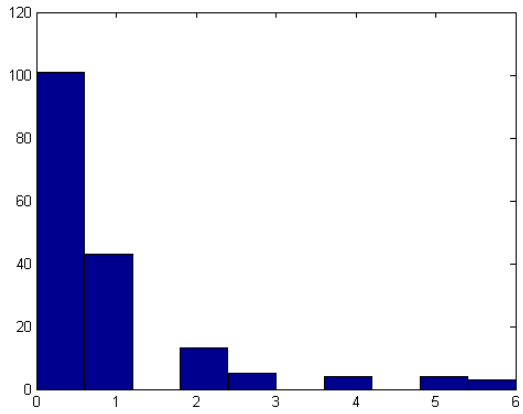

Figure 2: Histogram of the distribution of spectra in the database with respect to the number of deep minima common with the deep minima of the spectrum # 25.

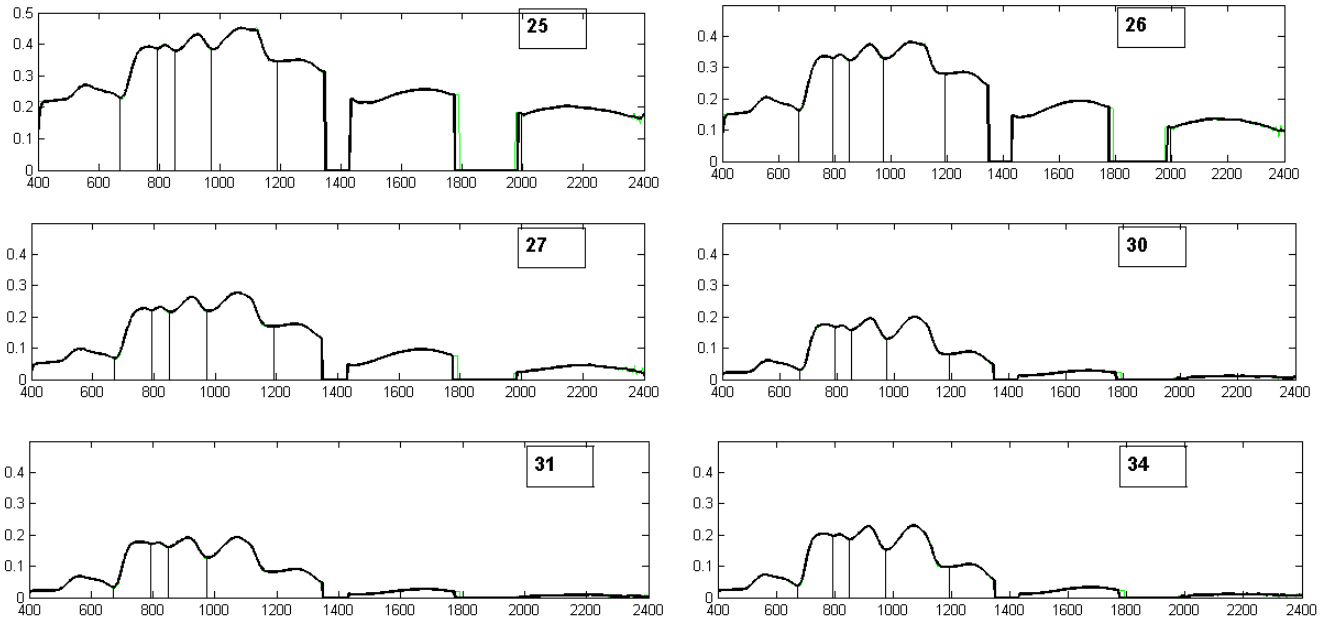

Figure 3: Tested spectrum # 25 (top left) and similar spectra. Spectra # 26 (top right) and # 34 (bottom right) have six common deep minima with spectrum # 25, spectra # 27 (center left), 30 (center right) and 31 (bottom left) have five deep minima common with the spectrum # 25.

Identification of the spectrum # 38: Figures 4 and 5 illustrate the results from the identification of the spectrum # 38 against the full database using only deep minima features. This spectrum was emitted by wet mud at an animal compound. Figure 4 displays the distribution of spectra in the database with respect to the number of deep minima common with the deep minima of the spectrum # 38. From the histogram we see that three spectra # 1, 5 and 13 have four common deep minima with the spectrum # 38. The spectrum # 38 and three similar spectra are displayed in Fig. 5. In addition, the locations of the features are marked in the plots. These three spectra were emitted by

the spring and river water measured in different locations.

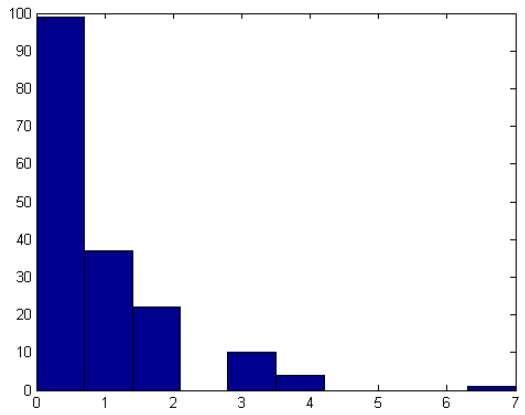

Figure 4: Histogram of the distribution of spectra in the database with respect to the number of deep minima common with the deep minima of the spectrum # 25.

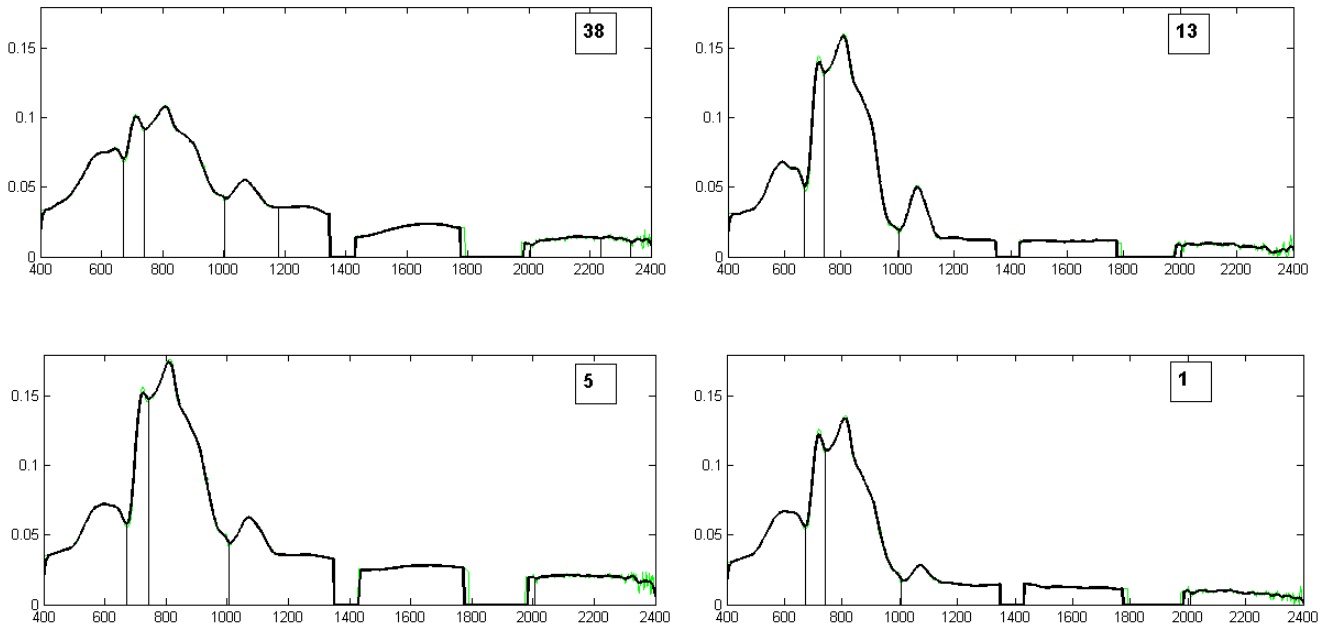

Figure 5: Tested spectrum # 39 (top left) and similar spectra. Spectra # 13, (top right) #5 (bottom left) and # 1 (bottom right) have four common deep minima with spectrum # 39.

Carton is similar to hey: Figure 6 illustrates the result from the comparison of the features in spectrum # 60 (carton) with the features of spectrum # 49, which was emitted by hey at an animal compound. In addition to the coincidence of two major absorption bands (deep minima), they have in common one shallow minimum and one flat interval. This results from the strong cellulose component inherent in both materials.

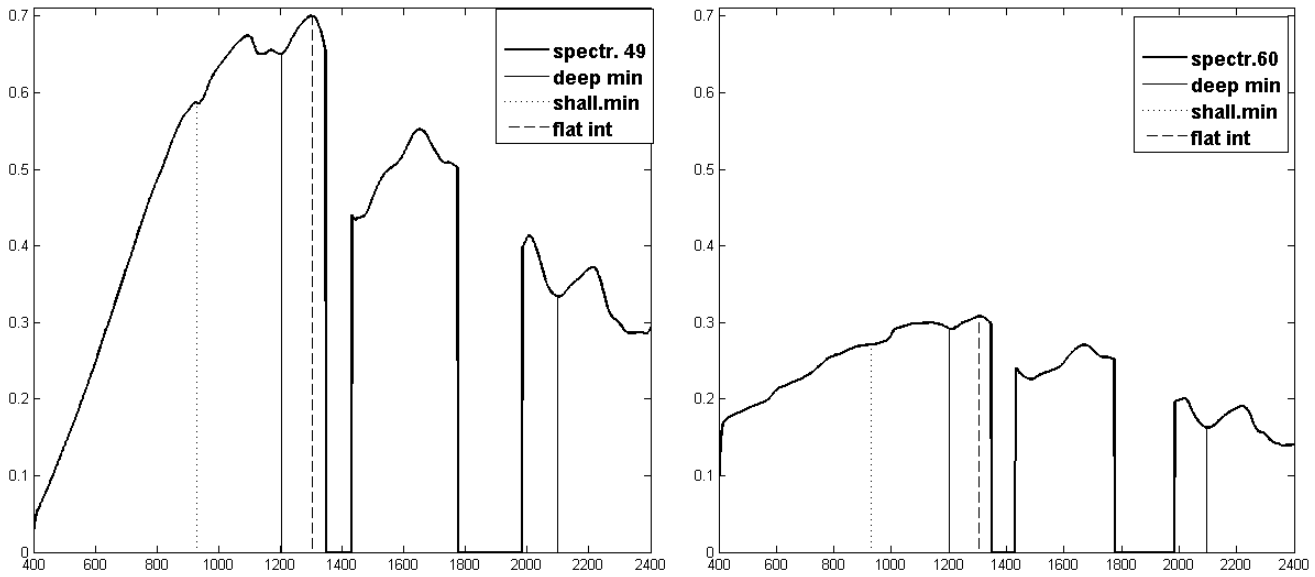

Figure 6: Tested spectrum # 49 (left) and similar spectrum # 60 (right).

5 Conclusions

The algorithm, which was described in this paper, efficiently reveals the spectra in a reference database, which are related to the spectrum presented for identification. The extent of the relation is interactively defined by the user. If the presented spectrum belongs to the reference database then the algorithm correctly identifies it and finds all the spectra originating from the same material as well as the spectra belonging to a similar group of materials. The scheme can be utilized as an initial step in solving the unmixing problem. Namely, it can reveal potential endmembers prior to a qualitative analysis.

Typically, when spectra are recorded in natural environment, they are corrupted by noise. Deep minima locations, which present the major absorption bands, remain stable and visible

even when strong noise is present. This is not the case for other sets of features. However, these are useful later for the identification of relatively clean spectra and for the exact search of a certain spectrum in a given database. Other physical features can be chosen either to replace or enhance the physical features presented in this paper.

The algorithm was implemented in Matlab and in ENVI-IDL.

6 Appendix

6.1 Extraction of characteristic features

Here is the description how features 1-4 are found:

Location of minima: We set two parameters T_1 and $T_2 < T_1$. First, we find all the points $\{k_c\}$, where the sequence $\{DY(k)\}$ changes the sign from $-$ *to* $+$: $DY(k_c - 1) < 0,\ DY(k_c + 1) > 0$. Let k_c be a marked point. Let $k_{cl} \leq k_c$ be the point such that for all $k_{cl} \leq k \leq k_c,\ DY(k) < 0$. Let $k_{cr} \geq k_c$ be the point such that for all $k_c < k \leq k_{cr}\ DY(k) > 0$. Denote by $V_l \stackrel{\Delta}{=} \{k : k_{cl} \leq k < k_c\}$ and $V_r \stackrel{\Delta}{=} \{k : k_c < k \leq k_{cr}\}$ the left- and right-side vicinities of the point k_c, respectively. We calculate $M_l \stackrel{\Delta}{=} \max_{k \in V_l} Y(k)$, $M_r \stackrel{\Delta}{=} \max_{k \in V_r} Y(k)$, $M \stackrel{\Delta}{=} \max(M_r, M_l) - Y(k_c)$ and $m \stackrel{\Delta}{=} \min(M_r, M_l) - Y(k_c)$.

Then, we sort the points in the following way:

- If $M < T_1$ then we discard the point k_c.
- If $M > T_1$ and $m > T_2$ then we mark the point $k_c = k_{dm}$ as a deep minima.
- If $M > T_1$ but $m < T_2$ then we mark the point $k_c = k_{sm}$ as a shallow or one-side minima.

Location of near-horizontal flat intervals: We set two parameters: a small T_3 and a natural number N_1. We find intervals of length no less than N_1 samples such that

$0 < |DY(k)| < T_3$ within these intervals. The central points k_f in these intervals are marked as the locations of near-horizontal flat intervals.

Location of inflexion points: We set a parameters T_4. First, we find all the points $\{k_q\}$, where the second difference $\{D^2Y(k)\}$ changes the sign $-to+$ or $+to-$: $D^2Y(k_q-1) < 0$ and $D^2Y(k_q+1) > 0$ or $D^2Y(k_q-1) > 0$ and $D^2Y(k_q+1) < 0$. Then, we sort the points in the following way: Let k_q marks "$-to+$" point. Let $k_{ql} \leq k_q$ be the point such that for all $k_{ql} \leq k \leq k_q$, $D^2Y(k) < 0$. Let $k_{qr} \geq k_q$ be the point such that for all $k_q < k \leq k_{qr}$, $D^2Y(k) > 0$. Denote by $V_w \stackrel{\Delta}{=} \{k : k_{ql} \leq k \leq k_{qr}\}$ and $V_n \stackrel{\Delta}{=} \{k : k_q - 5 \leq k \leq k_q + 5\}$ the wide and narrow vicinities of the point k_q, respectively. We calculate $M_w \stackrel{\Delta}{=} \max_{k \in V_w} |(DY(k)|$ and $m_n \stackrel{\Delta}{=} \min_{k \in V_n} |(DY(k)|$. We discard the point k_q if $M_w < T_4$ (near flat interval) or if $m_n > T_4$ (near vertical interval). Otherwise, the point k_q is marked as an inflexion point.

Thinning the features arrays: We set a parameter N_2 to be a natural number.

- If several deep minima locations are located inside an interval of length N_2, we retain only one point, where the value of Y is the smallest. Other points inside the interval are discarded.
- The same is done for shallow minima points and for the indicators of flat intervals.
- Similar operation is performed for the inflexion points but here points with the smallest $|DY|$ values are retained.

Sorting the features arrays: We consider the deep minima as the most significant features, next in significance are the shallow minima, which are followed by flat intervals and inflexion points, respectively. This rating does not affect the output of the algorithm. It affects only the structure of the reference table to be formed. Consequently, we carry out the following procedures:

We set a parameter N_3 to be a natural number.

- If samples from an inflexion point lies a minimum point or an indicator of a flat interval within distance of N_3, then, this inflexion point is discarded.
- If samples from an indicator of a of flat interval within distance of N_3 lies a minimum point then this indicator is discarded.
- If samples from a shallow minimum within distance of N_3 lies a deep minimum point then this shallow minimum point is discarded.

Finally, a spectrum $\#n$ has a reduced set of features.

6.2 Example of the identification of a spectrum in a small database

We consider the identification of the spectrum # 39 against the small database # 36–41. We extract the features (only deep minima in this example) from this spectrum and arrange them into the row

$$\mathbf{X}_{39} = (670\ 1009\ 1171\ 1444\ ).$$

Then, we reduce the reference table $\mathbf{T}$ by retaining only rows γ such that $|x_n(k) - t(\gamma, 1)| < I$, where $I = 10$ is the length of the tolerance interval. The reduced reference table is

$t(i,1)$	S.36 $t(i,2)$	S.37 $t(i,3)$	S.38 $t(i,4)$	S.39 $t(i,5)$	S.40 $t(i,6)$	S.41 $t(i,7)$
669	1	0	1	1	1	1
1008	0	0	1	1	0	0
1168	0	0	0	1	0	1
1444	0	1	0	1	0	1
$C(39,j)$	1	1	2	4	1	3

Table 3: Reduced reference table $\mathbf{T}_{39}$ for the database of spectra # 36–41.

The histogram in Fig. 7 displays the distribution of spectra from the reference database (spectra # 36 to # 41) according to the number of common features (deep minima) with the

presented spectrum # 39, which was emitted by wet mud at an animal compound. We can see from the histogram in Fig. 7 that three spectra have only one common feature with spectrum # 39, spectrum #38 (wet mud) has two common features and spectrum #41 (wet meadow) has tree common features with the spectrum # 39. We display last two spectra in Fig. 8.

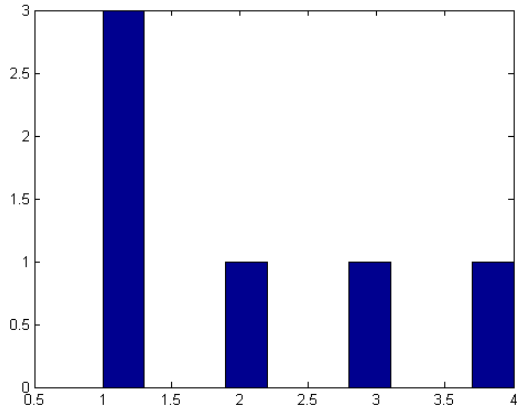

Figure 7: Distribution of spectra # 36 to # 41 according to the number of common features (deep minima) with the presented spectrum # 39.

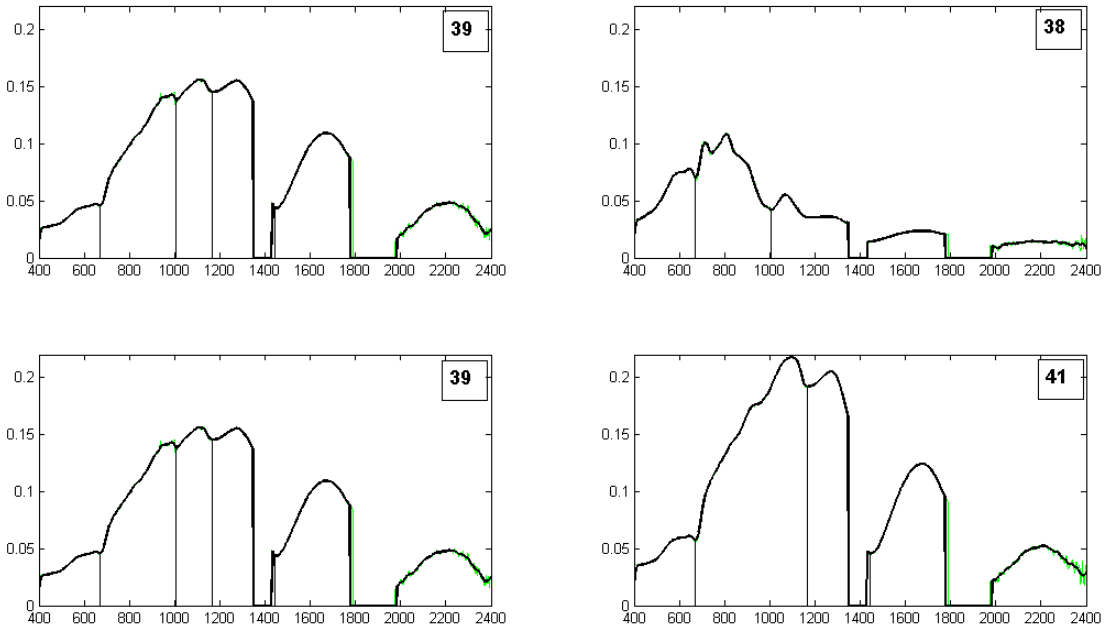

Figure 8: Spectrum # 39 (top left and bottom left) tested against the reduced database and similar spectra. Spectrum # 38 (top right) has two common deep minima with spectrum # 39, spectrum # 41 (bottom right) has three deep minima common with the spectrum # 39.

6.3 Shape preserving filtering

The centered B-spline of first order is the characteristic function of the interval $[-1/2, 1/2]$. The centered B-spline of order p is the convolution $M^p(x) = M^{p-1}(x) * M^1(x) \quad p \geq 2$. Note that the B-spline of order p is supported on the interval $(-p/2, p/2)$. It is positive within its support and symmetric about zero. The B-spline M^p consists of pieces of polynomials of degree $p-1$ that are linked to each other at the nodes such that $M^p \in C^{p-2}$. Nodes of B-splines of even order are located at points $\{k\}$ and of odd order, at points $\{k+1/2\},\ k \in \mathbb{Z}$.

The explicit representation of the B-spline is

$$M^p(t) = \frac{1}{(p-1)!} \sum_{k=0}^{p-1} (-1)^k \binom{p}{k} \left(t + \frac{p}{2} - k\right)_+^{p-1}, \quad t_+ \stackrel{\Delta}{=} (t + |t|)/2.$$

Shifts of B-splines form a basis in the spline space. Namely, any spline $S^p(x)$ of order p has the following representation: $S^p(x) = \sum_{k\in\mathbb{Z}} c_k M^p(x-k)$.

Assume that $f(x)$ is a smooth function and the data $g_k \stackrel{\Delta}{=} f(k) + e_k,\ k \in \mathbb{Z}$, are available, where $\{e_k\}$ is the random noise array. Then the spline $S^p(x) = \sum_{k\in\mathbb{Z}} g_k M^p(x-k)$ smoothes the data $\{g_k\}$, while preserving the shape of the function $f(x)$ ([15]).

The values of the spline at grid points are $S^p(n) = \sum_{k\in\mathbb{Z}} g_k M^p(n-k)$. They are produced by filtering the data array $\{g_k\}$ with the low-pass filter, which impulse response (finite and symmetric) consists of the values of B-spline at grid points. This filtering smoothes the data $\{g_k\}$, while preserving the shape of the function $f(x)$.

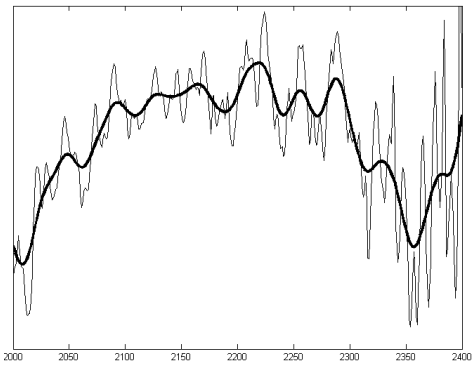

Figure 9: A fragment of the smoothed spectrum.

References

[1] Kruse, F. A., Lefkoff, A. B., Boardman, J. W., Heiehbrecht, K. B., Shapiro, A T., Barloon, P. J. and Goetz, A .F. H., The Spectral Image Processing (SIPS) - Software

for Integrated Analysis of AVIRIS Data, Summaries of the 4th Annual JPL Airborne Geosciences Workshop, JPL Publication **92-14**, pp. 23–25, 1992.

[2] Kruse, F. A., Lefkoff, A. and Dietz, J. B., Expert System-based Mineral Mapping in North Death Valley, California/Navad, Using the Airborne Visible/Infrared Imaging Spectrometer (AVARIS), Remote Sens. Environ., **44(2)**, pp. 309–336, 1993.

[3] Kruse, F. A., Lefkoff, A. B., Boardman, J. W., Heiehbrecht, K. B., Shapiro, A T., Barloon, P. J. and Goetz, A .F. H., The Spectral Image Processing System (SIPS) - Interactive Visualization and Analysis of Imaging Spectrometer Data, Remote Sens. Environ. **44**, pp. 145–163, 1993.

[4] Carvalho Jr., O. A. and Meneses, P. R., Spectral Correlation Mapper (SCM): An Improvement on the Spectral Angle Mapper, In: Ninth JPL Airborne Earth Science Workshop. JPL Publication **00-18**, pp. 65–74, 2000.

[5] Carvalho Jr., O. A., Carvalho A. P., Meneses, P. R. and Guimaraes, R. F., Spectral Identification Method(SIM): A New Classifier Based On The ANOVA and Spectral Correlation Mapper (SCM) Methods, 2001.

[6] Carvalho Jr.,O. A., Guimares R. F., Carvalho A. P., Silva, N. C. Martins, E.S., Gomes, R. A. T. , Vegetation mapping in the Parque Nacional, Brasilia (Brazil) area using advanced spaceborne thermal emission and reflection radiometer (ASTER) data and spectral identification method (SIM), in: Remote Sensing for Environmental Monitoring, GIS Applications, and Geology V. Edited by Ehlers, Manfred, Michel, Ulrich, Proceedings of the SPIE, **5983**, pp. 45–55, 2005.

[7] Davis, 1973, Statistic and analysis in geology, New York, John Willey & Sons, Inc., 1973.

[8] Souza, G. S., Intruducao aos Modelos de Regressao Linear e Nao-Linear, Brasilia: EMBRAPA-SPI/EMBRAPA-SEA, 1998.

[9] Steel, R. G. D., Torrie, J. K., Principle and procedure of statistics, second edition, New York, MacGraw-Hill, 1980.

[10] Vieira, S., Introducao a BioStatistica. 5th ed. Rio de Janeiro: Campus, 293, 1988.

[11] Clark, R. N. and Swayze, G. A., Mapping Minerals, Amorphous Materials, Environmental Materials, Vegetation, water, Ice and Snow, and Other Materials: The USGS Tricorder Algorithm, In: Summaries of the Fifth JPL Airborne Earth Science Workshop, JPL Publication **95-1, v.1**, p.39–40.

[12] Clark, R. N. and Swayze, G. A. King, T. V. V., Imaging Spectroscopy: A Tool for Earth and Planetary System Science Remote Sensing with the USGS Tetracorder Algorithm, Journal of Geophysical Research, 2001.

[13] Chavez P.S.J., Guptill, C., and Bowel,J.A.,(1984)Image rocessing techniques for thematic mapper data, in proceeding of the American Society of photogrametry Conference, 728–752. Washington, 1984.

[14] Jensen, J., Introductory Digital Image Processing. Prentice Hall, New Jersey, pp.97–100, 1986.

[15] Kvasov, B. I., Methods of shape-preserving spline approximation. World Scientific, 2000.