A Dimensionality Reduction Method for THz-TDS Signals via the Recursive Projective Splits Based on PCA

Weixin Xie[1,2*], Jihong Pei[1], and Jing Li[1,2]
[1]Intelligent Information Institute, Shenzhen University, Shenzhen, 518060, China
[2]School of Electronic Engineering, Xidian University, Xi'an, 710071, China
*E-mail: wxxie@szu.edu.cn

Abstract: Terahertz pulsed imaging delivers THz-TDS signals of a high dimensionality, which raises the difficulties and computations of high dimensional data process. Inspired by the applications of the projective split in "space time" physics, we apply the projective splits on THz-TDS signals and develop a new dimensionality reduction method for THz-TDS signals. In this method, THz-TDS signals are represented as vectors in a vector space of high dimension. By addition and multiplication, the vector space generates a geometric (or Clifford) algebra of the same dimension. A projective split can factorize the geometric algebra of high dimension into the geometric algebras of lower dimension. Thus, vectors of THz signals in the vector space of high dimension can similarly relate to vectors in the vector space of lower dimension. The projective splits are recursively employed and linearly map the vector space of high dimension into a sequence of sub-spaces step by step. In each step, the Principle Component Analysis (PCA) which explores statistical inherence is performed on vectors in each sub-space, and the homogenous vector of the projective split is determined by the eigenvector of the maximum principal component of PCA. In the vector space of lower dimension, as vectors related to THz-TDS signals from different substances are distant from each other, the application of substance classification and substance identification based on the relative THz-TDS signals can be easily worked out. Experiments are presented and the performance of the method is demonstrated.

Keywords: THz-TDS, geometric algebra, projective splits, dimensionality reduction

1. Introduction

THz time-domain spectroscopy (THz-TDS) is one of the important detecting techniques in the THz research field, by which the time-dependent electric field of THz signals are measured [1-3]. Each THz-TDS signal records a temporal response of the THz reference pulse and can be represented by a pulse waveform, from which the THz spectrum is obtained by the Fourier transform. The technique of terahertz spectroscopy is very attractive because many materials have an absorption band, which is called a fingerprint. Typically, substances are identified based on those waveforms, focusing on the waveform properties, such as peaks, slopes and shifts [4-9]. One THz-TDS signal usually contains 512 samples or more, which is necessary to be processed properly and effectively using the signal processing techniques especially for high dimensional signals. Recently, the component spatial pattern analysis by Fukunaga, etc., [10] is proposed to identify substances based on the THz spectrum at seven specified frequencies. However, how to choose the featured frequencies still needs further studies.

In the present work, THz-TDS signals are represented as vectors in a high-dimensional vector space. In our previous papers[11,12], vectors of THz-TDS signals are analyzed using the language of geometric algebra, an unified mathematical language based on geometric (or Clifford) algebra for physics and engineering. The analysis shows the projective property exists in THz-TDS signals originated from substances with the same thickness. As proposed by Hestenes [13-15], relations among geometric algebras of different dimensions can then be interpreted geometrically as "projective splits", and vector spaces of different dimensions can also be related in an algebraic coordinate-free form. Accordingly, we propose a novel dimensionality reduction for THz-TDS signals via the projective splits. In the method, to differentiate substances based on their statistical inherence, the projective splits are determined by the eigenvector of the maximum principal component, resulting from the principle component analysis (PCA) performed on THz-TDS signals. Employing the projective splits recursively, the dimensionality of the original vector space is reduced. Substances are classified and identified using general methods with the vectors in the resulting vector space which is of much lower dimension. The magnitude of the outer product of two vectors is used to measure their distance. Experiments which demonstrate the feasibility of the method are performed.

2. Basis of the Geometric Algebra [13-15]

Let v_n be an *n*-dimensional vector space over the reals R . The geometric algebra is $g_n = g(v_n)$ generated from v_n by defining the geometric product for all vectors. For vectors **a** and **b** in $g_n = g(v_n)$, the geometric product **ab** can be decomposed as:

$$\boldsymbol{ab} = \boldsymbol{a} \bullet \boldsymbol{b} + \boldsymbol{a} \wedge \boldsymbol{b} \tag{1}$$

where the inner product $\boldsymbol{a}\bullet\boldsymbol{b}$ is defined as $\boldsymbol{a} \bullet \boldsymbol{b} = (\boldsymbol{ab} + \boldsymbol{ba}) / 2$, and the outer product $\boldsymbol{a} \wedge \boldsymbol{b}$ is defined as $\boldsymbol{a} \wedge \boldsymbol{b} = (\boldsymbol{ab} - \boldsymbol{ba}) / 2$. The inner product is scalar-valued in consequence of the contraction axiom.

The projective geometry within geometric algebra is represented by adopting the standard identification of points in projective space P_{n-1}, or as vectors in v_{n-1} , with rays in the vector space v_n . Let $\mathbf{x}$ and $\mathbf{e}_0$ be vectors in $\boldsymbol{v}_n$, then, for fixed $\mathbf{e}_0$ with $\mathbf{e}_0{}^2 \neq 0$, the function $\mathbf{x} \wedge \mathbf{e}_0$ is a linear mapping of v_n into v_{n-1} . The projective mapping relating each ray $\{\lambda\mathbf{x}\}$ in v_n to a unique vector x' in v_{n-1} is defined by the following relation:

$$\mathbf{xe}_0 = \mathbf{x} \bullet \mathbf{e}_0 + \mathbf{x} \wedge \mathbf{e}_0 = \mathbf{x}_0 (1 + \mathbf{x}') \tag{2}$$

where $\mathbf{x}_0 = \mathbf{x} \bullet \mathbf{e}_0 \in R$, obviously $\mathbf{x}' = \mathbf{x} \wedge \mathbf{e}_0 / \mathbf{x} \bullet \mathbf{e}_0 \in v_{n-1}$. The vector $\mathbf{x}$ in v_n is relative to $\mathbf{e}_0$ amounts to a representation of the "point" $\mathbf{x}'$ in v_{n-1} by "homogenous coordinates".

3. Analysis of THz-TDS Signals

3.1 Signals in the THz-TDS Transmission System

In the typical THz-TDS system, we assume that the sample under a transmission-mode measurement has parallel and polished surfaces, and the angle of incidence of the incoming T-ray beam is normal to the surfaces. The raw time-domain electric field data from the sample $E_s(t)$ and those of the reference $E_r(t)$ are measured. Their Fourier transforms are $\tilde{E}_s(f)$ and $\tilde{E}_r(f)$ respectively, where the superscript $\sim$ denotes complex number and f is frequency. Suppose N is the number of frequencies sampled in the THz band, then the transfer function of the sample in discrete form is [16,17]:

$$\tilde{T}_i = \tilde{E}_s(f_i)/\tilde{E}_r(f_i) \quad (3)$$

The complex refractive index $\tilde{\mathbf{n}}$ is intrinsic for materials in optics and spectroscopy. It is frequency-dependent and is a complex number for every frequency sampled in the THz band. It can be inferred from [16,17] that two major variables, the complex refractive index $\tilde{\mathbf{n}}$ and the thickness *l* of the sample, govern the transfer function of the THz-TDS system.

3.2 The Projective Property of the THz-TDS Signals

In this paper, we would only give a brief analysis on $\ln\left|\tilde{T}_i\right|$ and only deal with the geometric algebra with positive signatures. More detailed analysis can be referred to our papers[11,12]. Let v_N be an *N*-dimensional vector space over real numbers with an orthonormal basis $\{\mathbf{e}_1, \mathbf{e}_2, ...\mathbf{e}_N\}$, and denote its corresponding geometric algebra by $g(v_N)$.

Then the scalar-valued real 1D THz spectrum signal with *N* frequencies sampled, $s = [s_i]_{1\times N} = [ln\left|\tilde{T}_i\right|]_{1\times N}$ can be embedded into an *N*-dimensional vector space v_n as a real-valued vector as $s = \sum_{i=1}^{N}(ln\left|\tilde{T}_i\right|)\mathbf{e}_i$, corresponding to the THz-TDS measurement. Peculiar to the Euclidean axiom depend on the fact that the square of a vector is a positive scalar, it is the consequence in [11,12] that:

$$\forall i = 1,2,....N, \hat{n}_i^A = \hat{n}_i^B \Leftrightarrow s_l^A \wedge s_l^B = 0 \Leftrightarrow \left|s_l^A \wedge s_l^B\right| = 0 \quad (4)$$

Therefore, the projective property exists for vectors of THz-TDS signals. That is, for vectors s_l^A and s_l^B, representing different THz signals, can be corresponding to the samples of the same substance, if and only if $s_l^A \wedge s_l^B = 0$, or equivalently $s_l^A \approx s_l^B$ in the projective geometry. The tangential distance is defined as $d(s^A, s^B) = \left|s_l^A \wedge s_l^B\right| / \left|s_l^A\right|\left|s_l^B\right|$, to measure the difference of the THz-TDS signals.

4. Linear Dimensionality Reduction via Projective Splits Based on PCA

As the projective split idea first explicitly formulated and applied to physics in[14], vectors of THz signals in the *n*-dimensional vector space can similarly relate to vectors in the (*n*-1)-dimensional vector space. We first analyze the dimensionality reduction of THz-TDS signals via the projective splits, and then present the method in detail.

4.1 Projective Splits of THz-TDS Signals

Because vectors of THz-TDS signals have the projective property, for vectors representing different THz signals through the samples of the same thickness but of distinct materials, they could be identified as different points in the projective geometry via the projective splits. On the contrary, for vectors of THz signals through samples of the same material and the same thickness, they would be relative to the unique point via the projective splits. Therefore, it is possible for THz signals to be decomposed via the projective splits while the discriminatory information is still remained in the vector space of one less dimension.

Given $\boldsymbol{\mu}_n \in v_n$, a new basis $\{\mathbf{u}_1, \mathbf{u}_2, \cdots \mathbf{u}_{n-1}, \mathbf{u}_n\}$ can be constructed for v_n. For an arbitrary vector $s^{(n)} \in v_n$ corresponding to a THz-TDS signal, it can be expressed as $s^{(n)} = \sum_{i=1}^{n} s_i^{(n)} \mathbf{e}_i^{(n)} = \sum_{i=1}^{n-1} a_i^{(n)} \mathbf{u}_i^{(n)} + a_n^{(n)} \boldsymbol{\mu}_n$. Then it can be projected down to the unique vector $s^{(n-1)}$ in v_{n-1} with respect to $\boldsymbol{\mu}_n$ as:

$$s^{(n-1)} = s^{(n)} \wedge \boldsymbol{\mu}_n / s^{(n)} \bullet \boldsymbol{\mu}_n = \sum_{i=1}^{n} s_i^{(n)} e_i^{(n)} \wedge \boldsymbol{\mu}_n / s_0 = \sum_{i=1}^{n-1} \frac{a_i^{(n)}}{s_0} \mathbf{u}_i^{(n)} \wedge \boldsymbol{\mu}_n = \sum_{i=1}^{n-1} s_i^{(n-1)} \mathbf{e}_i^{(n-1)} \qquad (5)$$

where $s_0 = s^{(n)} . \boldsymbol{\mu}_n \in R$ and the homogeneous coordinates $s_i^{(n-1)} = a_i^{(n)} / s_0, i = 1, 2, \cdots n-1$.

Furthermore, each THz-TDS signal can be regarded as a vector in v_n. It can be projected down to the corresponding space v_{n-1} via the projective split with respect to the vector μ_n. Recursively using this dimensionality reduction, it can be mapped into a much lower dimension space. That is, given a sequence of vectors $\boldsymbol{\mu}_n, \boldsymbol{\mu}_{n-1}, \cdots, \boldsymbol{\mu}_2$, the dimensionality of an *n*-dimensional vector $s^{(n)} \in v_n$ can be reduced recursively via the projective splits as below:

$$s^{(n)} \xrightarrow{\mu_n} \frac{s^{(n-1)}}{s^{(n)} . \boldsymbol{\mu}_n} \xrightarrow{\mu_{n-1}} \frac{s^{(n-2)}}{s^{(n)} . \boldsymbol{\mu}_{n-1}} \cdots \xrightarrow{\mu_3} \frac{s^{(2)}}{s^{(3)} . \boldsymbol{\mu}_3} \xrightarrow{\mu_2} \frac{s^{(1)}}{s^{(2)} . \boldsymbol{\mu}_2}$$

where, $s^{(k)} = \sum_{i=1}^{k} s_i^{(k)} \mathrm{e}_i^{(k)} = \sum_{i=1}^{k} a_i^{(k+1)} \mathrm{e}_i^{(k)} / (s^{(k+1)} \bullet \mu_{k+1}), k = 1, 2 \cdots, n-1.$

Such that, the projective splits are employed recursively on vectors of THz-TDS signals and linearly maps the vector space of high dimension into a sequence of sub-spaces step by step. During each step, the projective property could still exists, based on the assumption that none distinct relative points during the decompose process could be linearly related to each other. Therefore by employing the projective splits recursively, THz signals were dimensionality reduced and their projective properties are remained in the relative points in the final subspace. That also means THz signals can be classified and identified using general methods on the basis of vectors in the resulting vector space which is of much lower dimension.

4.2 PCA of THz-TDS Signals

Meanwhile, how to choose μ_n in each project split becomes the key point. To explore statistical inherence in THz-TDS signals, we perform principal components analysis (PCA) on the data here.

PCA can be used to find linear combinations of the variables called principal components in a multivariate data set corresponding to orthogonal directions maximizing variance in the data [3]. Linear transforms are useful both for noise extraction and for representing the information in the data in a lower dimensional space (that is, using fewer coefficients). Linear transformations commonly used for the processing of spectroscopic data are variable selection, Fourier transform, windowed Fourier transform, wavelet transform, and principal-component analysis (PCA) (or Karhunen–Loève transform). The transforms, such as the variable selection and the wavelet transform, embody an a priori assumption about the shape of the functions used to generate basis vectors [18, 19]. The Karhunen–Loève transform, or PCA, employs basis vectors that are built from the statistical properties of the data set to be analyzed.

If M THz signals containing N frequencies sampled each are gathered in an $M \times N$ matrix $S = [s_1, s_{2,} \cdots, s_M]^T = [s_{ki}]_{M\times N}$ with the mean vector $\mathbf{m}_s$ and the covariance matrix C_S, it is always possible to find a set of N normal eigenvectors as C_S is real and symmetric. Let $\mathbf{u}_i$ and λ_i, i=1,2,…, be the eigenvectors and the corresponding eigenvalues of C_S, arranged in descending order. Let U be a matrix whose columns are formed from the eigenvectors of C_S, ordered so that the first column of C_S is the eigenvector corresponding to the largest eigenvalue, and the last column corresponding to the smallest one respectively. Then using the Hotelling transform, the matrix S can be mapped into S' as:

$$S' = (S - \boldsymbol{m}_S)\mathbf{U} \tag{6}$$

where $\mathbf{U}$ is an unitary matrix ($\mathbf{U}^T = \mathbf{U}^{-1}$, and T signifies the transpose).

The mean of S''is zero, and the covariance matrix is $C_{S'} = \Lambda$, which is a diagonal matrix whose elements along the main diagonal are the eigenvalues of C_S. So the elements of columns of S' are uncorrelated and the matrix C_S and $C_{S'}$ have the same eigenvalues. The matrix S can be recovered correspondingly by:

$$S = S'\mathbf{U}^T + \boldsymbol{m}_S \tag{7}$$

PCA can explore statistical inherence in THz-TDS signals. THz signals themselves do not have a natural geometric structure, but only a high dimensional implicit representation. Hence in this case, PCA can be seen as a way of inferring a low dimensional explicit geometric feature space that best captures the structure of the data. The analysis vectors (principle components, PCs) are then seen to be columns of U, which are defined as the directions of maximum variance in the data. In this way, the information in the signals is maximally compressed into the transform coefficients. It is worth noting that higher PC coefficients may actually contain the information that is required to differentiate between classes. Naes and Mevik[20] showed that in some situations the discriminatory information may actually be contained in PCs with a small variance. In this case, the relevant information for the classification task would be contained in directions with smaller variance. Therefore, the eigenvector of the maximum principal component is chosen as the “splitting vector” $\boldsymbol{\mu}_n$ in the method.

More generally, PCA can be used to obtain the regression estimate $\hat{S}$ by replacing U with $\bar{U}_i$, where the bar denotes the reduced matrix composed of the *i* eigenvectors corresponding to the *i* largest eigenvalues. Then the signals reconstructed by using $\bar{U}_i$ is:

$$\hat{S} = S'\bar{\mathbf{U}}_i^T + \boldsymbol{m}_S \tag{8}$$

and the mean square error is $e_{ms} = \sum_{j=i+1}^{N} \lambda_j$.

Thus, PCA is optimal in the sense that it minimizes the mean square error between the original signals and their approximations. Signals may in fact lie in a lower dimensional subspace even if no individual feature is constant, which corresponds to the subspace not being aligned with any of the axes. The PCA is nonetheless able to detect such a subspace. If the eigenvalues beyond the *i*-th are small we can regard the data as being approximately *i*-dimensional, which means that the features beyond the *i*-th are approximately constant and the data has little variance in these directions. In such cases it can make sense to project the data into the space spanned by the first *i* eigenvectors. It is possible that the variance in the dimensions we have removed is actually the result of noise, so that their removal can improve the representation of the data in some cases. Hence, performing PCA can also be regarded as denosing. We set the threshold of the eigenvalue, *MINLATENT*, as in our experiments and we show that it is feasible.

4.3 Dimensionality Reduction of THz-TDS signals via the Projective Splits based on PCA

Based on the analysis in section 4.1, the dimensionality of THz signals can be reduced recursively via the projective splits. The projective splits are employed recursively and linearly map the vector space of high dimension into a sequence of sub-spaces step by step. In each step, the principle component analysis (PCA) which explores statistical inherence is performed on vectors in each sub-space, and the homogenous vector of the projective split is determined by the eigenvector of the maximum principal component of PCA. Details of the method are presented in the following.

For THz-TDS signals, $S^{(N)} = [s_1^{(N)}; s_2^{(N)}; \cdots, s_M^{(N)}]$, where *M* is the number of THz-TDS signals and *N* is the number of frequencies sampled, the dimensionality of THz-TDS signals can be recursively reduced to a desired dimension *mdim* using the following method via the projective splits:

Step1. Let $n = N, MINLATENET = 10^{-3}$, and initialize $s_k^{(n)} = \sum_{i=1}^{n} s_{ki}^{(n)} \mathbf{e}_i^{(n)} \in v_n, k = 1, 2, \cdots M$ for each THz-TDS signals;

Step2. Perform PCA analysis on $[s_{ki}^{(n)}]_{M \times N}$ and obtain the eigenvalues in decreasing order and the corresponding eigenvectors;

Step3. Set the eigenvector of the maximum principal component as $\boldsymbol{\mu}_{n,}$ and construct the new coordinates $\{\mathbf{u}_1, \mathbf{u}_2, \cdots \mathbf{u}_{n-1}, \boldsymbol{\mu}_n\}$, obtain $s_k^{(n)} = \sum_{i=1}^{n-1} a_{ki}^{(n-1)} \mathbf{u}_i^{(n)} + a_{kn}^{(n)} \boldsymbol{\mu}_n$;

Step4. Map each $s_k^{(n)} \in v_n$ into $s_k^{(n-1)} \in v_{n-1}$ via the projective split using $s_k^{(n-1)} = \sum_{i=1}^{n-1} s_k^{(n-1)} \mathbf{e}_i^{(n-1)} = \sum_{i=1}^{n} a_k^{(n)} \mathbf{e}_i^{(n-1)} / (s_k^{(n)} \bullet \boldsymbol{\mu}_n)$, and obtain the homogeneous coordinates $[s_{ki}^{(n-1)}]_{M\times(n-1)}$;

Step5. Compute $d_{jl} = \left| s_j^{(n-1)} \wedge s_l^{(n-1)} \right| / \left(\left| s_j^{(n-1)} \right| \left| s_l^{(n-1)} \right| \right)$ for $j, l = 1, 2, ..., M$, as the measurement of distance between two vectors in v_{n-1} ;

Step6. Perform PCA analysis on $[s_{ki}^{(n-1)}]_{M\times(n-1)}$. If the eigenvalues beyond the *i*-th are smaller than *MINLATENT*, then set *n*=max(*i*, *mdim*) and remove the columns beyond the *i*-th column; elsewise set *n*=*n*-1;

Step7. Continue steps 3 to 6, until *n* is equal to *mdim* and then END.

5. Experiments on THz-TDS Signals

Using the method presented here, vectors representing THz-TDS signals in high dimensional vector space can be mapped into vectors in a lower vector space. To demonstrate the method can be potentially useful in the identification and classification of the materials, we perform experiments on the experimental THz-TDS signals and THz data from the free database of RIKEN[21]. In these following experiments, the dimensionality reduction method is applied with *mdim* =3.

5.1 Results of Dimensionality Reduced THz-TDS Signals

We obtain eight THz-TDS signals and also their related reference signals from five distinct samples using our THz-TDS system in the transmission model. The materials are marked as CARD, OSA, PTFE, RB, and WB. Fig. 1 (a) shows the amplitudes of the transfer functions for these eight pairs of signals. These waveforms are quite noisy and fluctuated. It is difficult to classify three materials, which are WB, RB and OSA, from each other as their waveforms are quite similar.

The frequency range considered is 0.2~0.6 *THz*, and the sampling number in each signals is 32. Hence the data size of these THz-TDS signals is 8×32. Using the method, the signals are decomposed via the projective splits and their dimensionality is reduced from 32 to 31, and then to three. As shown in Fig. 1 (b), signals are clustered and signals from the same material are closed to each other. It verifies the feasibility of the method in the dimensionality reduction of THz-TDS signals. Fig. 1 (c) shows the results of dimensionality reduction of signals in RGB color bars, which also shows the potential of the method in the visualization of the THz-TDS signals.

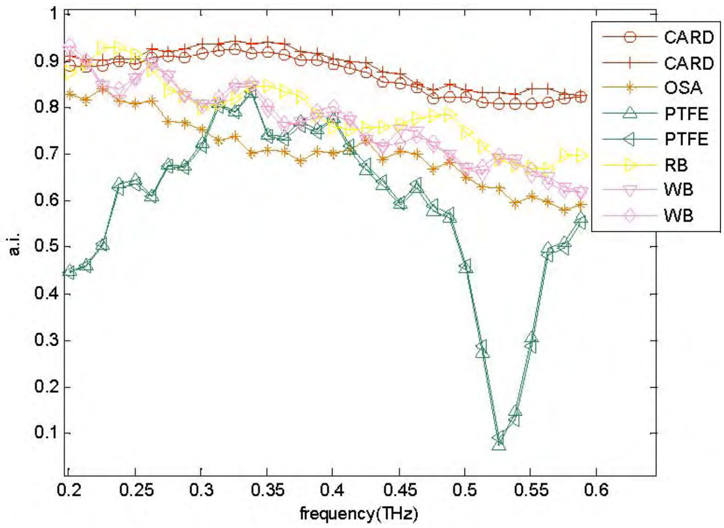
The dimensionality reduction method is also applied to the THz data from the RIKEN database. Fig. 2.(a) shows five different raw THz signals in the RIKEN database with file names labeled as

“PDC206”, “PDC207”, “PDC208”, “PDC211”, “PDC212”. These waveforms are quite similar and their peak values are also very close to each other. The signals are decomposed via the projective splits and their dimensionality is reduced from 1058 to 1057, and then to three. Fig.2 (b) plots out the points in the three-dimensional vector space, which are the processed results of the dimensionality reduction method. And Fig.2 (c) shows the results in RGB color bars. From these results, it is shown that THz-TDS signals originated from different samples can be easily classified. Especially for signals in Fig.2 (a), which could hardly be differentiate from the waveform peaks, it is obvious that points in the 3-dimensional vector space related to these raw signals distribute distantly from each other.

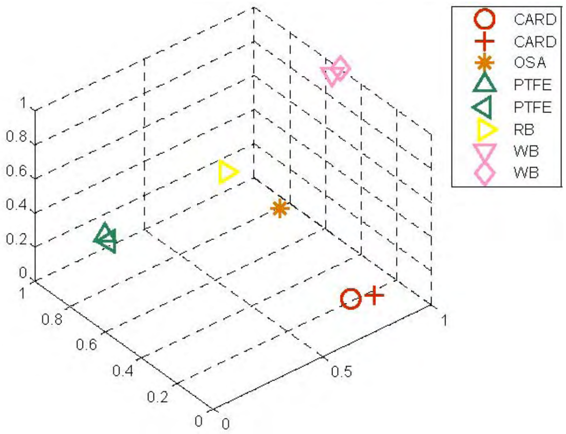
5.2 Substances Identification on basis of Their THz-TDS Signals

Using the method, vectors representing THz signals in high dimensional vector space can be mapped into vectors in a lower vector space. To demonstrate the method can be potentially useful in the materials’ classification, we apply the dimensionality reduction method on the THz signals, and then, in the final low dimensional vector space, magnitudes of the outer product (the tangential distance [11,12]) of the related vectors are calculated, and those materials whose related vectors are the closest to each other are identified as the same.

Raw signals in Fig.1 (a) and Fig.2 (a) are grouped as Group1 and Group2 respectively. Four simulations are performed in each group adding Gaussian white noises to signals with different parameters (SNR>10dB). Each simulation is run 50 times and 50 noisy signals are obtained for each raw signal. The dimensionality reduction method is performed on these noisy signals. Table 1 shows the accurate rate of identification over 50 runs for each simulation. The accurate rates of identification in simulations for both Group1 and Group2 are all 100%. Simulation experiments show the high performance of the method on the identification of substances on the basis of their THz signals.



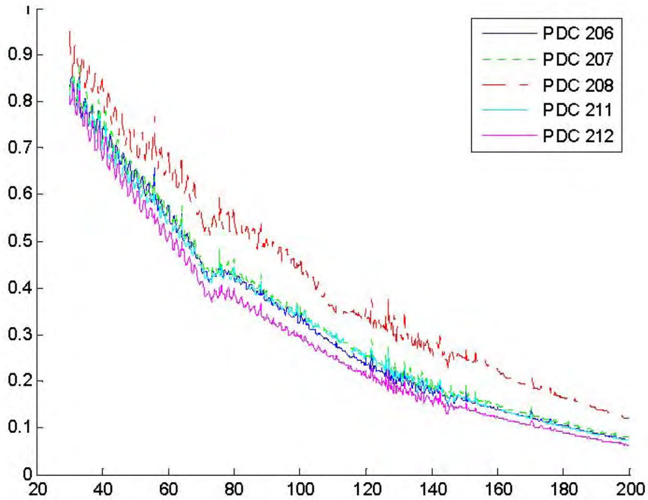
(a)



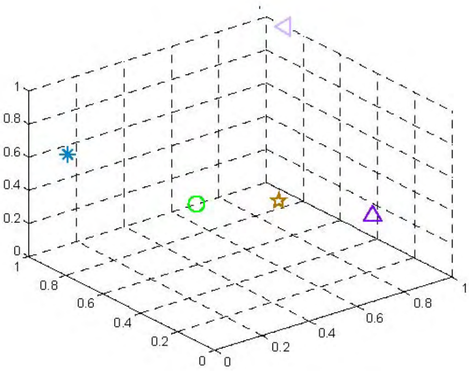
(b)

(c)

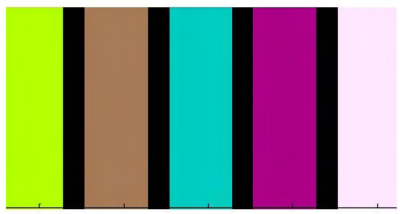
Fig.1 Results of Dimensionality Reduced experimental THz-TDS Signals



(a)



(b)



(c)

Fig.2 Results of Dimensionality Reduced experimental THz-TDS Signals

		Simulation 1	Simulation 2	Simulation 3	Simulation 4
Parameters of noise added	Mean* (%)	10	5	1	1
	Variance σ^2	10^{-6}	10^{-6}	10^{-6}	10^{-4}
Accurate rate of identification over 50 runs for each group (%)	Group1	100	100	100	100
	Group2	100	100	100	100

*the mean is some percent of the maximum value of the differences of the raw signals' amplitudes

Tab. 1 The accurate rate of identification

6. Conclusions

Terahertz pulsed imaging delivers THz-TDS signals of a high dimensionality. The aim of this paper is to study properties of THz-TDS signal using Geometric algebra and develop new tools and techniques for THz-TDS signals. The analysis of THz-TDS signals using Geometric algebra shows vectors of THz signals have the projective property. Inspired by the applications of the projective split in "space time" physics, we apply the projective splits on THz-TDS signals and develop a dimensionality reduction method. Using the method, THz-TDS signals can be linearly mapped into a space of lower dimension. Experiments demonstrate the feasibility and efficiency of our method on the applications of identification and differentiation of the substances with their THz transmission spectra.

Acknowledgement

This work is supported by National Natural Science Funds of China (No. 60672153), Doctoral Subject Funds of China (No. 20060590001), and Science and Technique Project of Shenzhen (China). The THz data we used in this paper is from the open THz database of the Tera-phtonics Laboratory, RIKEN Sendai.

References

[1] D. M. Mittleman, S. Hunsche, etc., "T-ray tomography", *Optics letters*, 22(12), 904-906, (1997).

[2] W. X. Xie, J. H. Pei, "Review of Terahertz signal processing and analysis", *ACTA Electronica sinica*, 35(10) , 1973-1978, (2007).

[3] J. H. Pei, Y. Hu, W. X. Xie, "PCA-based visualization of terahertz time-domain spectroscopy image", *Proceedings of the SPIE*, 6787, 67871M1-7, (2007).

[4] Naftaly, M.; Miles, R.E., "Terahertz Time-Domain spectroscopy for material characterization", *Proceedings of*

the IEEE, 95(8), 1658 - 1665, (2007).

[5] D. Dragoman, M. Dragoman, "Time-frequency signal processing of terahertz pulses", *Applied optics*, 43(19), 3848-3853, (2004).

[6] T. D. Dorney, R.G. Baraniuk, D. M. Mittleman, "Material parameter estimation with terahertz time-domain spectroscopy", *J. Opt. Soc. Am. A*, 18(7), 1562-1571, (2001).

[7] S. Gorenflo, U. Tauer, etc., "Dielectric properties of oil-water complexes using terahertz transmission spectroscopy", *Chemical physics letter*s, 421, 494-498, (2006).

[8] H. B. Liu, X. C. Zhang, "Dehydration kinetics of D-glucose monohydrate studied using THz time-domain spectroscopy", *Chemical Physics Letters*, 429, 229–233, (2006).

[9] Pupeza, R. Wilk, M. Koch, "Highly accurate optical material parameter determination with THz time-domain spectroscopy", *Optics express*, 15(7), 4335-4350, (2007).

[10] K. Fukunaga, Y. Ogawa, etc, "Application of terahertz spectroscopy for character recognition in a medieval manuscript", *IEICE Electronics Express*, 5(7), 223-228, (2008).

[11] W. X. Xie, J. Li, J. H. Pei, "A Clifford Algebra Analysis Method of THz-TDS Images", *9th International Conference on Signal Processing, ICSP 2008, Beijing, China,* (2008).

[12] W. X. Xie, J. Li, J. H. Pei, "An Analysis of THz-TDS Signals using Geometric Algebra", *Photonics and Optoelectronics Meetings, POEM 2008, Wuhan, China,* (2008).

[13] D. Hestenes, "Universal Geometric Algebra", *A Quarterly Journal of Pure and Applied Mathematic*, *Simon Stevin*, 62, 253-274, (1988).

[14] D. Hestenes, "Spacetime algebra", *Gordon and Breach, New York*, (1966), (1987).

[15] D. Hestenes, "The design of linear algebra and geometry", *Acta Applicandae Mathematicae*, 23, 65-93, (1991).

[16] W. Withawat, B.M. Fischer, D. Abbott, "Material thickness optimization for transmission-mode terahertz time-domain spectroscopy", *Optics express*, 16(10), 7382-7396, (2008).

[17] J.Z. XU, X.C. Zhang, "Tarahertz Science Technology and Application", *Peking University* Press, (2007).

[18] J.H. Peị, P.L. Ye and W.X. Xie, "Optimal wavelet analysis for THz-TDS pulse signals", *Photonics and Optoelectronics Meetings*, POEM 2008, Wuhan, China, (2008).

[19] Sillas Hadjiloucas, Roberto K. H. Galvão, John W. Bowen, "Analysis of spectroscopic measurements of leaf water content at terahertz frequencies using linear transforms", *Opt. Soc. Am. A*, 19(12), 2495-2509, (2002).

[20] T. Naes, B. H. Mevik, "Understanding the collinearity problem in regression and discriminate analysis", *J. Chemom*, 15, 413-426, (2001).

[21] THz database Web : http://www.riken.jp/THzdatabase/ (Tera-phtonics Laboratory, RIKEN Sendai, date of access)