Robust identification of concealed dangerous substances using THz imaging spectroscopy

Helle E. Nystad, Magnus W. Haakestad, and Arthur D. van Rheenen
Norwegian Defence Research Establishment (FFI), P. O. Box 25, NO-2027 Kjeller, Norway

ABSTRACT

False alarm rates must be kept sufficiently low if a method to detect and identify objects or substances is to be implemented in real life applications. This is also true when trying to detect and identify dangerous substances such as explosives and drugs that are concealed in packaging materials. THz technology may be suited to detect these substances, especially when imaging and spectroscopy are combined. To achieve reasonable throughput, the detection and identification process must be automated and this implies reliance on algorithms to perform this task, rather than human beings. The identification part of the algorithm must compare spectral features of the unknown substance with those in a library of features and determining the distance, in some sense, between these features. If the distance is less than some defined threshold a match is declared. In this paper we consider two types of spectral characteristic that are derived from measured time-domain signals measured in the THz regime: the absorbance and its derivative. Also, we consider two schemes to measure the distance between the unknown and library characteristics: Spectral Angle Mapping (SAM) and Principal Component Analysis (PCA). Finally, the effect of windowing of the measured time-domain signal on the performance of the algorithms is studied, by varying the Blackman-Harris (B-H) window width. Algorithm performance is quantified by studying the receiver-operating characteristics (ROC). For the data considered in this study we conclude that the best performance is obtained when the derivative of the absorbance is used in combination with a narrow B-H window and SAM. SAM is a more straight-forward method and requires no large training data sets and tweaking.

Keywords: THz (differential) spectroscopy, principal component analysis, spectral angle mapping, receiver-operating characteristics

1. INTRODUCTION

Apart from a few very specialized applications, such as high-power laser-induced plasma sources and accelerator-based sources, THz sources tend to be relatively low power: tens of μW to may be tens of mW. Because of strong absorption by water molecules, the range of this low-power radiation is limited to order-of-magnitude of 1 m. Initial hopes of applying the new technology to, for instance, route clearing during military operations were quickly dashed. Instead, attention was directed towards shorter range applications such as scanning persons and packages for dangerous and/or illegal substances, such as explosives and drugs. Common for these substances is that they have spectral fingerprints in the 0.1 – 10 THz range, which many of the THz sources and detectors cover. An added advantage of the technology in these types of applications is the transparency of many common packaging materials, such as plastic, cardboard, and cloth, in this frequency range. This makes it possible to detect and identify concealed materials. The development of an imaging capability makes it possible to visualize hidden object, shapes with different optical properties than their background, helping to identify suspicious objects, such as weapons. By combining the spectroscopic and the imaging capability detection of suspicious objects and subsequent substance identification seems in reach. However, there are a number of obstacles that have to be overcome before a reliable “identifier” is realized. Some of these difficulties, which are of course widely known, are listed in [1]. We have already mentioned the strong absorption of THz radiation by water vapor, cluttering the measured spectrum with its absorption lines for lower humidity values and possibly obscuring spectral fingerprints, absorption lines that widen and deepen for higher humidity values rendering detection and identification impossible. A substance’s surface roughness causes scattering of the THz radiation resulting in a low-pass filtering of the received THz signal, effectively limiting the bandwidth and hence the capability to identify spectral features at higher frequencies. Materials that conceal the substance of interest may in addition to increased scattering due to roughness of its surface also add spectral features, for instance when the weave of a piece of cloth has a structure size comparable to the THz wavelength between 0.03 and 3 mm. The added features clutter the spectral signature and make

identification more difficult. Except for postal scanners, which may rely on measuring in transmission mode, most applications must rely on reflection-mode measurements where "misalignment" between the THz beam and the object of interest may result in diffuse rather than specular reflection, resulting in a much weaker received THz signal, reducing the signal-to-noise ratio significantly. A reliable or robust identification scheme must be able to cope with widely varying measurement conditions. In addition, in general, a measure of specificity is required: ideally, the scheme should be able to distinguish between different substances with similar spectral features.

Obviously, all these issues have been realized before and there is a vast body of literature [2-17] on identifying substances by comparing their spectra. Spectral features have been measured in numerous wavelength bands corresponding to the energies of the transitions of interest. The purpose of this work is to compare different schemes for comparing spectra and find some objective measure to rank them. Equally importantly, we are interested in finding what the limitations are of the different schemes.

In general a scheme consists of transforming the measured raw data, a time-domain THz signal in our case into a spectral characteristic, for instance the absorbance. Then this spectral characteristic has to be compared to known spectral characteristics. This comparison requires a measure of similarity or distance and a threshold so that a match or not-a-match may be declared. We have chosen to study the effect of three aspects of this general scheme: (i) windowing of the time-domain data before Fourier transforming them, (ii) using the derivative of the absorbance (dA/df) rather than the absorbance spectrum (A) itself, and (iii) comparing the performance of Spectral Angle Mapping (SAM) and Principal Component Analysis (PCA) to classify the spectral characteristic. To be able to compare the performance of the different schemes we use ROC, where true-positive rates are plotted as a function of false-positive rates for different detection decision thresholds.

In the next five sections the analysis considerations are detailed. Then the experiment is presented, followed by a discussion of the results from our analyses.

2. WINDOWING

Before applying a Fourier transform of the time-domain signal preprocessing may be applied in the form of windowing. The multiplication of the time-domain signal by a window function is equivalent with sliding a window function over the Fourier transformed signal in the frequency domain. This is effectively a smoothing operation. We use a Blackman-Harris window whose window half-width was varied, from 8 to 30 ps. A narrower window implies stronger smoothing, reducing the spectral resolution. A wider window gives a noisier spectrum, possibly a disadvantage when differentiating the absorbance.

3. CHOICE OF SPECTRAL CHARACTERISTIC

Scattering of the THz signal either by the rough surface of the substance to be detected or by the material concealing it adds frequency dependence to the absorption spectrum, effectively reducing the signal bandwidth but also possibly hampering matching the spectrum to a library spectrum by introducing a background absorption in addition to the absorption lines. However, this added frequency dependence is much weaker (polynomial) than that of an absorption line. Therein lays the opportunity to distinguish the background from the signal: by taking the derivative. Therefore, we propose to study the derivative of the absorbance, dA/df. Differential spectroscopy is not new, and has also been applied to THz spectra [1, 4, 5]. Broader absorption features may be difficult to locate on the frequency axis. The differentiated spectrum has a zero crossing which is more sharply located on the axis, providing more specificity. These considerations have been presented earlier [1].

4. SPECTRAL ANGLE MAPPING (SAM)

The simplest and most intuitive method for comparing spectra is SAM. A measured spectrum is viewed as a vector and this vector's dot-product with library vectors (spectra) is calculated and then normalized by the lengths of the two vectors. In fact, one calculates the cosine of the angle between the vectors. Identical vectors point in the same direction,

the angle between them is zero and the cosine equals 1. When the correlation between two vectors is much less, their angle will be quite different from zero and the cosine significantly less than 1. The cosine of the angle is a direct correlation measure. When a threshold is defined then all correlations larger than the threshold will be declared a match and all others not-a-match. The only "intervention" is the choice of the threshold value.

5. PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is considered to be a simple and powerful technique for reducing the dimensionality of complex datasets while retaining most of the information [7]. Rather than storing single copies of a sample spectrum for each substance in PCA a larger set of measurements are performed, under different experimental condition, that act as a training data. What is sought is a transformation of the data set into components that maximize the variance. Hopefully a limited number of components then accounts for most of the variance. Once this linear transformation is found new data, the real observations, may be transformed in the same way and the distance, in some sense, of this new data, in principal component space, to the training data may be determined. As in SAM a threshold determines if the new data does or does not belong to the group, i. e. is classified as a match or not-a-match. In this method there are a number of "interventions": (i) size and variability of the training data set, (ii) a decision on how many PCs to take into account, (iii) how to define the distance in PC space, and (iv) the choice of the threshold or rather maximum distance.

6. RECEIVER OPERATOR CHRACTERISTICS (ROC)

ROC is a tool that is used to compare the performance of classifiers. A ROC is a plot of the true-positive rate (TPR) as a function of the false-positive rate (FPR) and is generated by sweeping the threshold value from the minimum to the maximum value and for each value counting the number of true and false positive matches that are found. For small threshold values, all data have larger correlation values than the threshold value and all true and false matches are detected, i. e. both the TPR and the FPR are 1. For maximum threshold value, all the data falls below the threshold and neither a false-positive nor a true positive is detected: TPR = FPR = 0. For intermediate values of the threshold, both true-positives and false-positive will be detected. For a good classifier the TPR > FPR and the ROC tends to bend towards the upper left-hand corner of the FPR-TPR space, where FPR = 0 and TPR = 1. The further the ROC is from the diagonal, the better the classifier is. This then provides a means to compare classifiers.

7. EXPERIMENT

The THz setup is based on a fiber-coupled time-domain spectroscopy system pumped by 100-fs pulses at 780 nm wavelength from a frequency-doubled Er-doped fiber laser [18]. THz images are acquired by mounting a sample holder on an x–y stage, which is scanned through the beam, with step size 2 mm for the training and reference data and 5 mm for the "real" data, while the transmitted THz waveform is captured. In this way a THz spectrum (after Fourier transform) is acquired for each stage position (pixel). A schematic of the setup is shown in Fig. 1. The distance between the emitter and detector modules is 31 cm and the sample holder has room for 3 x 3 sample pellets, with diameter 32 mm and thickness up to 4.2 mm. Fig. 1 (inset) shows the labeling of the sample positions. Teflon (25 μm average particle size) was used as a binder material, which was mixed with tartaric acid, lactose, or RDX and then pressed into pellets using a 2 ton press in two minutes. The bottom row of the sample holder (position 7–9) was used for reference measurements, using a metal plate (position 7), no sample (position 8), and a pure Teflon sample (4 mm thickness, position 9). All measurements were performed in ambient air (21–26 ^{o}C, 10–50% relative humidity). The signal at each position of the x–y stage (pixel) was measured with a time window of 60 ps and a scan speed of 1 ps/s, with a sample rate of 32 Hz. Figure 2(a) shows reference spectra for an open beam (air) and blocked beam (noise). All spectra were calculated from the time-domain signals by first applying a Blackman-Harris window with 8 or 15 or 30 ps half-width, centered in time at the signal peak, and then calculating the Fourier transform (FFT). The reference spectra indicate a bandwidth of about 2.5 THz and a peak signal-to-noise ratio (SNR) of about 60 dB. Figure 2(b) shows the absorbance for samples containing RDX (10%, 3.5 mm thickness), tartaric acid (10%, 4.0 mm thickness), and lactose (10%, 4.2 mm thickness). These samples were used as reference samples in the spectral library. The part of the spectrum spanning the

frequency range 0.1 to 1.5 THz was used in the correlation calculations, as the SNR is high in this frequency range (SNR > 40 dB for an open beam). Although there are several water vapor absorption lines in this wavelength range [2], we did not perform any numerical removal of water lines in the data processing. The location of the water lines was used to verify the calibration of the frequency axis in our measurements.

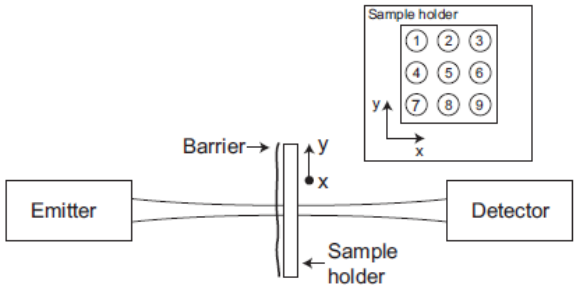

Figure 1. Experimental setup. Two fiber-coupled photoconductive antennas act as emitter and detector modules, which are separated by 31 cm. A sample holder, with transverse dimensions 15 x 15 cm, is scanned through the THz beam. Inset: Sample holder with labeling of samples indicated: 1 - 10% RDX, 4 mm, 2 - 10% Lactose, 4 mm, 3 - 10% Tartaric acid, 4 mm, 4 - 5% Tartaric acid, 4 mm, unground, 5 - 10% Tartaric acid, 1 mm, 6 - 5% Tartaric acid, 4 mm.

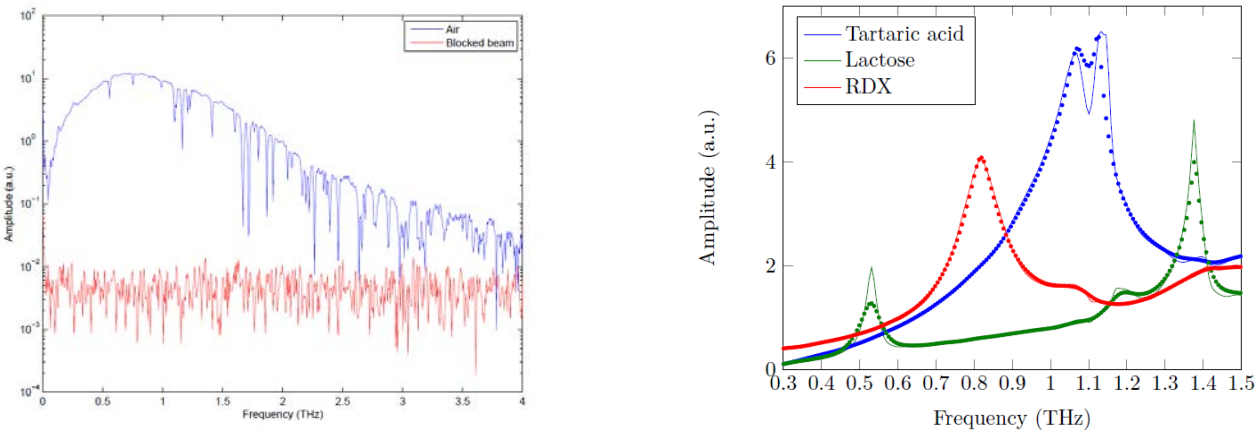

Figure 2. (left) Reference spectra of air and noise (blocked beam), after windowing of time-domain data, 30 ps window width. The maximum SNR > 60 dB and the many water absorption lines are clearly visible. (right) Absorbance the three substances considered in this study: tartaric acid (blue), lactose (green), and RDX (red) after applying two different B-H window widths: 30 ps (solid line) and 15 ps (dots).

A total of five images were recorded for this analysis: a high-spatial resolution image and four low-spatial resolution images: one without any cover and the three other with respectively plastic, paper, and cloth cover. The high spatial resolution image provided roughly 5000 spectra that were used both as a training set for the PCA method and three spectra were chosen to represent tartaric acid, lactose and RDX as reference spectra for SAM. Each of the low-resolution images consisted of roughly 900 pixels or spectra. Other than B-H windowing no preprocessing is applied on the time-domain data, and spectra are not corrected for water vapor absorption lines or background scattering effects.

8. ANALYSIS RESULTS

8.1 Spectral Angle Mapping

As an example of the SAM analysis we show the spectral correlation (SAM) results for one of the images, without barrier, with the tartaric acid reference spectrum in Figure 3 for two Blackman-Harris window widths (30 ps and 15 ps) and both the absorbance and its derivative.

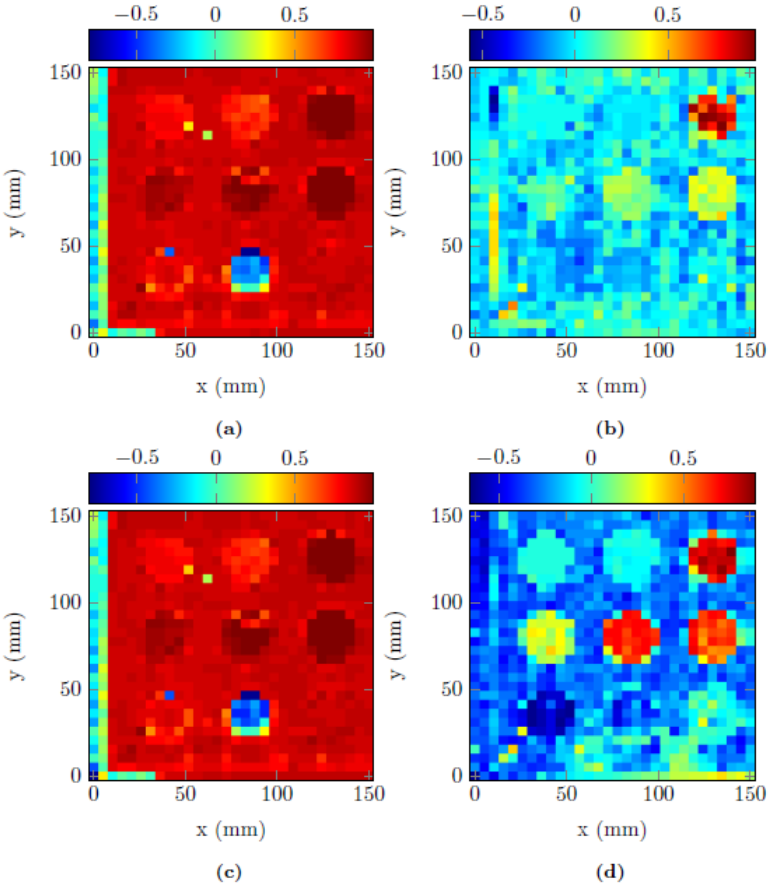

Figure 3. Spectral correlation of tartaric acid with an image taken without any barrier using (a) absorbance *A* and window 30 ps, (b) *dA/df* and window 30 ps, (c) absorbance A and window 15 ps, and (d) *dA/df* and window 15 ps. The color bar above each image indicates the correlation scale.

Even though one can observe a contrast between the pellets in the middle row and the background and also between the upper right-hand pellet and the background, this contrast is not very large and the width of the B-H window does not seem to affect the result, compare Fig 3a and 3c. The derivative of the absorbance is more sensitive, providing a stronger contrast even for the wider window (Fig. 3b), and this is even enhanced when the window width is reduced to 15 ps. All the four pellets that have some tartaric acid in them clearly jump out, sample positions 3 - 6. Looking at the images in Fig. 3 clearly the choice of the threshold value is decisive for the true-positive versus false-positive rates. By sweeping the threshold value from its minimum to its maximum and counting both the false- and true-positives the FPR and TPR for each of these thresholds may be calculated and plotted, as is done in Figure 4. A random classifier would show a diagonal ROC, and a good classifier “keeps itself” as far as possible from this diagonal, preferably only occupying the upper left-hand corner of the graph. If we first look at the upper row in Fig. 4, the data suggests that a wider B-H window gives a more reliable classification when the absorbance and SAM are considered. For the derivative of the absorbance, the bottom row in Fig. 4, the reverse is true: a narrower B-H window helps improve the SAM classification.

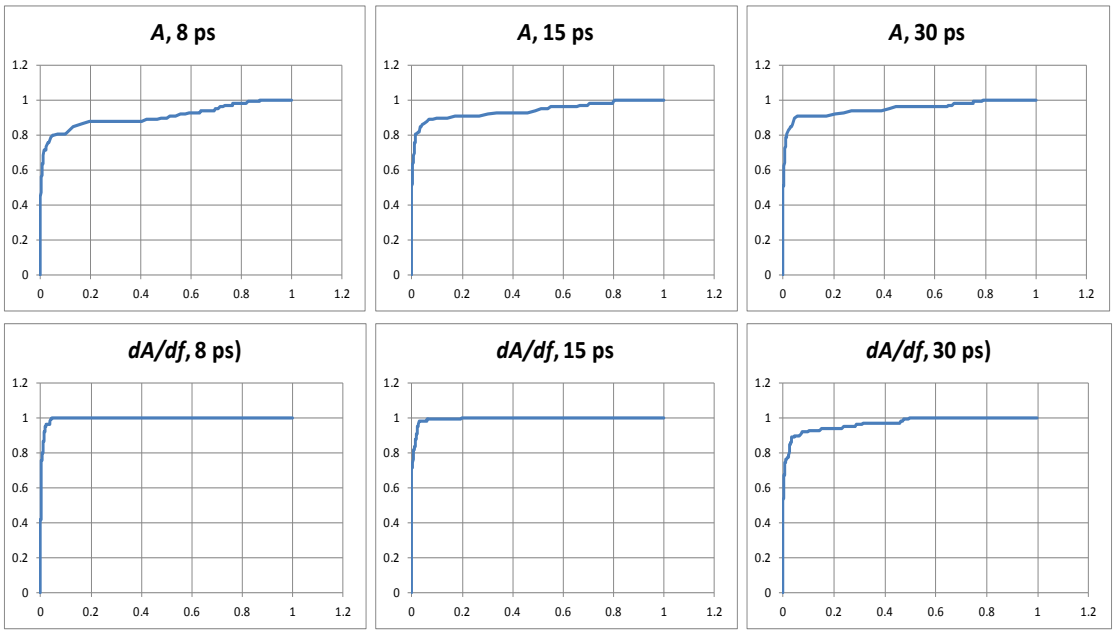

Figure 4. ROCs for the correlation of a reference tartaric acid spectrum with each of the spectra in a THz image, taken without concealing barrier. The TPR and FPR are plotted along the horizontal and vertical axis, respectively. The top row shows the ROCs for the absorbance and three different B-H window widths, whereas the bottom row shows the ROCs for the derivative of the absorbance, again for three different window widths.

A way to quantify the performance of the classifiers is to calculate the area under the ROC and subtract the area under the diagonal: this is an aggregate measure of how far the classifier is from a random classifier. These areas are presented in Table 1. When the absorbance is used the ROC area increase with increasing B-H window width until the half-width reaches one half the length of the time-domain signal which is 60 ps. For the derivative of the absorbance narrower window widths are an advantage, until they become to small: too much smoothing decreases the spectral sensitivity. Based on these observations we decided to limit the window half-widths between 8 and 30 ps.

Table 1. Areas under the ROCs, when spectral angle mapping is used on an image without any concealing barriers and the reference spectrum is tartaric acid.

	4 ps	8 ps	15 ps	30 ps	45 ps
A	0.309	0.407	0.441	0.450	0.447
dA/df	0.488	0.495	0.494	0.469	0.215

The ROCs may also be plotted in a different way, where the distance of the ROC to the diagonal is plotted versus the threshold value. Such a plot helps identifying what threshold value would give the classification result. Figure 5 shows such plots for the six cases presented in Fig. 4. Most eye-catching in Fig. 5 is the difference between the curves for the absorbance and those for its derivative. The former have a very narrow range of optimal threshold values, whereas the latter have broad ranges suggesting that the particular choice of the threshold value is not as critical. A threshold value close to one suggests that there is a very strong requirement for spectral similarity when trying to match the absorbance spectra, much stronger than for a match of the derivatives of the absorbance.

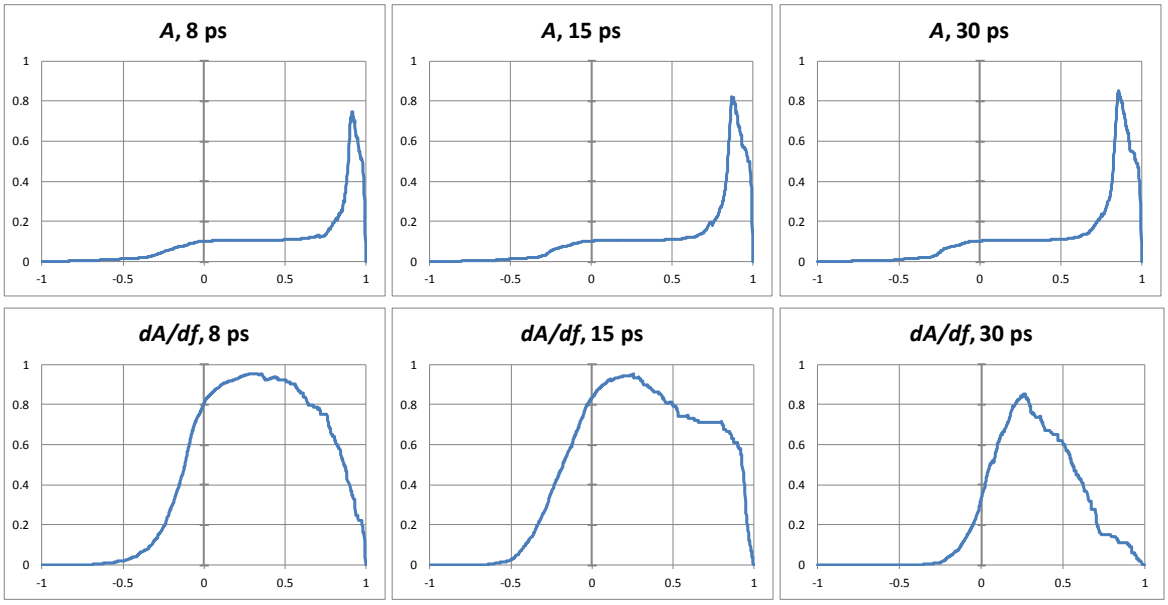

Figure 5. Plots of TPR – FPR (vertical axis) versus the threshold value (horizontal axis) for the same six cases as in Figure 4.

The same analysis may be performed on the other two substances, lactose and RDX, as well as for the three barrier materials considered here, paper, plastic, and cloth. This analysis is summarized in Fig. 6 where we show data similar as in Table 1

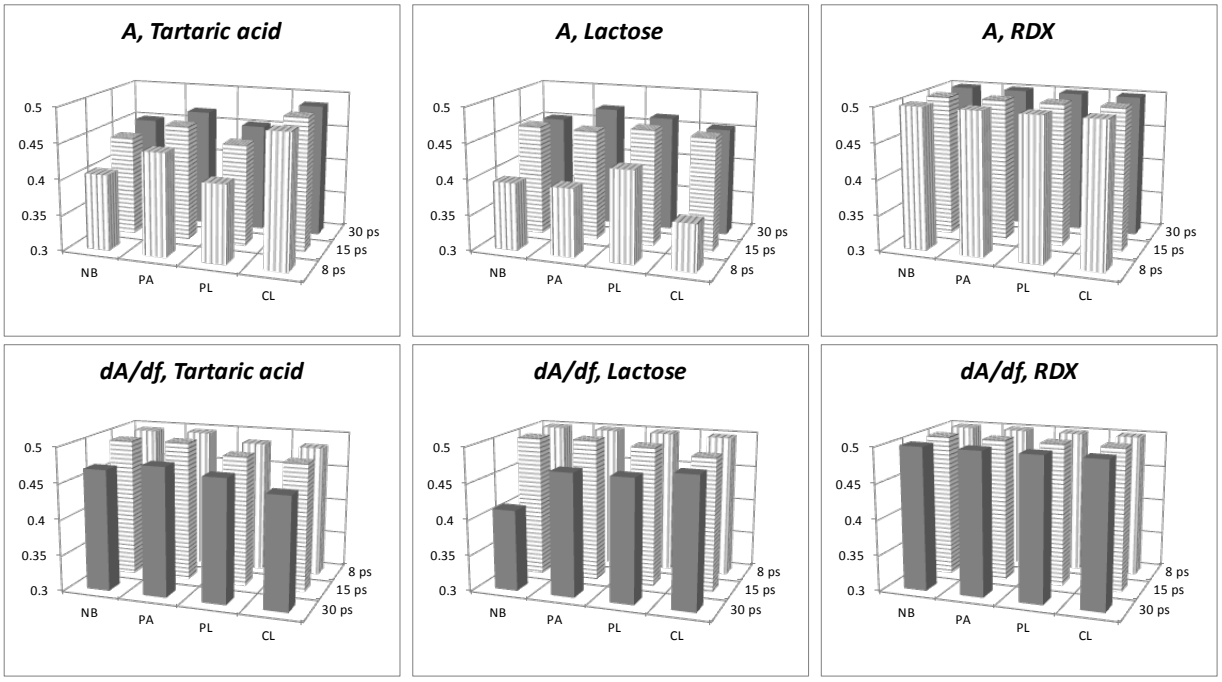

Figure 6. Area under the ROC for two spectral characteristics, absorbance (*A* , top row) and its derivative (*dA/df* , bottom row) and three substances: Tartaric acid (left column), Lactose (middle column), and RDX (right column). In each of the subplots the B-H window width is varied and different concealing barriers are used: NB = no barrier, PA = paper, PL = plastic, and CL = cloth. Note the reverse order on the time axis in the bottom row subplots.

Figure 6 reinforces the observations that were made above: when the absorbance is used as the spectral characteristic a better performance of the SAM classifier is obtained when the B-H window is wider and this is just the opposite when the derivative of the absorbance is used – here a narrower window gives a better result. Apparently, the increased noisiness that is the effect of a wider window on the absorbance spectrum is not a hindrance. Increased smoothing, which is the effect of a narrower window, of the absorbance before taking the derivative helps the classifier in this case. In general, both schemes work well, with the derivative absorbance yielding better results. Interestingly, the very best results are obtained when looking for RDX, with very low FPR and high TPR.

8.2 Principal Component Analysis - Training

In SAM it suffices to store one spectrum for each substance, a "gold standard". The first step in PCA is to collect a large set of training data, generally measurements of the substances of interest under widely varying conditions (humidity, signal strength, sample preparation, concealing barriers, etc.). In this study the training set is limited to the high-spatial resolution image, with a spectrum at each pixel site, that we took of the samples without a barrier. These measurements are used to find a transformation into principal components that account for the variance in the data, ordering the components according to their decreasing variance. A plot of the fraction of the variance accounted for by each component is used to decide how many principal components are needed to account for a sufficient fraction, e. g. 90 %, of the total variance. Such plots are presented in Fig 7.

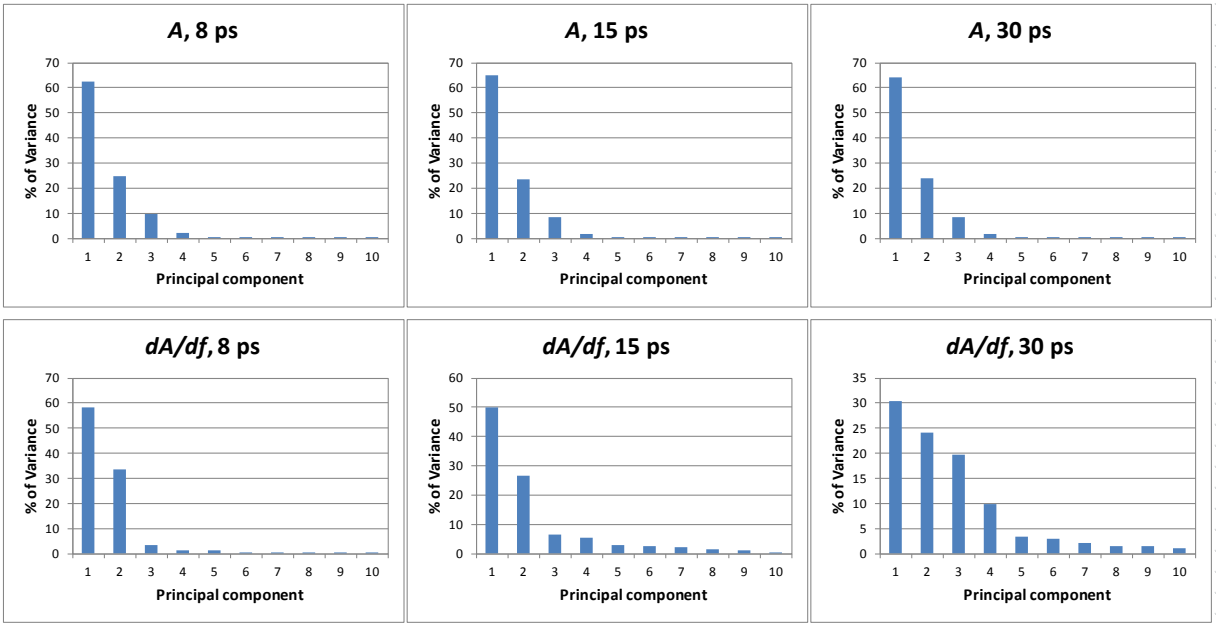

Figure 7. Plots of the percentage that each principal component contributes to the total variance. The same cases are considered as before: two spectral characteristics, absorbance A (top row) and its derivative dA/df (bottom row) as well as three B-H window widths namely 8, 15, and 30 ps corresponding to the left, middle, and right columns, respectively. Note the different vertical scales of the right two bottom subplots.

When the absorbance is considered as the spectral characteristic (see top row in Fig. 7), the effect of varying the B-H window width is relatively small: in all three cases the first three principal components account for 97% of the total variance. This suggests that it suffices to consider the first three principal components. Considering the derivative of the absorbance a clear effect of the windowing is observed, with increased window width requiring more principal components to account for the total variance. The summed contribution to the total variance of the first three principal components is 95%, 83%, and 74% for the three window widths, 8, 15, and 30 ps suggesting that more than three components should be used. We observe that more smoothing, i. e. using a narrower window, reduces the complexity of the reduced data set when the derivative is considered. For comparison purposes we chose to consider the first 10 principal components when analyzing the target images.

When the original data is projected onto *N* principal components it is expected that the data will form clusters in *N*-dimensional space. The number of clusters is determined by the number of substances in the training set, which in our case is three. Figure 8 shows the projection onto the first two components for the six cases considered. Except for the case where *dA/df* is used as the spectral characteristic with a 30-ps-window width, the grouping is quite obvious, even with only two principal components considered.

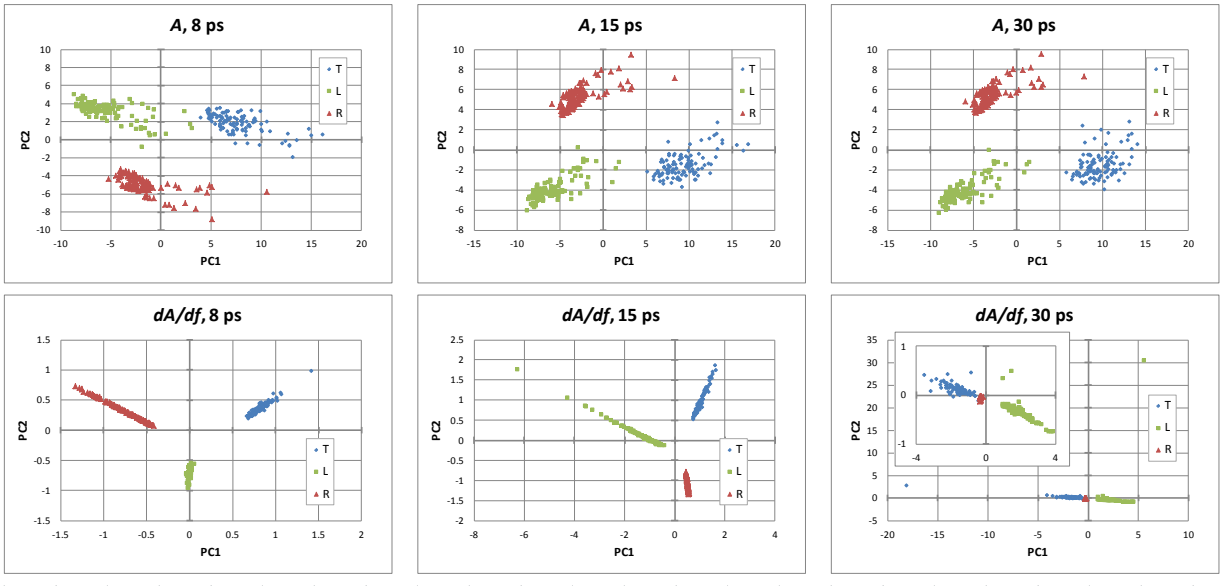

Figure 8. Score plots, projections of the original data onto the first two principal components for the two spectral characteristics absorbance A (top row) and its derivative dA/df (bottom row) as well as for the three B-H window widths 8, 15 and 30 ps, shown in the left, middle and right column, respectively. The inset in the lower right-hand subplot is a reduced-scale version of the whole plot, when the outliers are removed.

After the training procedure has been concluded the new spectra to be analyzed are projected onto the same principal components and a decision must be made to which cluster (substance group), if any, the new spectrum belongs. Each cluster is characterized by its mean value and standard deviation and the location of each new spectrum will be characterized by how many standard deviations it is removed from the clusters. The new spectrum will be classified as being part of a cluster when the distance is less than a given number of standard deviations.

8.3 Principal Component Analysis - Analysis

Using the transformations into principal components that were established using training data new images of the sample holder, both with and without a concealing cover (paper, plastic, cloth) were analyzed. It should be realized that this transformation must be calculated for each combination of spectral characteristic and B-H window width, i. e. six times in this study. Rather than considering a simple threshold value between -1 and 1 as for the case of SAM, in PCA a cluster of similar spectra is described as an ellipsoid in the *N*-dimensional principal component space. Now the size of the ellipsoid is increased from a minimum value to a maximum value and the TPR and FPR are calculated. The maximum size is determined by the largest distance between the spectra to be classified and the clusters determined from the training set. Outliers greatly extend this distance scale. Similar to Fig. 4 we show the ROCs for the six configurations we study in Fig. 9. These ROCs are the result of classifying tartaric acid spectra in an image taken of the sample holder when it is not covered. We observe that B-H window width does not affect the ROC when absorbance is used as the spectral characteristic when varied between 8 and 30 ps, see top row of Fig. 9. The ROCs are slightly worse than the ones in the top row of Fig. 4. This quantified by integrating the ROCs similarly to the procedure described in Section 8.1 yielding 0.344, 0.345, and 0.350 for the three window widths. This compares unfavorably with the numbers in the second row of Table 1. When we look at the bottom row of subplots in Fig. 9, where *dA/df* is used as the spectral

characteristic, we observe that the ROCs are better for the two narrower windows but significantly worse for the widest window. Integration of the three ROCs yields: 0.471, 0.464, and 0.277 for the three window widths. The first two compare reasonably well with the third row in Table 1, being only slightly lower. As may be seen in Fig. 8, bottom right subplot, the tartaric acid cluster has one extreme outlier whose distance to the center of the cluster is very large. This forces large steps in the ellipsoid size when calculating TPR and FPR, possibly resulting in a coarse ROC.

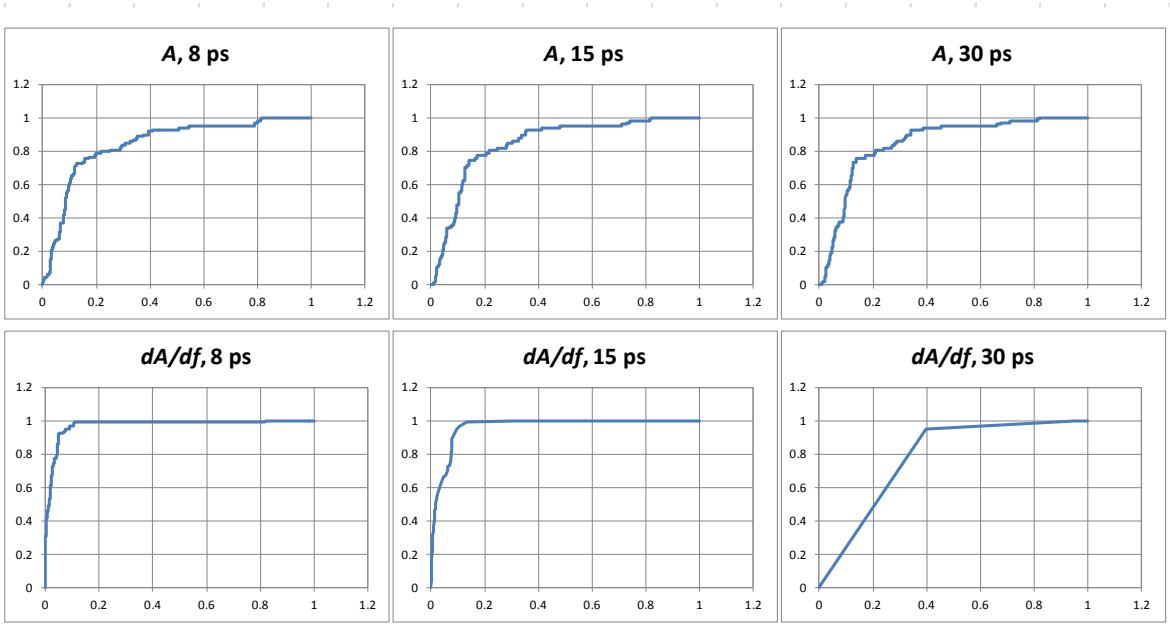

Figure 9. ROCs for classification of tartaric acid in an image of the sample holder without a concealing cover. Similar to Fig. 4, but here however, PCA was used rather than SAM.

Determining an objective value for the threshold, i. e. the maximum size of the ellipsoid describing a cluster, may not be straight forward, as alluded to above. A procedure similar to the one described for SAM gives results that are presented in Fig. 10. Comparing this figure with Fig. 5 we see that the maximum values of the curves in the top row are no as high as those in Fig. 5, however there is a certain range for reasonable threshold values suggested by the relative broadness of the tops. This is in contrast to the results in Fig. 5. However, it is difficult to compare the logarithmic threshold scale here with the linear one in Fig. 5. The bottom subplots (*dA/df*) show peaks that are narrower and higher, at least for the shortest window widths. The narrowness implies an increased sensitivity of the classifier to a proper choice threshold, but a good threshold value will give high TPR and low FPR, suggested by the high value of the maxima.

This analysis was also performed for the other two substances, lactose and RDX, and also on the other images that were recorded of the samples covered with paper, plastic or cloth. To allow for comparison of the performance of the different schemes the area spanned by the ROC and the diagonal (FPR = TPR) was calculated for each situation. The results are plotted in the bar graphs in Fig. 11. Clearly, the detection performance is rather variable, with disappointing performance when trying to detect lactose and RDX using the absorbance as the spectral characteristic and good performance when trying to detect tartaric acid and RDX using the derivative and appropriately narrow B-H windows.

9. COMPARING SAM AND PCA

When comparing the performance of PCA and SAM, i. e. comparing Figs. 11 and 6, it seems clear that SAM is better than PCA. However, this view should be modified. The performance of PCA is very dependent upon the quality of the training set – whether it contains most of the variations one may encounter in real-life measurement situations, as opposed to laboratory measurements. Comparing the left column in Fig. 11 to the other two, we observe that overall detection of tartaric acid is better than that of either lactose or RDX. There are, however, many more spectra of tartaric

acid in the training set, by a factor of four, than either of the other two substances. With the lack of training data for the detection of lactose and RDX the sensitivity to these two substances may be much worse. But this also points out an inherent difficulty with PCA, the fuzzy definition of a sufficiently good training data set. SAM does not seem to suffer from such a requirement: the detection of all three substances shows much less variation.

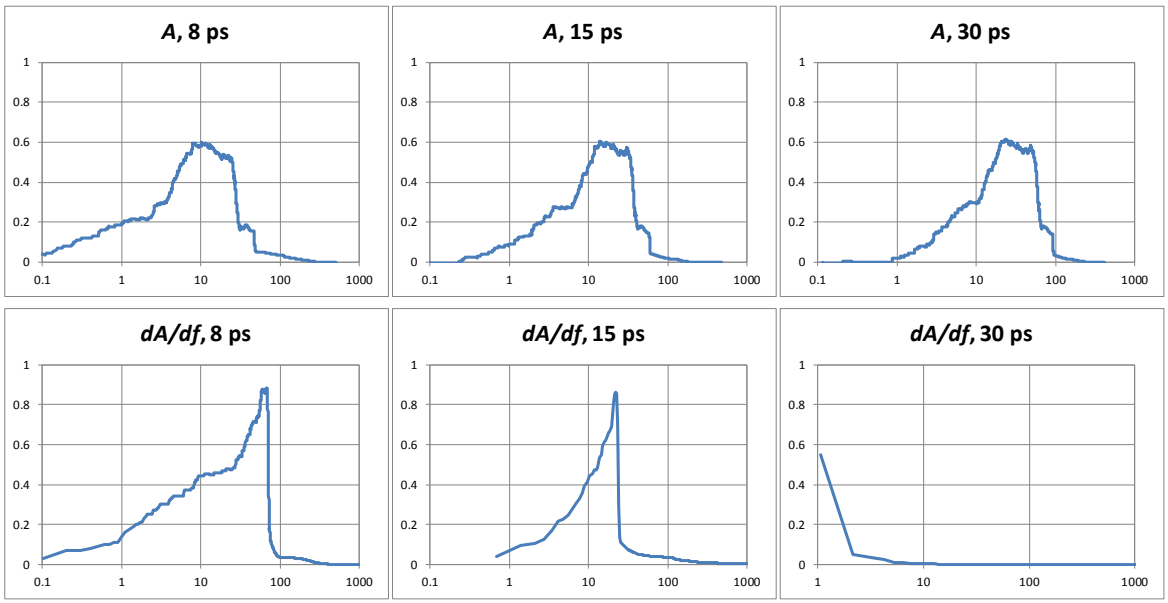

Figure 10. Plots of TPR – FPR (vertical axis) versus the threshold value (horizontal axis) for the same six cases as in. As Fig. 5, but here for PCA rather than SAM. Note that the horizontal scale, the threshold axis, is logarithmic.

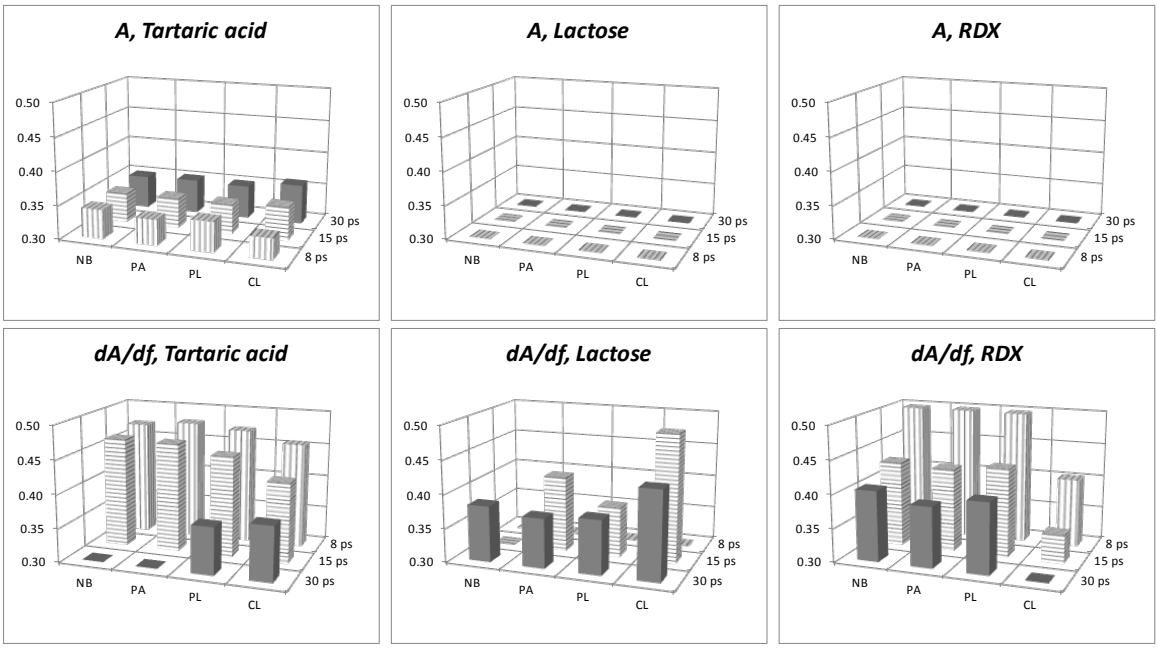

Figure 11. Area under the ROC for two spectral characteristics , three substances, and 3 B-H half-window widths. As Fig. 6, but here for PCA rather than SAM. For simple comparison, we have chosen to keep the vertical scales in all subplots: the

areas under the ROCs for (*A, Lactose*) and (*A, RDX*) are less than 0.3. The order of the window-width axes are reversed in the bottom-row subplots

The reliable detection of RDX, using the derivative of the absorbance and narrow B-H windows and PCA, is surprising in light of the limited number of RDX spectra in the training set. Also surprising is performance difference between the two spectral characteristics. Whereas there was hardly any difference in performance when SAM was used, it is much worse for the absorbance than for its derivative in combination with PCA.

Overall using the derivative of the absorbance seems to have the advantage when tartaric acid, lactose, and RDX are to be detected, whether used with SAM or PCA.

10. CONCLUSIONS

There are many approaches that could be useful to identify substances according to their measured spectra. In this study, where we look at THz spectra, with possible applications in the field of security such as mail scanning, parcel scanning, and person scanning, where many dangerous and/or illegal substances have characteristic spectral fingerprints. In security applications especially, but also in most other applications, false alarm rates must be balanced with throughput, sensitivity, and specificity. These requirements imply automated detection, with software identifying spectral matches and hence the sought-after substances. This study investigates some of the degrees of freedom associated with the process of spectral matching with the purpose of identifying the most reliable or robust method. The dimensions that were varied in this study are: (i) the spectral comparator – SAM or PCA, (ii)spectral characteristic – absorbance or its derivative with respect to frequency, (iii) the width of the Blackman-Harris window used in preprocessing the raw time-domain data – 8, 15, or 30 ps, (iv) the substance to be detected – tartaric acid, lactose, or RDX, and (v) the type of substance cover – none, paper, plastic, or cloth. Performance comparison is undertaken by using ROCs and ultimately by the area enclosed by the diagonal and the ROC. The closer this area is to 0.5, the maximum possible, the better the performance of the classification scheme.

Based on the experiments and analyses performed here we conclude that the absorbance derivative is a better spectral characteristic than the absorbance, especially when used in conjunction with a sufficiently narrow B-H window.

Performance of PCA is very dependent on the test cases, both number and variety, presented during training. Where PCA was reasonably successful detecting tartaric acid, it was much less successful detecting lactose and RDX, and this can probably be explained by larger fraction of tartaric acid spectra in the set of training data.

SAM is a much more intuitive method than PCA, where distance between spectra is quantified in terms of the cosine of the angle between vectors, resulting in a natural range of threshold values between -1 and 1. In addition, SAM requires only one sample spectrum of a substance in its library, rather than a large set of training data. This simplicity makes the method very attractive. It certainly did not perform worse than our implementation of PCA.

In addition to the dependence on training data of a certain volume and variability there is the aspect of the choice of threshold in PCA. This threshold is closely related to the distance measure employed. Most commonly the Mahalanobis distance is used, which has a degree of freedom in the choice of dimensionality. We have studied this aspect in some detail [19], and concluded that when the absorbance is used as the spectral characteristic three dimensions works best. When fewer dimensions are used in the distance measure too little of the variance is accounted for and the ROCs are worse. Using more than three dimensions yields a decomposition that is too sensitive to exceptional spectral behavior at the cost of normal behavior. The situation is different when the derivative of the absorbance is used as the spectral characteristic. In this case, the ROCs were best when two dimensions were used. Apparently, the sensitivity to exceptional behavior is larger when taking the derivative, something that one would intuitively expect. This discussion also highlights the complexity of the PCA, which requires tweaking of analysis parameters to optimize performance.

REFERENCES

[1] van Rheenen, Arthur D. and Haakestad, Magnus W., "Robust identification of concealed dangerous substances by spectral correlation of Terahertz transmission images", IEEE Transactions on Terahertz Science and Technology, vol. 5, pp. DOI: 10.1109/TTHZ.2015.2400224, March (2015).

[2] van Exter, M., Fattinger, C., and Grischkowsky, D., "Terahertz time-domain spectroscopy of water vapor," Optics Letters, vol. 14, pp. 1128–1130, Oct. (1989).

[3] Platte, F., and Heise, M., "Substance identification based on transmission THz spectra using library search", J. Molecular Structure Volume: 1073 Special Issue: SI Pages: 3-9 (2014).

[4] van Rheenen, Arthur D. and Haakestad, Magnus W., "Terahertz Imaging Spectroscopy - Towards Robust Identification of Concealed Dangerous Substances, presented at IRMMW & THz, Tucson, September (2014).

[5] There are many tutorials on PCA available on the internet, as an examples we mention L. I. Smith, (2002) (http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf), and J. Shlens (2014) (http://arxiv.org/pdf/1404.1100.pdf)

[6] Chan, W. L., Deibel, J., and Mittleman, D. M., "Imaging with terahertz radiation," Rep. Prog. Phys., vol. 70, pp. 1325–1379, Jul. (2007).

[7] Brigada, D. and Zhang, X.-C., "Chemical identification with information-weighted Terahertz spectrometry," IEEE Transactions on Terahertz Science and Technology, vol. 2, pp. 107–112, Jan. (2012).

[8] Tonouchi, M., "Cutting-edge terahertz technology," Nature Photonics, vol. 1, pp. 97–105, Feb. (2007).

[9] El Haddad, J., Bousquet, B., Canioni, L., and Mounaix, P., "Review in terahertz spectral analysis," TRAC - Trends in analytical chemistry, vol. 44, pp. 98–105, (2013).

[10] Fischer, B., Hoffmann, M., Helm, H., Modjesch, G, and Uhd Jepsen, P., "Chemical recognition in terahertz time-domain spectroscopy and imaging," Semicond. Sci. Technol., vol. 20, pp. S246–S253, (2005).

[11] Shen, Y. C., Lo, T., Taday, P. F., Cole, B. E., Tribe, W. R., and Kemp, M. C., "Detection and identification of explosives using terahertz pulsed spectroscopic imaging," Applied Physics Letters, vol. 86, paper no. 241116, (2005).

[12] Chen, J., Chen, Y., Zhao, H., Bastiaans, G. J., and Zhang, X.-C., "Absorption coefficients of selected explosives and related compounds in the range of 0.1-2.8 THz," Optics Express, vol. 15, pp. 12 060–12 067, Sep. (2007).

[13] Ellrich, F., Torosyan, G., Wohnsiedler, S., Bachtler, S., Hachimi, A, Jonuscheit, J., Beigang, R, Platte, F, Nalpantidis, K., Sprenger, T., and Hübsch, D., "Chemometric tools for analysing terahertz fingerprints in a postscanner," in 37th Int. Conf. on Infrared, Millimeter, and THz Waves, Wollongong, Australia, Sep. (2012).

[14] Wu, H., Heilweil, E. J., Hussain, A. S., and Khan, M. A., "Process analytical technology (pat): Quantification approaches in terahertz spectroscopy for pharmaceutical application," Journal of Pharmaceutical Sciences, vol. 97, pp. 970–984, Feb. (2008).

[15] Watanabe, Y., Kawase, K., Ikari, T., Ito, H., Ishikawa, Y., and Minamide, H., "Component analysis of chemical mixtures using terahertz spectroscopic imaging," Optics Communications, vol. 234, pp. 125–129, (2004).

[16] Kemp, M. C., "Explosives detection by terahertz spectroscopy–a bridge too far?" IEEE Transactions on Terahertz Science and Technology, vol. 1, pp. 282–292, Sep. (2011).

[17] van Rheenen, A. D. and Haakestad, M. W., "Detection and identification of explosives hidden under barrier materials - what are the THz technology challenges?" Proc. SPIE, vol. 8017, p. 801719, (2011).

[18] Ellrich, F., Weinland, T., Theuer, M, Jonuscheit, J., and Beigang, R., "Fibercoupled Terahertz spectroscopy system," Techn. Messen, vol. 75, pp. 14–22, (2008).

[19] Nystad, Helle E., "Comparison of Principal Component Analysis and Spectral Angle Mapping for Identification of Materials in Terahertz Transmission Measurements", Master's thesis, Norwegian University of Technology and Science, January (2015).