

AFRL-RY-HS-TR-2008-0005



USE OF EIGENVECTOR-GENERATED SCATTER PLOTS IN CLUSTERING IMAGE
DATA

*Jerry Silverman, Charlene Caefer
Infrared Sensor Technology Branch
80 Scott Drive
Hanscom AFB, MA 01731

*Solid State Scientific Corporation
27-2 Wright Road
Hollis, NH 03049

Interim Technical Report
28 July 2008

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

AIR FORCE RESEARCH LABORATORY
Sensors Directorate
Electromagnetics Technology Division
80 Scott Drive
Hanscom AFB MA 01731-2909

DTIC COPY

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the Electronic Systems Center Public Affairs Office for the Air Force Research Laboratory Electromagnetic Technology Division and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (<http://www.dtic.mil>).

AFRL-RY-HS-TR-2008-0005 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signature//

BRET KREH
Program Manager

//signature//

LUIGI SPAGNUOLO
Branch Chief
Infrared Sensor Technology Branch

//signature//

MICHAEL N. ALEXANDER
Technical Advisor
Electromagnetic Technology Division

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE (DD-MM-YYYY) 28-07-2008			2. REPORT TYPE INTERIM REPORT		3. DATES COVERED (From - To) 1 Oct 06 – 29 Sep 08	
4. TITLE AND SUBTITLE Use of eigenvector-generated scatter plots in clustering image data					5a. CONTRACT NUMBER	
					5b. GRANT NUMBER	
					5c. PROGRAM ELEMENT NUMBER 61102F	
6. AUTHOR(S) *Jerry Silverman, Charlene E. Caefer					5d. PROJECT NUMBER 2305	
					5e. TASK NUMBER HS	
					5f. WORK UNIT NUMBER 01	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Infrared Sensor Technology Branch, 80 Scott Drive, Hanscom AFB, MA 01731 *Solid State Scientific Corporation, 27-2 Wright Road, Hollis, NH 03049					8. PERFORMING ORGANIZATION REPORT	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Electromagnetics Technology Division Sensors Directorate Air Force Research Laboratory 80 Scott Drive Hanscom AFB MA 01731-2909					10. SPONSOR/MONITOR'S ACRONYM(S) AFRL/RYHI	
					11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-RY-SH-TR-2008-0005	
12. DISTRIBUTION / AVAILABILITY STATEMENT APPROVED FOR PUBLIC RESEASE; DISTRIBUTION UNLIMITED						
13. SUPPLEMENTARY NOTES Supported by AFOSR under Air Force Task 2305BN00. Clearance number, ESC PA 08-0038						
14. ABSTRACT Over the last few years we have been analyzing state-of-the-art spectral /temporal data of many events. Our goal was to develop specific techniques to classify and identify events based on these measurements. While the techniques evolved from one data type, we focus in this paper on the technique itself and its potential efficacy when applied to other data types. We use a Singular Value Decomposition (SVD) technique to cluster like events by forming a scatter plot from the first two eigenvectors. An evaluation of this approach using real data as well as simulations is given. A novel technique is introduced to assess cluster stability in the absence of ground truth. Results are presented along with the effects of misalignment of data samples, compression, training sets, and classifiers. The overall methodology is quite powerful and has remarkable noise immunity.						
15. SUBJECT TERMS scatter plots clustering eigenvector Singular Value Decomposition (SVD), 2-d image data						
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON BRET KREH	
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified	SAR	27	19b. TELEPHONE NUMBER (include area code) N/A	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

TABLE OF CONTENTS

I. Introduction.....	1
II. Generation and clustering of scatter plots.....	2
III. Alignment and compression of data samples.....	9
IV. Training sets and classifiers.....	13
V. Concluding remarks and future work.....	17
References.....	18
List of Acronyms.....	19

LIST OF FIGURES

Figure 1. Two samples of one data type (top) and two of another (bottom).....	3
Figure 2a 1. Scatter plot of example 1 based on first two eigenvectors (horizontal axis is first component).....	4
Figure 2b 1. Corresponding scatter plot after normalizing each data sample as described in text.....	4
Figure 3a 1. A scatter plot of a random 50% sub-sample.....	6
Figure 3b 1. A second random result.....	6
Figure 3c 1. Superposition of all 100 SP's.....	6
Figure 3d 1. The (10,3,3) cluster pattern of all 100 SP's displayed as described in text.....	6
Figure 4a 1. Type one sample image with Gaussian noise.....	8
Figure 4b 1. Type two sample image with Gaussian noise.....	8
Figure 4e.1 Display of cluster patterns at Dmax of one sigma.....	8
Figure 4f.1 Display of cluster patterns at Dmax of three sigma.....	8
Figure 5a 1. Sample of type 1.....	9
Figure 5b 1. Sample of type 2.....	9
Figure 5c 1. Sample of type 3.....	9
Figure 5d 1. Noise-added sample of type 1.....	9
Figure 5e 1. Noise-added sample of type 2.....	9
Figure 5f 1. Noise-added sample of type 3.....	9
Figure 6a 1. Original scatter plot of a data base of three types plotted as faint, (type 1), medium (2) and bright (3). Metrics of three pairs are 5.8, 12.6, 7.2 for pairs 1-2, 1-3,2-3 respectively.....	10
Figure 6b 1. Noise- added version with metrics of 3.7, 6.9, 3.8.....	10
Figure 6c 1. Noise-added version compressed to top 25% of data; metrics now are 3.4, 6.3, 3.6.....	10
Figure 6d 1. Noise-added version compressed to top 10% of data; metrics are 2.9, 5.2, 3.4.....	10
Figure 7a 1. Scatter plot of example 1 with one pixel random jitter. Compare Fig. 3b.....	12
Figure 7b 1. Scatter plot of example 1 with two pixel random jitter.....	12
Figure 7c 1. Scatter plot of example 2 with two pixel random jitter. Compare Fig. 6a. Metrics are 4.7, 12.6, 4.2.....	12
Figure 7d 1. Scatter plot of example 2 with three pixel random jitter. Metrics are 3.6, 8.9, 3.1.....	12

Figure 8a 1. Nearest neighbor classifier with underlying training set. Dark blue segment in upper right corner is region which is further from each center than any of the inter-center distances and is a region of non-decision..... 14

Figure 8b 1. Axial-oriented ellipses classifier based on same training set. Dark blue region outside any 3σ ellipse is taken non-decision as is the overlap region of types 1 and 2..... 14

Figure 8c 1. Rotated ellipses classifier version of 8b. 14

ACKNOWLEDGEMENTS

We would like to acknowledge the engineering teams from SSSC and AFRL responsible for designing and building the spectral-temporal sensors. Additionally special thanks go to the personnel responsible for all the data collects. Thanks to Dr. Jonathan Mooney for sharing his clear vision on the SVD and to Dr. Stanley Rotman, Dr. Glen Healey, and Shawn Higbee for technical discussions. We would like to thank Clement Wong, Karen Duseau, and Pearl Yip for software support. Finally we would like to thank Dr. Donald Silversmith of AFOSR for supporting this work under Air Force Task 2305BN00.

I. Introduction

Working on a project of repeated field-measured samples of events reducible to the time evolution of spectral profiles, we have developed novel methods for clustering and developing training sets of such 2D-image data. Our methods are generic rather than geared toward specific applications such as face recognition^{1,2} or video summarization³ or pattern recognition^{4,5,6} and should be useful in numerous applications which involve unsupervised classification^{7,8,9}.

Specifically, we are addressing the following scenario. One acquires a substantial data set of samples of several types of “image data” of the same generic kind. The x and y axis coordinates of each sample could be positional coordinates, other physical entities such as time or wavelength, or a numerically-indicated feature value. The third (z) coordinate designates the image sample. The x, y data values are arranged in a long vector with the corresponding alignment along z maintained: we have not used a reduced “feature vector” but rather the original gray scale values.

We use a simple and pictorial method of generating 2D scatter plots from the eigenvectors of an SVD of the data. Similar samples can be clustered to check the veracity of the “ground truth” designations. Alternately, unsupervised classification/clustering in the absence of ground truth as well as the development of classifiers viewing the sample set as laboratory signatures is demonstrated.

Our paper is organized as follows. In Section II, we use a simple real-data example and additional simulation based on the example to illustrate the basic methodology, show the need for data normalization, and introduce a novel technique for assessing cluster stability. Introducing a second real-world example with a larger data base, we discuss the issues of data compression and alignment in Section III and development and use of a classifier in Section IV. In Section V, we offer final comments and suggestions for further extensions.

II. Generation and clustering of scatter plots

To implement our approach, an SVD is performed on the M by N matrix formed by N column vectors (N samples) of the x, y data values of each sample arranged as an M -length column vector as described by Lee and Hayes¹⁰. A scatter plot (SP) point for each sample is formed from the first two eigenvectors of the N by N right singular matrix \mathbf{V}^T of the SVD decomposition¹⁰. To realize the significance of our data alignment, imagine an alignment along the z -axis of two distinct images: 10 identical samples of one image type and 10 of another. Note that the z -profiles for each data point $D(x,y)$ are step functions, in effect 'identity' profiles. The corresponding eigenvectors would reflect this in having identical values for the same image type. This tendency persists for real data, especially for the first two eigenvectors, and hence scatter plots of the first versus second eigenvector (one data point per sample) tend to cluster like sample types. The significance and enhancement of such scatter plots form the chief focus of this paper. Note that the present work, along with Reference 10, are the rare exceptions in using the N by N right singular matrix to cluster data. Typically, the M by M left singular \mathbf{U} matrix, which are the principle components of a KLT transform, are used in pattern recognition applications.(ref. 9, p.417)

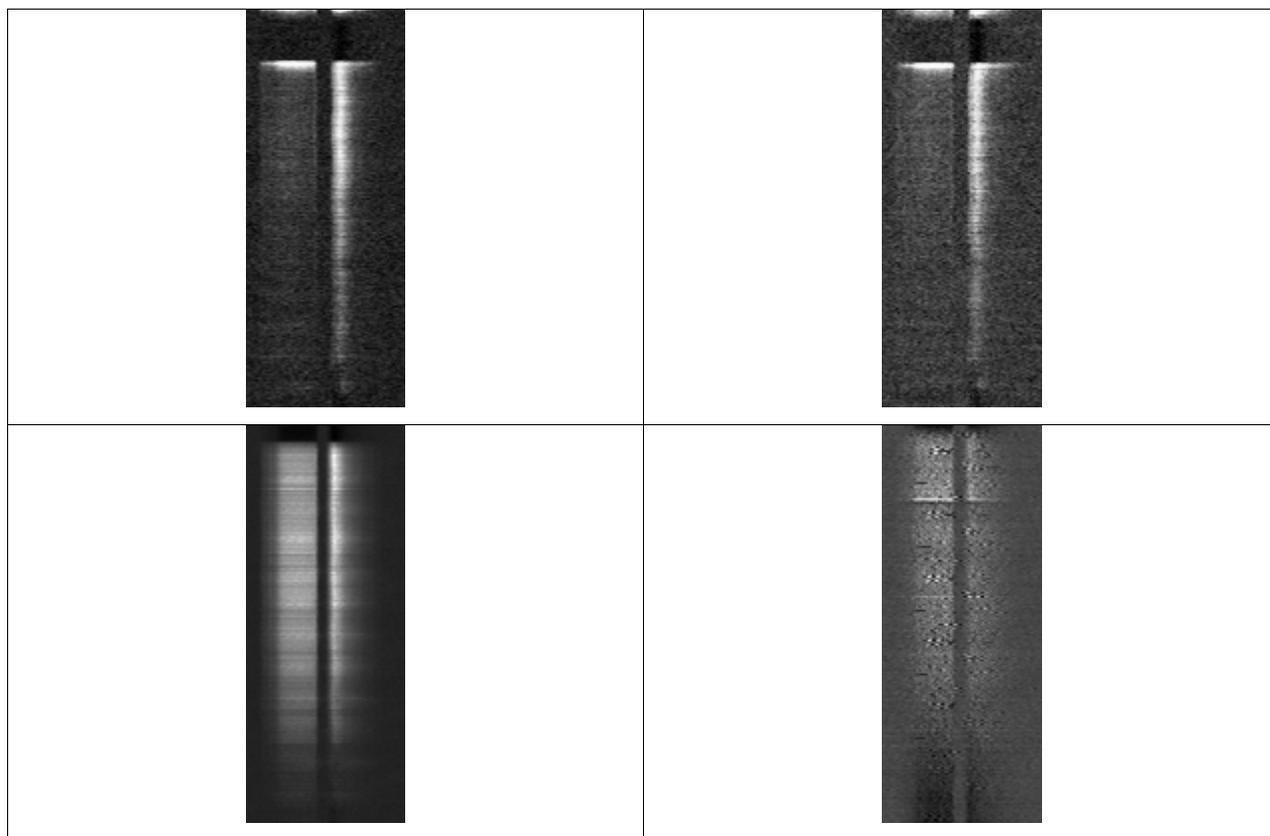


Figure 1. Two samples of one data type (top) and two of another (bottom).

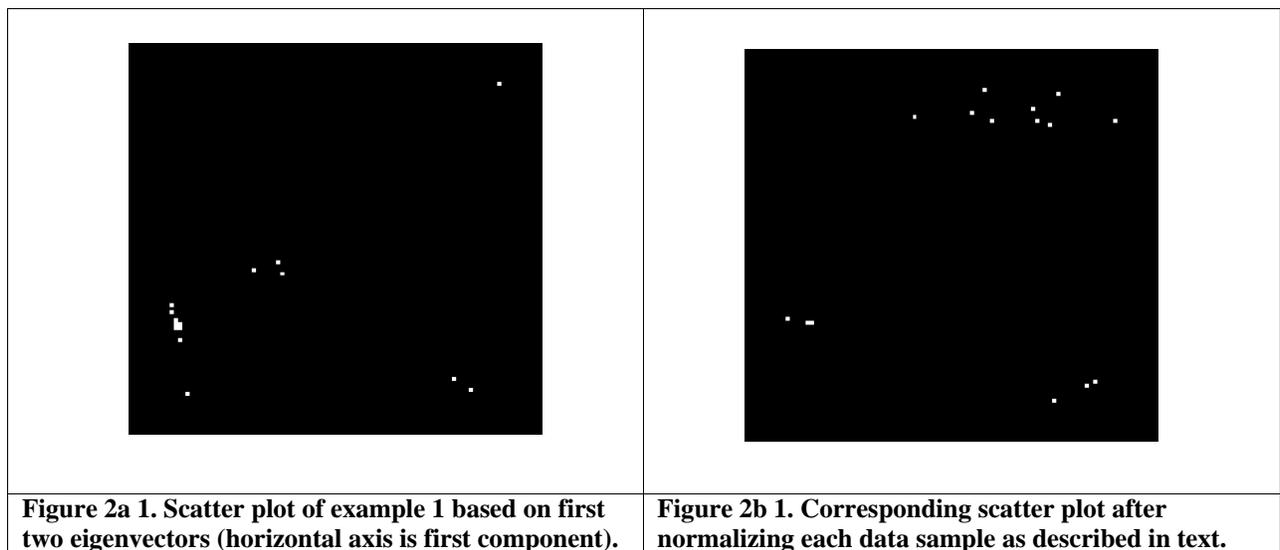
The first four figures are based on a real data set taken with laboratory camera instruments. Figure 1 shows two samples each of the two basic data types. For organization convenience, hereafter, the temporal spectral data will be re-arranged in (x,y) , but the original z -alignment is retained. The data base consists of 10 samples of type 1 and 6 samples of type 2 (image sizes 80 by 200) and the SVD of the 16,000 by 16 matrix is computed. A scatter plot of the first two eigenvectors, each component value scaled to an integer from 0 to 100, is shown in Fig. 2a. The anticipated groupings are a cluster of 10 samples of type 1 and 2 clusters of 3 samples each of type 2, the 3/3 split stems from a different angle of measurement direction. However, an anomalous magnitude of one type 2 image (upper right corner) perturbs the result and compresses the dynamic range of the rest of the scatter plot, which raises the need for normalization as described next.

The issue of normalization is critical; otherwise the SP's tend to be dominated by magnitude

effects. Of course, if one is dealing with an application where the key differentiations might be in the magnitudes, one would refrain from any normalization. Underlying our methods of normalization is the assumption that two data samples differing by an additive constant and/or a multiplicative constant represent the same data type. Our preferred normalization then consists of the following steps carried out independently on each data sample:

1. Determine the average profile as averaged over one of the coordinates and fix the DC level by setting the minimum of this profile at zero.
2. Adjust the multiplicative constant by dividing by the area under this adjusted average profile. In some cases, step 3 below improves the result.
3. Adjust the degree of variance about this average profile to a preset variance.

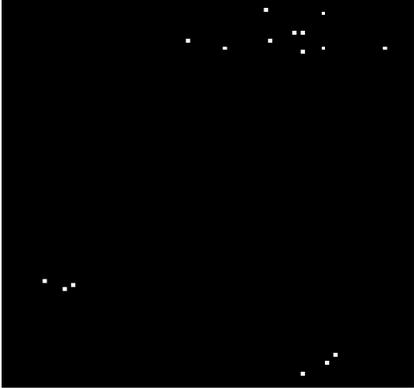
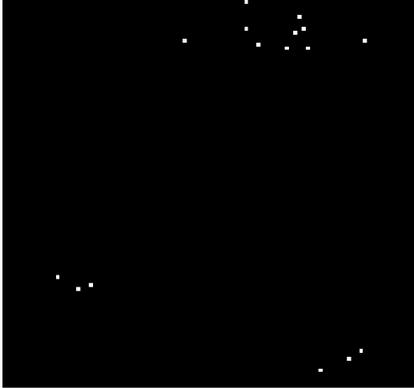
All three steps are similar to Duda and Hart's normalization method (ref. 9, p.215) of subtracting the mean and dividing by the standard deviation. Figure 2b shows the normalized scatter plot with the anticipated division of 10, 3, 3. All remaining SP's in the paper will be shown in the normalization just described. Other types of data might require different normalization.



A key feature of the present method resides in assessing cluster pattern stability. Suppose our ground truth information on the 16 data samples of Figures 1 and 2 were not available, i.e. we

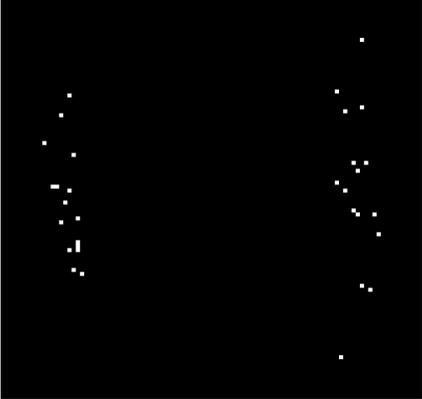
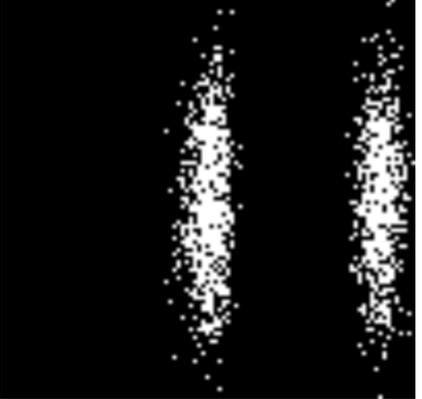
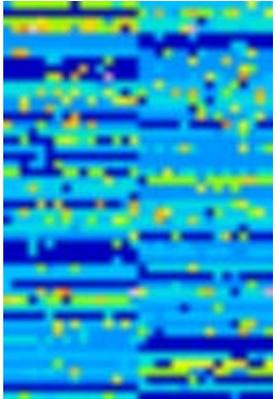
were faced with unsupervised classification. Can we extract from the data itself a stable and justified clustering? In our answer to this question, we use a variation of the standard clustering algorithm: “Agglomerative Single-linkage Algorithm”^{7,8}. This is a conservative approach that avoids excessively fine clusters and leaves high confidence in the remaining differentiations. Either from other available sources or from the process described below, one establishes maximum distances D_{max1} and D_{max2} for each component at a given SP integer scale. Pairs of points separated by D_{max} or less in each component are assigned to the same cluster. Each point in a final cluster has at least one link within the values of D_{max} to another point in the cluster, often referred to as a “friend-of-a-friend approach”, which allows for chain-like clusters.

We use a Monte-Carlo technique of random sub-sampling to estimate appropriate values of D_{max} ^{11,12}. A random address in x and y is chosen and the corresponding data point from the full data set is used until some sub-sample size is reached (we use 50% and find that allowing or precluding repeated selections makes little difference). The SP of this sub-sample is generated and the random process is repeated to generate N scatter plots—typically 50 to 100. Figures 3a and 3b show two of the random SP’s for the present example and Fig. 3c shows the superposition of 100 such SP’s. The latter suggests that the 10, 3, 3 cluster configuration is the finest supported by the data precision. We can demonstrate the stability of this clustering in more quantitative fashion. From the scatter in the 100 SP’s, one can estimate an average sigma (over all 16 sample points) in each component. The values of D_{max} is set to 3 times this estimated sigma in each component. We examine the constancy of the clustering pattern over the 100 SP’s. Fig.3d displays the result of applying our clustering algorithm to each of the 100 SP’s with these D_{max} . The vertical co-ordinate is SP run number and the horizontal co-ordinate designates one of the 16 data samples; an integer value is assigned to each distinct cluster. The color display indicates that over the 100 runs the 6,3,3 cluster configuration persists.

	
<p>Figure 3a 1. A scatter plot of a random 50% sub-sample.</p>	<p>Figure 3b 1. A second random result.</p>
	
<p>Figure 3c 1. Superposition of all 100 SP's.</p>	<p>Figure 3d 1. The (10,3,3) cluster pattern of all 100 SP's displayed as described in text.</p>

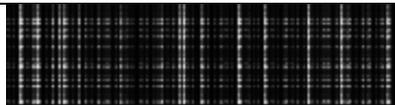
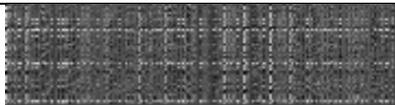
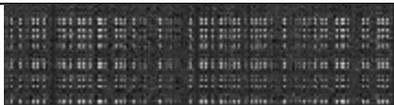
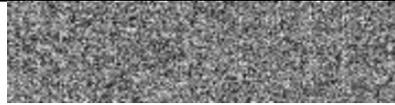
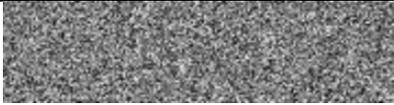
A simulation further supports this technique of assessment and its build-in immunity against over fine clustering. Using one sample of each of the two image types in our example, we add random Gaussian noise to create 16 noisy variations of each type (data set of 32 samples). The noise sigma is roughly 50 times the spatial sigma of the images and, as the two examples in Fig. 4a and 4b indicate, the observer can no longer distinguish the two patterns (compare Figure 1). The full-data SP is given in Fig. 4c and the superposition of 50 random sub-samples shown in Fig. 4d. The scatter from the 50 sub-sample SP's yields estimated sigmas of 3 and 16 (on the 100 integer scale of the SP) for each component, respectively. The difference in the sigma values is not unexpected since there is a larger variability in one component direction as compared to the other, even between samples as displayed in the SP shown in Fig. 4c. The cluster patterns for the 50 runs are displayed for values of Dmax of one sigma in Fig. 4e and three sigma in Fig. 4f. The latter shows the expected group of two clusters of 16 each over all runs; while the former shows too fine sub-clusters within the same type which vary from run to run. The still more

fundamental simulation of samples, which are Gaussian noise variants of one sample type, has also been done and results (not shown) are as expected: a stable single cluster at values of D_{\max} of three times the scatter-estimated sigmas. Hence, one has some confidence that a valid clustering can be extracted from the data itself. While in the remainder of the paper we will only show SP's based on the full data, our conclusions about cluster configurations will have been supported by examining the set of sub-sampled variations and clustering stability through the sub-sampled ensemble.

	
<p>Figure 4a 1. Type one sample image with Gaussian noise.</p>	<p>Figure 4b 1. Type two sample image with Gaussian noise.</p>
	
<p>Figure 4c 1. Scatter plot of 16 noise variants of each type.</p>	<p>Figure 4d 1. Superposition of all 50 random sub-sampled SP's.</p>
	
<p>Figure 4e.1 Display of cluster patterns at Dmax of one sigma.</p>	<p>Figure 4f.1 Display of cluster patterns at Dmax of three sigma</p>

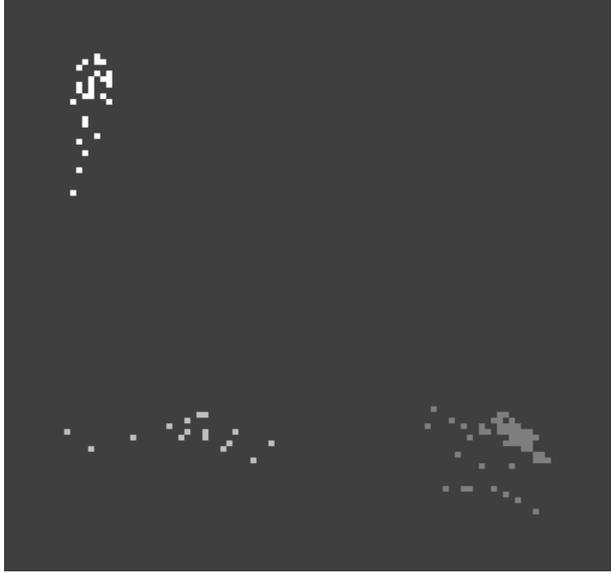
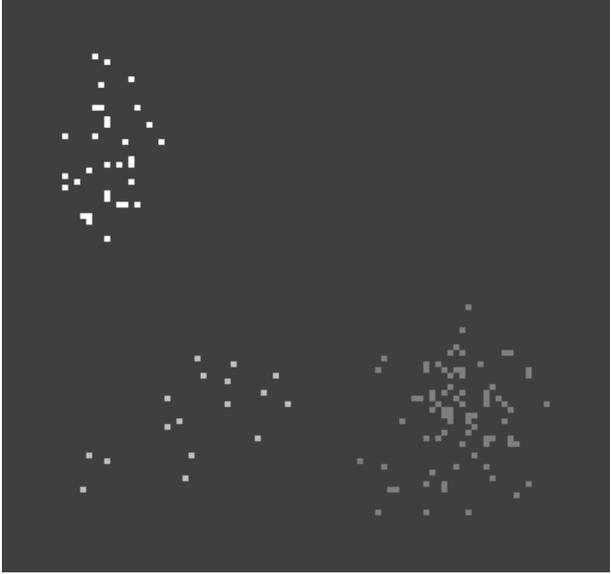
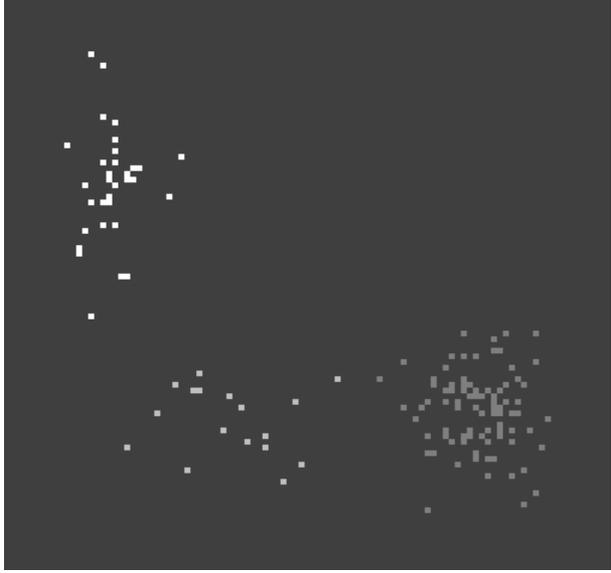
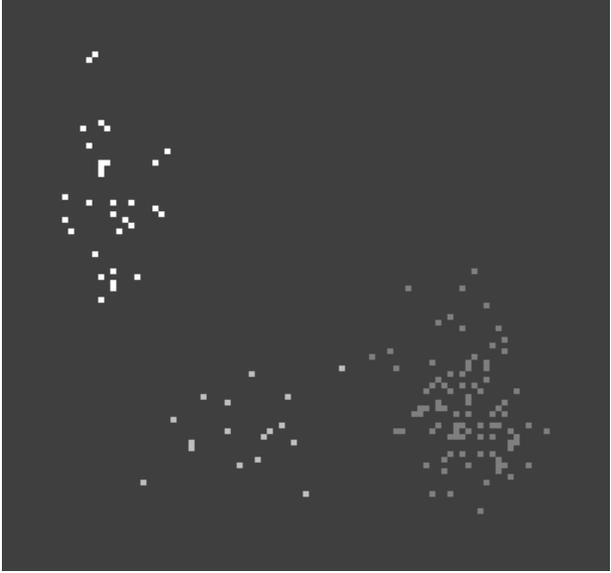
III. Alignment and compression of data samples

To address some subsidiary issues in this and the following section, we turn to a much larger data base of three sample types: 90 samples of type 1, 17 samples of type 2, and 32 samples of type 3. (These are also from real field measurements of time evolution of spectral profiles which have been re-arranged in (x-y).) In Figures 5a, 5b and 5c, we show a representative sample of each type. The inherent contrast spatial standard deviation is about 12 for the noise-free samples. For purposes of the discussion in this and the next section, we use as well a noisier version of our data base as shown in 5d, 5e and 5f. We have added Gaussian noise with a standard deviation of 80 to each sample to a point of non-recognition by an observer.

		
Figure 5a 1. Sample of type 1.	Figure 5b 1. Sample of type 2.	Figure 5c 1. Sample of type 3.
		
Figure 5d 1. Noise-added sample of type 1.	Figure 5e 1. Noise-added sample of type 2.	Figure 5f 1. Noise-added sample of type 3.

In Figure 6a we show the SP of the original data; a sharp distinction of the three sample types is evident. The remarkable noise-immunity of the technique is manifest in the SP of the noise-added samples, Fig. 6b, where the distinction is retained. We introduce a simple metric to characterize the quality of a given cluster pattern and to track changes in the SP quality. A metric value ratio, independent of the integer scale of the SP, is computed for each cluster pair (A and B) in which the numerator is the Euclidian distance of the cluster centroids, $ED(A,B)$, and the denominator is the geometric mean of the average distance among the elements of cluster A, $ID(A)$, and the corresponding value $ID(B)$ for cluster B. A cluster pair which largely overlaps will have a metric close to 1.0. Note that in contrast to metrics designed to evaluate the validity of a cluster pattern⁷ versus another pattern within the same SP (using a criterion of optimality), this metric is designed to track the changing quality of a particular clustering pattern as the SP changes due to noise, compression, or jitter as treated below.

$$M(A,B) = ED(A,B) / \text{sqrt} (ID(A) \times ID(B)) \quad (1)$$

	
<p>Figure 6a 1. Original scatter plot of a data base of three types plotted as faint, (type 1), medium (2) and bright (3). Metrics of three pairs are 5.8, 12.6, 7.2 for pairs 1-2, 1-3,2-3 respectively.</p>	<p>Figure 6b 1. Noise- added version with metrics of 3.7, 6.9, 3.8.</p>
	
<p>Figure 6c 1. Noise-added version compressed to top 25% of data; metrics now are 3.4, 6.3, 3.6.</p>	<p>Figure 6d 1. Noise-added version compressed to top 10% of data; metrics are 2.9, 5.2, 3.4.</p>

The present technique readily lends itself to data compression, which is of particular interest for large data size. Recall that the data arrangement which the eigenvectors are “summarizing” are profiles along the z-axis. Even $D(x,y)$ data points with strong intensity are important only to the

extent that they vary among the aligned images. Hence, our data point selection/compression process is based on the variance of each data point along the sample direction (z) . Specifically, we compute with respect to z:

$$V(x_0, y_0) = E\{ D^2(x_0, y_0, z) \} - (E\{ D(x_0, y_0, z) \})^2 \quad (2)$$

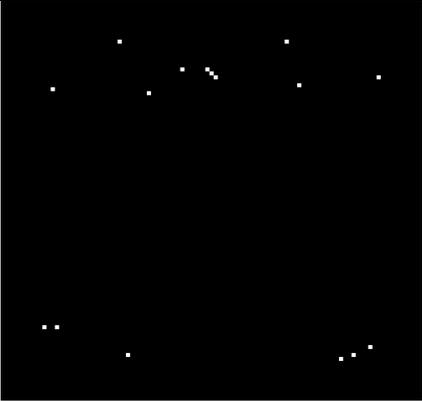
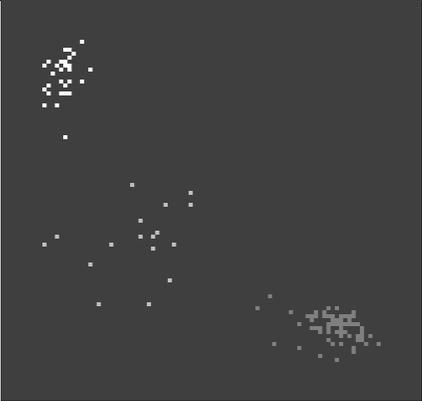
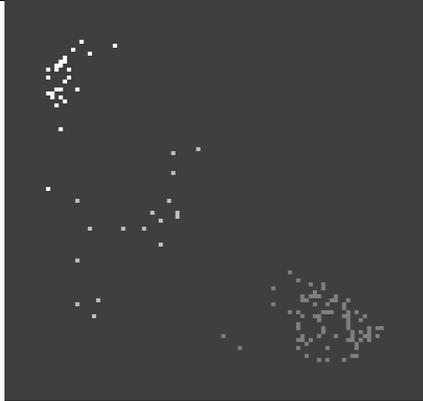
and if we wish to compress on the basis of x-coordinate (or y), we can add up the variances to give a final score:

$$S(x_0) = \sum_{y_i} V(x_0, y_i) \quad (3)$$

The SP's based on the top 25 and 10% (Figs. 6c and d) from Eq. (2) show only slight degradation at the 1-2 boundary as perceived by an observer and reflected in the metrics. This result is typical of many data sets we've reviewed.

We next consider the effect of data mis-alignment, i.e., registration error of our image set. This is important as the methodology depends on z-profiles of consistent identity. The effect is strongly dependent on the x,y spatial frequency of the data types i.e., image samples with high spatial frequency place much more severer demands on the alignment accuracy.

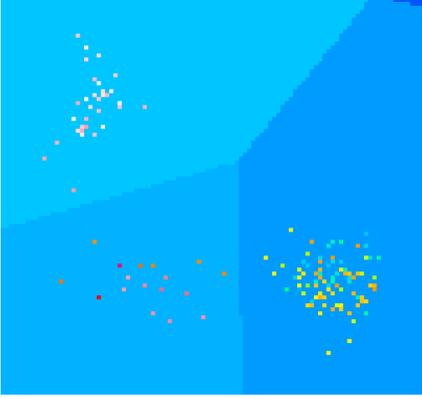
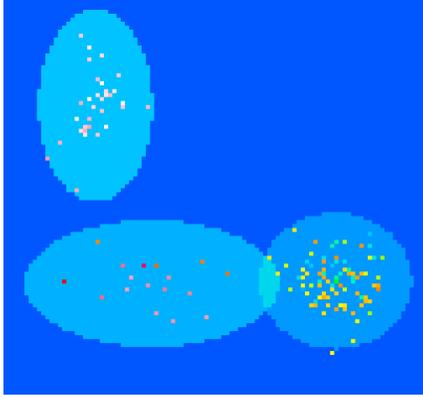
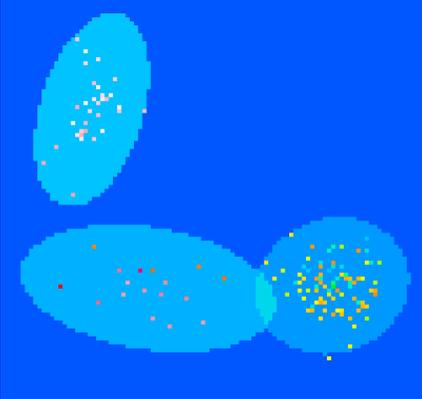
The data of the two examples used here are reasonably well-aligned (1-2 pixels) due to the nature of the data and control over laboratory measurements. However, we can simulate the effect of mis-registration by imposing a random jitter independently on each image. For a selected random jitter of N, each image is randomly and independently assigned a shift of 0, 1, 2,...up to N pixels in x and in y. For our initial example data set, a random shift of one pixel retains the 10, 3, 3 separation although the 10 samples of type 1 are much less compact (Fig. 7a, compare Fig. 2b); while, a shift of two pixels (Fig. 7b) loses the 3, 3 cluster division of the type 2 samples. The effect of jitter of two and three pixels applied to the noise-free version of our second data set (Figs. 7c and 7d) is similar to the effect of noise (compare to Fig. 6a,b). For high spatial frequency imagery or very noisy imagery, the effects of mis-alignment are expected to be far more severe.

	
<p>Figure 7a 1. Scatter plot of example 1 with one pixel random jitter. Compare Fig. 3b</p>	<p>Figure 7b 1. Scatter plot of example 1 with two pixel random jitter.</p>
	
<p>Figure 7c 1. Scatter plot of example 2 with two pixel random jitter. Compare Fig. 6a. Metrics are 4.7, 12.6, 4.2.</p>	<p>Figure 7d 1. Scatter plot of example 2 with three pixel random jitter. Metrics are 3.6, 8.9, 3.1.</p>

IV. Training sets and classifiers

An important goal in our original application is the generic one of using extensive field or laboratory measurements as a training set in order to identify an unknown. Towards that goal and as one possible approach, we present some preliminary work on using SP's as training sets to form classifiers.

As representative of several simulations and real data sets we have tested, we will employ our second example with noise added (as in Fig. 6b) as the training set. Figures 8a, b, and c show three possible classifiers with the underlying training set SP. Fig 8a is a nearest (Euclidian) neighbor (NN); Fig. 8b are axial-oriented ellipses (AE) and 8c are ellipses rotated to match the data correlation (RE). The ellipses are sized in axial radii at $3\sigma_1$ and $3\sigma_2$ where the sigmas refer to scatter of the SP data points about the ensemble means in the requisite directions.

	
<p>Figure 8a 1. Nearest neighbor classifier with underlying training set. Dark blue segment in upper right corner is region which is further from each center than any of the inter-center distances and is a region of non-decision.</p>	<p>Figure 8b 1. Axial-oriented ellipses classifier based on same training set. Dark blue region outside any 3σ ellipse is taken non-decision as is the overlap region of types 1 and 2.</p>
	
<p>Figure 8c 1. Rotated ellipses classifier version of 8b.</p>	

We have carried out some Monte-Carlo simulations as a preliminary assessment of trends with these classifiers. Rather than generating the SP anew from the training set and the unknown(s), as a more practical real-time operation, one can use the SVD of the training set to predict the position of the unknown(s) without generating an expanded new SP,

$$\mathbf{V}^T \sim \mathbf{W}^{-1} \mathbf{U}^T \mathbf{D}, \quad (4)$$

where \mathbf{W}^{-1} and \mathbf{U}^T are from the SVD of the training set and \mathbf{D} are the data values of the unknown(s). One can readily show that that the first two rows of \mathbf{V}^T , i.e. the two components which generate the SP position, depend only on the data of the unknown \mathbf{D} , the first two singular

values of \mathbf{W} , and the first two principle components of \mathbf{U} . The approximation requires that the values of the first two components of \mathbf{U} and \mathbf{W} from the SVD based on the original training set will be little changed by the addition of any one unknown. Over the surveyed examples from our in-house acquired data, we find we can predict positions to within one pixel on the integer scale of 100, i.e. 1%. However, for the Gaussian-noise dominated case of the present example, the SP's are more volatile and a new unknown with another sample noise-pattern can only be predicted to about 10%. Hence in our Monte-Carlo assessment of the three classifiers, we regenerate the full SP's.

This assessment consisted of *runs* as follows. The basic training set of 139 samples was extended by one "unknown" and a 140-sample SP was generated. The position in this SP of the unknown on one of the classifier templates of Fig. 8 registered either a detection, missed detection (dark blue region), or false alarm (wrong decision) for that unknown. The unknown was taken from a different standby Gaussian-noise version of the 139 data samples. Versions at standard deviations of 80, 100, and 120 were used. A *run* generated 139 SP's as each of the samples of a standby version was used as the unknown. Hence each of the runs gives a rate of detections (D), missed detections (MD), and false alarms (FA) averaged over 90 samples of type 1, 17 of type 2, and 32 of type 3.

While the absolute result values of our simulations are unlikely to reflect more realistic scenarios, such as newly measured samples at different times and conditions than those of the training set, we suggest that the relative performances of the classifiers are accurately conveyed by our simulations. Rather than burdening the reader with detailed tables we will summarize the overall trends.

For this example, where the "unknown" is one of the three data types, the NN classifier is markedly superior. A typical run at $\sigma=100$ gives respectively for the NN, AE, and RE classifiers: 99.3%,88.5%, 88.4% D; 0%, 10.1%, 10.8 MD; and 0.7%, 1.4%, 0.7% FA, respectively.

An average of two runs at the $\sigma =120$ noise level (50% higher than the level of the training sets) gives: 98.2%,70.2%, 67.6% D; 0%, 28.0%, 32.0 MD; and 1.8%, 1.8%, 0.4% FA. The FA rate of all three is similar with the RE slightly better but both the elliptical classifiers retain their low FA at the expense of a high rate of missed detections. The expected tradeoff between FA and MD

using the NN classifier with its large scale regions of hard decision is seen in other simulations with classifiers and data sets introducing unknowns of different types to those in the training set. One should mention with regard to the 3 sigma boundaries of the AE and RE classifiers that the sigmas in a particular application could be chosen to upper bound false alarms or to adapt to a CFAR constraint.

V. Concluding remarks and future work

We have presented a widely applicable and simple method of clustering a set of 2D-image data based on the scatter plots of the first two eigenvectors of an SVD of the data set. The data matrix transformed by an SVD is formed by arranging each data sample in a long vector with the correct alignment along z maintained: we have not used a reduced “feature vector” but the original gray scale values. Using a chosen clustering algorithm and statistical sub-sampled versions of the scatter plot, one can assess the most stable clustering configuration. A second example data set indicates the noise immunity of the technique and how training sets can be used to generate classifiers. Issues of alignment, jitter, and compression were also addressed.

Some natural extensions of our work are as follows. Corresponding features of a feature vector could be aligned along the z -axis rather than the gray scale values, which could make the alignment process less critical as in the video summarization case³ where histograms are used. All our techniques such as the generation of the SP, the use of sub-sampling to estimate cluster stability, and the design of classifiers are done on a component-independent basis and could readily be extended to three or more eigenvectors.

Finally, other kinds of data should be tested. Our methods have been developed from and applied to an extensive set of field measurements, all of which involve wavelength versus time data sets. In principle, one can apply the methodology to data samples of any dimension since the sample data is re-arranged in column form for the SVD matrix. The critical point is that alignment of corresponding points from sample to sample in the data description is attainable.

References

1. M.A. Turk and A.P. Pentland, "Face recognition using eigenfaces", *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 586-591 (1991).
2. P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection", *IEEE Transactions of Pattern Analysis and Machine Intelligence* 19(7), 711-720 (1997).
3. Y. Gong and X. Liu, "Video summarization using singular value decomposition", *IEEE International Conference on Image Processing*, Vol. 3, 362-369 (2001).
4. A. Ahmadi, S. Omatu, and T. Kosaka, "A PCA based method for improving the reliability of bank note classifier machines", *Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis*, 494-499 (2003).
5. B. Luo, R.C. Wilson and E.R. Hancock, "Object recognition by clustering spectral features", *IEEE International Conference on Image Processing*, Vol. 1, 429-432 (2002).
6. A.T. Targhi and A. Shademan, "Clustering of singular value decomposition of image data with applications to texture classification", *Visual Communications and Image Processing 2003*, SPIE Vol. 5150, 972-979 (2003).
7. A.K. Jain, M.N. Murty, and P.J. Flynn, "Data clustering: a review", *ACM Computing Surveys*, Vol. 31(3), 264-323 (1999).
8. R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York (1973).
9. K. Fukunaga, *Introduction to Statistical Pattern Recognition (2nd Edition)*, Academic Press, Inc., Boston (1990).
10. S.Lee and M.H. Hayes, "Properties of the singular value decomposition for efficient data clustering", *IEEE Signal Processing Letters*, Vol. 11(11), 862-866 (2004).
11. A. Ben-Hur, A. Elisseeff, and I. Guyon, "A stability based method for discovering structure in clustered data" *Pacific Symposium on Biocomputing*, World Scientific 6-17 (2002).
12. S. Higbee, private communication.

List of Acronyms

AE	Axial-oriented Ellipse
AFOSR	Air Force Office of Scientific Research
AFRL	Air Force Research Laboratory
CFAR	Constant False Alarm Rate
FA	False Alarm
MD	Missed Detection
NN	Nearest Neighbor
SP	Scatter Plot
SSSC	Solid State Scientific Corporation
SVD	Singular Value Decomposition