A Wavelet, Fourier, and PCA Data Analysis Pipeline: Application to Distinguishing Mixtures of Liquids

Münevver Kóküer,*,[†] Fionn Murtagh,[†] Norman D. McMillan,[‡] Sven Riedel,[§] Brian O'Rourke,[‡]
Katie Beverly,[‡] Andy T. Augousti,[§] and Julian Mason[§]

School of Computer Science, Queen's University Belfast, Belfast BT7 1NN, United Kingdom, Institute of Technology Carlow, Kilkenny Road, Carlow, Ireland, and Faculty of Science, Kingston University, Penrhyn Rd, Kingston, Surrey KT1 2EE, United Kingdom

Received September 9, 2002

Using a new optical engineering technique for the "fingerprinting" of beverages and other liquids, we study and evaluate a range of features. The features are based on resolution scale, invariant frequency information, entropy, and energy. They allow mixtures of beverages to be very precisely placed in principal component plots used for the data analysis. To show this we make use of data sets resulting from optical/near-infrared and ultrasound sensors. Our liquid "fingerprinting" is a relatively open analysis framework in order to cater for different practical applications, in particular, on one hand, discrimination and best fit between fingerprints, and, on the other hand, more exploratory and open-ended data mining.

1. INTRODUCTION

The work which we report on here makes use of a new (patented) technique for measuring the tensile and viscosity properties of any liquid. A laser-derived beam of light is directed into a drop as it builds up on a drop-head, grows, and eventually falls off through gravity. The light is reflected through the drop, and a trace is built up of its intensity over time (see Figure 1). The trace recorded for one drop is known as the tensiotrace, and this will be used to provide a unique fingerprint of the liquid. Figure 2 shows a typical tensiotrace recorded at 950 nm for water with all of the trace features labeled. The tensiotrace has been found to have very good discrimination potential for various classes of liquid. Other sensing modalities can be used— multiple simultaneous optical and near-infrared wavelengths, ultraviolet, ultrasound. In the studies reported on here, we use the optical/near-infrared and ultrasound modalities. Further background on this new technology for the fingerprinting of liquid content and composition can be found in refs 1−3.

2. DATA AND ANALYSIS REQUIREMENTS

The data were related to 614 optical/near-infrared modality traces. The entire set consists of 0.05 M copper sulfate at 850 and 660 nm, 0.05 M copper nitrate at 850 and 660 nm, 0.1 M copper sulfate at 850, 660, 555 and 470 nm, 0.1 M copper nitrate at the four wavelengths as above, and finally a mixture of 0.1 M copper nitrate and copper sulfate again at the same four wavelengths. Each trace has several samples. Examples of the traces are shown in Figure 3. Because the adjustment of the wavelength moves the drophead slightly, a water reference is taken for each trace as well.

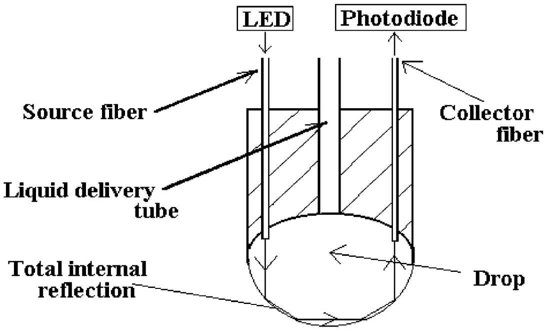

Figure 1. Light trace from source to detector.

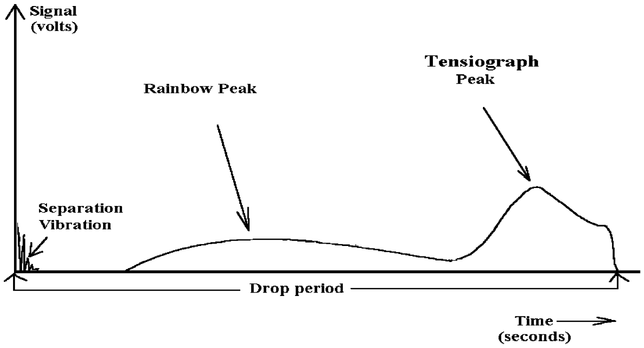

Figure 2. Typical tensiotrace showing the characteristic features.

The samples were chosen in order to determine the upper detection limit for the highly colored copper solutions and to determine the influence of color on tensiotrace details. Mixtures of nitrate and sulfate were made in order to determine whether trace details varied due to the presence of these widely found water pollutants and whether specific variations could be attributed to the individual pollutants.

The traces for 0.1 M copper nitrate at 660 nm are not representative, since the software package used for data acquisition assumes that when the signal falls to zero, the

* Corresponding author phone: +44 28 90335488; e-mail: m.kokuer@qub.ac.uk.

[†] Queen's University Belfast.

[‡] Institute of Technology Carlow.

[§] Kingston University.

10.1021/ci025601j CCC: $25.00 © 2003 American Chemical Society
Published on Web 01/23/2003

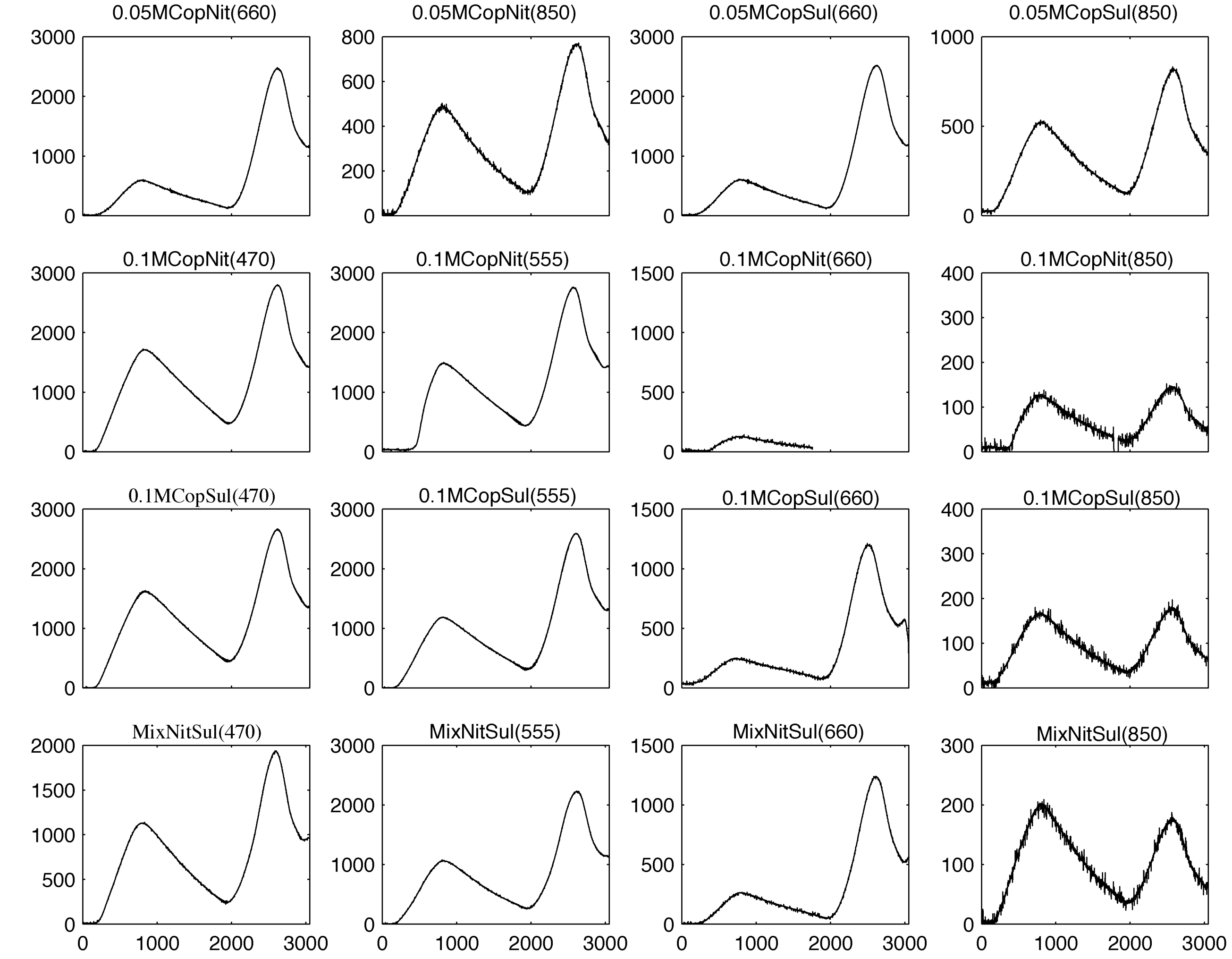

Figure 3. Optical/near-infrared modality traces for 16 liquids and mixtures.

drop has detached from the drophead. In these cases, the solutions were so highly absorbing that the signal falls to zero between the first and second main peaks, termed the rainbow and tensiograph peaks, and the trace is automatically stopped. The 0.1 M copper nitrate samples have also less runs in the infrared region as we were working right on the upper limit of detection (the software assumes when all light is absorbed by the sample that the drop has been detached and automatically starts acquiring data again). Where other runs are missing it is because the samples may have contained bubbles, or vibrations in the lab made the samples nonrepresentative.

In all cases shown here, each trace was defined by approximately 3100 values. To have the same length, all the traces were truncated at a common number of values (i.e. 3053) except 0.1 M copper nitrate samples at 660 nm which have only around 1700 data values for the above-mentioned reason. Note also that the (vertical, intensity) scales are different.

The development of the trace is a function of the buildup of the drop, until it falls off the drop-head. The development of the trace is also a function of the tensile and viscosity properties of the liquid. An important consideration in operational use is to exclude other unwanted functional dependencies, such as liquid volume, temperature, dust, electrostatic forces, and instability through jitter.

We want to effectively discriminate between different signals, based on overall shape of the signal. Therefore, background or continuum components will have to be discounted, as will the values in boundary (start, end of the trace) regions. The “effective” length of the trace is important. “Superimposed” peaks are important. Intensity scale is considerably less important and so should be relativized (or normalized out). In addition to having a robust and stable analysis procedure, a central concern is that the feature selection be computationally efficient and as far as possible automated.

The relativizing of background, and of intensity scale, is very well handled by the resolution scale components yielded by a wavelet transform method. So, too, is the robustness of the analysis.

In this paper we develop an innovative and very effective analysis pipeline consisting of the following analysis methods:

a. Wavelet transform, to select out informative constituent resolution scales. These scales are not overly noisy, nor overly smooth.

b. On the basis of these informative resolution scales, we apply a Fourier transform, to provide an invariant characterization of signal curve shapes. Together the Fourier-wavelet processing allows us to define a set of features that characterize each signal.

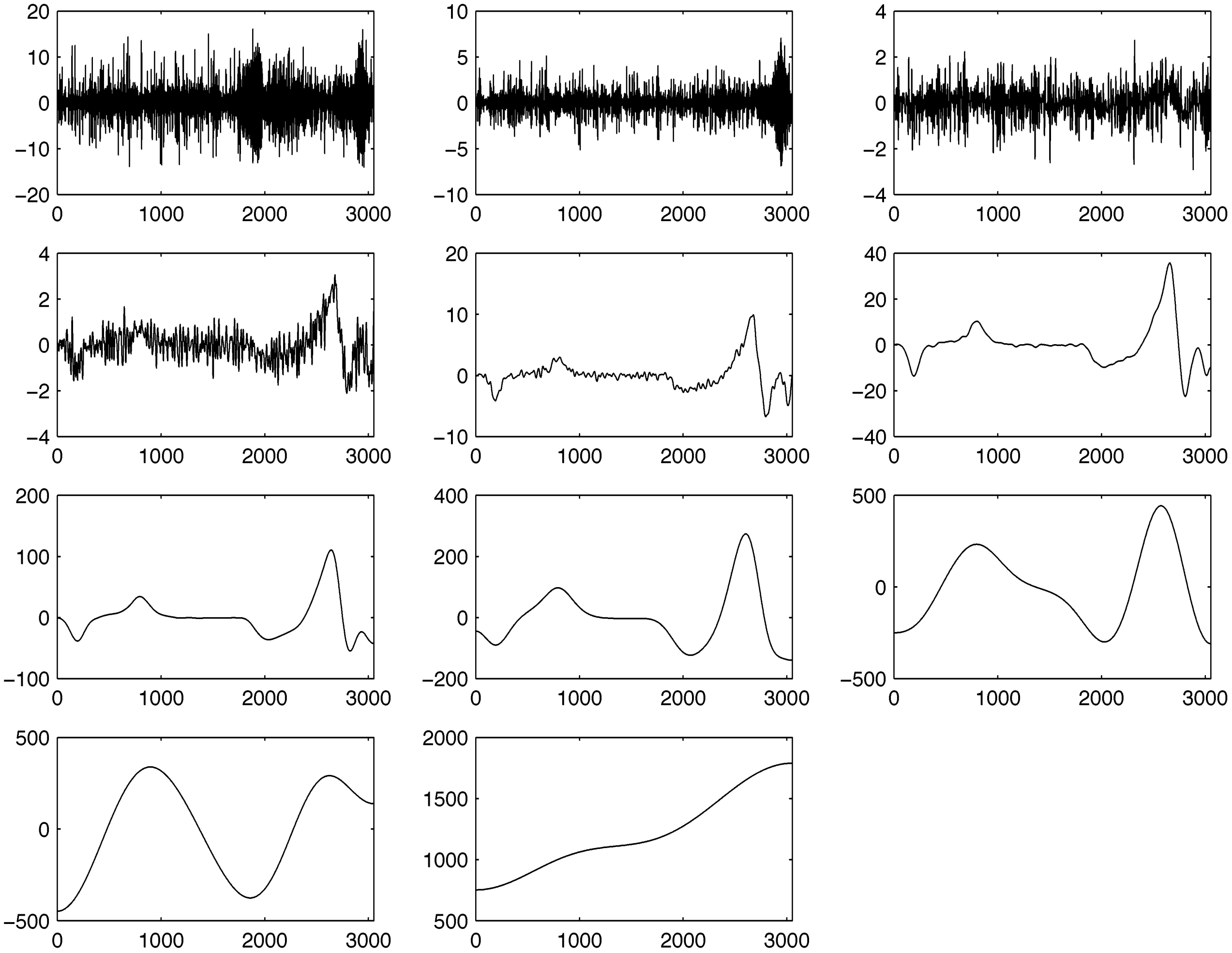

Figure 4. Wavelet transform of fifth (0.1 M copper nitrate at 470 nm) trace in Figure 3.

c. The frequency values resulting from the Fourier-wavelet processing are input to principal components analysis, to produce (i) the most informative latent variables (i.e., the principal components) in our signals and (ii) subsequently permit visualization of our results.

3. ANALYSIS DESCRIPTION AND DISCUSSION

3.1. Wavelet Transform. Figures 4−6 show an 11-level wavelet transform analysis for traces 5 (0.1 M copper nitrate at 470 nm), 9 (0.1 M copper sulfate at 470 nm), and 13 (a mixture of 0.1 M copper nitrate and 0.1 M copper sulfate at 470 nm), respectively. Wavelet resolution scale 1 is in the upper left in all cases, and the sequence of wavelet scales are arranged from upper left to lower right. The final, low middle, panel in all cases is the smoothed "residual" or DC component in the data. The residual or DC component is an overall background or continuum component. Note how all other panels show signals which move around an intensity of zero. In fact, in all such cases, the mean intensity value of the wavelet coefficients is, by construction, zero. It is important to realize that the pixel-wise addition of the wavelet component values will allow exact recreation of the input trace. Thus the addition of the signals shown in Figure 4 leads to exact reconstruction of the fifth (from upper left) trace in Figure 3.

The wavelet transform used is the redundant à trous transform, using a B_3 spline scaling function which results in a wavelet function very similar to a so-called Mexican hat function. Boundaries of the signal are handled using a "mirror" approach. The 11 levels used in all cases are determined by the user: we chose this as the maximum number of levels compatible with the 2-fold dilation of the scaling and wavelet functions at each level ($2^{11} < 3053$). Further background on, and applications of, this transform can be found in ref 4.

The wavelet scales provide information on the data which is related to features at varying resolution. The information obtained by the resolution scales 1−6 and 10 to 11 correspond to high frequency and low frequency information, respectively (see Figures 4−6). As such, these may not provide useful discriminating features and therefore were omitted from further analysis. We found that resolution scales 7−9 bring useful insights to the analysis.

3.2. Fourier Transform. A simple characterization of the resolution scale is in terms of energy or variance. This is very closely related also to the entropy of the resolution scales.[5] Murtagh et al.[6] used variance as a potentially good descriptor of each resolution scale. Here, to exploit the shape of the resolution scale better, instead, a feature vector is used to represent each resolution scale. The calculation of this feature vector can be based on various methods. Here, we applied the Fourier transform to each of those three resolution scales mentioned above and used the magnitudes to compose the feature vector representing each scale. Using the mag-

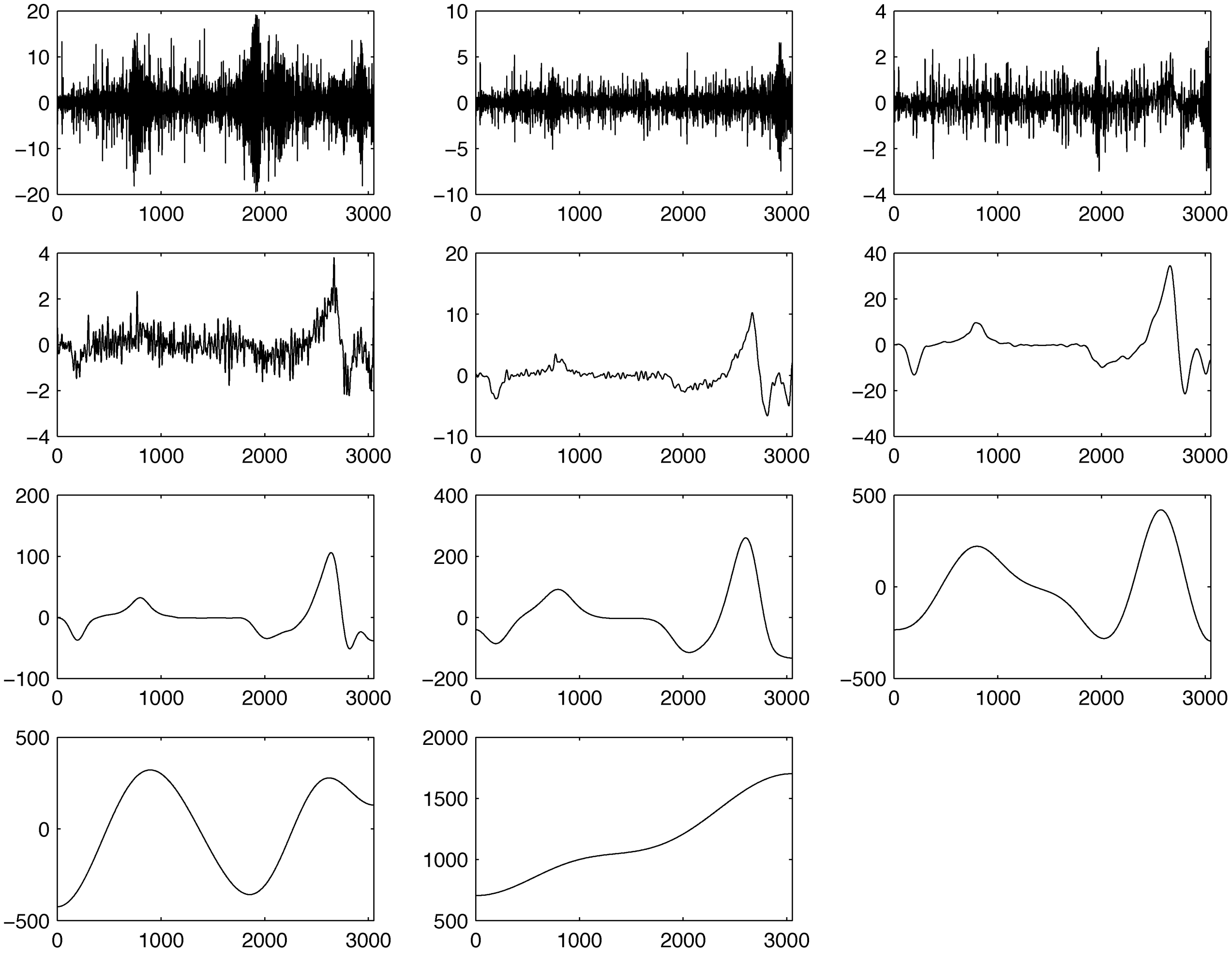

Figure 5. Wavelet transform of ninth (0.1 M copper sulfate at 470 nm) trace in Figure 3.

nitudes of the Fourier transform is of benefit due to the property of being shift invariant. Thus, the shift of the signal, i.e., the different starting and ending positions, in the tensiotrace (and therefore also the wavelet scales), is handled by the transform very well. To smooth out the edges, before the Fourier transform, the signal was multiplied by the Hamming window.

In particular, we used magnitudes of the first 20 Fourier coefficients to form the feature vector for scale 7, 15 coefficients for scale 8, and 10 coefficients for scale 9. The number of coefficients used was chosen according to the frequency range of each resolution scale. Figure 7 shows absolute values of the first 20 coefficients of the Fourier transform when applied on the selected wavelet scales (i.e. scales 7, 8, and 9) for 0.05 M copper nitrate at 660 nm, 0.1 M copper nitrate at 470 nm, 0.1 M copper sulfate at 470 nm, and a mixture of 0.1 M copper nitrate and 0.1 M copper sulfate at 470 nm. These three feature vectors were joined together to compose the final feature vector for each of the 614 traces. Final feature vectors were visualized by using principal components analysis.

3.3. Principal Component Analysis. Figure 8 shows the principal plane. Due to the many overlapping positions, we have labeled the locations with liquid names. Although throughout the analysis pipeline we took into account the reference water samples as well, we did not plot them in the principal plane for the sake of clarity.

In Figure 8, the traces of 0.05 M copper nitrate and 0.05 M copper sulfate at 660 nm are in the top left corner, away from the others, while the 0.05 M copper nitrate and 0.05 M copper sulfate at 850 nm are in the middle overlapping each other. 0.1 M copper nitrate and 0.1 M copper sulfate at 470 nm are toward the left side of the figure, and a mixture of traces formed from these traces are placed close to them to their right. 0.1 M copper nitrate and 0.1 M copper sulfate at 555 nm are in the left corner and in the middle on the left side, respectively. A mixture of 0.1 M copper nitrate and 0.1 M copper sulfate at 555 nm is located close to the 0.1 M copper sulfate at 555 nm.

The first principal component projection of the traces of 0.1 M copper sulfate and 0.1 M copper nitrate at 660 nm are in the middle and top left corner, respectively. Note that the latter samples are truncated in the original tensiotraces and do not have the same length as the others. A mixture of them at 660 nm are also positioned just under the samples of 0.1 M copper nitrate at 660 nm in the middle top.

The other traces, namely 0.1 M copper sulfate and 0.1 M copper nitrate at 850 nm, and a mixture of them are very closely positioned on the right side of the figure.

The solutions copper nitrate and copper sulfate are highly colored and absorb light strongly in the 660 nm and 850 nm regions. The absorbance falls until at 470 nm it reaches a negligible value.

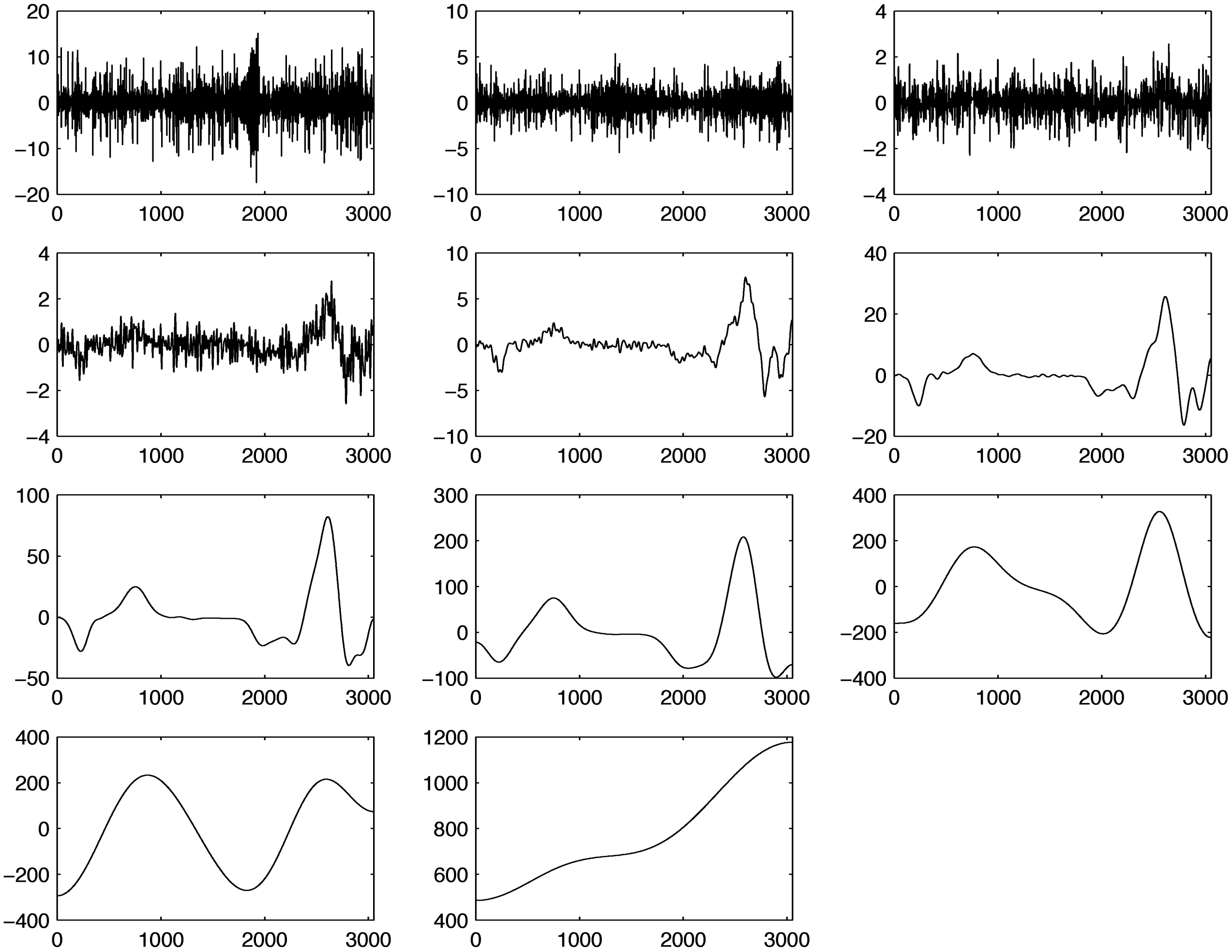

Figure 6. Wavelet transform of 13th (a mixture of 0.1 M copper nitrate and 0.1 M copper sulfate at 470 nm) trace in Figure 3.

In the region where light is strongly absorbed, it is this light absorbance which causes the predominant changes in the tensiotrace. In the principal component analysis, it can be seen that the values for copper nitrate, copper sulfate, and the mixture of them at 660 nm and 850 nm are very similar. It is possible that the light absorbance “masks” other changes to the traces in this region.

At 555 nm, there is still a small degree of light absorbance, but the principal component analysis shows significant differences. In this region, changes to the tensiotrace arise both from absorbance changes and other physical properties of the liquid (e.g. density, surface tension, and refractive index). The differences in the principal plane are larger because of this double contribution.

At 470 nm, the light absorbance is negligible, but differences within the principal plane indicate the changes to the tensiotrace which arise from the physical properties.

3.4. Conclusion on Analysis. Some open issues to be addressed are as follows.

Denoising of the traces, which can be carried out very well in wavelet space, may be an option to consider. However from Figure 3, we do not see any need for it, given these data. Noise modeling is comprehensively supported in the MR/1 software environment.[7]

The first principal component dominates (94% of the variance) so a single tensiograph indicator is easily proposed.

Overall, our wavelet transform, Fourier transform, and eigen-analysis approach seems to respect all of the desired properties of our analysis and to provide a flexible and sensitive environment for interpretation and later specification of indicators.

3.5. The Ultrasound Modality. To see how this approach works with a different sensor modality, this time we make use of a small data set resulting from the ultrasonic sensor. The data related to 20 traces from Coke, Diet Pepsi, Fosters lager, Heineken lager, Hofmeister lager, water, a mixture of 50% Heineken and 50% water, a mixture of 90% Heineken and 10% water, and a mixture of 50% Fosters and 50% Hofmeister. Each trace has two samples except water which was represented by four samples. The examples of the traces are shown in Figure 9. In all cases shown here, there were 3000 values.

Figure 10 shows the principal plane. In this figure, the water traces are in the top left corner, well away from the others; only Diet Pepsi traces are relatively closer but still well separated. Coke traces are down in the middle. Heineken traces are in the top right corner. The other beers, namely Fosters and Hofmeister, are very closely positioned on the right side of the figure. The first principal component projection of the traces of mixture of 50% Heineken and 50% water are positioned almost halfway between the two traces, i.e., Heineken and water. The mixtures of 90% Heineken and 10% water traces are almost overlapping with

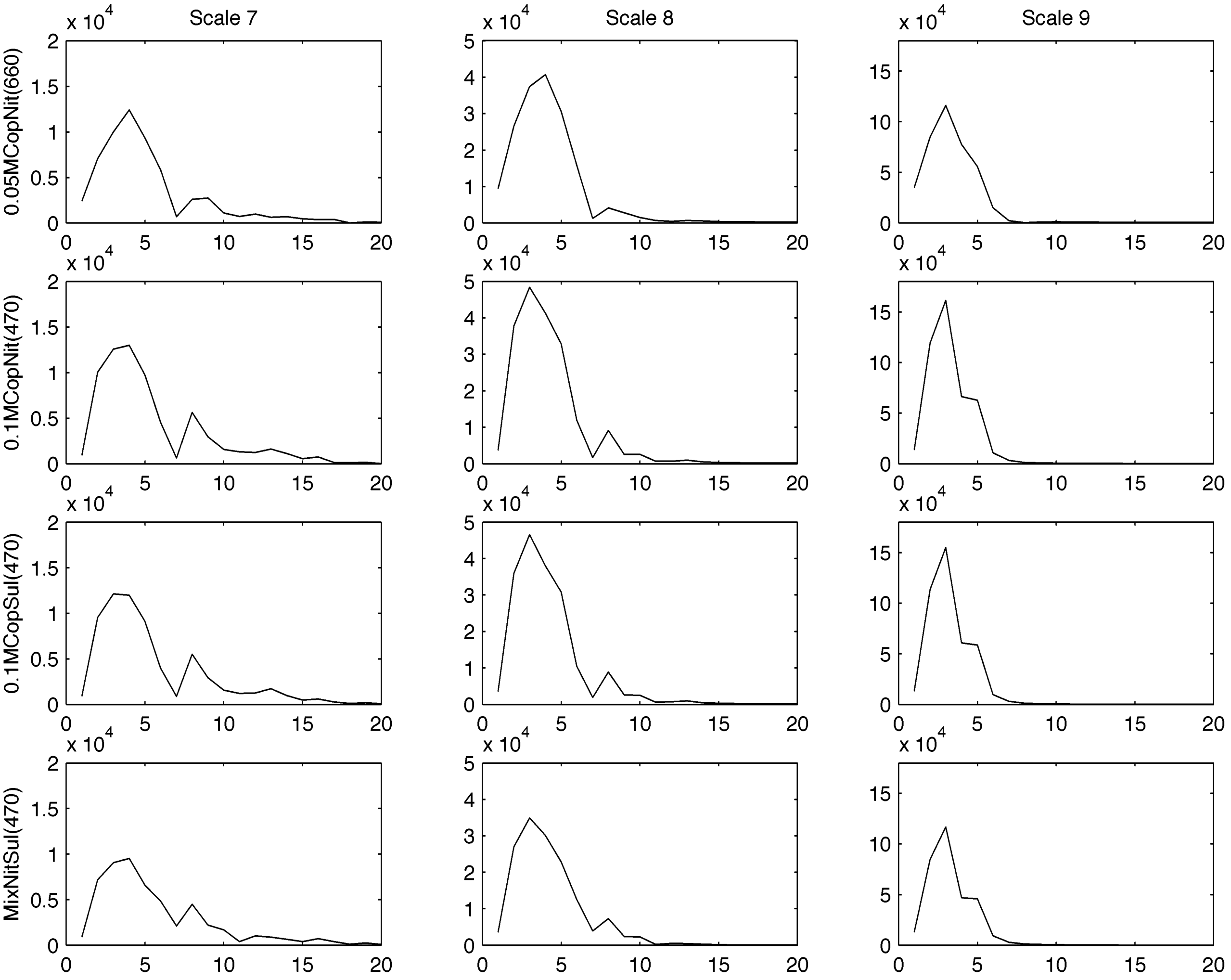

Figure 7. The absolute values of the Fourier coefficients of scale 7, 8, and 9.

Heineken traces. A mixture of 50% Fosters and 50% Hofmeister traces are also very closely located to the two beers from which this mixture is composed.

3.6. Underlying Physical Variables. The tensiograph sensor relies on changes in physical properties which are brought about by the chemical properties. There are a number of such physical and optical properties including viscosity, color, turbidity, density, surface tension, and so on. An innovative information analysis of measurement systems has been described in Tienan et al.[8] A pivotal aspect of this analysis is to arrive at the minimal number of independent underlying physical variables which characterize the liquids under investigation.

The analysis pipeline described in this paper has a very clear answer to this question of the inherent number of physical variables which describe this sensor system: in Figure 8 and elsewhere, we see that principal component 1 accounts for over 94% of the variance, and cumulatively with principal component 2 accounts for over 99%. Therefore, based on the feature set used, we can conclude that two latent variables account for the information content of our data sets.

4. THE RELATIONSHIP BETWEEN ENERGY, NOISE, ENTROPY, AND SCALE

In this section, we will present a short review of the relationship between energy and entropy, in the context of signal and noise at varying resolution scales. Its relevance is (i) if we wish to consider energies as an alternative to the Fourier magnitudes and (ii) if it is necessary to noise-filter the signals prior to analysis.

Our objective in this section is to show how very close are such concepts as energy and entropy, and how they are related to noise and scale. Because of this tight linkage it is not infeasible to use, e.g., variance instead of energy. That somewhat different vantage points could be adopted for our methodology is a favorable aspect, in that it suggests that our methodology is quite robust and stable for different input data sets.

A possibility is to consider that the entropy of a signal is the sum of the information at each scale of its wavelet transform,[4] and the information of a wavelet coefficient is related to the probability of it being due to noise. Let us look at how this definition holds up in practice. Denoting h the information relative to a single wavelet coefficient, we define

$$H(X) = \sum_{j=1}^{l} \sum_{k=1}^{N_j} h(w_{j,k}) \qquad (1)$$

with $h(w_{j,k}) = -\ln p(w_{j,k})$. l is the number of scales, and N_j is the number of samples in band (scale) j. For Gaussian noise, we get

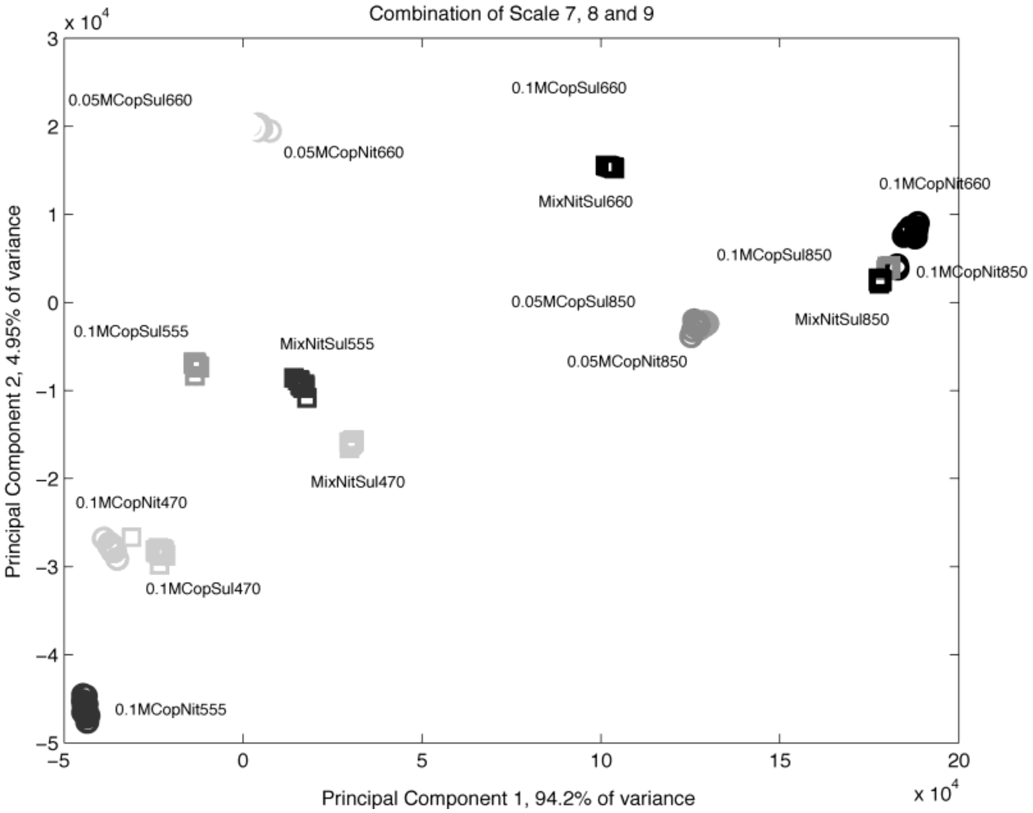

Figure 8. Principal component analysis−principal plane−of the multiresolution features (scales 7, 8, and 9) defined from the liquids.

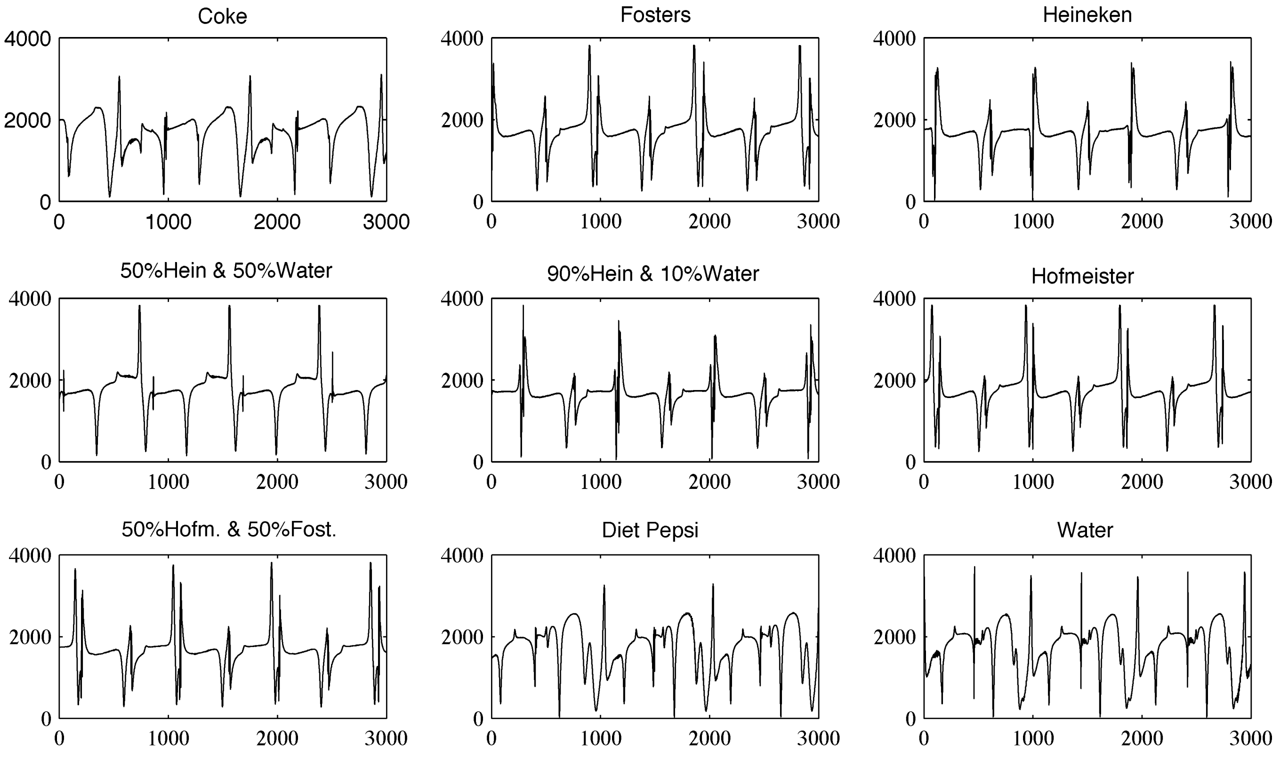

Figure 9. Ultrasound modality traces for nine beverages and mixtures.

$$h(w_{j,k}) = \frac{{w_{j,k}}^2}{2{\sigma_j}^2} + \text{const} \qquad (2)$$

where σ_j is the noise at scale j. Without loss of generality we ignore the constant. Then eq 2 in words is as follows: entropy is proportional to energy divided by noise.

Equation 1 holds if the wavelet coefficients are statistically independent, which should imply that our approach is limited to an orthogonal or biorthogonal transform. However, this limitation may be addressed through the use of the so-called cycle-spinning algorithm[9] (also named translation-invariant algorithm), which consists of performing the process of

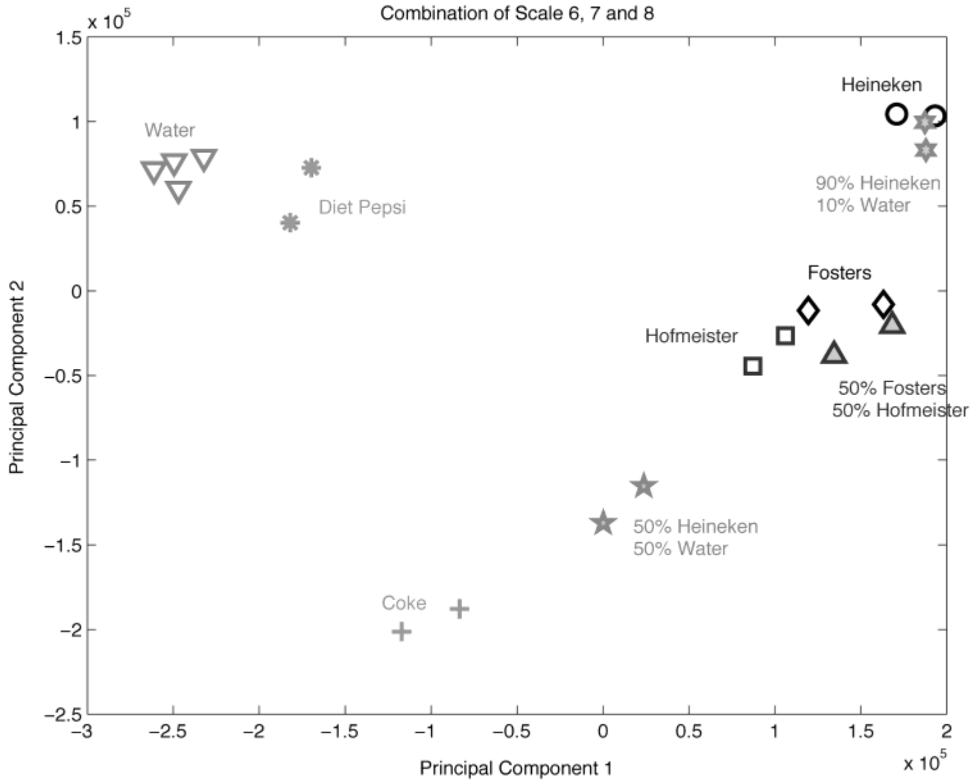

Figure 10. Principal component analysis of the traces (scales 6, 7, and 8) in the plane of principal components 1 and 2.

"transform", "denoise", and "inverse transform" on every orthogonal basis corresponding to versions of the data obtainable by combinations of circular left-right translations.

5. CONCLUSION

Our goals regarding the liquid characterization are 2-fold. First, in supervised classification, we will seek to find the best match between liquid fingerprints and to discriminate between them. Second, in unsupervised classification, a more suggestive or "browse mode" of analysis is pursued which is often termed data mining. These two objectives are very relevant for any new technology: we target specific discrimination and other problems to be solved using this new technology; and we also are open to new, serendipitous uses of the technology.

The processing pipeline described in this article consists of the use of a redundant transform to relativize signal background and to extract useful resolution scale components; a Fourier transform to yield frequency information and thereby particular shape features; and principal components analysis to provide visualizations of the signal data sets.

In addition, we discussed possibilities for enhancing this analysis in order to handle noisy input signals.

The processing pipeline has proven robust and stable over a range of sensor modalities. Visual light and ultrasound modalities were used in this article. The analysis techniques are also scalable over input signal lengths.

The output visualization provides a demonstrably valid representation of our data. We have shown how such visualization helps greatly in regard to our objectives of data interpretation and decision support.

ACKNOWLEDGMENT

This work was supported by the Fifth Framework project "Aqua-Stew−Water Quality Surveillance Techniques for Early Warning by Interface Sensors".

REFERENCES AND NOTES

(1) McMillan, N. D.; Finlayson, O.; Fortune, F.; Fingleton, M.; Daly, D.; Townsend, D.; Mcmillan, D. D. G.; Dalton, M. J. The fibre drop analyser: a new multianalyser analytical instrument with applications in sugar processing and for the analysis of pure liquids. *Measurement Sci. Technol.* **1992**, *3*, 746−764.

(2) McMillan, N. D.; Lawlor, V.; Baker, M.; Smith, S. From stalagmometry to multianalyser tensiography: the definition of the instrumental, software and analytical requirements for a new departure in drop analysis. In *Drops and Bubbles in Interfacial Research*; Möbius, D., Miller, R., Eds.; Elsevier: 1998.

(3) McMillan, N. D.; Riedel, S.; McDonald, J.; O'Neill, M.; Whyte, N.; Augousti, A; Mason, J. A Hough transform inspired technique for the rapid fingerprinting and conceptual archiving of multianalyser tensiotraces. *Irish Machine Vision Conference Proceedings*; 1999; pp 330−346.

(4) Starck, J. L.; Murtagh, F.; Bijaoui, A. *Image and Data Analysis: The Multiscale Approach*; Cambridge University Press: 1998.

(5) Starck, J. L.; Murtagh, F. Multiscale entropy filtering. *Signal Processing*; 1999; pp 147−165.

(6) Murtagh, F.; Starck, J. L.; McMillan, N. D.; Campbell, J. G. Intelligent data modeling based on the wavelet transform and data entropy. *Data Analysis: Scientific Modeling and Practical Applications*; Gaul, W., Opitz, O., Schader, M., Eds.; Springer-Verlag: 2000; pp 273−284.

(7) MR/1 and MR/2 Multiresolution Software Environment 1999 (Multi Resolutions Ltd., http://www.multiresolution.com).

(8) Tienan, K.; Riedel, S.; McMillan, N.; Kennedy, D.; Augousti, A.; Mason, J.; Doyle, G. Design from chaos-applying data entropy methods for complex system design. In *Opto-Ireland 2002: Optics and Photonics Technologies and Applications*; Glynn, T. J., Ed.; Proc. SPIE Vol. 4876, SPIE: Bellingham, 2002; in press.

(9) Donoho, D. L.; Coifman, R. R. Translation-invariant de-noising. *Wavelets and Statistics*; Antoniadis, A., Oppenheim, G., Eds.; Springer-Verlag: 1995.

CI025601J