Talanta 87 (2011) 46–52

Contents lists available at SciVerse ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Wavelets and genetic algorithms applied to search prefilters for spectral library matching in forensics

Barry K. Lavine[a,*], Nikhil Mirjankar[a,*], Scott Ryland[b], Mark Sandercock[c]

[a] Department of Chemistry, Oklahoma State University, Stillwater, OK 74078, USA
[b] Florida Department of Law Enforcement Regional Crime Laboratory, 500 West Robinson Street, Orlando, FL 32801, USA
[c] Royal Canadian Mounted Police Forensic Laboratory, 15707-118th Avenue, Edmonton, Alberta, T5V 1B7, Canada

ARTICLE INFO

Article history:
Received 5 August 2011
Received in revised form
20 September 2011
Accepted 21 September 2011
Available online 6 October 2011

Keywords:
Spectral pattern recognition
Wavelets
Search prefilters
PDQ database
Forensic paint analysis

ABSTRACT

Currently, the identification of the make, model and year of a motor vehicle involved in a hit and run collision from only a clear coat paint smear left at a crime scene is not possible. Search prefilters for searching infrared (IR) spectral libraries of the paint data query (PDQ) automotive database to differentiate between similar but nonidentical Fourier transform infrared (FTIR) paint spectra are proposed. Applying wavelets, FTIR spectra of clear coat paint smears can be denoised and deconvolved by decomposing each spectrum into wavelet coefficients which represent the sample's constituent frequencies. A genetic algorithm for pattern recognition analysis is used to identify wavelet coefficients for underdetermined data that are characteristic of the model and manufacturer of the automobile from which the spectra of the clear coats were obtained. Even in challenging trials where the samples evaluated were all the same manufacturer (Chrysler) with a limited production year range, the respective models and manufacturing plants were correctly identified. Search prefilters for spectral library matching are necessary to extract investigative lead information from a clear coat paint smear; unlike the undercoat and color coat paint layers, which can be identified using the text based portion of the PDQ database.

1. Introduction

Paint samples often are recovered from hit-and-run collisions where damage to vehicles or injury or death to a pedestrian has occurred. The Royal Canadian Mounted Police has shown that automobiles can be identified by comparing the color, layer sequence and chemical composition of each individual layer of recovered automotive paint [1,2]. To make these comparisons possible, a comprehensive database called the paint data query (PDQ) database was developed, as well as a means of searching and retrieving information from it [3–6]. Currently, PDQ contains over 16,000 samples that correspond to over 60,000 individual paint layers, representing the paint systems used for most domestic and foreign vehicles marketed in North America. Each year approximately 500 samples are painstakingly collected, analyzed, and added to the PDQ database.

Automotive paints [7] consist of three or four layers: a clear coat over a color coat, which in turn is over one or more undercoats. With the exception of the clear coat, each layer contains pigments and fillers, and all layers contain binders (the glue that holds the layer together). Automotive manufacturers tend to use unique combinations of pigments and binders in each layer of paint. It is this unique combination that allows forensic scientists to determine the manufacturer, model, and year for a vehicle from a paint chip left at a crime scene.

The analytical method used to identify automotive paint relies on the selective absorption of infrared light by the components in the paint. For infrared analysis, each paint layer is separated, placed between two diamonds [8] and subjected to the infrared light beam of an infrared spectrometer. A transmission IR spectrum of each layer is thereby obtained. The comparison of the IR spectrum of each paint layer in a paint system (clear coat, color coat, and undercoats or primers) to spectra in the paint database allows the manufacturing plant at which the paint system was applied, and therefore the model of the vehicle, to be identified.

PDQ contains information about the physical attributes (i.e. color and layer sequence) and the chemical composition of each layer of the original manufacturer's paint system. It also contains digital libraries of IR spectra for each of those layers. The text-based portion (physical attributes) of PDQ is comprised of two components: (1) the data which contains the complete layer sequence's colors, general binder compositions, extender pigment (inorganic filler) compositions, and sourcing information on known paint systems, and (2) search and retrieval software to generate the hit list. To use the program, the forensic scientist first translates the chemical formulation of each of the recovered paint chip's layers from its IR spectrum into specific text codes. Next, the scientist constructs

* Corresponding authors. Tel.: +1 405 744 5945; fax: +1 405 744 6007.
E-mail address: bklab@chem.okstate.edu (B.K. Lavine).

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.09.039

Table 1
Clear coat paint data set.

	Plant	Code	Vehicle model	Number of spectra
1	Bramalea, Canada	BRA	[a]Intrepid, Concorde, LHS and 300 M (1999)	25
2	[b]St. Louis, USA	STL	Dodge Ram Trucks (1999–2000) Chrysler/Plymouth SUV's (1999)	21
3	Jefferson North Plant, USA	JFN	Jeep Cherokee (1999)	13
4	Sterling Heights, USA	STH	Dodge Stratus (1999)	9
5	Saltillo, Mexico	SAL	Dodge Ram Trucks (1999)	12
6	Newark, USA	NEW	Durango SUV (1999)	11

[a] Chrylser Intrepid, Chrysler Concorde, Chrysler LHS and Chrysler 300 M are mechanically similar vehicles as they are interrelated models. Both the Concorde and Intrepid are built on the same identical (LH) platform.
[b] St. Louis plant has two distinct production lines: Dodge Ram Trucks (North Plant) and Chrysler/Plymouth SUV's (South Plant)).

a search query based on the color, chemical text codes, and layer sequence of the unknown paint chip left at the crime scene. The software searches the database, comparing all records of make, model and years having a paint system similar to the coded (text based) information being searched. The final step in this process is confirming the database hits by manually comparing IR spectra of each unknown paint layer against the library spectra of each layer identified in the database hit list. Topcoat color is compared to topcoat color charts to narrow down the hit list so that only those manufacturers known to have used a similar topcoat color in the years indicated by the database search are reported.

Modern automotive paints use thinner undercoat and color coat layers protected by a thicker clear coat layer; consequently, a clear coat paint smear will often be the only layer of paint left at the crime scene. Modern clear coats applied to any painted metal parts have only one of two possible formulations (i.e., they are coded as either acrylic melamine styrene, or acrylic melamine styrene polyurethane). There are no inorganic fillers or color with which to further discriminate a clear coat sample. In these cases, the PDQ database cannot be used to identify the motor vehicle because the PDQ search relies heavily on the relatively large variations in color and the chemical formulation (inorganic filler) used. Direct searching of FTIR spectra in the database does not exist, and commercial spectral search algorithms have not been able to distinguish subtle but significant features in the spectra such as shoulders, unique shapes and patterns and minor peaks. Most commercial search algorithms involve some form of numerical comparison between the spectrum of an unknown and each library spectrum [9]. However, these algorithms are unable to handle peak shifting and ignore peaks of low intensity, which may be informative [10,9]. The failure of the text based portion of PDQ to identify clear coat paint smears and the inability of the database to accurately search IR spectra are significant limitations to the current implementation of the PDQ database.

By using search prefilters [11,12], many of the problems encountered in spectral library searching of paint samples can be addressed. A prefilter is a quick test to identify library spectra that are dissimilar to the unknown. Prefilters allow for more sophisticated and correspondingly more time-consuming algorithms to be used for spectral matching since the size of the library can be culled down for a specific match. However, information contained in a search prefiliter should be based on the relationship between the composition of the clear coat layer and the manufacturer and model of the vehicle. The exceptionally high quality of the FTIR data in the spectral libraries associated with the PDQ database, and the comprehensiveness of this database, makes it an excellent source of data for the development and subsequent validation of search prefilters.

To develop search prefilters for spectral library matching, a two-step procedure has been developed as part of this study. First, IR spectra of clear coats in the PDQ library are preprocessed using wavelets [13,14] to enhance subtle but significant features in the data. Second, wavelet coefficients characteristic of the model and manufacturer of the vehicle are identified using a genetic algorithm (GA) for pattern recognition analysis [15–25] that employs both supervised and unsupervised learning to identify wavelet coefficients that optimize the separation of the spectra by manufacturer and model type in a plot of the two or three largest principal components of the data. Because principal components [26] maximize variance, the bulk of the information encoded by the wavelet coefficients selected by the pattern recognition GA is about differences between manufacturer and model type of the vehicle. This fitness criterion will reduce the size of the search space since it limits the search to coefficients whose PC plot shows separation of the samples on the basis of manufacturer or model type. In addition, the GA focuses on the classes and/or samples that are difficult to classify as it trains by boosting the relative importance of the classes and samples that consistently score poorly. Over time, the algorithm learns its optimal parameters in a manner similar to that of a neural network. The pattern recognition GA integrates aspects of artificial intelligence and evolutionary computations to yield a "smart" one-pass procedure for feature selection and pattern classification.

2. Experimental

IR spectra of the clear coat paint layers were collected using a BioRad 40A or BioRad 60 equipped with a DTGS detector. Each clear coat sample was between 3 μg and 4 μg, and was run from 4000 to 200 cm^{-1} between diamond windows [27,28]. Further details about the infrared and sampling conditions used can be found elsewhere [29].

IR spectra of paint samples from six Chrysler plants selected for pattern recognition analysis (see Table 1) were obtained from the PDQ database. Each plant (BRA, STL, JFN, STH, SAL, and NEW) was represented by at least 10 paint samples obtained from a variety of automobile parts (see Table 2). With the exception of the STL plant, all of the paint samples were from the same production year (see Table 1). For this study, only the clear coat layer from metal parts was used. The IR spectra obtained from the PDQ database were divided into a training set of 88 spectra (see Table 3) and a validation set of 3 spectra (see Table 4). The samples comprising the validation set were chosen by random lot.

Table 2
Automobile parts used in the data set.

Part	Number of samples
Roof	68
Hood	9
Fender	10
Door	2
Hatchback	1
Trunk	1

Table 3
Training set.

	Plant	Number of spectra
1	Bramalea (BRA)	23
2	St. Louis (STL)	21
3	Jefferson North Plant (JFN)	13
4	Sterling Heights (STH)	9
5	Saltillo (SAL)	11
6	Newark (NEW)	11

Table 4
Validation set.

Sample (PDQ number)	Manufacturing plant
M0057OT2	Bramalea (BRA)
W0001OT2	Bramalea (BRA)
P0093OT2	Saltillo (SAL)

Each clear coat IR spectrum was normalized to unit length. The wavelet packet tree (WPT) was applied to each normalized spectrum using the MATLAB Wavelet toolbox 3.0.4 (MathWorks, Natick, MA). WPT [30] was implemented by iteratively passing each spectrum through pairs of wavelet filters. This filtering process was allowed to continue until the required level of decomposition was achieved to give what is known as a wavelet packet tree.

Wavelet coefficients from all nodes in the tree for each clear coat spectrum were organized as a data vector. For pattern recognition analysis, each wavelet coefficient was autoscaled such that it had a mean of zero and a standard deviation of one. Autoscaling removed inadvertent weighting of the coefficients that would otherwise have occurred due to differences in magnitude among the wavelet coefficients, thereby ensuring that each wavelet coefficient had an equal weight in the pattern recognition analysis.

2.1. Genetic algorithm for pattern recognition analysis

Software to implement the pattern recognition GA was compiled in JAVA. The number of chromosomes in the initial population ϕ was set to 10,000. A large number of chromosomes were used because each spectrum was represented by 1944 points or 16,362 wavelet coefficients. The chromosomes comprising the initial population (i.e., the population at generation 0) were generated at random to minimize potential bias. Each chromosome represents a potential solution to the feature selection problem, i.e., a unique subset of wavelet coefficients. During each generation, the chromosomes are sent to the fitness function for evaluation. Each chromosome is assigned a value by the fitness function, which is a measure of the quality of the proposed feature subset for the classification problem. The fitness function of the pattern recognition GA (known as PCKaNN) emulates human pattern recognition through machine learning to score the principal component plots and thereby identify a set of wavelengths or wavelet coefficients that optimize the separation of the classes in a plot of the two or three largest principal components of the data.

To facilitate the tracking and scoring of the principal component plots, class and sample weights, which are an integral part of the fitness function, are computed (see Eqs. (1) and (2)) where CW(c) is the weight of class c (with c varying from 1 to the total number of classes in the data set). $SW_c(s)$ is the weight of sample s in class c. The class weights sum to 100, and the sample weights for the objects comprising a particular class sum to a value equal to the class weight of the class in question.

$$\mathrm{CW}(c) = 100\frac{\mathrm{CW}(c)}{\sum_c \mathrm{CW}(c)} \tag{1}$$

$$\mathrm{SW}(s) = \mathrm{CW}(c)\frac{\mathrm{SW}(s)}{\sum_{s\in c}\mathrm{SW}(s)} \tag{2}$$

Using PCKaNN, a PC plot is generated for each chromosome after the subset of features in the chromosome is extracted. The plot is scored using the k-nearest neighbor (K-NN) classification algorithm. For a given sample point, Euclidean distances are computed between it and every other point in the PC plot. These distances are arranged from smallest to largest, and a poll is taken of the point's k-nearest neighbors. For the most rigorous classification, k is equal to the number of samples in the class to which the point belongs. The sample hit count (SHC), or the number of like nearest neighbors is $0 \le \mathrm{SHC(s)} \le K_c$. ($K_c$ is a user defined term and refers to the number of k-nearest neighbors for class c). The fitness is computed using Eq. (3). Class and sample weights are adjusted during each generation through a process called boosting. Further details about PCKaNN can be found elsewhere [31–35].

$$F(d) = \sum_c \sum_{s\in c} \frac{1}{K_c} \times \mathrm{SHC}(s) \times \mathrm{SW}(s) \tag{3}$$

To understand how a principal component plot is scored, consider a data set with two classes, which have been assigned equal weights. Class 1 has 50 samples, and class 2 has 10 samples. At generation 0, the samples in a given class have the same weight. Thus, each sample in class 1 has a sample weight of 1, whereas each sample in class 2 has a weight of 5. Suppose a sample from class 2 has as its nearest neighbors 8 class one samples. Hence, $\mathrm{SHC}/K = 0.8$, and $(\mathrm{SHC}/K_c) \times \mathrm{SW} = 0.8 \times 5$, which equals 4. By summing $(\mathrm{SHC}/K_c) \times \mathrm{SW}$ for each sample, each principal component plot can be scored. One advantage of using this procedure to score the principal component plots is that a class with a large number of samples will not dominate the analysis because of the class weights.

The fitness function of the pattern recognition GA is able to focus on samples and classes that are difficult to classify by boosting their weights over successive generations. In order to boost, it is necessary to compute both the sample-hit rate (SHR), which is the mean value of SHC/K_c over all feature subsets produced in a particular generation (see Eq. (4)), and the class-hit rate (CHR), which is the mean sample hit rate of all samples in a class (see Eq. (5)). ϕ in Eq. (4) is the number of chromosomes in the population, and *AVG* in Eq. (5) refers to the average or mean value. During each generation, class and sample weights are adjusted by a perceptron (see Eqs. (6) and (7)) with the momentum, P, set by the user, g + 1 is the current generation, and g is the previous generation. Classes with a lower class hit rate are boosted more heavily than those classes that score well.

$$\mathrm{SHR}(s) = \frac{1}{\phi}\sum_{i=1}^{\phi}\frac{\mathrm{SHC}_i(s)}{K_c} \tag{4}$$

$$\mathrm{CHR}_g(c) = \mathrm{AVG}(\mathrm{SHR}_g(s) \ : \forall_{s\in c}) \tag{5}$$

$$\mathrm{CW}_{g+1}(s) = \mathrm{CW}_g(s) + P(1 - \mathrm{CHR}_g(s)) \tag{6}$$

$$\mathrm{SW}_{g+1}(s) = \mathrm{SW}_g(s) + P(1 - \mathrm{SHR}_g(s)) \tag{7}$$

Boosting is crucial for the successful operation of the pattern recognition GA because it modifies the fitness landscape by adjusting the values of the class and sample weights. This helps to minimize the problem of convergence to a local optimum. Hence, the fitness function of the pattern recognition GA changes as the population is evolving towards a solution.

In this study, the Hopkins statistic [36] has been combined with PCKaNN to incorporate transverse learning into the feature selection process. Features are selected to optimize clustering using all of the data points (training set and prediction set samples via the Hopkins statistic) and to create class separation using only the labeled

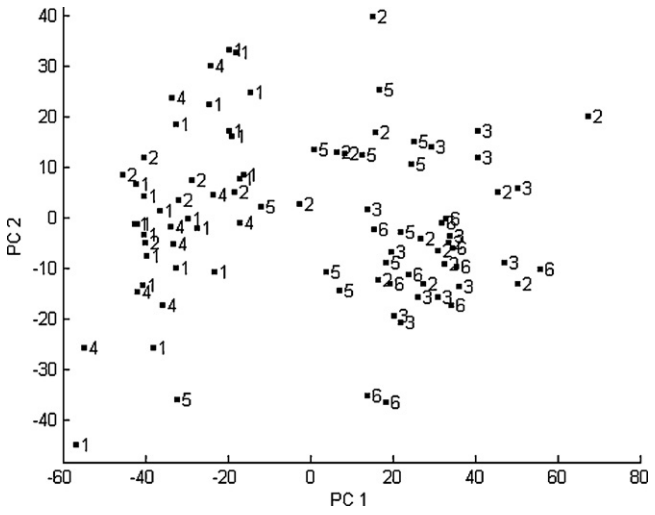

Fig. 1. Plot of the two largest principal components of the 88 IR spectra and 1944 points that comprise the training set. Each spectrum is represented as a point in the plot (1 = BRA, 2 = STL, 3 = JFN, 4 = STH, 5 = SAL, and 6 = NEW).

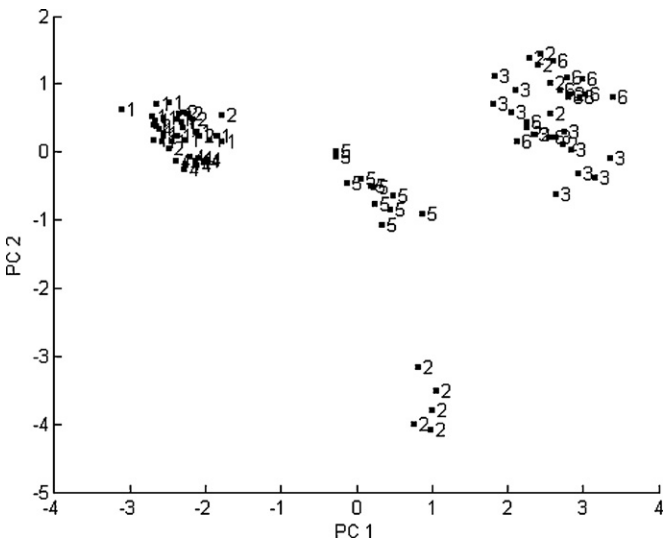

Fig. 2. Plot of the two largest principal components of the 88 IR spectra from the training set and 8 wavelengths identified by the pattern recognition GA. Each spectrum is represented as a point in the plot (1 = BRA, 2 = STL, 3 = JFN, 4 = STH, 5 = SAL, and 6 = NEW).

data points (training set samples via PCKaNN). The advantage of this approach to the one used in support vector machines [37] is that transverse learning is used not only to predict future data, but also to identify the truly informative features in the data set, thereby ensuring that a more reliable classification of the data is obtained. By varying the contribution of PCKaNN and the Hopkins statistic to the scoring of the feature subsets, it is possible to tune the fitness function of the pattern recognition GA, enabling it to explore the structure of a data set, and to uncover hidden relationships in the data by uncovering the truly informative features in the data. For underdetermined data sets (large number of features per sample and a small number of samples per class), this approach to feature selection is preferred because it will perform better than a learning model developed from a set of features whose selection is based solely on the dichotomization power of the features for samples with known responses. Further details about the use of transverse learning to select features and the coupling of the Hopkins statistic to PCKaNN can be found elsewhere [38].

3. Results and discussion

The initial focus of the pattern recognition analysis was the training set data. The overall goal of this study was to differentiate IR spectra of clear coats by manufacturing plant. The first step in the study was to apply principal component analysis (PCA) to the normalized and autoscaled IR spectral data. PCA is a powerful method for uncovering hidden relationships in complex multivariate data sets. Using this procedure is analogous to finding a new coordinate system that is better at displaying the information in the data than axes defined by the original measurement variables. The new coordinate system is linked to variation in the data. The basis vectors of this new coordinate system are the principal components of the original data. Each principal component is a linear combination of the original measurement variables. Often, only the two or three largest principal components are necessary to explain most of the information present in a spectral data set due to the large number of interrelated measurement variables.

For PCA, each sample was initially represented as a data vector, $x = (x_1, x_2, x_3, .x_j, \ldots x_{1944})$ where x_j is the absorbance of the jth point from the normalized IR spectrum. Fig. 1 shows a plot of the two largest principal components of the 88 spectra and 1944 points comprising the training set. Each spectrum is represented by a point in the map (1 = BRA, 2 = STL, 3 = JFN, 4 = STH, 5 = SAL, 6 = NEW). The overlap of the clear coats from the different manufacturing plants in the plot is evident.

The next step was feature selection. A genetic algorithm for pattern recognition analysis was used in the study to identify spectral features characteristic of the profile of each manufacturing plant. The pattern recognition GA identified features by sampling key feature subsets, scoring their principal component plots, and tracking clear coat samples or plants that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 300 generations, the GA identified 8 wavelengths whose PC plot showed clustering on the basis of Plant ID (see Fig. 2). Plant 5 is well separated from the other manufacturing plants, whereas Plants 1 and 4 are separated from each other, and Plants 3 and 6 are not completely separated from each other. Plant 2 (STL) appears to be composed of three distinct clusters indicative of three different types of paint samples. Two of these clusters overlap with Plants 1 and 4, and Plants 3 and 6, respectively. The principal component map of these 8 spectral features suggests that information about manufacturing plant is present in the IR spectra of the clear coats.

A PC plot of only the STL clear coat samples was generated to better understand the nature of the clustering. Fig. 3 shows a plot of the two largest principal components of the 21 IR spectra and 1944 points. Clustering of the paint samples in three distinct groups is again evident, which corresponds to the clustering shown by STL samples in the original six-class study. Each cluster has a distinctive IR spectrum (see Fig. 4). Group A corresponds to SUV's (Plymouth Voyager, Dodge Grand Caravan, and Chrysler town and country), whereas Groups B and C correspond to Dodge Ram trucks (1500, 2500, and 3500 for Group B and 1500 and 3500 for Group C). During the year 2000, Chrysler had made a change in the clear coat formulation used at the St. Louis North Plant. Group B samples fall under the BASF supplied Duraclear II clear coat and has a chemistry of acrylic, melamine, styrene, and polyurethane. Group C falls under the DuPont supplied Gen IV AW clear coat and has the chemistry acrylic, melamine, and styrene. From this PC plot, one can conclude that information about the model and specific production line can be obtained from an infrared spectrum of a clear coat paint smear.

STL clear coat samples were removed from the training set and the pattern recognition analysis was again repeated using the pattern recognition genetic algorithm for feature selection. Fig. 5

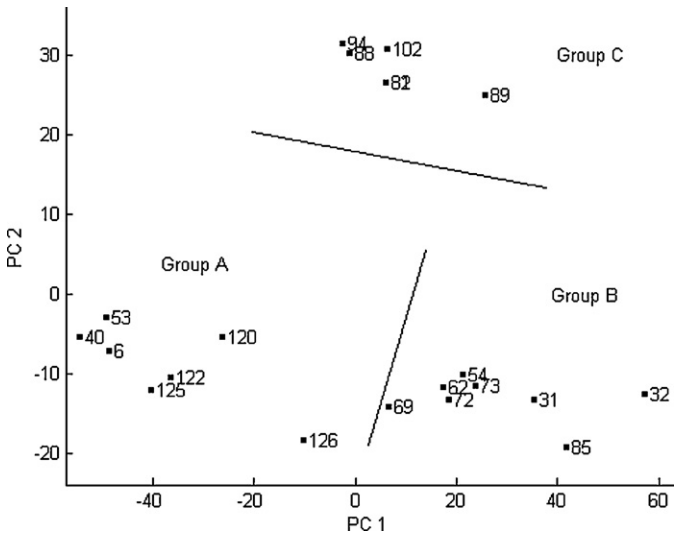

Fig. 3. Plot of the two largest principal components of the 21 IR spectra and 1944 points comprising the STL clear coat paint samples. Each spectrum is represented as by sample ID in the plot.

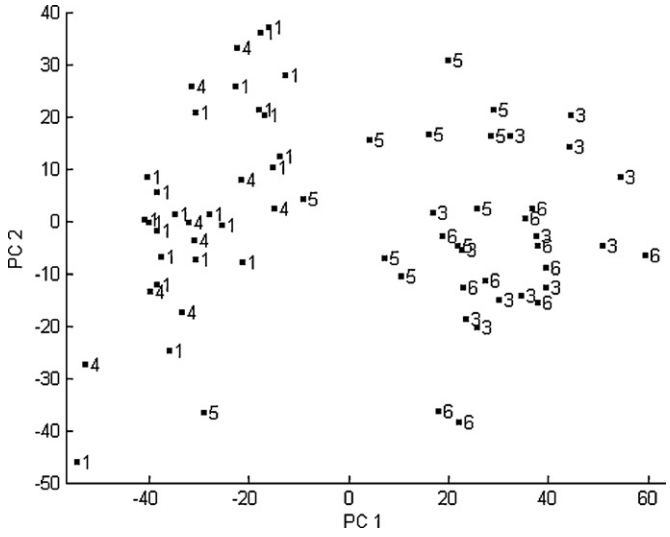

Fig. 5. Plot of the two largest principal components of the 67 IR spectra and 1944 points that comprise the training set used for prediction. Each spectrum is represented as a point in the plot (1 = BRA, 3 = JFN, 4 = STH, 5 = SAL, and 6 = NEW).

shows a plot of the two largest principal components of the 1944 features and the 67 IR spectra. Using the pattern recognition GA, 10 wavelengths were identified that contained information about the manufacturing plant of the paint samples. Fig. 6 shows a plot of the two largest principal components of the data developed from 10 wavelengths selected by the pattern recognition GA. The same trends observed in the PC plot for the larger training set (which contained STL samples) are again reported. Plant 5 is well separated from the other plants, whereas Plants 1 and 4 (BRA and STH) are separated but are adjacent to each other, and Plants 3 and 6 (JFN and NEW) overlap.

More powerful preprocessing methods are needed to extract information about manufacturing plant from the IR spectra of the clear coats to discriminate paint spectra from BRA, JFN, STH, and NEW. For this reason, the wavelet packet transform was applied to the IR data. Since the goal is deconvolution and not data

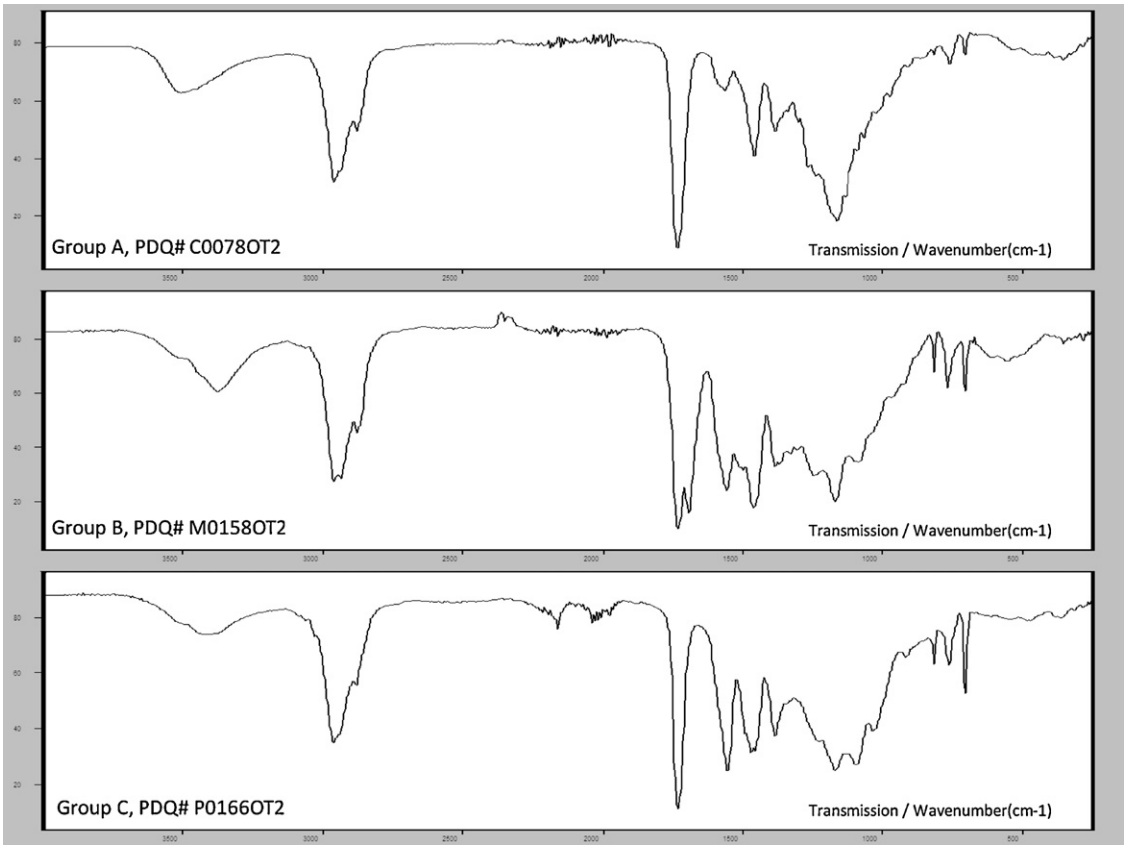

Fig. 4. Prototypical IR spectrum representative of each cluster for the STL clear coat paint samples. Variation between spectra in each cluster did not appear significant to the naked eye.

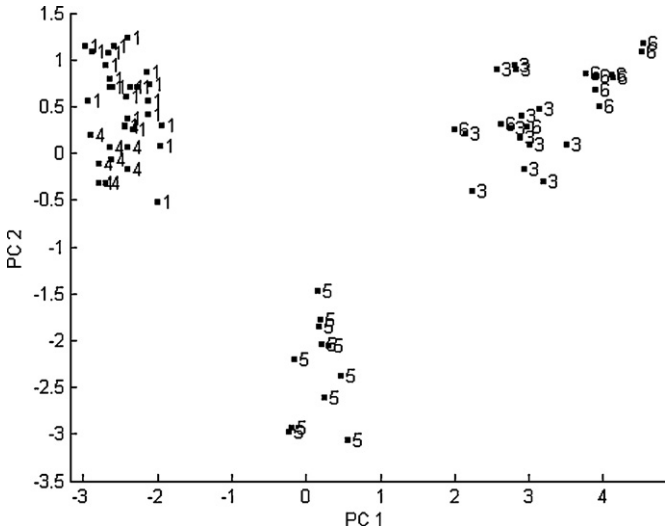

Fig. 6. Plot of the two largest principal components of the 67 IR spectra from the training set and 10 wavelengths identified by the pattern recognition GA. Each spectrum is represented as a point in the plot (1 = BRA, 3 = JFN, 4 = STH, 5 = SAL, and 6 = NEW).

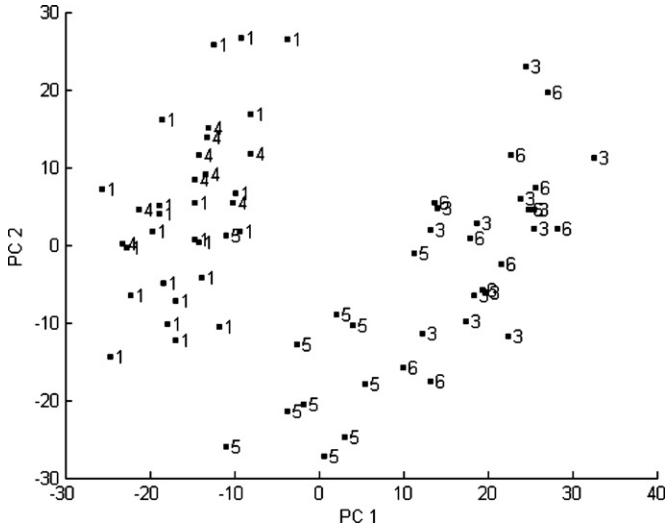

Fig. 7. Plot of the two largest principal components of the 67 wavelet transformed IR spectra and 16,362 wavelet coefficients that comprise the training set used for prediction. Each spectrum is represented as a point in the plot (1 = BRA, 3 = JFN, 4 = STH, 5 = SAL, and 6 = NEW).

compression, the number of wavelet coefficients used to initially represent each IR spectrum will be greater than the number of data points comprising each spectrum. To identify informative wavelet coefficients, it will be necessary to use the pattern recognition GA.

Using the wavelet packet transform, the Daubechies 12 mother wavelet up to the 8th level of decomposition was used to denoise and deconvolute each IR spectrum into 16,362 wavelet coefficients. Criteria used to select this mother wavelet were based on the ability of the mother wavelet to extract information about the manufacturing plant from the data, which can then be exploited using the pattern recognition GA. There was a decrease in the ability of the pattern recognition GA to correctly classify the spectra when other wavelets from the same family were used, e.g., Daubechies 6 or Daubechies 18 to denoise and deconvolute the data. This result can be explained by a well-known empirical rule often applied to guide the selection of suitable wavelets for denoising data. If the signal contains sharp peaks or discontinuities, the use of the Haar or other compact wavelets would be indicated, whereas for signals that comprise broad peaks, a smoother wavelet than the Haar should be used such as coiflets. If the signal lies between these two extremes, such as mid-IR spectra, then Daubechies are expected to give good results.

Fig. 7 shows a plot of the two largest principal components of the wavelet transformed infrared spectra. Each IR spectra was represented by 16,362 wavelet coefficients. Because this plot was uninformative, the pattern recognition GA was used to identify the so-called informative wavelet coefficients. The pattern recognition GA identified these coefficients by sampling key coefficient subsets of the data, scoring their PC plots, and tracking those classes and samples that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 100 generations, the GA identified 36 wavelet coefficients, which contained information characteristic of the manufacturing plant. Fig. 8 shows a plot of the two largest principal components of the 36 wavelet coefficients and the 67 clear coat spectra that comprised the training set. Separation of the IR spectra by manufacturing plant is evident.

A prediction set of 3 spectra was employed (see Table 4) to assess the predictive ability of the 36-wavelet coefficients identified by the pattern recognition GA. Because of the large number of features and small number of samples in the data set, there were concerns about overfitting the data, which is a problem with underdetermined data sets when feature selection methods are applied using supervised pattern recognition techniques. For this reason we used a smaller validation set – one that is less than 10% of the size of the training set. Employing transverse learning, the 3 prediction set samples were correctly classified in the principal component plot developed from the 67 training set samples and 36 wavelet coefficients. Fig. 9 shows the prediction set samples projected onto the PC map developed from the training set. Each projected sample lies in a region of the map near a paint sample with the same class label. Evidently, the pattern recognition GA can identify wavelet coefficients characteristic of the manufacturing plant of the clear coat samples.

Linear discriminant analysis (LDA) was also used to analyze the data. Loadings from the 36-wavelet coefficients identified by the pattern recognition GA were computed for each principal component. Nine-wavelet coefficients with loadings less than 0.05 on both the first and second principal components of the data were

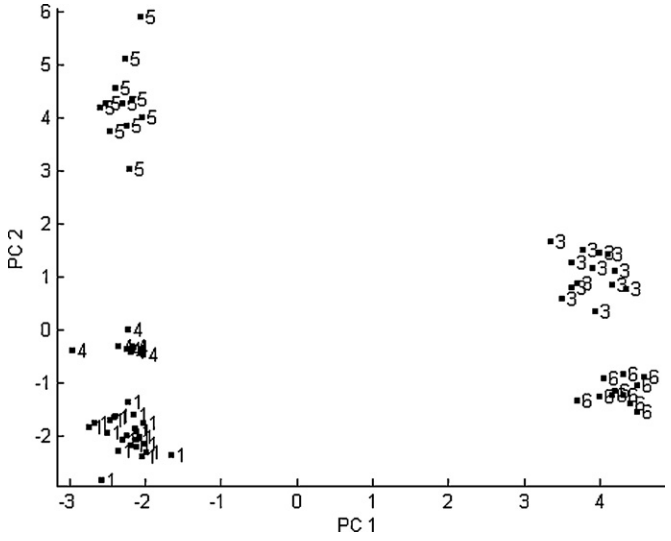

Fig. 8. Plot of the two largest principal components of the 67 wavelet transformed IR spectra from the training set and 36 wavelet coefficients identified by the pattern recognition GA. Each spectrum is represented as a point in the plot (1 = BRA, 3 = JFN, 4 = STH, 5 = SAL, and 6 = NEW).

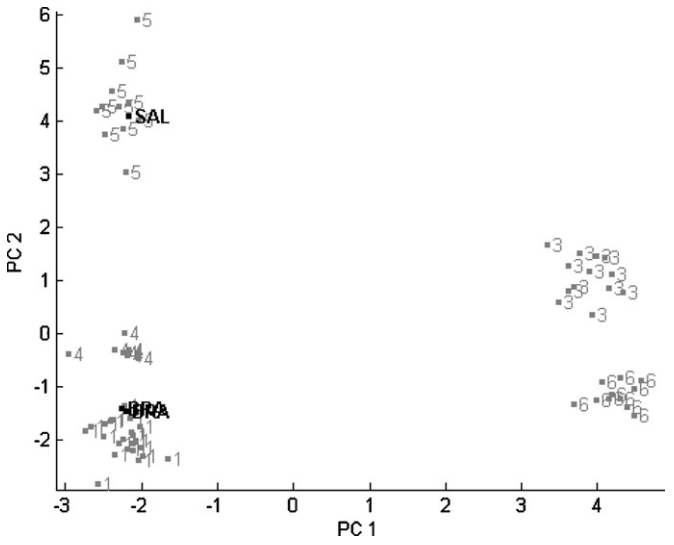

Fig. 9. Projection of prediction set samples onto the PC map developed from the 67-wavelet transformed IR spectra and 36-wavelet coefficients identified by the pattern recognition GA is shown. Each spectrum in the training set (1 = BRA, 3 = JFN, 4 = STH, 5 = SAL, and 6 = NEW) and prediction set (BRA and SAL) is represented as a point in the plot. All projected samples lie in a region of the map near paint samples with the same class label.

discarded as these coefficients did not contain information about the manufacturing plant. When LDA was applied to the 67 training set samples and 27 wavelet coefficients, every sample in the training set was correctly classified. The linear discriminant was also able to correctly classify all the samples in the prediction set. The results obtained from this pattern recognition study suggest that IR spectra of clear coats can be used to characterize paint smears by manufacturing plant and production line.

4. Conclusions

A two-step procedure for spectral pattern recognition is proposed. First, a wavelet packet tree is used to decompose each spectrum into wavelet coefficients that represent both high and low frequency components of the signal. Second, a genetic algorithm for pattern recognition analysis is used to identify those wavelet coefficients that contain information about the class label of the samples. The proposed two step procedure is well suited for the development of search prefilters based on chemical information that have the potential to facilitate spectral library matching.

Acknowledgements

The authors would like to express their thanks to Mr. Denis Lafleche and Ms. Tamara Hodgins of the Royal Canadian Mounted Police Forensic Laboratory for their assistance in retrieving automotive paint data, and for providing technical information about Chrysler automotive paint systems.

References

[1] J.L. Buckle, D.A. MacDougal, R.R. Grant, Can. Soc. Forens. Sci. J. 30 (1997) 199–212.
[2] N.S. Cartwright, P.G. Rodgers, Can. Soc. Forens. Sci. J. 9 (4) (1976) 145–154.
[3] A. Beveridge, T. Fung, D. MacDougall, in: B. Caddy (Ed.), Forensic Examination of Glass and Paint Analysis and Interpretation, Taylor and Francis, NY, 2001, pp. 220–233.
[4] N.S. Cartwright, L.J. Cartwright, E.W.W. Norman, R. Cameron, W.H. Clark, D.A. MacDougal, Can. Soc. Forens. Sci. J. 15 (3/4) (1982) 105–115.
[5] A. Hobbs, Trace Evidence Symposium, August 13–16, Clearwater Beach, FL, 2007.
[6] B. Christy, Trace Evidence Symposium, August 13–16, Clearwater Beach, FL, 2007.
[7] G. Fettis (Ed.), Automotive Paints and Coatings, VCH Publications, New York, 1995.
[8] F.T. Tweed, R. Cameron, J. Deak, P.G. Rodgers, Forens. Sci. 4 (1974) 211–218.
[9] S.R. Lowry, D.A. Huppler, C.R. Anderson, J. Chem. Inf. Comput. Sci. 25 (1985) 235–241.
[10] J.C.W. Bink, H.A. Van'T Klooster, Anal. Chim. Acta 150 (1983) 53–59.
[11] G.W. Small, Anal. Chem. 59 (7) (1987) 535A–546A.
[12] C.P. Wang, T.I. Isenhour, Appl. Spectrosc. 4192 (1987) 185–194.
[13] Walter, S. James, A Primer on Wavelets and their Scientific Applications, Chapman & Hall/CRC, New York, 1999.
[14] B.B. Hubbard, The World According to Wavelets, 2nd ed., A. K. Peters, Natick, MA, 1998.
[15] B.K. Lavine, J. Ritter, A.J. Moores, M. Wilson, A. Faruque, H.T. Mayfield, Anal. Chem. 72 (2) (2000) 423–431.
[16] B.K. Lavine, D. Brzozowski, J. Ritter, A.J. Moores, H.T. Mayfield, J. Chromatogr. Sci. 39 (12) (2001) 501–506.
[17] B.K. Lavine, D. Brzozowski, A.J. Moores, C.E. Davidson, H.T. Mayfield, Anal. Chim. Acta 437 (2) (2001) 233–246.
[18] B.K. Lavine, C.E. Davidson, A.J. Moores, P.R. Griffiths, Appl. Spectrosc. 55 (8) (2001) 960–966.
[19] B.K. Lavine, A. Vesanen, D.M. Brzozowski, H.T. Mayfield, Anal Lett. 34 (2) (2001) 281–294.
[20] B.K. Lavine, C.E. Davidson, A.J. Moores, Chemom. Intell. Lab. Instrum. 60 (1) (2002) 161–171.
[21] B.K. Lavine, C.E. Davidson, A.J. Moores, Vib. Spectrosc. 28 (1) (2002) 83–95.
[22] B.K. Lavine, C.E. Davidson, C. Breneman, W. Katt, J. Chem. Inf. Comp. Sci. 43 (2003) 1890–1905.
[23] B.K. Lavine, C.E. Davidson, D.J. Westover, J. Chem. Inf. Comp. Sci. 44 (3) (2004) 1056–1064.
[24] G.A. Eiceman, M. Wang, S. Pradad, H. Schmidt, F.K. Tadjimukhamedov, B.K. Lavine, N. Mirjankar, Anal. Chim. Acta 579 (1) (2006) 1–10.
[25] J. Karasinski, L. White, Y. Zhang, E. Wang, S. Andreescu, O.A. Sadik, B. Lavine, M.N. Vora, Biosens. Bioelectron. 22 (11) (2007) 2643–2649.
[26] I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, 1986.
[27] P.G. Rodgers, R. Cameron, N.S. Cartwright, W.H. Clark, J.S. Deak, E.W.W. Norman, Can. Soc. Forens. Sci. J. 9 (1) (1976) 1–14.
[28] P.G. Rodgers, R. Cameron, N.S. Cartwright, W.H. Clark, J.S. Deak, E.W.W. Norman, Can. Soc. Forens. Sci. J. 9 (2) (1976) 49–68.
[29] P.G. Rodgers, R. Cameron, N.S. Cartwright, W.H. Clark, J.S. Deak, E.W.W. Norman, Can. Soc. Forens. Sci. J. 8 (3) (1976) 103–111.
[30] C. Foo-tim, L. Yi-zeng, G. Jumbin, S. Xue-guang, Chemometrics—From Basics to Wavelets, vol. 164, Wiley Interscience, NY, 2004.
[31] B.K. Lavine, C.E. Davidson, in: S. Brown, R. Tauler, R. Walczak (Eds.), Comprehensive Chemometrics, vol. 3, Elsevier, Oxford, NY, 2009.
[32] B.K. Lavine, M. Vora, J. Chromatogr. A 1096 (2005) 69–75.
[33] B.K. Lavine, C.E. Davidson, W.T. Rayens, Anal. Lett. 7 (2004) 115–131.
[34] B.K. Lavine, A.J. Moores, in: K. Siddiqui, D. Eastwood (Eds.), Pattern Recognition, Chemometrics, and Imaging for Optical Environmental Monitoring, SPIE (Proceeding Series), vol. 3854, 1999, pp. 103–112.
[35] B.K. Lavine, A. Moores, Buletin Kimia 12 (2) (1997) 73–86.
[36] R.G. Lawson, P.C. Jurs, J. Chem. Inf. Comput. Sci. 30 (1990) 36–41.
[37] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, NY, 1995.
[38] B.K. Lavine, K. Nuguru, N. Mirjankar, J. Chemom. 25 (2011) 116–129.