Research Article

Automatic Identification of Artistic Pigments by Raman Spectroscopy Using Fuzzy Logic and Principal Component Analysis

M. Castanys, M. J. Soneira, and R. Perez-Pueyo

Departament de Teoria del Senyal i Comunicacions (TSC), Universitat Politècnica de Catalunya (UPC), C/ Sor Eulàlia de Aniziu s/n, Campus Nord D5, 08034 Barcelona, Spain

Received 15 September 2006; Revised 26 October 2006; Accepted 27 October 2006

Recommended by Marta Castillejo

This work offers an automatic identification system of Raman spectra of artistic pigments. The proposed methodology is based on a fuzzy logic system, and uses principal component analysis to reduce redundancies in data and the correlation operator as an index of similarity between two Raman spectra. Moreover, as sometimes pigments are used in mixtures by artist, the designed system is able to recognize binary mixtures of pigments on the basis of their Raman fingerprints.

Copyright © 2006 M. Castanys et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

The identification of materials used in artworks is of great importance for conservation, restoration, and comprehensive study of our historical and cultural heritage. Raman spectroscopy is a well-established tool in the investigation of a wide range of archaeological and art historical artefacts. It has proved particularly useful for identifying pigments [1–6]. Its great popularity stems from its ability to provide definitive identification due to its high specificity; its applicability “in situ” allowing the analysis directly on unprepared sample and its nondestructive behavior. This analytical technique is based on the Raman effect that provides chemical and structural information of almost any material allowing its identification. When monochromatic light encounters matter, most of the scattered light has the same wavelength as the incident light. However, a small fraction of the scattered light is shifted in a different wavelength by the molecular vibrations and rotations in the sample. The representation of this shifted light is named Raman spectrum, and contains many sharp bands characteristics of the sample, allowing its identification without ambiguity. This identification is usually made by a user by visual inspection of the Raman spectrum. There are two strategies to make a visual analysis, one consists of comparing the whole unknown spectrum with the known patterns spectra and the other is based on comparing the wavenumber positions of the Raman bands of the unknown spectrum with those of the reference spectra. In any case, thus are a time consuming process and an imprecise process. Moreover, some problems like noise and fluorescence interference inherent to the own acquisition of the spectra, or the possibility to detect pigment mixtures make this comparison difficult and dependent on user’s experience. On effect, artists often paint with pigment mixtures and apply transparent coatings, which complicate spectroscopic signatures and confound protocols for on-painting identification. So, it is useful and desirable to develop methods to identify spectra in an automated way, that is, that do not depend on the subjective assessment of user.

The above commented identification strategies in a conventional visual process are equally valid but there are some differences between them. While the first method involves only one step, which is the comparison of the whole spectrum of the analyzed sample with those of the patterns, the second involves two steps. First, it is necessary to locate the wavenumber position of the Raman bands, this can be made automatically [7] or not. And second, once the Raman bands are detected, the identification of the unknown spectrum is possible by searching the coincidence with the Raman bands of the reference spectra in an automated way [8, 9]. Both approaches to the problem of the automatic identification have been dealt with fuzzy logic which has demonstrated being a
powerful tool with a large number of applications in particular to signal processing [10, 11]. Although both points of view are possible and useful for identifying, the comparison of the whole spectra could be more suitable when the spectra have many bands and/or common bands to each other. Furthermore, when pigments to identify have many common Raman bands, it could be difficult to assess if the analyzed spectrum is representative of one pigment or it corresponds to a mixture of pigments. The comparison of the whole spectrum can overcome this situation. So, the work described in this paper offers an automatic system to identify the pigment or pigments in binary mixtures, to which the measured spectrum can correspond based on fuzzy logic. This signal processing tool has demonstrated to be useful to deal with the uncertainty present in the measured spectrum. The methodology proposed follows the guidelines of visual comparison but automates the decision-making process so that, subjective errors will be avoided. The system checks which spectrum or spectra of the pattern's library of Raman spectra are the most similar to the unknown spectrum. This method of comparison can be more or less laborious based, among other things, on the spectral database and also on the quality of the measured spectrum. Hence, it is desirable and even necessary to define a suitable database of reference pigments taking into account the available data about the analyzed artefact. For instance, the type of painted work of art such as easel paintings, wall paintings, and so forth, supposed date or author of the artwork, can define the Raman spectral library of patterns, well suited for enhancing the efficiency of the identifier system. Furthermore, in order to reduce redundancies in data which may be important as the library database increases, principal component analysis (PCA) is used [12].

This paper is organized as follows: in the second section a brief description of the used mathematical tools is explained; in the third section the proposed fuzzy rule-based identification system is described; in the fourth section the results obtained when apply the proposed system to different Raman spectra are shown and discussed; and in the fifth section the conclusions are summarized.

2. THEORY

It is well known that a Raman spectrum can provide a large amount of information about a sample, and as the majority of signals, it can be divided into two parts: the information signal and the noise. The signal is the part which contains the desired information and which allows a sample characterization, whereas the noise is the unwanted information which does not report any specific trait of the analyzed material. It must be stood out that the consideration of what is signal and what is noise depends on the analysis which will be made, then in the case of pigment identification the useful information is extracted from the Raman bands position. Noise in Raman spectroscopy can be mainly classified into three categories: shot noise, which is the result of the statistical nature of light, fluorescence, and cosmic rays. These three types of noise are inherent on each Raman measure, but taking some precautions and some software and hardware procedures, they can be reduced. Hence, one assumes that the Raman spectra which will be processed will be collected in the optimal conditions to reduce noise as much as possible. As well as, some signal treatment must be done to ensure the success of the analysis which is in our case the identification of the analyzed pigment.

As it has been mentioned, the identification methodology is based on the comparison between spectra, so it is necessary that the spectra were stored in a compatible way. Note that a Raman spectrum could be read as a vector of \( N \) points where each of its coordinates, \( e(i) \), corresponds to the Raman intensity for the wavenumber \( v_i \). Then, the interpolation step ensures that the same coordinate for each spectrum corresponds to the Raman intensity at the same wavelength, and that each of them has the same number of points. Another arrangement that results useful is to normalize the spectra in order to reduce the impact of measurement conditions. Some characteristics of the Raman bands depend on instrumental conditions as the spot of the laser, its wavelength, or the time of exposure, which can change for different measurements and which are then not specific of the analyzed sample. The normalized spectra maintain the relative relation between their Raman bands and their Raman intensity values ranging from 0 to 1. Finally, the fluorescence is a radiated phenomenon, generated by many materials, which underlies the measured spectra. Several techniques involving hardware and software have been devised to minimize its presence in order to resolve and analyze the Raman spectra. In this work, the most popular polynomial fitting method, available in the software LabSpec of our laboratory, has been used.

Due to the fact that each spectrum is a vector of \( N \) points, with \( N \) being of the order of 1000, we decide to reduce its dimension by means of a data reduction tool called the principal component analysis (PCA). The most important use of this chemometric technique is to represent the \( N \)-dimensional data in a smaller number of dimensions, usually two or three, without loss of information [13– 15]. With this data reduction, the searching identification algorithm will be more efficient and less tedious. The PCA is one of the most used multivariate procedures because it is easy to interpret and permit an explanation for the maximum variability of initial distribution. In fact, PCA is a mathematical method of reorganizing information in a data set of samples. What PCA does is to discover new variables called principal components (PCs) which account for the majority of the variability of the data. Then the scores associated with significant PC can be used to represent the spectra. This permits to represent the data in 2-dimensional plots in which one can observe for example groupings of objects or outliers, and define the structure of a data. In some, PCA is a way of identifying patterns and expressing the data in such way so as to highlight their similarities and differences.

The aim of this research is to find out to which pigment corresponds the measured Raman spectrum. As it has been mentioned, the most frequently methodology to identify unknown Raman spectra is based on the comparison of the analyzed spectrum with some well-known spectra, called patterns, searching which of them is the most similar to it. In
order to make this process systematic and not depending on the subjectivity of any user, in this work the comparison is made by a mathematic operator which quantifies the similarity between spectra in an automatic and objective way.

One standard mathematic operator to estimate the degree of similarity between two spectra is the correlation coefficient \([16]\). Once the Raman spectrum of the analyzed pigment has been collected, \(e(i)\), it is compared with each spectra of the library, \(p(i)\). The correlation coefficient between the two spectra is computed as follows:

\[
    C_{ep} = \frac{\sum_k [(e(i) - m_e) \cdot (p(i) - m_p)]}{\sqrt{\sum_k (e(i) - m_e)^2 \cdot \sum_k (p(i) - m_p)^2}},
\]

where \(m_e\) and \(m_p\) are the mean values of \(e\) and \(p\), respectively.

Note that the higher the coefficients are, with the maximum equal to one, the more similar the spectra are. Nevertheless, even though spectra have been treated as in the preceding paragraph we have described, they inevitably contain some noise which alters the information and makes the correlation analysis nondefinitive. That is, even if we compute the correlation between two spectra coming from the same material, the result could not be the maximum value. So, the results obtained by the correlation coefficient must be interpreted, that is, evaluate if the value is high enough to consider that the two spectra correspond to the same material. This interpretation is made by means of fuzzy logic.

Fuzzy logic aims at exploiting the tolerance of imprecision and uncertainty in order to have a close resemblance to human decision making. Fuzzy logic appeared in the sixties by Lofti Zadeh and as the result of searching a mathematical option to overcome the inflexibility of the classic logic which only contemplates the possibility that one can belong or not to a set. Fuzzy logic is subtended by the theory of fuzzy sets which assigns to each element a degree of membership to it in a range of 0 to 1. Then, in fuzzy logic, the truth of any statement becomes a matter of degree \([17]\). So, a fuzzy set is a set without a crisp, clearly defined boundary, and it can contain elements with only a partial degree of membership. Fuzzy logic is concerned with the formal principles of approximate reasoning, with precise reasoning viewed as a limiting case. In this sense, fuzzy logic incorporates an alternative way of thinking, which allows modeling complex systems using a higher level of abstraction originating from human knowledge and experience. A fuzzy logic system, Figure 1, maps nonlinearly crisp inputs into crisp outputs. It is unique in that it is able to simultaneously handle numerical data and linguistic knowledge by means of a group of rules which models the proposed problem \([18]\). These rules are expressed as a collection of if-then statements which may be provided by experts or can be extracted from numerical data. The fuzzifier assigns to every numerical input value a degree of membership to each input fuzzy sets through the membership functions associated to them. This first step is needed in order to active rules which are in terms of linguistic variables, with fuzzy sets associated to them. The inference engine provides the way in which rules are combined and by means of different logical operators modifies the defined output fuzzy sets. Finally, the defuzzifier maps output fuzzy sets into crisp numbers applying mathematical mechanisms in order to obtain the final result.

Hence, in our frame, Fuzzy logic appears as the best tool for modeling in a mathematical language and by means of a Fuzzy system the human perception of similarity.

### 3. FUZZY RULE-BASED IDENTIFICATION SYSTEM

As we had seen in a previous section, in order to identify a pigment, we will apply the correlation coefficient to compare the unknown spectrum to each of the selected patterns. The problem of the noise in the measures will make the value of correlation difficult to interpret. Moreover, it is known that a correlation equal to 1 corresponds to two signals which are identical, and a value of 0 corresponds to two completely different signals (uncorrelated signals), but what happens with the others values? What do we have to interpret with a correlation of 0.23? The interpretation of all the possible values will be made by means of a fuzzy logic system. The crisp input value will be the correlation coefficient computed between the unknown spectrum \((e)\) and one of the patterns \((p_i)\), the if-then rules model the human reasoning employed in a visual comparison, and finally the output of the system will be the identification or not of the analyzed pigment (see Figure 2).

On the other hand, in a practical situation, the complexity of pictorial layer is a problem to be solved and mixtures of pigments are the first step for a comprehensive view of artworks. So, the Raman trace of two or more pigments could appear in a measured Raman spectrum. The mixtures could then confound searching protocols for on-painting identification. Then, we propose a system to identify artistic pigments including cases where various pigments are mixed.

The proposed design is then based on two fuzzy logic systems (see Figure 3). The first step is to compare the unknown spectrum with each of the patterns; the result of this system is a list of patterns which are candidates to be the unknown...
pigment. Then, if there is only one candidate, the unknown spectrum can be identified or not. But, if there is more than one candidate, a new library is made by mixing them. The second step is to compare the analyzed Raman spectrum with each of these new patterns. The final result would be that the analyzed spectrum was identified as a mixture, as a single pigment, or simply nonidentified.

The first system design follows the guideline proposed in [7]. The crisp inputs are the correlation coefficients, \( C_{epi} \), computed between the unknown spectrum \( (e) \) and each of the reference Raman spectra \( (p_i) \) stored in the selected library. For this input linguistic concept, three fuzzy sets “low,” “medium,” and “high” have been chosen. Each of them is defined by a particular membership function that is to say, the fuzzy set “low” has a trapezoidal membership function while the membership functions associated with the fuzzy sets “high” and “medium” are cosine functions (see Figure 4(a)).

The two threshold values called \( \xi_1 \) and \( \xi_2 \), in Figure 4(a), divide the universe of discourse (input space range) into three different zones. As we have mentioned, the reference library will be constructed depending on the pigments that we want to analyze and/or taking into account the available information about the sample. The identification process depends on this selection, with the more similar patterns are, the more difficult the identification is. In order to overcome this dependence, the bounds of each fuzzy set are computed from the maximum correlation coefficient between patterns as follows: \( \xi_1 = 2 \times C_{max} - 1 \) and \( \xi_2 = C_{max} \), where \( C_{max} \) is the maximum correlation coefficient between the reference Raman spectra (it determines which are the more similar patterns). So, they determine the degree of membership of the input correlation coefficient to the fuzzy sets; high, medium or low.

As we have said before, two output fuzzy sets labeled with the linguistic concepts “identified” and “nonidentified” (see Figure 4(b)) have been defined. The output crisp variable is a single number obtained by the defuzzifier process by means of the centroid computation of the final output fuzzy set. This output value is called “degree of similarity” \( g_{epi} \) and quantifies the likeness of the two compared spectra.

The fuzzy rules define the way to obtain the final crisp values from the input data, in our identification problem, there are only four rules to be considered.

**Rule 1.** If \( C_{epi} \) is low, then the spectrum \( e \) is nonidentified with \( p_i \).
1. If \( C_{p_1 e_2} \) is low, then the spectrum \( e_2 \) is nonidentified with \( p_1 \).

2. If \( C_{p_1 e_2} \) is medium, then the spectrum \( e_2 \) is nonidentified with \( p_1 \).

3. If \( C_{p_1 e_2} \) is medium, then the spectrum \( e_2 \) is identified with \( p_1 \).

4. If \( C_{p_1 e_2} \) is high, then the spectrum \( e_2 \) is identified with \( p_1 \).

Rule 2. If \( C_{e \pi} \) is medium, then the spectrum \( e \) is nonidentified with \( p_1 \).

Rule 3. If \( C_{e \pi} \) is medium, then the spectrum \( e \) is identified with \( p_1 \).

Rule 4. If \( C_{e \pi} \) is high, then the spectrum \( e \) is identified with \( p_1 \).

The rules 2 and 3 which have the same antecedent but different consequent are used to solve the cases where the values of correlation are ambiguous. For example, when two spectra have some Raman bands in common, they are similar but not identical, so the output of the system must be nonidentified even if a medium value for its correlation coefficient is obtained. Another example could be when the analyzed spectrum is compared with its corresponding pattern, but the measure is strongly affected by noise, and then the correlation coefficient is not as high as it is supposed to be. In such case, although the correlation is medium too, the unknown pigment must be identified with its corresponding pattern.

The implication of rules is made using the product operator and a fuzzy set is obtained for each one. Every output fuzzy set is aggregated, by means of the sum, into a single output fuzzy set. Finally, using the centroide method, the degree of similarity between \( e \) and \( p_i (g_{\text{nepi}}) \), is calculated. The system
has been scheduled for considering each pattern a candidate with which the degree of similarity is higher than 5/10. With the purpose to evaluate the reliability when a reference spectrum is considered by the system as a candidate to identify the unknown spectrum, we define a new variable named degree of security or certainty ($g_s$). The definition of this degree has been obtained empirically and it is calculated as $g_s(g_{nepi}) = 100(g_{nepi}/8 - 1/4)$. Next, in order to clarify how this first system works, an application example is shown. An unknown Raman spectrum is compared with the pattern spectrum of the anatase pigment. The spectra and the evaluation of the four rules are shown in Figure 5.

The design of the second fuzzy logic system is based on the first one. Nevertheless, in this system, the library is not initially defined, but it changes depending on which candidates have been found by the first system. As we have mentioned, the reference spectra of the new library are obtained mixing the candidates two by two. So, in this first approach, the global system can only identify binary mixtures that are the most common mixtures made by artists. The extension of the system in order to be able to identify mixtures of more than two pigments will be a future work. The system runs like the first one, that is, it has the same variables and fuzzy sets, as well as the same rules but it uses the new spectral library. Then, if the first system considers C candidates for the identification, in the second system the used library will be formed by $C(C - 1)/2$ patterns.

4. EXPERIMENTAL RESULTS

Let we see some experimental results of the proposed system. The Raman spectra analyzed had been measured in the Raman laboratory of the UPC. This laboratory incorporates optical fibre which allows Raman measurements to be made far from the Raman analyzer without extraction of any sample. Its main instrumentation is a laser (monochromatic source), a monochromator, a CCD (charge-coupled device) detector, and a computer (see Figure 6). The output of the laser (which can be a He-Ne, Argon, or IR laser) is guided by the excitation optical fibre through the optical head to the analyzed material. The radiation scattered by the material, which contains the Raman signal, is guided through the collection fiber to the monochromator where it is spectrally split and then detected to the CCD. The spectrum is stored and visualized in the PC which, furthermore, controls the Raman equipment. On the other hand, the chosen library for testing the system has been made selecting some of the more usual pigments found in paintings, which are the anatase, the ultramarine blue, the atzurite, the vermillion, the Naples yellow, the rutile, and the chromium yellow. In Figure 7 it is shown the representation of each of there projection in the space of PCA (PC1 versus PC2). Note that the closer they are, the more similar they are.

The first example represents the identification of the unknown Raman spectrum represented in Figure 8(a). The first fuzzy system returns as candidates, the anatase and rutile pigments. By means of the application of the second system, the analyzed pigment is identified as a mixture of anatase and rutile with a similarity degree equal to 8, 5/10 and a security
degree of 81%. This mixture could appear in artwork analysis if, for example, a white point painted originally with anatase has been later repainted with the most modern white pigment as the rutile.

In the second example, three candidates have been determined by the first fuzzy logic system: ultramarine blue, chromium yellow, and Naples yellow. The second system which uses a library formed by the binary mixtures of these three candidates identifies the analyzed pigment as a mixture of ultramarine blue and chromium yellow. In this case the similarity degree with the mixture is 6.3/10, and the security degree is equal to 53.7%. It can be noted that the degree of similarity is lower than the first example. It is because, as it can be seen in Figure 9, the intensity of the Raman band of the analyzed spectrum which corresponds to the fundamental Raman band of the chromium yellow is lower than the one of the pattern. This difference in the intensities is not decisive because its position is the same, and this is what contains the Raman information for identifying.

It must be pointed out that if in a measured spectrum there is no Raman band of a pigment which is in a mixture, it is not possible to identify it neither in a visual way nor in automatic way. Obviously, if there is no Raman fingerprint, the identification is impossible.

5. CONCLUSIONS

This work presents a fuzzy logic system like a useful and valid system to recognize automatically experimental Raman spectra following the guidelines of the visual identification. The
obtained results show that fuzzy logic offers a versatile and flexible technique, easily adapted to the requirements of the problem to solve. With the proposed system, a nonexpert on Raman spectroscopy is able to identify the Raman fingerprint of any pigment, overcoming ambiguous situations which could be produced by noise or pigments mixtures. Note that, even if the system is robust, it is required that the experimental conditions of measurement should be as suitable as possible. Finally, it must be pointed out that the system has been designed in MATLAB and has a very short time of processing.

ACKNOWLEDGMENT

This work was supported by the CICYT project reference TEC 2005-07239 of the Spanish government.

REFERENCES