Automating the classification of Vis-NIR reflectance spectra in the Mobartech Project
--Manuscript Draft--

Manuscript Number:	
Full Title:	Automating the classification of Vis-NIR reflectance spectra in the Mobartech Project
Article Type:	Regular Article
Section/Category:	Accelerators Physics and Physics of Cultural Heritage
Corresponding Author:	Letizia Bonizzoni, Ph.D. Universita degli Studi di Milano Milano, Lombardia ITALY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Universita degli Studi di Milano
Corresponding Author's Secondary Institution:	
First Author:	Jacopo Melada
First Author Secondary Information:	
Order of Authors:	Jacopo Melada
	Letizia Bonizzoni, Ph.D.
	Marco Gargano
	Emanuela Grifoni
	Nicola Ludwig
Order of Authors Secondary Information:	
Funding Information:	Regione Lombardia — Not applicable
Abstract:	The present work presents the preliminary outcome of one of the tasks inside the MOBARTECH project, that is the creation of the prototype for a compact FORS spectrophotometer capable of automatically identifying pigments and dyes. On the road toward this goal, we concentrate on the automation of pigment identification, as aspect of great interest which would allow an in line analysis and could be useful also for unskilled operators. In this paper, we evaluate high order derivative approximation pre-processing and features extraction methods, strictly related to the way the database is built; we have chosen to use a dynamic database to be able to elaborate data and give an answer in real time, independently from the experimental set up, with an interface suitable for both non-skilled and expert users. The aim of the present research is to evaluate how spectra pre-processing affects the response and the ability of the search algorithm to correctly identify an unknown spectrum, in order to develop a totally automated pipeline for spectra identification or to prefer an interactive framework.
Suggested Reviewers:	Noël Richard University Poitiers, XLIM, CNRS UMR 7252, F-86000 Poitiers, France noel.richard@xlim.fr He published several papers about related subjects
	Haida Liang Nottingham Trent University haida.liang@ntu.ac.uk Her research deals with optical techniques and cultural heritage materials
	Alice Plutino Universita degli Studi di Milano alice.plutino@unimi.it

Opposed Reviewers:	

Automating the classification of Vis-NIR reflectance spectra in the Mobartech Project

Jacopo Melada[1], Letizia Bonizzoni[2*], Marco Gargano[2], Emanuela Grifoni[2], Nicola Ludwig[2]

1 Dipartimento di Scienze della Terra "Ardito Desio", via Mangiagalli 34, 20133 Milano, Italy

2 Dipartimento di Fisica "Aldo Pontremoli", Università degli Studi di Milano, Via Celoria 16, 20133 Milano, Italy

*letizia.bonizzoni@mi.infn.it

Abstract

The present work presents the preliminary outcome of one of the tasks inside the MOBARTECH project, that is the creation of the prototype for a compact FORS spectrophotometer capable of automatically identifying pigments and dyes. On the road toward this goal, we concentrate on the automation of pigment identification, as aspect of great interest which would allow an in line analysis and could be useful also for unskilled operators. In this paper, we evaluate high order derivative approximation pre-processing and features extraction methods, strictly related to the way the database is built; we have chosen to use a dynamic database to be able to elaborate data and give an answer in real time, independently from the experimental set up, with an interface suitable for both non-skilled and expert users. The aim of the present research is to evaluate how spectra pre-processing affects the response and the ability of the search algorithm to correctly identify an unknown spectrum, in order to develop a totally automated pipeline for spectra identification or to prefer an interactive framework.

Keywords: Pigment, Vis-NIR spectra, automated spectral identification, spectral similarity

1. Introduction

The MOBARTECH project, funded by the European Regional Development Fund of Regione Lombardia, is meant to develop "a technological, interactive and participatory mobile platform for the study, conservation and valorisation of historical and artistic heritage" that integrates cultural, social and creative skills; the involved fields include information technology, diagnostics, conservation and restoration of cultural heritage, intelligent logistics systems, public interaction and infotainment technologies. Among the goals of the project, a mobile laboratory, also allowing real-time data analysis. With this aim, we are developing a smart spectrophotometer able to automatically classify pigments on the bases of acquired Vis-NIR reflectance spectra. Indeed, the acquisition of reflectance spectra for the identification of pigments in artworks is a common procedure in conservation studies [1,2]. The main advantages rely on its ease of use, cost, effectiveness, portability, non-invasiveness, and amount of information acquired [3]. Nevertheless, the identification of the pigment is currently mainly subjective and require the work of a trained expert. The automation of this process is thus of great interest for unskilled operators (e.g. art historians) and educational purposes as well as simplify the work of experts.

Several steps are required to reach this goal: from deciding how to assemble the instrument, to the setting up of a large and reliable database. The spectra to be classified, in fact, have to be

compared with reference spectra through a spectral library search algorithm, a widely used method for automated spectral identification. The automatic identification of pigments and dyes, based on the reflectance spectra comparison and exploiting spectral library search algorithms, involves the measurement of the similarity, i.e. the distance, between two spectral reflectance curves in the Vis-NIR by comparing the unknown spectrum with a reference spectra database, where the distance function must possess the properties of reflexivity, non-negativity, and symmetry.

There is an extensive literature describing the properties and evaluating the performances of spectral distance measurements [4–7] . In this work, we apply some of the most widespread methods present in scientific literature considering that the comparison of distance measurement methods currently used for the spectral matching of pigments and dyes is outside the scope of this work. The aim of the present research is to evaluate how spectra pre-processing affects the response and the ability of the search algorithm to correctly identify an unknown spectrum, in order to explore the feasibility of developing a totally automated pipeline for spectra identification or to prefer an interactive framework. This process critically depends on the characteristics of both the spectral library and the matching algorithm. A different possibility for classification is the use of machine learning classification models. To sum up, given several training examples associated with desired outcome (labels), the process consists in finding the relationship between the data and labels, using solely the training examples, so to predict the outcome for new test samples. To perform this task, it is necessary to build a model, or predictor, which is typically a function with adjustable parameters. The problem of pigment identification based on reflectance spectra via machine learning can be described either as multi-label classification or multiclass classification, according to the structure used for the classifier and the library. In the last years, several researches have been devoted to this specific thread, mainly applied to hyperspectral imaging. A supervised multi-label classification approach was compared to a deep neural network and to an unmixing strategy for automatic pigment identification [8]. The effect of binders, priming, and substrate were taken into consideration in a paper proposing the use of a hyperspectral unmixing algorithm, able to identify pigments exploiting a Multilayer Perceptron neural network [9]. The implementation of an automatic classification method for hyperspectral images using MaxD endmembers extraction and spectral angle mapper (SAM) for endmember classification can also be considered [10]. What is common for all these algorithms is the need of a pre-processing of the spectra. In the present work, we evaluate high order derivative approximation pre-processing methods; based on the obtained results, the structure of the reference spectra library and the development of an interactive Graphical User Interface will be decided, as they are strictly related. It should be noted that, while the data set for machine learning is normally a closed one (i.e. the class of the unknown spectra must be present in the training set and the adding of new classes necessitate of a retraining of the model), we aim to use a dynamic database so to be able to elaborate data and give an answer in real time, independently from the experimental set up.

2. Towards the automated identification

According to our experience, an algorithm for spectrum identification should avoid model retraining, hyper parameter tuning, and issues associated with class unbalance; maintenance and modularity should be well-thought-out. The presence of many classes and a small number of samples per class is possible, depending on the architecture of the reference database. Moreover, it is mandatory to analyse in depth the limitations and potentialities in detecting pigments alteration or degradation. The optimum methodology for the automatic pigment identification requires an algorithm minimizing the uninformative effects of binders, primers, varnishes, and substrate, while hopefully enhancing the information about conservation status. In the beginning of the present research, the importance of a correct and reasoned compilation of the reference reflectance spectra database for the automatic recognition has become obvious. Before further elaborations, it is also necessary to refine data allowing to extract significant components from spectra intrinsically affected by noise. We initially tested the most common smoothing algorithm, selecting as the most suitable one the Savitzky–Golay filter [11], and then using the spectra derivatives. Numerical differentiation of experimental signal is, in fact, often used to ease the detection and location of components of spectra [12]. Differentiation unavoidably degrades the signal to noise ratio and thus some form of smoothing is required in conjunction to differentiation.

3. Method: the building of the model

All the data presented were processed with ad hoc scripts in Python 3.8.

In order to ensure the possibility of using the same algorithm and the same spectral library even with spectra acquired from different spectrometers and having, in principle, a different number of channels each spectrum is firstly interpolated in order to have 300 measurement points between 400 and 1000 nm, this means a spectral resolution in wavelength of 2 nm which is realistic with many gratings based portable low cost spectrometers. Moreover, the reduction of the number of points in the spectrum allows to speed up the computational effort. The interpolation was done using Piecewise Cubic Hermite Interpolating Polynomial in order to preserve abrupt changes in reflectance values [13]. After this preliminary step, we focused on the evaluation of the search algorithm that computes the shortest spectral distance (i.e. the most similar spectrum) between unknown and database spectra suggesting possible match. Spectra were smoothed using the Savitzky-Goaly filter with a window size of 21 datapoint and a second order polynomial [14]. The derivative was approximated using the finite difference approximation method [12]. The derivative of the spectra was normalized (unity-base normalization) in the range [0,1].

We tested 6 different distance measures methods generally applied to pattern recognition in hyperspectral imaging and spectral matching to verify which pre-processing method should be the better performing one. Namely, we applied Euclidean distance (EUC) and root-mean-square deviation (RMS) [4] , spectral angle mapper (SAM) [15], spectral correlation mapper (SCM) [16],

spectral information divergence (SID) [17] and Kullback-Leibler pseudo-spectral divergence (KLPSD) [18,19].

4. Preliminary results

Our reference database is composed by 549 mock-ups realized used as pure or combined in mixture and layers and with linseed oil as binder applied on a cloth-lined board (Fabriano) in either linseed oil or acrylic binder. Among these, 140 are pure pigment layers and the others are mainly mixture with white lead in different proportions; mixture of two or three different pigment, and some are spread with the glazing technique (*velature*). To achieve our purpose, we set the problem as multiclass classification: each spectrum in the database has a single label. This is valid both for pure pigments (e.g., a pure azurite spectrum is labelled as "azurite") and for mixtures (e.g., a spectrum of azurite mixed with white lead is labelled as "azurite + white lead" without considering the proportions).

We first performed tests on a set of widely used pigments showing common features (single or multiple bands and s-shape), to verify whether the identification improves following automatic pre-processing. Then we tested the same procedure on real pigment layers belonging to a 16th century painting previously analysed with long-established methods able to determine the mixtures and layers composition [20], comprehending mixtures and layered pigments. Each single dataset is described in the following subsections.

In this case, the reference database and the datasets were both acquired using an Ocean Optics spectrophotometer. The device is connected to a co-axial optical fiber mounted in a reflection probe older in the 45°/45° geometry with a spot of 2 mm of diameter. The light source was a halogen lamp (D65, HL2000, Ocean Optics) for both database and the real painting case study (par. 4.2), while for the simple dataset (par. 4.1) a Xenon arc lamp was used.

4.1 The datasets

We considered two different data set to check different aspects of the issue. To achieve our purpose, we set the problem as multiclass classification: each spectrum in the database has a single label. Labelling of the spectra was done by a trained expert.

4.1.1. The basic dataset

A simple dataset was firstly considered, to evaluate whether the chosen automatic pre-processing feature extraction methodology improves spectral classification. The pure pigment considered were azurite (Zecchi 107/I), malachite (Zecchi 104/II), cadmium red (Zecchi, 047), yellow ochre (Maimeri, 784), and red ochre (Maimeri, 782). They were considered as plane ground powders and layered in different combinations: oil on canvas, oil on paper, oil on panel, egg tempera on paper, fresco. For each test layer, we acquired 10 spectra in different positions to get a good statistic on data. In Fig. 1, the acquired raw spectra are reported: they were interpolated to get 300 measurement points between 400 and 1000 nm (see section 3), but not filtered. Please note that the Xe arc lamp peaks

are visible: we did not suppress them on purpose, so to better simulate the real situation in which we want our system to work in real time on supposed defected spectra. The independence from the acquisition system was also tested, as dataset spectra were acquired with the already mentioned Xe light source while the database spectra were acquired exploiting a halogen lamp.
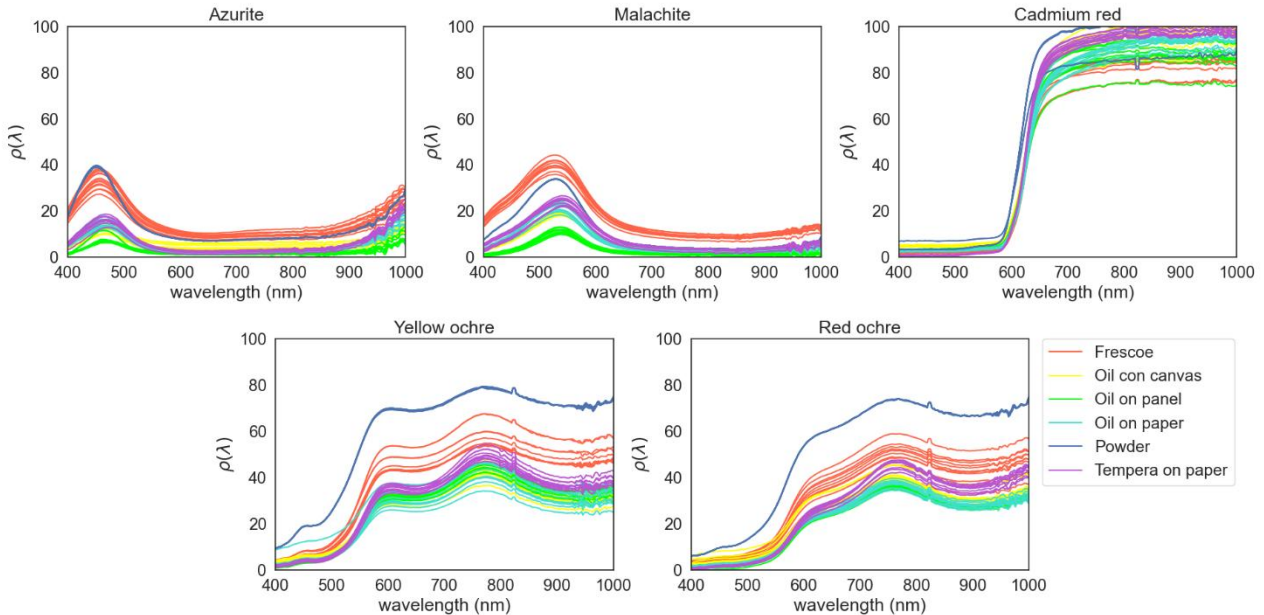

Fig.1: Unfiltered interpolated spectra from the basic dataset, the peak at 827 nm is due to Xe-arc lamp source.

4.1.2 Case study

A real case study was also tested, starting from the spectra acquired on the painting "The Martyrdom of St. Catherine" by Gaudenzio Ferrari (16th century) held at the Pinacoteca di Brera in Milan [20] [Fig 1S]. About 75 analysis point were considered for FORS analyses, which results were synergistically compared with XRF and imaging techniques. In this work, to test our automatic classifier, we selected 18 unambiguously identified spectra avoiding complex mixtures and uncertainly labelled pigments as detailed in section 4.2.2.

4.2 Results

4.2.1 The basic dataset

Since the basic dataset consist of a limited set of spectra it offers the best condition to evaluate the effect of different binders and substrates in the spectral features, and to observe how the different pre-processing methods and the different distance measurements can influence the recognition of the spectra.   Results of the classification algorithm obtained exploiting different ordered derivatives are summarised in Table 1, where the accuracy score is reported: the use of this parameter is widespread in describing the algorithm performance [21] .

Table 1: accuracy score reported for the basic dataset, varying the distance applied and the derivative considered.

	Euclidean	RMS	SAM	SCM	SID	KLPSD
Raw data	0.38	0.38	0.66	0.63	0.67	0.37
First derivative	0.87	0.87	0.91	0.92	0.88	0.75
Second derivative	0.83	0.83	0.86	0.89	0.79	0.72
Third derivative	0.87	0.87	0.87	0.89	0.85	0.90
Fourth derivative	0.87	0.87	0.87	0.87	0.87	0.87
Fifth derivative	0.87	0.87	0.80	0.80	0.76	0.80

As can be seen from Table 1, the different distance measurements have a different performance in correctly identifying the unknown spectrum. However, this result has a strong dependence to the amount of the unknown spectra and the number and types of spectra in the spectral library. More interesting is that the different performance of the measures is less evident on the unity-based derivatives of the spectra. For example, KLPSD performs better using the third derivative, while SCM using the first derivative. We highline that the lack of recognition of some spectra is linked on the type of pigment used, but also on its binding medium and layering substrate. The spectra of azurite powder and cadmium red are those worst classified as, despite their spectra present the same trend as the relative layering, they are distorted with respect to the spectra in the reference database library, all spread with binders. As far as green pigments are concerned, the problems for misclassification are instead due to the notorious difficulty of their classification through FORS as the spectra of copper-based green pigments are very similar [22]. Therefore, malachite from our dataset is often identified as brochantite. Finally, often some pigments with a different name and use, due to extraction method, historical period or field of application, such as iron oxide pigments, are difficult to discriminate, having the same chromophore and thus very similar spectral signature, barely discriminated from a search algorithm based on the spectral point-to-point distance (Fig. 2).
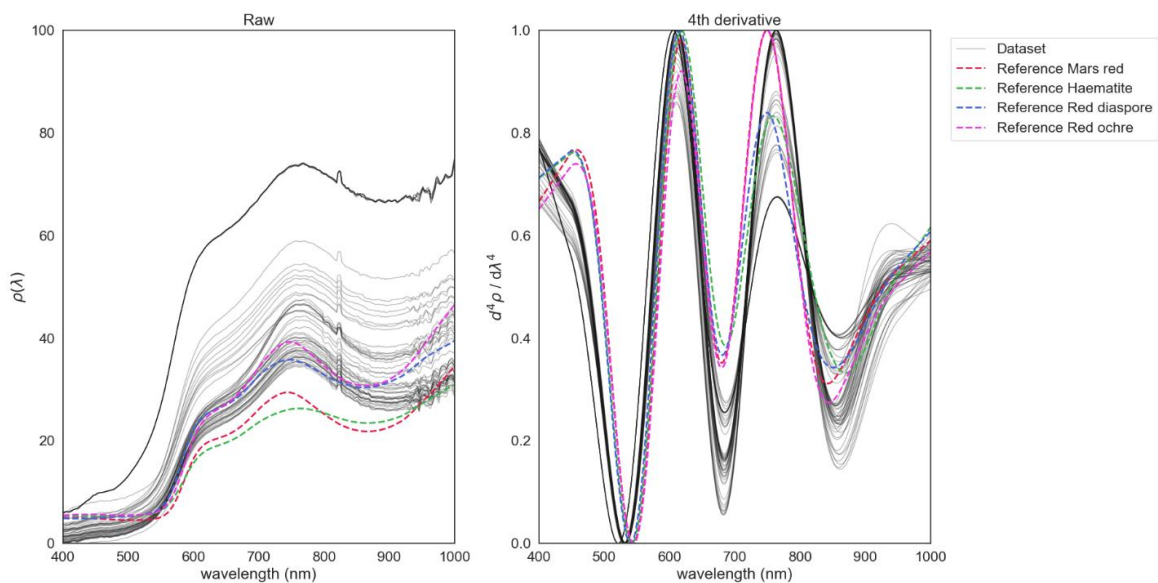

Fig. 2: Raw spectra and fourth derivative for iron oxides based pigments
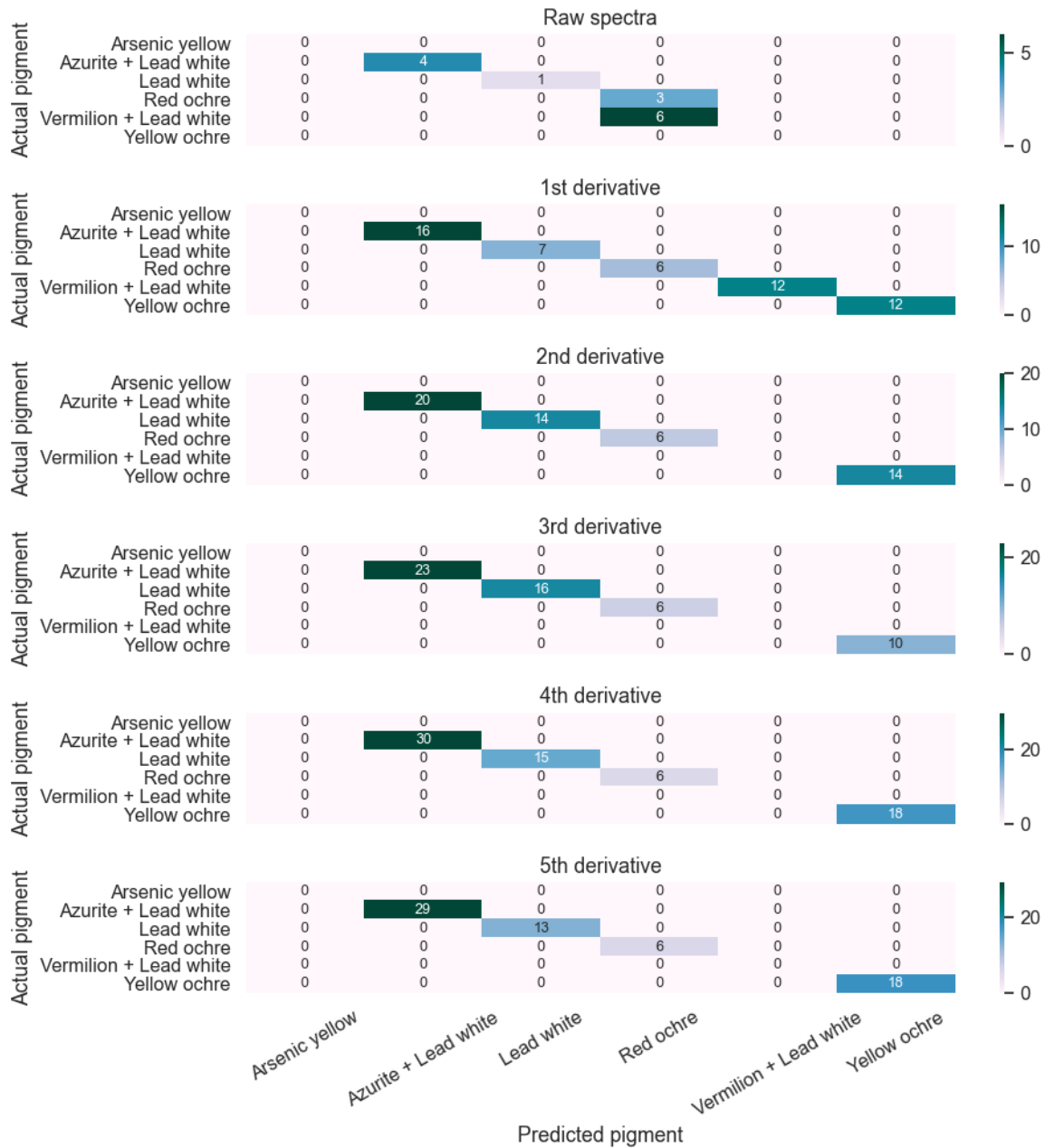
4.2.2. Case study

Fig. 3: confusion matrices for the selected of pigment layers from Gaudenzio Ferrari painting.

In the previous section, only a small set of pigments and mixtures were studied (basic dataset). In the present section, we report the result of an automatic classification algorithm for spectra acquired in a previous published work [20]. In that work, the pigments have been identified combining multiple techniques allowing to compare the result of the classification. The present section aims to highlight and discuss some common problems that may be encountered during real case study applications, first of all the alteration or degradations of pigments, binders or varnishes that affect the reflectance spectra shape. We choose a selection of the 80 points considered, namely those resulting in lead white layering, blue, yellow and red areas.

Fig. 3 shows the confusion matrices for all these selected areas. In order to assure readability, we provide the results considering all the distance measures aggregating the spectra based on their class. It is possible to notice that, globally, the fourth derivative is able to correctly identify the greatest number of spectra.

As an example, in Fig. 4 we show the spectra corresponding to white areas for Gaudenzio masterpiece where all of them have been identified as white lead on the basis of FORS and XRF spectra. We first tried to verify if this white pigment can be automatically recognized by the proposed method. As it is known, the kind of white pigment in quite important for an indirect dating and the detection of restored areas. Indeed, in real paintings, almost no pure white areas are present, and this makes bot manual and automatic identification of white spectra more difficult.

The raw spectra of Gaudenzio white areas present a lower spectral reflectance factor with respect to reference ones. This is due both to the aging of the binder, the varnishes and the effect of soiling, and to the presence of small quantities of other pigments. In general, as can be observed by comparing the spectra of whites reported in Fig 3 with the spectra of red ochers in Fig 2, the main uncertainty and difficulty in identification lies in the almost complete absence of spectral features of reflectance spectra from white pigments, which can be distinguished only by the different rise in blue-green and yellow wavelengths. In this case, even the presence of small noise signals is amplified in the higher order derivatives and the effect of yellowing due to the aging of pigments and/or binders is very evident. Despite this limit, the spectra are correctly recognized using any distance with the third and fourth derivative. On the opposite, raw spectra are not correctly classified, while using the first and second derivatives, the performance is very different depending on the metrics, with SCM recognizing them all, while EUC and RMS identifying the white lead as lithopone.
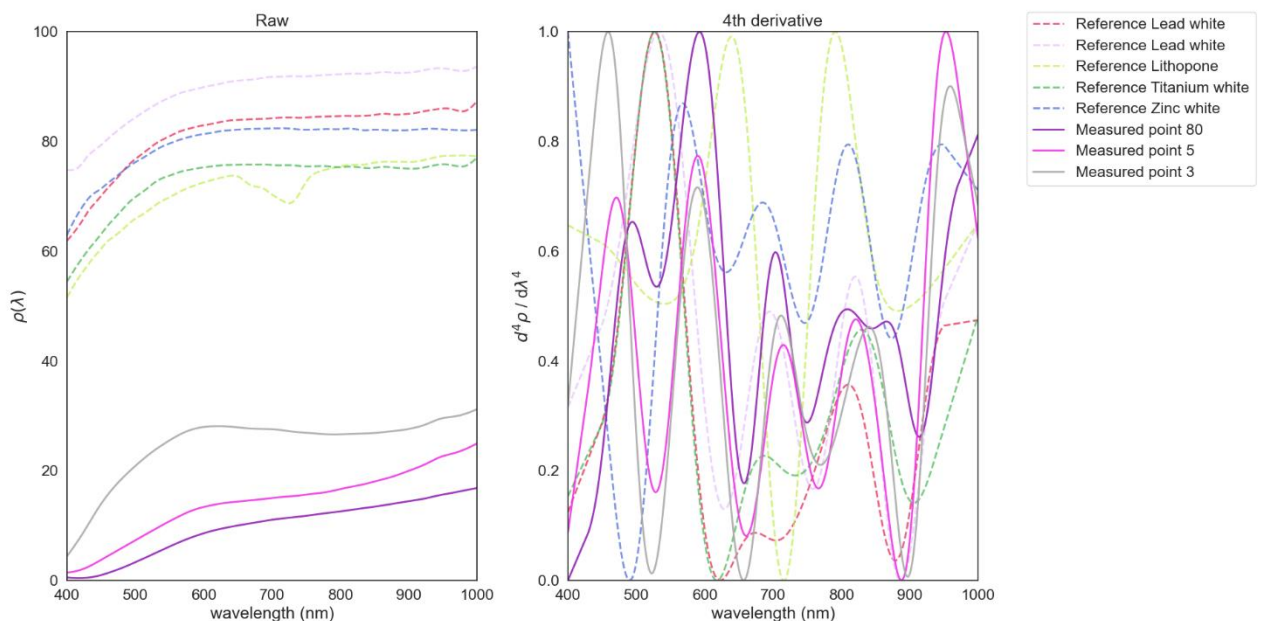
Fig. 3: white lead FORS spectra from Gaudenzio masterpiece previously manually identified on the basis of both FORS and XRF spectra.

Different is the situation for blue pigment spectra which were recognized as a mixture of azurite and white lead in the already quoted work [20] by comparing FORS and XRF results. In this case, the order 3, 4 and 5 derivatives allow the recognition of the detected spectra as a mixture of azurite and white lead, while using the second derivative the correct classification is accomplished only by SCM, SID and KLPD, while EUC, RMS and SAM recognize the spectral signatures of azurite, but not its mixture with white lead. Red spectra show major variability: red ochre was always correctly identified in all the derivative orders while vermilion red mixed with lead white was identified only using the second derivative. Yellow ochre was always identified while arsenic yellow was never identified.

4.3 A prototype of User Interface

Taking into consideration the issues discussed above, to achieve our goal of creating a framework to help the identification of pigment and dye spectra using reflectance spectrophotometry also by non-skilled users, we have developed a User Graphic Interface with Tkinter frontend (see fig. 2S in supplementary materials). It allows real-time manipulation and recognition of spectra, facilitating the traditional user-based comparative approach in conjunction with the automatic methodology based on spectral matching algorithms. For each preprocessing step (interpolation, smoothing, derivation and recognition), It is possible to choose among some statistical and graphic solutions that guide and help the operator to correctly elaborate and recognize the spectra. There is therefore the possibility to process the spectrum following a guided path, for instance by choosing the smoothing degree and the order of derivative. The comparison with the reference library database spectra takes place by confronting all the distance measurements proposed so far, with the possibility, based on the operator's choices, to add new ones or remove some. After the identification ends, it is not proposed a single identified spectrum, but the 10 spectra having the greatest similarity; for each selected similarity measure, the system removes duplicates if in the spectral library there are more spectra belonging to the same class. Finally, there is the possibility to access, download and locally alter the reference library to obtain a quick visual comparison. As for the automatic search, it is possible to dynamically work on a sub selection of the library spectra based on hue, binding media and artistic technique or period of usage. This design allows to advanced users to better manage and customize the whole procedure.

5. Conclusions

What is clearly appearing from this preliminary work, is that totally automated spectral identification based on a similarity measure is problematic for many factors (e.g. intrinsic difference in noise and resolution of different instrument, low characteristic of Vis-NIR reflectance spectra as many pigment present the same chromophores, completeness of the reference spectral library, choose of the similarity measure and the preprocessing methodologies...). With this awareness, we are planning a dynamic method, suitable for both non-skilled and expert users.

Future work includes the refinement of the search algorithm that will be no longer based on a simple point-to-point distance, but will exploit a similarity learning model, while maintaining the

characteristics of dynamic matching. Furthermore, we are planning to add the possibility of sub selecting the reference library in relation to other types of analysis often complementary to FORS analysis, such as XRF, Raman and colorimetric ones, or to any historical or conservative information about the painting under study.
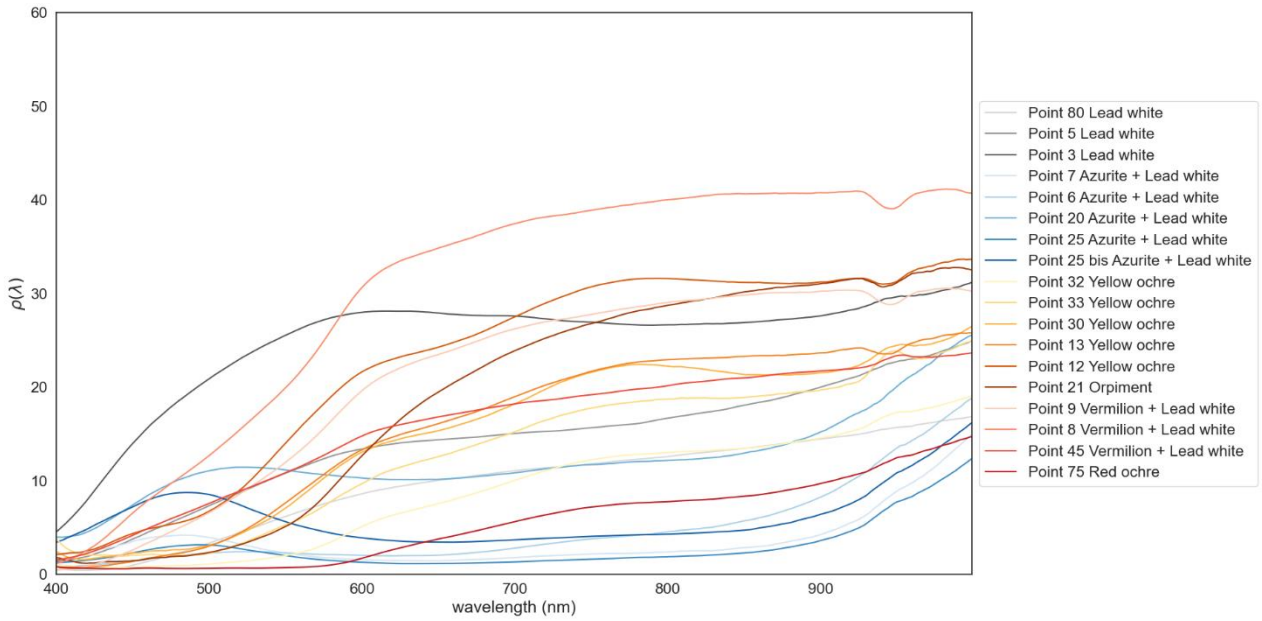
6. Acknowledgments

Declarations

The authors declare that they have no conflict of interest.

Supplementary Materials

Fig.1S: FORS spectra from ref. 20 considered in the present work
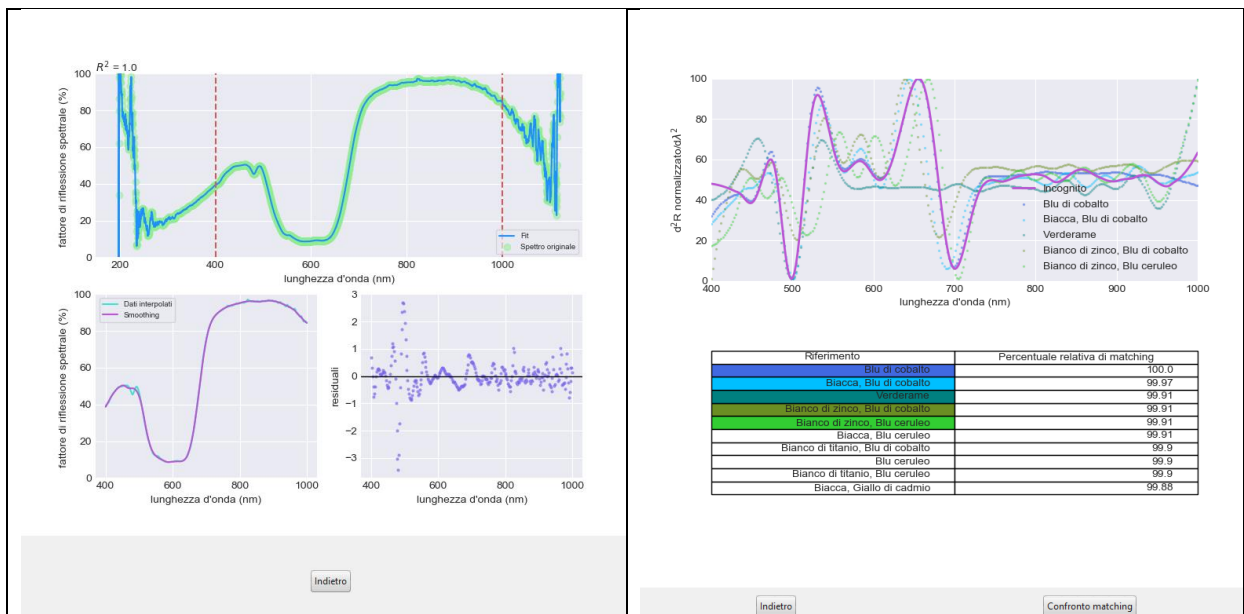
Fig.2S: User Graphic Interface with Tkinter frontend

Reference

1. M. Picollo, M. Bacci, A. Casini, F. Lotti, S. Porcinai, B. Radicati, and L. Stefani, in *Optical Sensors and Microsystems: New Concepts, Materials, Technologies*, edited by S. Martellucci, A. N. Chester, and A. G. Mignani (Springer US, Boston, MA, 2000), pp. 259–265.

2. M. Gargano, N. Ludwig, and davide pandini, Journal of the International Colour Association **8**, (2012).

3. L. Bonizzoni, S. Bruni, A. Galli, M. Gargano, V. Guglielmi, N. Ludwig, L. Lodi, and M. Martini, Microchemical Journal **126**, (2015).

4. H. Deborah, N. Richard, and J. Y. Hardeberg, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **8**, 3224 (2015).

5. V. Kumar, J. K. Chhabra, and D. Kumar, INFOCOMP Journal of Computer Science **13**, 38 (2014).

6. J. Li, D. B. Hibbert, S. Fuller, and G. Vaughn, Chemometrics and Intelligent Laboratory Systems **82**, 50 (2006).

7. F. van der Meer, International Journal of Applied Earth Observation and Geoinformation **8**, 3 (2006).

8. N. Rohani, E. Pouyet, M. Walton, O. Cossairt, and A. K. Katsaggelos, in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), pp. 3217–3221.

9. B. Grabowski, W. Masarczyk, P. Głomb, and A. Mendys, Journal of Cultural Heritage **31**, 1 (2018).

10. T. Kleynhans, D. W. Messinger, and J. K. Delaney, Microchemical Journal **157**, 104934 (2020).

11. C. Ruffin and R. L. King, in *IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS'99 (Cat. No.99CH36293)* (1999), pp. 756–758 vol.2.

12. F. Tsai and W. D. Philpot, in *Remote Sensing for Geography, Geology, Land Planning, and Cultural Heritage* (International Society for Optics and Photonics, 1996), pp. 200–211.

13. F. N. Fritsch and R. E. Carlson, SIAM J. Numer. Anal. **17**, 238 (1980).

14. (n.d.).

15. YUHAS, Roberta H, GOETZ, Alexander FH, and BOARDMAN, Joe W., Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop. 147 (1992).

16. DE CARVALHO, O. Abilio and MENESES, Paulo Roberto, Summaries of the 9th JPL Airborne Earth Science Workshop (2000).

17. Chein-I Chang, in *IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS'99 (Cat. No.99CH36293)* (1999), pp. 509–511 vol.1.

18. A. Plutino, N. Richard, H. Deborah, C. Fernandez-Maloigne, and N. G. Ludwig, Color and Imaging Conference **2017**, 141 (2017).

19. H. Deborah, N. Richard, M. Ö. Úlfarsson, J. A. Benediktsson, and J. Y. Hardeberg, in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium* (2019), pp. 1104–1107.

20. M. Gargano, L. Bonizzoni, E. Grifoni, J. Melada, V. Guglielmi, S. Bruni, and N. Ludwig, Journal of Cultural Heritage (2020).

21. A. Tharwat, Applied Computing and Informatics **17**, 168 (2020).

22. L. Bonizzoni, M. Gargano, N. Ludwig, M. Martini, and A. Galli, Appl. Spectrosc., AS **71**, 1915 (2017).