Method Article

PCA-based method for managing and analyzing single-spot analysis referenced to spectral imaging for artworks diagnostics

Giacomo Marchioro, Claudia Daffara*

Dept. of Computer Science, University of Verona, Strada le Grazie 15, 37134, Verona, Italy

ABSTRACT

Artworks diagnostics is based on the joint use of several nondestructive techniques to acquire complementary information on the materials. A common practice in the field is to perform the analyses with single-spot analytical techniques, e.g. spectroscopy-based, after a preliminary screening of the artwork with full-field imaging-based techniques. We present a method and its practical implementation for fusing and analyzing data collected using analytical systems that acquire single spot measurements mapped to spectral imaging stacks. The fused dataset of single-spot and imaging observations is analyzed using principal component analysis (PCA). The effectiveness of the method for artworks diagnostics is shown on spectroscopy and imaging datasets of an ancient canvas painting. The results of the PCA analysis on the final fused dataset are compared against the PCA analysis performed on the original datasets from single-spot and imaging measurements taken separately. We propose two practical implementations of the procedure, one based on using graphical user interface (GUI) and open-source GIS software (QGIS), the other one based on an open-source Python module, named SPOLVERRO, specifically developed for this project and released on a public repository. The method allows conservation scientists to analize effectively the heterogeneous datasets acquired in a diagnostic campaign.

- single-spot spectroscopy data are referenced on imaging data.
- the sampling area of each spectroscopy spot is used for extracting and averaging the respective imaging data values.
- the final matrix is analyzed using PCA for extracting further information.

Specification Table

Subject Area:	Environmental Science
More specific subject area:	Diagnostics of Cultural Heritage, multi-modal data fusion
Method name:	PCA-based method for managing and analyzing single-spot analysis referenced to spectral imaging for artworks diagnostics
Name and reference of original method:	NA
Resource availability:	https://github.com/giacomomarchioro/spolverro

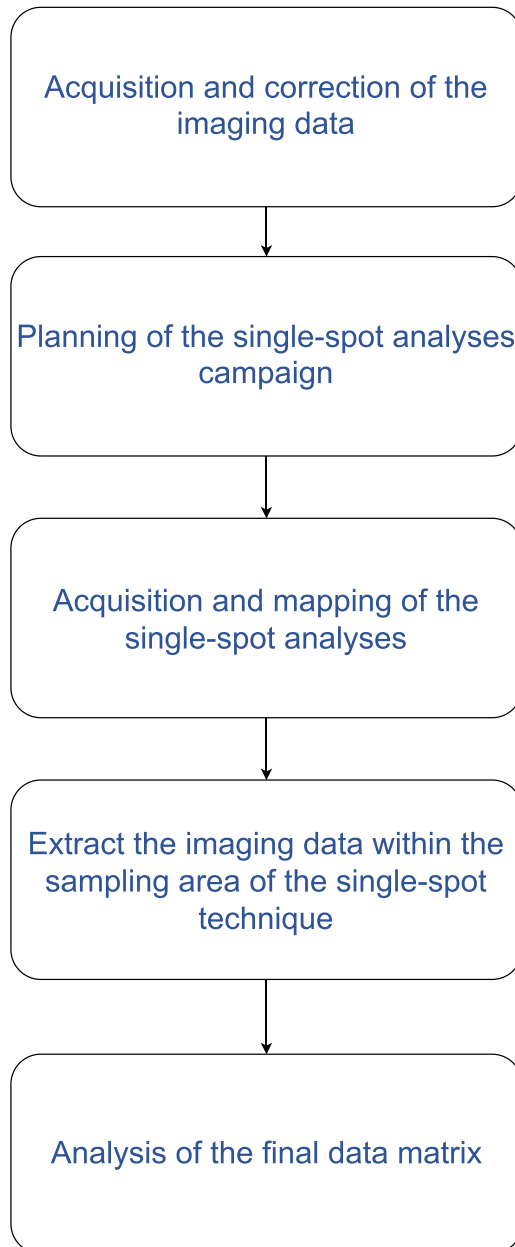## Method details

Background

During the diagnostic campaign different nondestructive techniques are used for investigating the artwork [1–4]. We can distinguish between systems that are able to acquire images (imaging systems) and systems that are able to acquire only a single spot for each measurement, e.g. spectroscopy systems. In this method we call the latter single-spot systems, where we mean by "spot" the sampling area, i.e. the area and the associated volume of the artwork from which most of the analytical signal is collected. Even though in recent years researchers have been developing many spectroscopy scanning systems able to scan relatively large areas (in the order of the square metre), these systems are often expensive and not commercially available. Furthermore, in case of large artworks (such as frescos and large canvas paintings) the time required for scanning the entire surface is often prohibitive. In the current praxis, imaging techniques comprising multispectral and hyperspectral cameras are used for exploring large areas, while, instead, hand-held and portable spectrometers are often used in a limited number of selected points for gathering further information. The management of the different data collected using the various techniques is crucial for taking full advantage of the information contained. But actually, conservators and conservation scientists often lack the required computer literacy for combining and analyzing heterogeneous datasets.

In this context, we propose a method and its practical implementation for fusing and analyzing the data collected from analytical single-spot systems mapped to spectral imaging stacks. The procedure was originally developed to support the analysis of spectroscopy and imaging data routinely acquired during the diagnostics of ancient artworks in the museums of Venice [5], then it was standardized and, here, released as available resource.
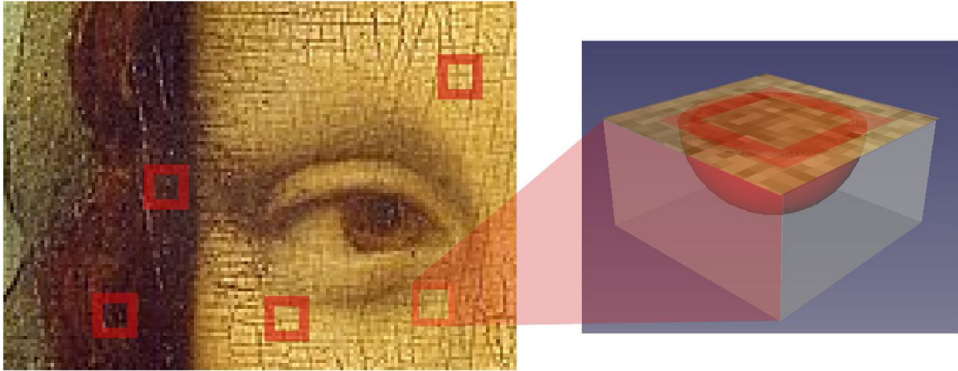
The need to perform the method in a reproducible way, for example, by using programming languages to write all the steps of the analysis as reproducible script, may limit the adoption of the method to specific users. Hence, in parallel with the implementation of the method using scripts we show an alternative practical solution based on using software with graphical user interface (GUI). Geographical information systems (GIS) can be used effectively in the management of the data collected during the diagnostic campaign on artworks [6]. Because of the wider user community, GIS are generally stable and mature and can be used as an alternative to *ad hoc* solutions for Cultural Heritage [7]. Cultural Heritage operators and scholars using GIS are facilitated by the GUI, however, because of the many functionalities of this software, they often need to be guided step by step.

General procedure

The general procedure to design data management and analysis in artwork diagnostics can be summarized in the steps in Fig. 1. In the first phase, imaging systems are used for exploring a large area or the entire area of the artwork. If the acquired images are distorted, it is preferable to perform the correction of the distortion before proceeding with the following steps. In the second phase, the

Acquisition and correction of the imaging data

Planning of the single-spot analyses campaign

Acquisition and mapping of the single-spot analyses

Extract the imaging data within the sampling area of the single-spot technique

Analysis of the final data matrix

Fig. 1. A diagram illustrating the steps of the method.

Fig. 2. Illustration of sampling area and associated experimental surface for the single-spot analysis. In the application shown, the side of the outlined internal square is approximately 3 mm; the sampling area is typical of XRF handheld spectrometers.

conservation scientist performs a first analysis of the data and plans together with conservators and art historians further investigations based on using, for example, single-spot systems as portable X-ray fluorescence (XRF), FT-IR, and Raman spectrometers. During the acquisition, the position of the measurements performed with the single-spot technique is recorded and eventually mapped to the imaging data. At this stage, the heterogeneous data collected from imaging and single-spot techniques can be fused in a single dataset. Based on the position of the referenced single-spot measurement and on the estimated sampling area of the single-spot technique (Fig. 2), the software extracts the pixels of the imaging technique enclosed in the boundary of the spot and computes the average of the pixel values belonging to the same multispectral channel. The result is a data matrix $p \times n$, where p indicates the number of rows representing the observations, equal to the number of spots collected, and n indicates the columns containing the values of each single-spot analysis and the respective averaged multispectral imaging values. In the last phase, the final data matrix can be analyzed using multivariate analysis. In this method, we propose principal component analysis (PCA) for extracting meaningful information from the data fusion process and validating the procedure.

Software needed

The general procedure described in the previous section has been implemented by taking advantage of available free and open-source tools and the developed code is made available as a Python module, that we called SPOLVERRO. For implementing the software we used Python 3.6 programming language, together with third-party open-source scientific libraries to make the workflow automated and more reproducible. The Python Shapefile Library (PyShp v.2.1) is used for reading the shapefile vector data format, while raster data (i.e. imaging data) are processed using the Rasterio library (v. 0.10) [8], which is based on the Geospatial Data Abstraction Library (GDAL) [9] by the Open Source Geospatial Foundation. For the analysis of the data we used the NumPy and the SciPy library, together with the scikit-learn library [10] for computing the PCA. The Matplotlib library [11] is used for plotting the outputs.

Regarding the alternative implementation of the method based on GIS software, we choose QGIS (version 2.16 Nødebo or greater), a free and open-source multiplatform software with a wide community of users able to provide support. QGIS can be used for performing the first four steps of the procedure (Fig. 3).

Hardware needed: imaging and single-spot instruments

To apply the method it is needed at least one imaging technique and one single-spot technique for performing the measurements. The method presented is more effective when multispectral or
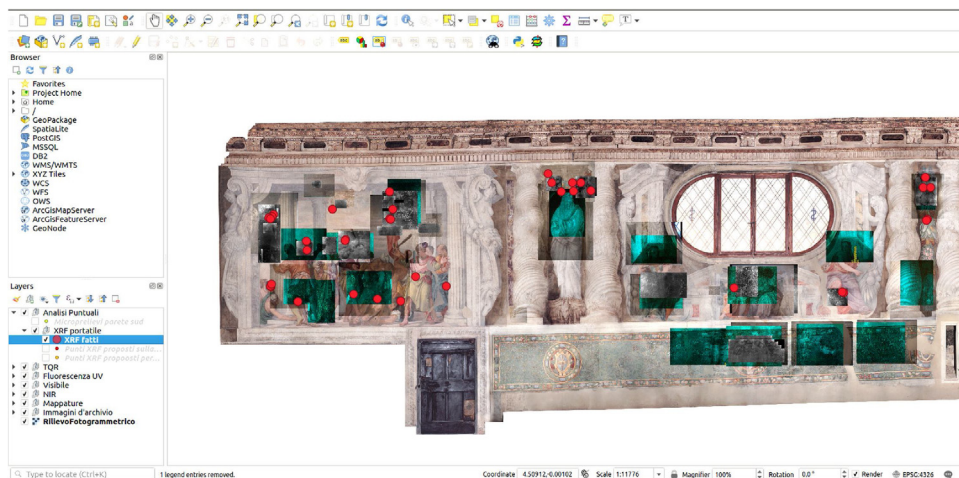
Fig. 3. A screenshot of a QGIS project, where XRF measurements are mapped to multiple raster images over a large mural painting by Veronese (Venice, San Sebastiano church).

hyperspectral imaging systems are combined with portable or handheld spectrometers, as shown. However, the method can in principle be applied also with systems that record a limited number of bands or variables. The method, as reported here, assumes that the imaging technique has a spatial resolution equal or higher than the single-spot technique. However, it can be applied also with imaging techniques with lower spatial resolution assuming a certain homogeneity of the artwork material.

Step 1 – acquisition and correction of the imaging data

The first step consists in acquiring the imaging data. It is preferable to correct any distortion of the images to ensure a correct referencing of the acquisitions collected using the single-spot techniques. The distortions can be corrected using specific software, otherwise, if ortho-photos are available as ground truth, the images can be registered manually using control points and the "Georeferencer Plugin" of the QGIS software. If the imaging systems output the images in a format not readable by QGIS, the GDAL drivers can be used for converting the multispectral or hyperspectral image stack to a suitable format e.g. geoTIFF or .lan. If we assign a projection when importing the data in GIS, we must be sure to use a planar projection e.g. EPSG: 32632 and not to select ellipsoidal measurements.

Step 2 – planning of the single-spot analyses campaign

After the acquisition of the imaging dataset a first inspection of the images collected is propaedeutic to the successive step. In this case, false-colour composite imaging is often helpful for identifying together with the conservators and art historians the problematic areas (for example, addressing question such as detecting original or restored areas of the artwork) and guiding the following sampling using single-spot analyses. There are many strategies for facilitating this process, for example, a vector layer in the GIS software can be created for storing the positions where the analysis will be performed. This planning step is very important in case the single-spot instrument is not at disposal for the duration of the entire diagnostic campaign and hence the time (and cost) of measurements must be optimized.

Step 3 – acquisition and mapping of the single-spot analyses

The position where the single-spot measurements are acquired must be recorded accurately. If mechanical positioning systems or optical tracking systems are available, they can be used for

assigning automatically the position to each spectrometry measurement relative to the images coordinate reference system. Otherwise, GIS software can be used for creating a new vector layer and the operator can reference manually the single spot-measurements to the imaging data previously imported. It is a good practice to take photos during the acquisition and establish a unique ID for each measurements to ensure the correct referencing of the single points. In this way, we can store the position of the points (for instance, as a shape file) and later join the corresponding single-spot analysis data. This can be done by exporting the spectroscopy data in a common format e.g. csv and by importing it as an attribute table, which can be joined with the shape file using the built-in QGIS "join feature" into the layer properties and the ID as join field. The same procedure can be carried out using the Python Shapefile Library.

*Step 4 – extract the imaging data within the sampling area of the single-spot technique*

For creating a new matrix containing both the single-spot analyses and the corresponding signal recorded with the imaging techniques in the same position, for each location of the single-spot analyses the corresponding imaging values must be extracted. For performing this task, the sampling area of the spectroscopy technique used must be previously estimated. The imaging data within the sample area are extracted and averaged for each spectral channel or band of the imaging technique. This procedure can be implemented programmatically using Python and Rasterio. In this case, we approximate the sampling area to a square of side equal to the diameter of the sampling area. Knowing the corresponding physical dimension of the pixel in the artwork surface, a grid of points equivalent to the sampling area is calculated. For each point the nearest raster pixel for each channel is extracted. Otherwise, using QGIS, a similar procedure can be carried out using the "Point sampling tool" plugin or the Saga function "Raster Values to Points" (in QGIS v. > 3.14).

Step 5 – analysis of the final data matrix

The final data matrix is composed of a number of rows corresponding to the number of measurements performed with the single-spot technique. Each row is the joint vector comprehending the single-spot analysis data and the corresponding averaged imaging values extracted within the sampling area for each spectral channel or band. In other words, each row is an observation while each column represents a variable. Different preprocessing of the data can be carried out before applying multivariate analysis to the data. However, it might be reasonable to exclude noisy channels before applying further processing. In this method, we suggest using Principal Component Analysis after the standardization of the dataset. The standardization is performed by removing the mean and by scaling to unitary variance the fused dataset. The processing can be carried out via scripts using scikit-learn [10], otherwise, for non-Python users, using available free and open-source tools with GUI such as Gnumeric. Eventually, clustering techniques can be used to identify clusters in the score plots obtained with the PCA analysis.

Method validation

The method has been validated on datasets of an ancient large canvas painting (called *telero*), masterpiece by the venetian artist Carpaccio, collected during the Iperion [12] project as presented in our publication [5]. In the case study, the multispectral imaging technique is combined with the XRF spectroscopy single-spot technique. A multispectral cube has been collected using a Vis-NIR scanner (16 bands in the visible and 16 bands in the near infrared up to 2250 nm) while a set of 47 spot analyses has been performed using a hand-held XRF spectrometer over the red areas of the painting. To validate the method we first performed the steps 1–4 of the procedure, then we tested the effectiveness of the method for analyzing the two datasets separately and after the fifth step of the procedure that fuses the two datasets together. Once extracted the values for each spot we standardized the Vis-NIR dataset and the XRF dataset separately, we performed the PCA on each dataset, and we evaluated the results against the PCA computed on the two datasets fused in a single representation after the standardization. Previously, each observation was categorized into two groups, one comprising the
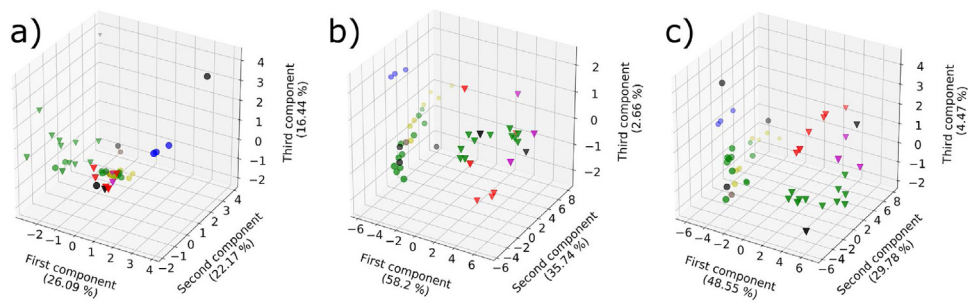
Fig. 4. Score plots of the first three principal components of the PCA performed on the XRF spectroscopy dataset (a), on the Vis-NIR imaging dataset; and on the joint XRF–Vis-NIR dataset obtained using our method (c).

various mixtures of pigments and labelled using colour depending on the attributed pigment, the other one using different markers representing data collected in a compromised area (circle marker) or not (triangle marker).

The results are shown in Fig. 4 depicting the score plots of the first three principal components of the PCA performed on the datasets. The score plot of the XRF dataset (Fig. 4a) highlights some clusters with different mixture of pigments. The score plot with the first three principal components of the dataset obtained extracting the values from the Vis-NIR multispectral cube (Fig. 4b) shows two clusters separated by the first principal component, one representing the compromised pictorial layers and the other one relative to the data collected on the other areas. However, in this case the clusters relative to the pigment mixture are not clear. Finally, the score plot obtained using the joint dataset matrix (Fig. 4c) clearly distinguishes the data collected on the damaged areas from the others. Furthermore, we can identify different clusters depending on the pigments used in the pictorial layer.

Hints

In general, we must consider that even if the data is referenced on the artwork surface the signal may come from different depths of the pictorial layer, depending on the properties of the materials and the radiation used. Hence, the information retrieved must be considered relative to an unknown portion of volume rather than to the sampling area. The contribution of artwork stratigraphy materials to the collected signal is not easy to determine and generally it varies depending on the radiation wavelength. This issue concerns both the use of analytical single-spot techniques and multispectral imaging techniques.

In step 4, the user may average the imaging data on an area greater than the effective sampling area of the single-spot technique. In general, the size of the area to be averaged should take into account also the error in mapping the single-spot measurement to the raster image.

In step 4, it can be useful to calculate, besides the average, also the standard deviation for each multispectral channel. This can give an insight into the homogeneity of the sampled area. In artworks diagnostics this issue may be crucial due to the high heterogeneity of the materials. If the standard deviation is high, errors in the process of mapping the single-spot measurements to the imaging data may have great influence on the final result.

In step 4, the user should know how the software computes the extraction of the values from the image stack in the sampling area. In particular, it must be clarified if the values are interpolated or are extracted from the nearest pixel. In our case, we extracted the values to the nearest pixel.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

References

[1] H. Liang, Advances in multispectral and hyperspectral imaging for archaeology and art conservation, Appl. Phys. A-Mater. 106 (2012) 309–323, doi:http://dx.doi.org/10.1007/s00339-011-6689-1.

[2] K. Janssens, G. Van der Snickt, F. Vanmeert, S. Legrand, G. Nuyts, M. Alfeld, L. Monico, W. Anaf, W. De Nolf, M. Vermeulen, Non-invasive and non-destructive examination of artistic pigments, paints, and paintings by means of X-ray methods, Top. Curr. Chem. 374 (2016) 81, doi:http://dx.doi.org/10.1007/s41061-016-0079-2.

[3] M. Alfeld, L. de Viguerie, Recent developments in spectroscopic imaging techniques for historical paintings-a review, Spectrochim. Acta B 136 (2017) 81–105, doi:http://dx.doi.org/10.1016/j.sab.2017.08.003.

[4] C. Daffara, S. Parisotto, D. Ambrosini, Multipurpose, dual-mode imaging in the 3–5 μm range (MWIR) for artwork diagnostics: a systematic approach, Opt. Laser Eng. 104 (2018) 266–273, doi:http://dx.doi.org/10.1016/j.optlaseng.2017.10.006.

[5] N. de Manincor, G. Marchioro, E. Fiorin, M. Raffaelli, O. Salvadori, C. Daffara, Integration of multispectral visible-infrared imaging and pointwise X-ray fluorescence data for the analysis of a large canvas painting by Carpaccio, Microchem. J. 153 (2020), doi:http://dx.doi.org/10.1016/j.microc.2019.104469.

[6] W. Schmid (Ed.), GRADOC - Graphic Documentation Systems in Mural Painting Conservation. Research Seminar, Rome 16–20 November 1999, ICCROM, Rome, 2000.

[7] A. Amat, C. Miliani, B.G. Brunetti, Non-invasive multi-technique investigation of artworks: a new tool for on-the-spot data documentation and analysis, J. Cult. Herit. 14 (2013) 23–30, doi:http://dx.doi.org/10.1016/j.culher.2012.02.015.

[8] S. Gillies, et al., Rasterio: Geospatial Raster I/O for Python programmers, (2013) . https://github.com/mapbox/rasterio.

[9] F. Warmerdam, The geospatial data abstraction library, in: B. Hall, M.G. Leahy (Eds.), Open Source Approaches in Spatial Data Handling, Springer-Verlag, Berlin Heidelberg, 2008, pp. 87–104, doi:http://dx.doi.org/10.1007/978-3-540-74831-1.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[11] J.D. Hunter, Matplotlib: a 2D graphics environment, Comput. Sci. Eng. 9 (2007) 90–95, doi:http://dx.doi.org/10.1109/MCSE.2007.55.

[12] IPERION CH Integrated Platform for the European Research Infrastructure ON Cultural Heritage, Molab Facility, (2019) . (Accessed 6 September 2019) http://www.iperionch.eu/molab/.