Invited paper

Emeline Pouyet, Neda Rohani, Aggelos K. Katsaggelos, Oliver Cossairt and Marc Walton*

Innovative data reduction and visualization strategy for hyperspectral imaging datasets using t-SNE approach

https://doi.org/10.1515/pac-2017-0907

Abstract: Visible hyperspectral imaging (HSI) is a fast and non-invasive imaging method that has been adapted by the field of conservation science to study painted surfaces. By collecting reflectance spectra from a 2D surface, the resulting 3D hyperspectral data cube contains millions of recorded spectra. While processing such large amounts of spectra poses an analytical and computational challenge, it also opens new opportunities to apply powerful methods of multivariate analysis for data evaluation. With the intent of expanding current data treatment of hyperspectral datasets, an innovative approach for data reduction and visualization is presented in this article. It uses a statistical embedding method known as t-distributed stochastic neighbor embedding (t-SNE) to provide a non-linear representation of spectral features in a lower 2D space. The efficiency of the proposed method for painted surfaces from cultural heritage is established through the study of laboratory prepared paint mock-ups, and medieval French illuminated manuscript.

Keywords: ChemCultHerit; data reduction and visualization; illuminated manuscript; multivariate analysis; t-distributed stochastic neighbor embedding; visible hyperspectral imaging.

Introduction

Background and research aim

The analytical identification and classification of colorants augments our understanding of artistic practice and informs how works of art are conserved. In this context, multi- or hyperspectral reflectance imaging has become a workhorse technique for the characterization of painted materials [1–4]. This success may be attributed to the merits of the technique: it is non-invasive, relatively inexpensive and allows for the wide field imaging of an artwork in under a few minutes. Particularly, visible hyperspectral imaging (HSI) (400–900 nm) has proven to be a powerful method to map the distribution of colorants across a painted surface. However, the contiguous set of 2D images collected through the visible range produces a 3D data cube containing millions of spectra that poses an analytical and computational challenge. Therefore, extracting meaningful information from such large spectral volumes requires an efficient and robust approach to undertake unsupervised and/or supervised reduction of the multivariate data [5].

Article note: A special issue containing invited papers on Chemistry and Cultural Heritage (M.J. Melo, A. Nevin and P. Baglioni, editors).

***Corresponding author: Marc Walton,** Center for Scientific Studies in the Art, Northwestern University, 2145 Sheridan Rd., Evanston, IL 60208, USA, e-mail: marc.walton@northwestern.edu
Emeline Pouyet: Center for Scientific Studies in the Art, Northwestern University, 2145 Sheridan Rd., Evanston, IL 60208, USA
Neda Rohani, Aggelos K. Katsaggelos and Oliver Cossairt: Electrical Engineering and Computer Science, Northwestern University, 2145 Sheridan Rd., Evanston, IL 60208, USA

Data reduction is expected to facilitate visualization and classification of the data as well as to reduce computation time. Thus it requires that the multivariate structure of the original dataset is maintained in the lower-dimensional representation in order for the visualization and the resulting coordinates of the reduced space to have meaning to the user.

As described in more depth below (Section "Previous work"), most multivariate data reduction methods in cultural heritage studies rely on either principal component analysis (PCA) or minimum noise fraction (MNF) analyses [5–13]. As well-established techniques for data reduction, their effectiveness for high-dimensional dataset of non-linear mixtures of pure reflectance spectra present several limitations. First, both methods assume that the underlying data manifold is linear [14]; an erroneous assumption as it is well established through the Kubelka-Munk (KM) theory of reflectance that pigment mixtures absorb and scatter light non-linearly [3, 15, 18]. Moreover, for these techniques, dimensionality reduction is obtained by discarding the components with the lowest variance which does not always accurately model chemical variation of the high-dimensional dataset [16, 17]. Finally, the visualization and segmentation of high-dimensional data is compromised when examining only several selected 3D projections, as is a common practice.

To overcome these limitations, this work proposes a novel technique for data reduction called t-distributed stochastic neighbor embedding (t-SNE) [16]. To highlight the advantages of this innovative approach for cultural heritage, the results obtained on mock-up paint samples are compared to two established data reduction and segmentation methods, namely (i) PCA combined with K-means clustering, (ii) and MNF combined with pure pixel index (PPI) endmember extraction. The potential of the proposed approach is further examined for the study of a historical object: an illuminated page from a French medieval Book of Hours. These examples demonstrate the potential of t-SNE to classify, and map pure and mixed pigments in artistic material.

Previous work

Within the field of cultural heritage, data reduction and segmentation is often undertaken using statistical techniques borrowed from remote sensing applications [3]. For instance, Baronti et al. [6] and more recently Mounier et al. [11] applied PCA for detecting areas of paintings characterized by similar chemical composition or/and physical properties. The eigenvectors and eigenimages of the first principal components are used to identify and map spectral features characteristic of specific pure pigment or pigment mixture. In the context of illuminated manuscript and painting studies, two additional multivariate analysis protocols have been recently proposed [8, 9, 13]. The first approach uses an algorithm called the "hourglass paradigm" implemented in the ENVI software [9]. This approach can be broken into two steps. First, the reduction of dimensionality and denoising of the data is performed using MNF, where the data cube is transformed into a contiguous set of images with regularly increasing noise levels. In general, only the few first MNF images are kept for further data processing (~10 of them). Second, the reduced data are projected onto random unit vectors and the extreme pixels in each projection are tallied to record a per-pixel extremity-score which is directly relatable to pixel purity and is thus called a PPI [19, 20]. From the pixels presenting the highest purity, end-member spectra are extracted by manual clustering. These endmembers are dominated by a single chemical component based on the hypothesis that pixel spectra from pure components present more significant differences compared to mixed component spectra. The match between the reflectance spectra of the dataset and the mutually independent endmembers is finally estimated using the spectral angle-mapping (SAM) algorithm [12, 21]. This method computes an angle between the endmember considered as reference and each pixel spectrum, the smaller angles represent closer matches to the reference endmember spectrum. The number of pixels displayed as belonging to a specific end-member class is a function of a threshold applied during this classification step. A second approach [5] uses an inverse MNF analysis that aims to first denoise and then interrogate the dataset by PCA combined with (i) manual clustering of the data in the PC space, and (ii) iterative key set factor analysis (IKSFA). IKSFA is a method proposed by Malinowski [22] which seeks to find the minimum number of spectra required to reproduce the entire dataset through the characterization

of the most orthogonal spectra that typify the original data matrix. In IKSFA, the goal is to find a set of rows (columns) from the data which are most orthogonal to each other. In HSI, this method translates to finding the spectra of pure elements. As with pixel-purity, the IKSFA algorithm assumes that the purest spectra are more dissimilar than the corresponding mixture spectra on a per pixel basis and separates the spectra of pure elements from the spectra of mixtures by factor analysis. Comparable to the hourglass paradigm approach, the key spectra extracted using IKSFA are mapped using SAM.

Using these different approaches, classification and mapping are driven by the data themselves. However, pigment spectral libraries are used during the last step of data evaluation to correlate the extracted end-member spectra with pure pigment spectral signatures. For these different approaches, pigment identification is solely based on visual comparison of the absorption edge positions (in wavelengths) and spectral shape.

A third common approach to interrogating hyperspectral datasets of painted surfaces uses databases built a priori from materials expected to be within the painting. This approach relies on spectral libraries of pigments and binder computed from KM theory [15, 18, 23–26] that model the non-linear effects of spectral mixing. When the non-linear library (i.e. non-linear combinations of pigments reflectance spectra) is compared to the data, the best match is found using a least squares linear combination (LSLC) approach [18, 27]. This is a straightforward approach for pigment identification and mapping, but is limited by the extensive computation time and memory allocation necessary for imaging an entire painting. One study tackled this limitation by segmenting the 2D image space based on chromatic information; for each segment a limited number of representative spectra were then selected to perform pigment mapping pixel-wise. The spectra from the spectral library, in this case, that presented the closest linear fit are deemed the pigment combination in a given pixel [26].

t-SNE approach

t-SNE is an innovative technique for multidimensional scaling [16] inspired by earlier work on SNE [28, 29], that is particularly well suited to data reduction and visualization of high-dimensional datasets. SNE first converts the high-dimensional distances between datapoints into conditional probabilities that represent similarities. The similarity of datapoint x_j to datapoint x_i is expressed by the conditional probability $p_{j|i}$, that is, the probability that x_i would pick x_j as its neighbor if neighbors were picked proportionally to their probability density under a Gaussian centered at x_i. It is given by

$$p_{j|i} = \frac{\exp(-||x_i - x_j||^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-||-x_i - x_k||^2 / 2\sigma_i^2)} \tag{1}$$

where σ_i is the variance of the Gaussian that is centered on datapoint x_i. For the low-dimensional counterparts y_j and y_i of the high-dimensional datapoints x_j and x_i, the same expression for the computation of the conditional probability is utilized as in eq. (1), resulting in $q_{j|i}$.

If the points y_j and y_i correctly model the similarity between the high-dimensional datapoints x_j and x_i, the conditional probabilities $p_{j|i}$ and $q_{j|i}$ will be equal. Motivated by this, SNE aims to find a low-dimensional representation that minimizes the mismatch between $p_{j|i}$ and $q_{j|i}$. The Kullback-Leibler (KL) divergence is the natural measure of the similarity of two distributions. SNE minimizes the sum of the KL divergences overall datapoints, using a gradient descent approach. As the KL divergence is not symmetric, different types of errors are introduced. As an alternative to minimizing the sum of the KL divergence between the conditional distributions $p_{j|i}$ and $q_{j|i}$, a single KL divergence between a joint probability distribution P, in the high-dimensional space, and a joint probability distribution Q, in the low-dimensional space is minimized. The cost function in this case is given by

$$C = \mathrm{KL}(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{2}$$

where p_{ij} is given by

$$p_{ij} = \frac{\exp(-||x_i - x_j||^2 / 2\sigma_i^2)}{\sum_{k \neq l} \exp(-||x_l - x_k||^2 / 2\sigma_i^2)} \tag{3}$$

and q_{ij} is defined also using eq. (3). Minimization of the cost in eq. (2) is now referred to as symmetric SNE. It is mentioned here that alternative definitions of p_{ij} exist.

One of the problems present when mapping a high-dimensional to a low-dimensional space is the so called "crowding problem". It is due to the fact that the volume of a sphere centered on datapoint *i* scales as r^m, where *r* is the radius and *m* is the dimensionality of the sphere. So if the datapoints are approximately uniformly distributed in the region around *i* on the high-dimensional manifold, and we try to model the distances from *i* to the other datapoints in the 2D map, we have the "crowding problem", that is, the area of the 2D map that is available to accommodate moderately distant datapoints will not be nearly large enough compared with the area available to accommodate nearby datapoints. Therefore, in order to model the small distances accurately in the map, most of the points that are at a moderate distance from datapoint *i* will have to be placed much too far away in the 2D map.

In order to address the crowding problem, while a Gaussian distribution is utilized in the high-dimensional space to convert distances into probabilities as shown in eq. (1), the Student-t distribution which is a heavy-tailed distribution is utilized for the same purpose in the 2D space. That is the joint probabilities q_{ij} are now defined as

$$q_{ij} = \frac{(1+||y_i - y_j||^2)^{-1}}{\sum_{k \neq l} (1+||y_k - y_l||^2)^{-1}} \tag{4}$$

The Student-t distribution (with a single degree of freedom) shown in eq. (4) allows a moderate distance in the high-dimensional space to be faithfully modeled by a much larger distance in the map.

Thus by preserving the local data structure in the low-dimensional embedding, this approach represents a new opportunity for data reduction and visualization, through segmentation, of large non-linear HSI datasets.

Experimental

Samples

To test the feasibility of quantitative characterization of pigment mixtures based upon HSI, reference paints were analyzed. A set of mock-up paintings were prepared composed of 12 pure colors and 16 mixtures. There are 15 mixtures of two colors and one of three colors. Among the two-color mixtures, 11 are called tints and are a mixture of pure pigment and white color [namely lead white ($2PbCO_3 \cdot Pb(OH)_2$)], the four others being a mixture of two colored pigments. To build the mock-ups, the pigment weight ratio percentages were (i) 50/50 for the two pigments mixture, and (ii) 33/33/33 for three the pigments mixture. The pigments were then mixed with a gum arabic to bind the paint layers to a commercially primed canvas. The layers were applied thick enough to prevent the priming layer from interfering with the measured reflectance spectra.

Suffrages from a Book of Hours (Ms. 6.T.6) manuscript part of the Isabella Garner Museum collection were also analyzed. The decorative program of this book comprises a range of schools of French illumination, apparently executed in two campaigns in the mid-1400s and around 1500. At least three different "hands" or styles have been identified among the pages of the manuscript. Additional modern restoration by the 19th-century English illuminator and copyist of manuscripts, Caleb William Wing, were also visually

identified by scholars through the manuscript. To redefine artistic hands and practice among one or different workshops and modern add-in, the painting technique and the materials used for the manufacturing have been analyzed using a combination of in-situ techniques namely: HSI and X-Ray fluorescence (XRF). In order to focus this study on the efficiency of the proposed data treatment approach, the results of a single page of the full book are reported (page 34v).

Experimental

Hyperspectral images were recorded using a Resonon Pika II Pushbroom system in the 400–900 nm range with spectral resolution of 2 nm. The system was connected to a stage allowing the scanning of the entire width of the samples, with a pixel size of 125×125 μm^2 in the case of the historic object and 375×375 μm^2 for the mock-up paintings. During acquisition, the object was illuminated using two broad-spectrum tungsten halogen lamps placed at 45° of the object normal. A Lambertian reference reflector (Epson UltraSmooth Fine Art Paper, 13" × 19", A3+, 325 g/m^2) was used as a calibration target to convert the image cubes to diffuse reflectance. Hyperspectral acquisition was performed using the SpectrononPro software.

The macro XGLab's ELIO XRF imaging spectrometer system (MA-XRF) was used in combination with HSI for characterizing the pigments palette of the illuminated manuscript page. The instrument is equipped with a transmission Rh anode X-Ray tube, the polychromatic beam presenting an incoming angle of 63.5° prior to the sample plane. A collimator allowing a 1 mm diameter focused spot size at the surface of the object was used to acquire XRF maps at the surface of the book. The instrument was operated at 50 kV and 40 μA. The 2D surface rastering was executed with acquisition times of 1 s per point and with 1×1 mm^2 step size.

Data treatments

Principal component analyses (PCA) and subsequent k-means clustering were performed using the TXM-Wizard software [30]. Pixels with similar reflectance spectra were pooled in principal component space, effectively segmenting the image based on the variance in the recorded reflectance features into a pre-defined number of regions (k areas) consisting of pixels with a similar spectral signature. The number of clusters was defined as the number of pure pigments or mixtures present into the data to analyze, i.e. each paint composition being accounted as a single cluster. The first eight principal components represented more than 98 % of the variance of the total stacks, consequently the clustering was run on these eight components only (20 K replicates were chosen for obtaining an efficient segmentation of the data).

Matlab hyperMnf function [31] was used to perform the MNF transform on the data. Subsequently, the pixel purity index (PPI) algorithm hyperPpi was used to determine endmembers within the dataset (the number of "skewer" vectors to project data was fixed to 10^6 iterations). For each datapoint, we counted the number of times it was being picked as the extreme point in random projections. This count number for each point was used as the PPI index, the higher the index is, the higher is the probability for the point to be a pure pixel.

The Matlab implementation of the standard t-SNE method was used. The function performed symmetric t-SNE on the $N \times N$ pairwise Euclidean distance matrix to construct an embedding with two dimensions. The perplexity of the Gaussian kernel that is employed can be specified through perplexity, for this study it was fixed at 50. For both mock-up and historical sample datasets, the pixel number was downsized by a factor of 4; instead of binning that would lead to pixels signal averaging, one pixel over four was kept for data processing. This downsampling was necessary to allow reasonable computing time (typically shorter than 10 min for the full dataset) with standard computer resources.

Results and discussion

Mock-up pigment classification

PCA combined with K-mean clustering

To segment and classify the different pigment signatures of the mock-up samples, the PCA results have been combined with k-means clustering. The efficiency of this approach for data reduction and classification has been determined for the full dataset composed of pure pigments, tints (lead white) and colored pigment mixtures.

When the full dataset is used as an input for the reduction and segmentation processes, several limitations are observed. The pure Prussian blue [$Fe_4(Fe[CN]_6)_3$, ivory black (composed of about 10 % carbon, C, and 83 % calcium hydroxyapatite, $Ca_5(PO_4)_3(OH)$, along with smaller amounts of magnesium phosphate, $Mg(H_2PO_4)_2$ and calcium carbonate, $CaCO_3$] and ivory black mixed with lead white, all highly absorbing paints within the visible range, do not cluster separately (Fig. 1b,c). The low reflectance index of Prussian blue and ivory black pigments together with the absence of specific spectral features over the range studied prevent their differentiation by the proposed method (Fig. 2a). Whereas the add-in of white pigment allows for a slight increase of the reflectance value of the ivory black tint (Fig. 2a), these variations remain too low in comparison with the total dataset variance to allow for a specific differentiation of their fingerprints. Similarly, cobalt blue ($CoO \cdot Al_2O_3$) and cobalt blue mixed with lead white were not clustered separately (Fig. 1b,c); the add-in of lead white being barely observable as the reflectance intensity of the paint tint is almost unchanged compared to the pure pigment (Fig. 2b). More interestingly, the paint tint of ultramarine ($Na_8[Al_6Si_6O_{24}]S_n$, mineral lazurite) with lead white, is split in two different clusters (Fig. 1b,c). These two clusters being explained by a heterogeneous thickness of paint application through the area studied; for some part of the paint out the colored layer is thin enough to be partially transparent and thus presents a reflectance spectra resulting of the nonlinear mixture of the white preparation layer and the blue pigment. For this example, thickness effects present variations that were more easily clustered than changes in pigment composition.

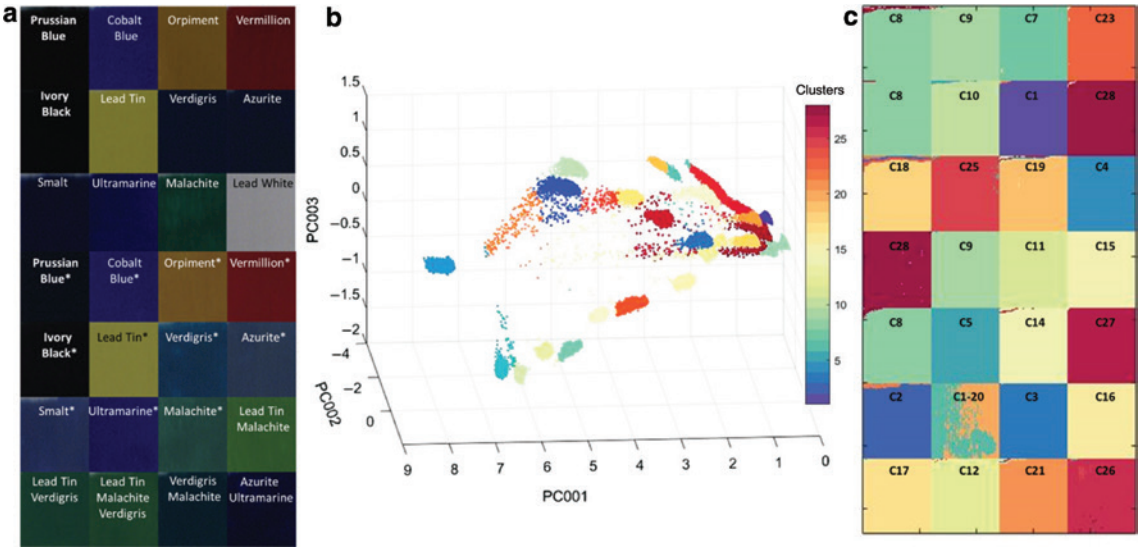


Fig. 1: Results of the PCA combined to K-mean clustering approach for the study of mock-up paint samples. (a) Visible picture of the mock-up samples analyzed, showing the color aspect of pure pigment and pigment mixture layers – pigment name followed by (*) corresponds to tint, i.e. mixture of coloring pigment with lead white; (b) datapoints represented into PC01, PC02 and PC03 space, pixels are colored based on their cluster assignment; (c) the same color scale is used to represent the data clustering into real sample space (the cluster number is provided for each paint layer to ease data interpretation).

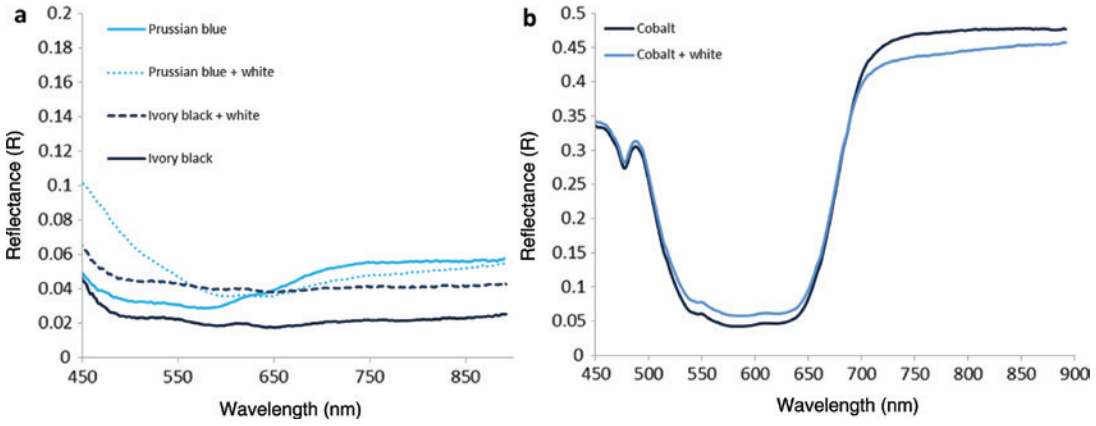


Fig. 2: Comparisons of the mean spectra. (a) Comparison of the mean spectra collected in the paint out area corresponding to Prussian blue and ivory black pure pigments, and tints; (b) comparison of the mean spectra collected in the paint out area corresponding to cobalt blue pure pigment and tint.

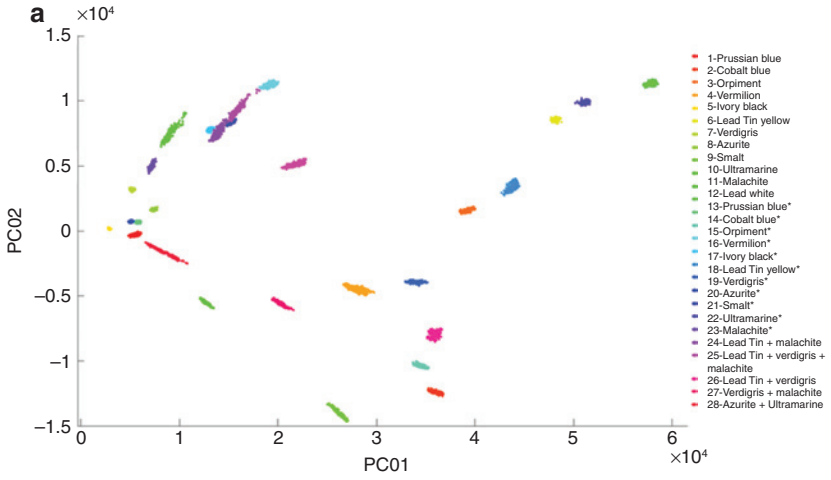
Based on these results, the approach represents a fast and easy solution for the segmentation of different pure pigments and pigment mixtures, respectively, with distinct spectral signature. However, small spectral variation in shape and intensity are difficult to visualize and extract when looking at the linearly reduced dataset. As previously outlined by focusing on describing the largest variance in the dataset, low-dimensional variations are harder to cluster with any segmentation methods – the approach being more sensitive to low signal to noise ratio (SNR) area, and paint layer heterogeneities containing pigments with strong absorption edges which will result in more variance in the spectral domain. Moreover, the segmentation step of the approach requires a good a priori knowledge of the number of pure pigments and mixtures present, as an expected number of clusters is required prior to the clustering. Thus, the data reduction does not provide a straightforward data representation to define the different spectral signatures present in the dataset.

Minimum noise function combined with PPI algorithm

Endmember extraction using MNF combined with PPI algorithm is an established approach to solve the unmixing problem of hyperspectral datasets, where MNF is used to perform dimensionality reduction to ease computational complexity as well as compact information in the transformed components to perform the endmember extraction.

Figure 3a presents the dataset in the two first component spaces identified by PPI. Figure 3b presents for each pure pigment box after PPI calculation (i) the index of the first appearance in the purity score of one single pixel of the pure pigment area, (ii) the maximum voting that refers to the purity score of the point selected as the most extreme, which corresponds to the first occurrence pixel and (iii) the sum of the purity score of all pixels.

Similarly to PCA results, by focusing on high variance between datapoints, smallest variance between spectra is statistically less represented in the unmixed dataset. As an example, none of the pure Prussian blue, ivory black, azurite [$2CuCO_3 \cdot Cu(OH)_2$] and malachite [$CuCO_3 \cdot Cu(OH)_2$] pigment spectra are identified as pure pixels after PPI calculation. On the contrary, lead white, lead tin yellow (Pb_2SnO_4, type 1), cobalt blue, smalt (a potash glass with color given by cobalt ions), vermilion (HgS) and verdigris [$Cu(C_2H_3O_2)_2 \cdot 2Cu(OH)_2$] which appear as maxima clusters in the first and second component values of the PPI projection (Fig. 3a), are identified as pure pixel between the 1st and 2222th projections of the PPI protocol (Fig. 3b), value of index of first appearance in PPI. Based on these results, defining the cut-off threshold value for the scores produced by the PPI to extract candidate pixel vectors for final selection of endmembers is not straightforward. It also calls into question the efficiency of a manual selection of the final set of endmembers by using the purity map index as a visualization tool. Indeed, some pigments seem to be more efficiently identified as single



b

Pure pigments	Index of first appearance in PPI	Maximum voTing	Sum of voTing
12-Lead white	1	3006	53 280
2-Cobalt blue	30	489	12 413
9-Smalt	62	297	7000
6-Lead Tin yellow	140	154	4280
3-Orpiment	543	35	839
4-Vermillion	583	31	628
10-Ultramarine	1398	5	49
7-Verdigris	2222	1	10
1-Prussian blue	*None*	*0*	*0*
5-Ivory black	*None*	*0*	*0*
8-Azurite	*None*	*0*	*0*
11-Malachite	*None*	*0*	*0*

Fig. 3: Results of the MNF combined with PPI approach for the study of mock-up paint samples. (a) Dataset represented in the two first space components of the PPI calculation; (b) table presenting the results of the PPI calculation for each pure pigment paint out: (i) index value of the first appearance as a pure component of a single pixel, (ii) purity score of the single pixel that first appeared as a pure component, (iii) sum purity score of all pixels.

component than others. Thus, using the sum of voting and maximum voting values presented in Fig. 3b, the determination of pure pigment as endmember seems compromised as only eight pure pigments over 12 present are identified as pure pixels. Moreover, when comparing the sum of voting per pigment boxes, the tints of vermilion, lead tin, orpiment, smalt and verdigris present a higher purity score value than their pure pigments. Similarly, mixtures of lead tin with either verdigris or malachite present a higher purity score values than their respective pure verdigris or malachite pigments.

With this example, the presence of highly reflective pigment seems to strongly affect the result of the unmixing, as instead of identifying pure pigment area, the PPI identify higher reflectance values as more extreme pigment signatures.

Whereas this approach is presented as an efficient approach for spectral unmixing, this example demonstrates that it shares similar limitations with the previous approach when used for data reduction and segmentation.

In particular to perform endmember identification from dataset three main limitations can be emphasized, the loss of non-liner relationship in the data reduced space, the loss of low-variance variation of the dataset, and the difficulty to identify rare events that can reflect a few pixels of the entire dataset [32]. These limitations ultimately determine the quality and efficiency of the approach to determine endmember spectra of pure pigment as observed for this dataset.

t-SNE

The function performs symmetric t-SNE on the N × N pairwise Euclidean distance matrix D to construct an embedding with two dimensions. To ease visualization, in the scatterplot each pixel of the dataset is represented in the RGB color domain using three channels of the HSI wavelength range: red is assigned to channel 118 (637.6 nm), green to channel 75 (547.8 nm) and blue to channel 32 (458.1 nm).

The results obtained using the t-SNE approach are presented in Fig. 4a. For comparison, a manual clustering of the dataset in the reduced data space is provided in Fig. 4b together with its result in the sample map space.

For studying mock-up systems, the technique performs an efficient dimension reduction of the dataset while keeping the local data structure (Fig. 4a). Pure pigments, tint of colored pigments mixed with lead white, and colored pigment mixtures appear as distinct pixel groups in the reduced space (Fig. 4b). Thus, the 2D representation of the dataset provides a straightforward visual segmentation and classification of

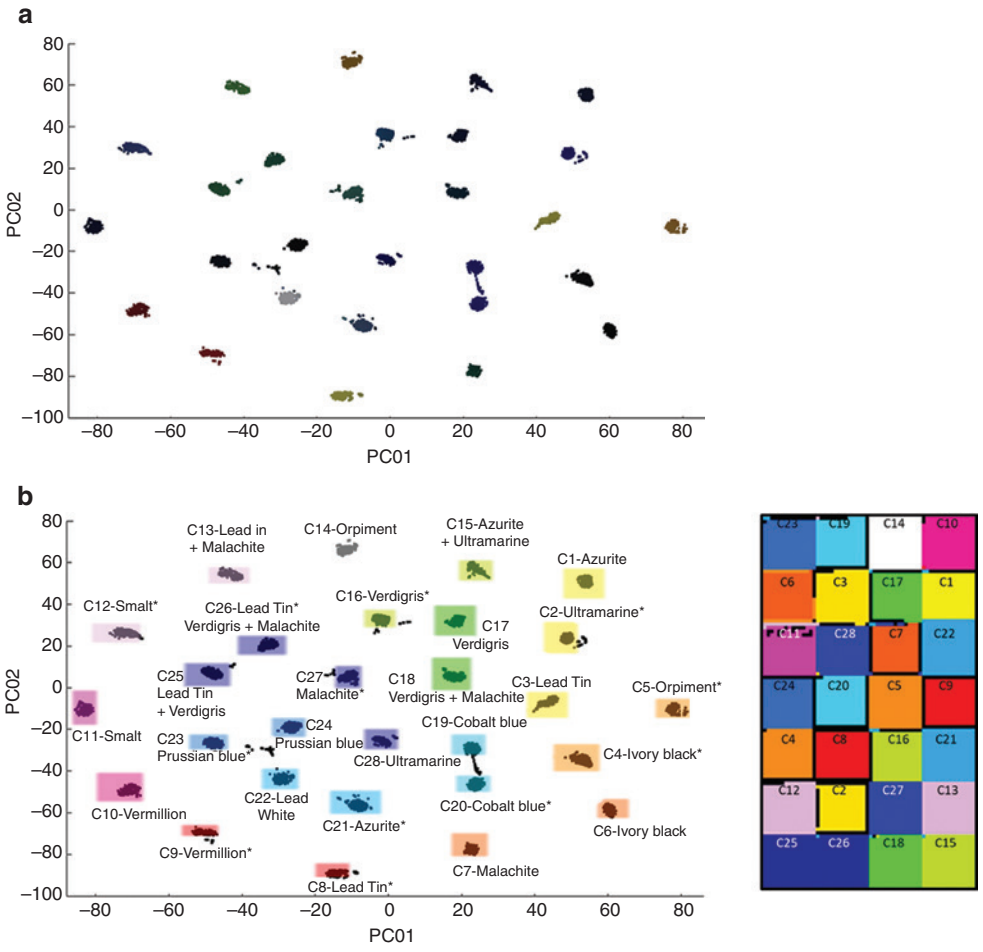


Fig. 4: Results of the t-SNE approach for the study of mock-up paint samples. Visualization of 60×105 pixels map from the mock-up HSI dataset produced by symmetric t-SNE approach (a) using RGB pixel values for labeling; and (b) combined with user-guided manual clustering – the same color scale is used to represent the result of data clustering into sample map space (the cluster number is provided for each paint layer).

the pigment and pigment mixture signatures present within the data cube. This data reduction not only preserves similarities within the dataset in a 2D reduced dimension but acts as an efficient method to spatially cluster spectra with similar signatures (Fig. 4b).

In comparison to PCA and MNF, t-SNE efficiently differentiates the pure pigments from their corresponding tints. As an example, by focusing on small pairwise distances, this approach successfully distinguishes spectra of Prussian blue and ivory black. Similarly, pure cobalt blue is efficiently separated from its mixture with lead white (Fig. 4a). Whereas, the distance of one cluster to another cannot be used as quantitative information on the similarities between pixel groups (as t-SNE provides a non-linear representation of those similarities), the presence of pixels centered between the cluster of pure cobalt blue pigment and cobalt blue mixed with lead white highlight the possibility that some pixels of the mixture contain pure pigment in higher concentration than in the rest of the mixture. Interestingly this effect is observed for a pure pigment that was not differentiated from its tint with the two other approaches.

t-SNE provides efficient data reduction with a clear visualization of the similarities between the spectral features of both pure pigment and pigment mixtures. As the clustering of the 2D representation of dataset is straightforward, this method offers the possibility of performing the data segmentation using user-guided manual clustering, rather than more complicated and computationally slow statistical algorithms. This drastically simplifies the approach compared to classical data analyses protocol, which requires segmentation after data reduction to extract meaningful information on the dataset.

Results of t-SNE approach for historical samples

t-SNE algorithm has been used to process hyperspectral dataset acquired on an illuminated manuscript page (detailed above) to assess the capabilities of this innovative approach in the case of historical painted materials.

Similarly to the work performed on the mock-up samples, symmetric t-SNE on the N × N pairwise Euclidean distance matrix D was performed to construct a data embedding in two dimensions. The original dataset contains 100 × 149 pixels encompassing the 240 wavelengths of the HSI spectra. For this dataset, the processing time represented 460 s in total. For each pixel its RGB color was assigned in the t-SNE 2D representation using three specific wavelengths: red to channel 118 (637.6 nm), green to channel 75 (547.8 nm) and blue to channel 32 (458.1 nm).

The results obtained using the t-SNE approach are presented in Fig. 5a. Here no prior information on the pigments was provided apart from the elemental mapping using MA-XRF method.

As presented for the mock-ups example, the 2D dataset was further segmented using the inherent 2D structure of the reduced dataset, Fig. 5b. A user-guided manual clustering was used to select and assign a characteristic color to a set of pixels presenting the highest similarities in the 2D space. This information was returned into the original image space where the different clusters appeared with different colors. In addition, the corresponding centroid spectrum was extracted. For more clarity, the cluster identification is provided by color hue, Fig. 5c–e.

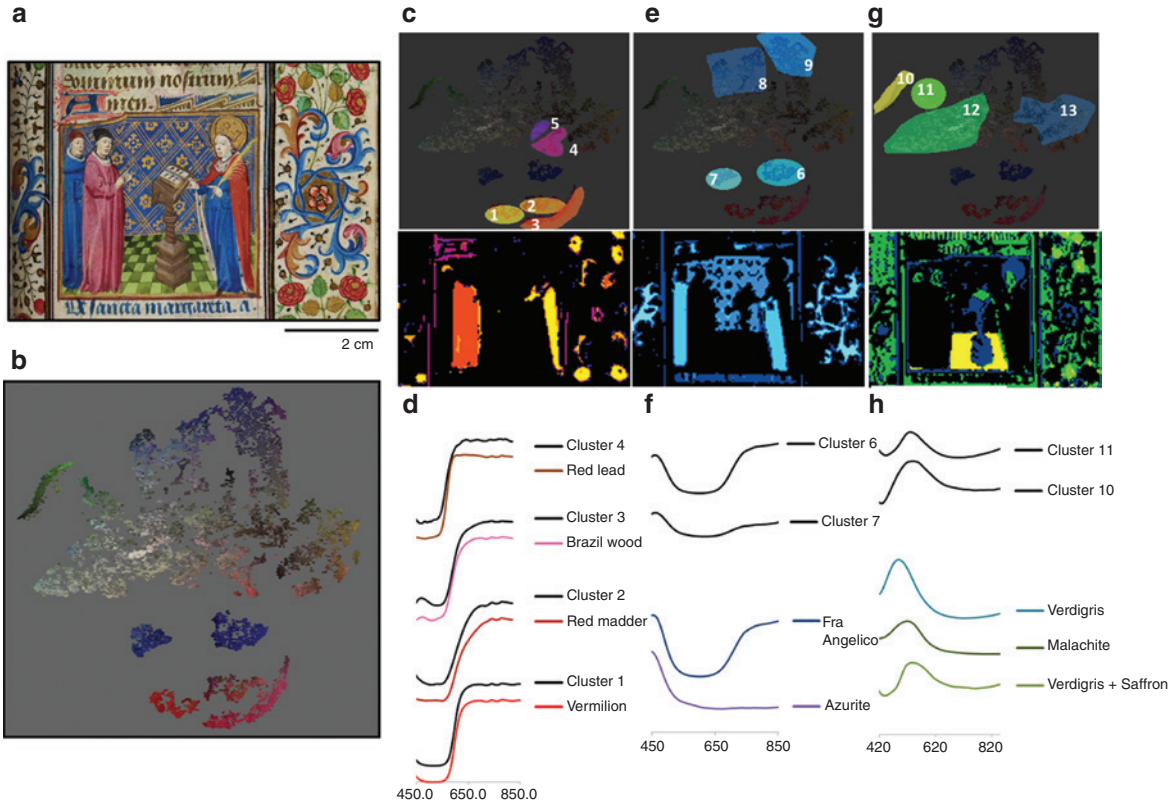


Fig. 5: Results of the t-SNE approach for the study of an illuminated page from a French medieval Book of Hours. (a) Visible picture of the area analyzed by t-SNE; (b) 2D representation of the algorithm result; (c) assignment of the red and pink clusters in the t-SNE component space (top) and transposed into data space (bottom); (d) centroids spectra of data cluster 1, 2, 3 and 4 compared to reflectance spectra of red lead and vermilion pigments, and brazil wood and red madder lakes; (e) assignment of the blue clusters in the t-SNE component space (top) and transposed into data space (bottom); (f) centroids spectra of data cluster 6 and 7 compared to reflectance spectra of azurite and Fra Angelico pigments; (g) assignment of the green clusters in the t-SNE component space (top) and transposed into data space (bottom); (h) centroids spectra of data cluster 11 and 10 compared to reflectance spectra of verdigris, malachite and mixture of verdigris and saffron.

Assignment of the red and pink clusters

Cluster 1 is present in both the mantle of the left character and red flowers present on both side of the main figure (Fig. 5c). The spectra for the corresponding area reveal shapes similar to semiconductors pigment, i.e. a sigmoid with a steeper rise around the inflection point [33]. For cluster 1, the inflection point being around 590–605 nm indicates the presence of vermillion, confirmed by the identification of mercury and sulfur by XRF analyses [8].

Cluster 4, present in the localized area of the white and red flower surrounded by gilding and blue decoration, present a similar semi-conductor fingerprint (Fig. 5c). However, with an inflection point around 565 nm, the presence of red lead (Pb_3O_4, mineral form minium) is proposed in this single flower [33]. This has been confirmed by the presence of high content of lead together with the absence of additional elements that could refer to other semi-conductor red pigments.

Cluster 2 is present in all vegetable dark red decoration observed on the side of the main figure (Fig. 5c). The first derivative of its representative spectra has a broad asymmetric band with a maximum centered at 590 nm. For both spectra a rise in reflectance in the red part of the spectra is also observed. XRF spectra from these sites show mostly the presence of calcium, which seems consistent with the presence of a red dye with chalk ($CaCO_3$) as possible substrate. Moreover the presence of a weak absorption structured in two bands at 510–515 and 540–545 nm, might suggest the presence of a red madder dye [8, 33].

Similarly, cluster 3 refers to vegetal pink decoration, with an XRF spectra dominated by Ca signal and a broad asymmetric band with a maximum centered at 590 nm with a constant absorption intensity in the red part (Fig. 5c). A single absorption band centered at 560 nm is observed for this cluster, pointing toward the use a pink dye namely brazilwood applied on a chalk substrate [8, 33].

Assignment of the blue clusters

Two clusters are identified among the blue areas: clusters 6, mostly present in the robe of the characters, and 7, mostly present in the blue vegetable decoration identified in the vicinity of the central scene (pink and light pink, respectively) (Fig. 5d).

Cluster 6 presents a good match with the mineral ultramarine ($Na_8[Al_6Si_6O_{24}]S_n$, mineral lazurite), with a reflectance spectrum presenting a maximum of absorption around 600nm and a transition to high reflectance around 700 nm. In the case of cluster 7, a decrease in reflectance between 700 and 900 nm is observed, that can be attributed to the presence of azurite pigment together with ultramarine. XRF spectra confirms the significant amount of Cu, consistent with the presence of azurite, together with dominant Al, Si and K signals, consistent with the presence of ultramarine.

These results can be interpreted as a combined use of both pigments to paint vegetable blue area of the illumination. Together with lead white, Pb being identified by XRF, different shades of blue can be achieved as observed in the RGB image of the illumination. In previous publications the application of a thin layer of ultramarine on top of azurite is proposed to create deeper blue hues, instead of a mixture of both pigments, as ultramarine glazing over azurite was a common technique during this time period [8, 34].

Two other clusters refer to blue area: clusters 9, identified in blue brocades in the central figure, and 8, identified at the edge of the blue vegetable decorations (red light and light blue, respectively – Fig. 5d). In the case of cluster 9, the presence of ultramarine is proposed based on the same spectral features described for cluster 6, however its presence as a single cluster is explained by the fact that the blue paint is applied on top of a gold leaf. The blue paint being relatively thin, the reflectance from the gold leaf underneath modified the signal of the pure pigment, mostly by increasing the reflectance intensity of these pixels between 600 and 900 nm. Similarly, cluster 8 refers to blue paint that diffuses into the parchment, here again the high and constant reflectance of the prepared parchment mostly flatten the reflectance spectra of the blue pixels leading to a distinct spectral fingerprint.

Assignment of the green clusters

The green garments are clustered in two clusters: cluster 10 and 11 (Fig. 5e). Cluster 11 is limited to the leaves painted within the edge section. Its characteristic spectrum presents a peak reflectance around 535–540 nm and a slow rise from 800 nm. Similarly cluster 10 presents a peak reflectance at 535 nm, however the peak is broader than cluster 10, accompanied by a maximum of absorption shifted toward higher wavelengths (760 nm) and the absence of reflectance rise after 800 nm. These spectral features point out the use of Cu-containing green such as verdigris, copper-resinate ($C_{19}H_{29}COOH$), or malachite, confirmed by the presence of Cu identified as the main element entering in the composition of the green paints by XRF technique [8]. The difference in peak broadening in the case of cluster 10 can be attributed to many different factors, however a mixture of verdigris with a yellow dye, e.g. type saffron, to achieve a different type of green can be proposed. Within cluster 11 both dark and light greens are present – they both present similar reflectance peaks however the overall intensity of the spectra in the darker area is lower suggesting the add-in of a dark pigment in that area (e.g. carbon black).

Two additional clusters are attributed to the gilded area (cluster 13), and to the parchment without colored paint layers (cluster 12).

In conclusion, the combination of t-SNE 2D representation with manual clustering allows for the segmentation of the data cube into pure pigments and pigment mixture clusters that confirm the first results obtained by XRF mapping. It represented a relatively fast and easy multivariate approach to extract and differentiate the different spectral fingerprints present into the dataset – the use of centroids spectra for each cluster together with an adapted database allow identifying the main pigment and dyes palette used by the artist to depict the illumination.

Conclusion

This study presents a new technique for the visualization of similarity between hyperspectral data that is capable of retaining the local structure while also revealing some important global structure of the data cube.

Our experiments on mock-up paint samples show that t-SNE outperforms existing state-of-the-art techniques for visualizing and segmenting pigment fingerprints in a reduced data space, in particular in the case of highly absorbing pigments. Three main advantages of the technique can be mentioned to clarify this result. First, instead of preserving the distances between widely separated datapoints, t-SNE preserves the distances between nearby datapoints and thus the similarities between datapoints rather than dissimilarities. Moreover, whereas PCA and MNF are hampered by non-linear feature relations, the non-linear approach proposed by t-SNE allows for the modeling of any curved manifolds within the dataset. Finally, for visualizing the structure of very large datasets, t-SNE uses random walks on neighborhood graphs to allow the implicit structure of all the data to influence the way in which a subset of the data is displayed. Consequently, the final results provide a straightforward visualization of the clustering of the dataset. This explains the fact that in this study t-SNE has been used alone without preprocessing, or further segmenting strategies – the clustering being selected manually based on the cluster information visually observed in the reduced dataset created.

This approach was successfully applied to a historical artifact to segment and identify paint composition. In particular, for the illuminated page studied, four red pigments and dyes have been identified as pure material: cinnabar, vermillion, red madder, pink Brazil wood. For the blue pigment, both pure ultramarine and mixture of ultramarine and azurite have been found. Finally two types of green were differentiated, both Cu-based pigments from which no further information can be extracted without further analytical techniques. These results provided pigment/paint mixture maps that have been correlated to XRF results, confirming the information extracted using t-SNE approach.

This technique represents a new opportunity for hyperspectral data reduction and visualization in the cultural heritage. Whereas it has been combined with manual clustering, some further development in terms

of data segmentation could be proposed, and an approach similar to k-means clustering and PPI could benefit the results.

More powerful than classical approaches for data reduction and classification, computation time and memory allocation drawbacks were faced when t-SNE was run using standard computer performances. That led to a loss of resolution of the original images (often by of a factor of 4). Strategies to speed up the calculation process using more powerful processors, or performing the calculation on previously clustered dataset can be envisaged for future works.

Acknowledgments: This collaborative initiative is part of NU-ACCESS's broad portfolio of activities, made possible by generous support of the Andrew W. Mellon Foundation as well as supplemental support provided by the Materials Research Center, the Office of the Vice President for Research, the McCormick School of Engineering and Applied Science and the Department of Materials Science and Engineering at Northwestern University. The authors would like to thank Jessica Chloros and Valentine Talland Associate Objects Conservator at the Isabella Gardner Museum (Boston, USA) for allowing and facilitating non-invasive analyses on the illuminated manuscript. Johanna Salvant and Noalle Fellah from NU-ACCESS Laboratory, Northwestern University (Chicago, USA) are gratefully acknowledged for their help in preparing mock-up samples.

References

[1] C. Fischer, I. Kakoulli. *Stud. Conserv.* **51**, 3 (2006).
[2] M. Kubik. "Hyperspectral imaging: a new technique for the non-invasive study of artworks", in *Physical Techniques in the Study of Art, Archaeology and Cultural Heritage*, D. Creagh, D. Bradley (Eds.), Vol. 2, pp. 199–259, Elsevier, Amsterdam (2007).
[3] H. Liang. *Appl. Phys. A*. **106**, 309 (2012).
[4] M. Alfeld, L. de Viguerie. *Spectrochim. Acta B*. **136**, 81 (2017).
[5] E. Catelli, L. L. Randeberg, B. K. Alsberg, K. F. Gebremariam, S. Bracci. *Spectrochim. Acta A*. **177**, 69 (2017).
[6] S. Baronti, A. Casini, F. Lotti, S. Porcinai. *Chemom. Intell. Lab. Syst.* **39**, 103 (1997).
[7] F. Daniel, A. Mounier, J. Pérez-Arantegui, C. Pardos, N. Prieto-Taboada, S. F.-O. de Vallejuelo, K. Castro. *Microchem. J.* **126**, 113 (2016).
[8] J. K. Delaney, P. Ricciardi, L. D. Glinsman, M. Facini, M. Thoury, M. Palmer, E. R. D. L. Rie. *Stud. Conserv.* **59**, 91 (2014).
[9] J. K. Delaney, J. G. Zeibel, M. Thoury, R. Littleton, M. Palmer, K. M. Morales, E. R. D. L. Rie. *Appl. Spectrosc.* **64**, 584 (2010).
[10] A. Mounier, F. Daniel. *Stud. Conserv.* **60**, S200 (2015).
[11] A. Mounier, G. Le Bourdon, C. Aupetit, C. Belin, L. Servant, S. Lazare, Y. Lefrais, F. Daniel. *Heritage Science* **2**, 24(2014).
[12] P. Ricciardi, J. K. Delaney, M. Facini, L. Glinsman. *JAIC* **52**, 13 (2013).
[13] P. Ricciardi, J. K. Delaney, M. Facini, J. G. Zeibel, M. Picollo, S. Lomax, M. Loew. *Angew. Chem. Int. Ed.* **51**, 5607 (2012).
[14] A. Mohan, G. Sapiro, E. Bosch. *IEEE Geosci. Remote Sens. Lett.* **4**, 206 (2007).
[15] P. Kubelka, F. Munk. *Z. Tech. Phys*. **12**, 593 (1931).
[16] L. V. D. Maaten, G. Hinton. *J. Mach. Learn. Res.* **9**, 2579 (2008).
[17] G. Licciardi, P. R. Marpu, J. Chanussot, J. A. Benediktsson. *IEEE Geosci. Remote Sens. Lett.* **9**, 447 (2012).
[18] H. Liang, K. Keita, B. Peric, T. Vajzovic. Pigment identification with optical coherence tomography and multispectral imaging, Proceedings of OSAV 2008, The 2nd Int. Topical Meeting on Optical Sensing and Artificial Vision, 33–42 (2008).
[19] J. W. Boardman. Automating spectral unmixing of AVIRIS data using convex geometry concepts, JPL, Summaries of the 4th Annual JPL Airborne Geoscience Workshop, 1: AVIRIS Workshop 11–14 (1993).
[20] J. W. Boardman, F. A. Kruse, R. O. Green. Mapping target signatures via partial unmixing of AVIRIS data, Summaries of the Fifth Annual JPL Airborne Earth Science Workshop, 1: AVIRIS Workshop 23–26 (1995).
[21] F. Kruse, A. Lefkoff, J. Boardman, K. Heidebrecht, A. Shapiro, P. Barloon, A. Goetz. *Remote Sens. Environ.* **44**, 145 (1993).
[22] E. R. Malinowski. *Anal. Chim. Acta* **134**, 129 (1982).
[23] P. Geladi, H. F. Grahn. *Multivariate image analysis*, John Wiley & Sons, Ltd, Hoboken, NJ, USA (1996).
[24] R. S. Berns, M. Mohammadi. *Color Res. Appl.* **32**, 201 (2007).
[25] R. S. Berns, M. Mohammadi. *Stud. Conserv.* **52**, 299 (2007).
[26] Y. Zhao, R. S. Berns, L. A. Taplin, J. Coddington. *Proc. SPIE*, **6810**, 1 (2008).
[27] G. Dupuis, M. Menu. *Appl. Phys. A Mater. Sci. Process.* **83**, 469 (2006).
[28] G. E. Hinton, S. T. Roweis. *Adv. Neural Inf. Process. Syst.* 857 (2003).
[29] J. Cook, I. Sutskever, A. Mnih, G. Hinton. *Artif. Intell*. 67 (2007).

[30] Y. Liu, F. Meirer, P. A. Williams, J. Wang, J. C. Andrews, P. Pianetta. *J. Synchrotron Radiat.* **19**, 281 (2012).
[31] A. A. Green, M. Berman, P. Switzer, M. D. Craig. *IEEE Trans. Geosci. Remote Sens.* **26**, 65 (1988).
[32] F. Chaudhry, C.-C. Wu, W. Liu, C.-I. Chang, A. Plaza. "Pixel purity index-based algorithms for endmember extraction from hyperspectral imagery", in *Recent Advances in Hyperspectral Signal and Image Processing,* C.-I. Chang (Ed.), Vol. 37, p. 29, Transworld Research Network, Kerala, India (2006).
[33] M. Aceto, A. Agostino, G. Fenoglio, A. Idone, M. Gulmini, M. Picollo, P. Ricciardi, J. K. Delaney. *Anal. Methods* **6**, 1488 (2014).
[34] A. Roy. *Artists' pigments: a handbook of their history and characteristics,* Vol. 2, National Gallery of Art, Washington, DC (1993).