Order from chaos: Analyzing quantitative hyperspectral imaging data of historical documents

M.E. Klein[1], B.J. Aalderink[1], R. Padoan[2], G. de Bruin[2], Th. A.G. Steemers[2]

[1] Art Innovation BV, Zutphenstraat 25, 7575 EJ Oldenzaal, The Netherlands
E-mail: info@art-innovation.nl; Phone: +31 541 570720; Web: www.art-innovation.nl

[2] Nationaal Archief, P.O. Box 90520, 2509 LM The Hague, The Netherlands
E-mail: roberto.padoan@nationaalarchief.nl; Phone: +31 70 3315400

Introduction

The Nationaal Archief (National Archives of the Netherlands) and Art Innovation cooperate in the development of the **Quantitative Hyperspectral Imaging** (QHSI) technique for the **non-destructive** examination of historical documents.

Applications of the QHSI technique

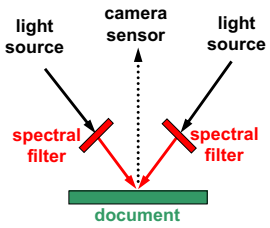
- Measuring the document condition quantitatively for an objective risk assessment
- Studying aging processes on original and sample documents (natural + artificial aging)
- Distinguishing and mapping of different writing materials on documents
- Enhancing the legibility of deteriorated text

Each QHSI measurement comprises a huge amount of data, which can be analyzed in numerous ways. For some applications, advanced mathematical algorithms are used to combine data of several QHSI measurements of the same of different documents. For an effective, time-efficient exploitation of the measurement data suitable **analysis workflows** have to be developed.

Analysis workflow for mapping areas with similar response

Step 1: Acquisition of calibrated data

During a measurement with the SEPIA hyperspectral instrument, the document is illuminated with **70 narrow spectral bands** in the near-UV, visible and near-infrared (wavelength range 365 to 1100 nm). For each band, a calibrated digital image is recorded. The entire series of calibrated images forms the so-called hyperspectral data cube, which contains the **calibrated reflectance curve** for each pixel.



Step 2: Real-colour image

From the 41 calibrated reflectance images in the visible wavelength range 380 – 780 nm, via the tristimulus values XYZ the CIELAB and other colour indices can be calculated for any illumination spectrum.

From the calculated indices local colour differences in the same document, between different documents and between two recordings of the same document (e.g. before and after an exhibition) can be detected and quantified (ΔE). This can be used to assess objectively the visual effects of document degradation and conservation treatments.

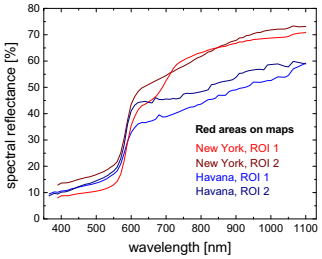
The real-colour images are calculated from spectral images for illuminant D65 and 2° observer. The arrows indicate red areas where Regions-Of-Interest (ROI's) were defined.

Upper: Johannes Vingboons, *View of New York*, ca. 1670, Nationaal Archief inv. no. 4.VELH619-14
Lower: Johannes Vingboons, *View of Havana,* ca. 1665, Nationaal Archief inv. no. 4.VELH 619-57

Step 3: Regions-of-interests (ROI's)

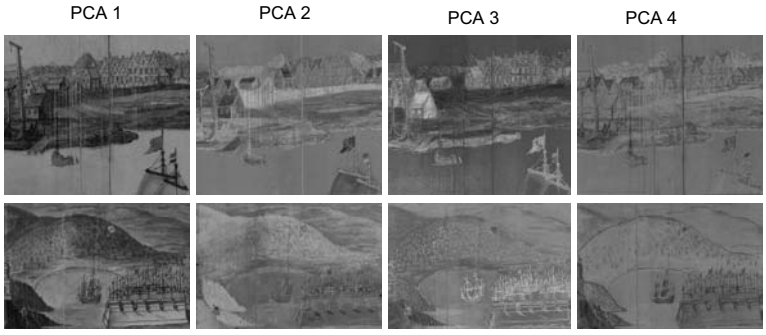
Using a graphic tool or automatic functions (e.g. thresholding), ROI's are defined. Care is taken that the spectral response within each ROI is fairly homogeneous, so that the mean spectral curve of the ROI is a good representation of its spectral characteristics.

The comparison of the spectral curves of two red areas in each of the Vingboon drawings indicates that two different red pigments or mixtures were used on the View of New York and other mixtures on the View of Havana.



Step 4: Principal Component Analysis (PCA)

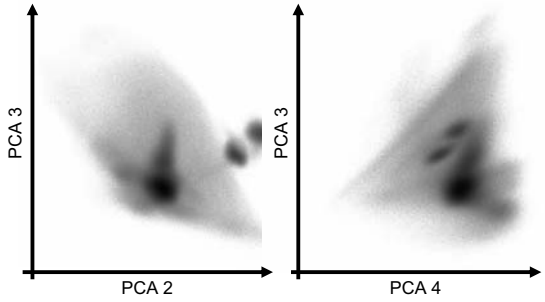
The Principal Component Analysis (**PCA**) **feature extraction** algorithm is a statistical method for extracting the relevant spectral data from the data cube and condense it into a small number of component images. For the Vingboon maps, the first seven PCA images represent the entire data set in a 7-D abstract feature space. The corresponding data reduction by a factor 10 facilitates further processing such as for classification. The third and fourth of the four shown PCA images show ink lines and underdrawings with high contrast.



Step 5: Scatter plots for pixel classification

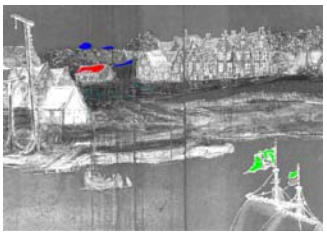
A **scatter plot** is 2D containing a point for each pixel. The (x,y)-coordinates of the pixel are given by the values of the pixel in, e.g., 2 selected PCA feature images. Scatter plots can help to detect groups of pixels with similar spectral characteristics, because these will form clusters that are seen as dark regions in the scatter plot.

The View of New York data is represented in two scatter plots for which the y-axis coordinate was taken from the PCA 3 image and the x-axis coordinate from the PCA 2 and PCA 4 images shown above.

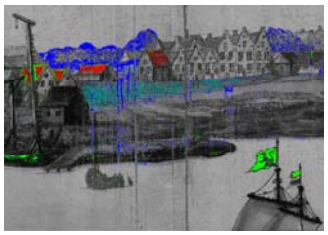


Step 6: Classification using a training set of labeled pixels

Using either the scatter plots or the PCA images directly, groups of pixels with homogeneous spectral characteristics are identified and labeled. On the PCA 3 image of the View of New York, two ROI's were defined in two different red and two different green areas (left image). Together with other marked areas, this labeled training set was used as the input for the so-called Quadratic Maximum Likelihood (QML classifier) to map pixels that are similar to either of the two red and green areas (right image).



QML



Using the same training set, the QML classifier was used to map pixels also in the View of Havana measurement. Identical artificial pixel colours in both maps mean that the spectral characteristics are very similar to those of the ROI marked with this colour in the training set.