This is the peer reviewd version of the followng article:

A novel proposal to investigate the interplay between the spatial and spectral domains in near-infrared spectral imaging data by means of Image Decomposition, Encoding and Localization (IDEL) / Ahmad, M., Vitale, R., Silva, C.S., Ruckebusch, C., Cocchi, M.. - In: ANALYTICA CHIMICA ACTA. - ISSN 0003-2670. - 1191:(2022), pp. 1-12. [10.1016/j.aca.2021.339285]

Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

03/11/2026 11:11

(Article begins on next page)

Journal Pre-proof



A novel proposal to investigate the interplay between the spatial and spectral domains in near-infrared spectral imaging data by means of Image Decomposition, Encoding and Localization (IDEL)

Mohamad Ahmad, Raffaele Vitale, Carolina S. Silva, Cyril Ruckebusch, Marina Cocchi

PII: S0003-2670(21)01111-9

DOI: https://doi.org/10.1016/j.aca.2021.339285

Reference: ACA 339285

To appear in: *Analytica Chimica Acta*

Received Date: 7 August 2021

Revised Date: 6 November 2021

Accepted Date: 14 November 2021

Please cite this article as: M. Ahmad, R. Vitale, C.S. Silva, C. Ruckebusch, M. Cocchi, A novel proposal to investigate the interplay between the spatial and spectral domains in near-infrared spectral imaging data by means of Image Decomposition, Encoding and Localization (IDEL), *Analytica Chimica Acta* (2021), doi: https://doi.org/10.1016/j.aca.2021.339285.

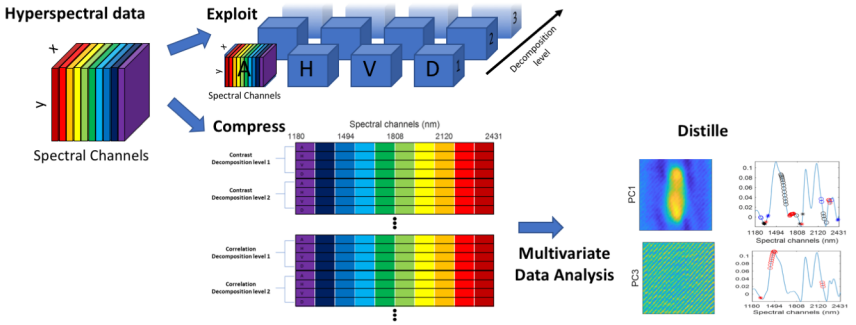
This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Published by Elsevier B.V.

Sample CRediT author statement

Mohamad Ahmad: Writing- Original draft preparation, Conceptualization, Methodology, Visualization, Software. **Raffaele Vitale**: Supervision, Software, Data curation. Writing- Reviewing and Editing. **Carolina S. Silva**: Investigation, Resources, Writing- Reviewing and Editing, Funding acquisition. ***Cyril Ruckebush****:* Supervision, Validation, Writing- Reviewing and Editing, Project administration. **Marina Cocchi**: Supervision, Validation, Software, Visualization, Writing- Reviewing and Editing, Project administration, Funding acquisition.

Journal Pre-proof



Journal Pre-proof

Journal Pre-proof

A novel proposal to investigate the interplay between the spatial and spectral domains in near-infrared spectral imaging data by means of Image Decomposition, Encoding and Localization (IDEL)

Mohamad Ahmad[1,2], Raffaele Vitale[2], Carolina S. Silva[3], Cyril Ruckebusch[2], Marina Cocchi[1*]

[1]Università di Modena e Reggio Emilia, Dipartimento di Scienze Chimiche e Geologiche, Via Campi 103, 41125 Modena, Italia

[2]Univ. Lille, CNRS, LASIRE, LAboratoire de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, Cité scientifique, F-59000 Lille, France

[3]Department of Food Sciences and Nutrition, University of Malta, Msida 2080, Malta

[*]Corresponding author: marina.cocchi@unimore.it

Abstract

The emergence of new spectral imaging applications in many science fields and in industry has not come to be a surprise, considering the immense potential this technique has to map spectral information. In the case of near-infrared spectral imaging, a rapid evolution of the technology has made it more and more appealing in non-destructive analysis of food and materials as well as in process monitoring applications. However, despite its great diffusion, some challenges remain open from the data analysis point of view, with the aim to fully uncover patterns and unveil the interplay between both the spatial and spectral domains. Here we propose a new approach, called Image Decomposition, Encoding and Localization (*IDEL*), where a spatial perspective is taken for the analysis of spectral images, while maintaining the significant information within the spectral domain. The methodology benefits from wavelet transform to exploit spatial features, encoding the outcoming images into a set of descriptors and utilizing multivariate analysis to isolate and extract the significant spatial-spectral information. A forensic case study of near-infrared images of biological stains on cotton fabrics is used as a benchmark. The stain and fabric have hardly distinguishable spectral signatures due to strong scattering effects that originate from the rough surface of the fabric and the high spectral absorbance of cotton in the near-infrared range. There is no selective information that can isolate signals related to these two components in the spectral images under study, and the

Journal Pre-proof

complex spatial structure is highly interconnected to the spectral signatures. *IDEL* was capable of isolating the stains, (spatial) scattering effects, and a possible drying effect from the stains. It was possible to recover, at the same time, specific spectral regions that mostly highlight these isolated spatial structures, which was previously unobtainable.

Keywords: spectral imaging, wavelet decomposition, near-infrared, image encoding, forensics, cotton, biological fluids, multivariate image analysis, spatial-spectral analysis

1. Introduction

Near-infrared (NIR) imaging has become a cheap, versatile and very attractive method in many fields of science and diverse industrial applications, for its ability to capture phenomena in both spectral and spatial domains. Examples of applications are remote sensing in agriculture [1], stain analysis in forensics [2,3] and foodstuffs quality control [4]. With hundred-thousands of pixels for which a full NIR spectrum can be registered, the information content available from a spectral imaging data set is potentially overwhelming. This issue is often amplified by the nature of the sample itself, due to the complexity of its chemical composition and/or physical structure [5]. In most studies, it is mainly information extracted from the spectral domain that is exploited by the chemometric analysis, while the spatial (structural, textural) information of the sample is often disregarded.

The analysis of highly scattering materials is challenging in the NIR range and limited by the difficulty to describe the chemical and physical properties of the sample separately [6]. In practice, the separation of absorption and scattering has been the subject of different spectroscopic studies [7,8], with most of them removing the effect of scattering on the measured spectra by adequate preprocessing. Nonetheless, with NIR images of highly diffusive samples, scattering and absorption are entangled by highly non-linear mechanisms, which cannot always be fully eliminated by applying scatter-correction techniques on the individual spectral pixels, without the consideration of the spatial domain. Indeed, dramatic changes in the scattering contribution to the spectral signal can be expected

from object borders and texture, which may fully dominate the spectral signal. Strong spectral interferences can be localized in the spatial domain by supervised or first principle modelling [9,10], but these methods require significant a priori information about the scattering behavior of the samples.

Most of the approaches to analyze NIR imaging data solely exploit spectral variation as the two spatial dimensions of the measured data cube are unfolded pixel-wise, ignoring spatial correlation. Still, one possible approach is multivariate image analysis (MIA), where the unfolded imaging data is augmented with pixel-neighbor information, to incorporate local-spatial information before it is analyzed with multivariate analysis tools, such as principal component analysis (PCA) or partial least squares (PLS) regression [11–13]. MIA has been originally proposed for RGB images [11,13] then extended to multi-channel images [14] and only recently to spectral images [11]. However, the number of neighboring pixels increases rapidly with the distance (or window size in pixels) from the center pixel at which to consider the neighborhood, and this applies to all spectral channels, making the data unmanageable in some cases. In this context, a parsimonious solution can be to employ multivariate curve resolution-alternating least square (MCR-ALS) using image processing constraints to take into account the spatial structure [15]. Nonetheless, this does require the data to strictly follow a bilinear model.

Image processing techniques (object detection [16], contrast enhancement [17], etc.) might be used to highlight some features of single images, i.e. at a given spectral channel or the mean image across all spectral channels [18], but disregarding the spectral domain will prevent chemical interpretation. Some work has also been done on image segmentation, with the integration of the spectral domain [19], as well as utilizing the spectral and spatial domain, interactively switching between the two modes [20]. The analysis of textural features in spectral imaging has also been explored, by using the wavelet transform (WT) [14, 21-26]. These analyses i) use a MIA-like approach, where the local spatial information is extracted by WT, and 2D-WT sub-images are then analyzed by

multivariate analysis [14, 21-23], either on each single sub-image [21, 22] or on the entire sets [14, 23-24]; ii) exploit 3D-WT on the imaging data cube [25] or iii) fuse the 2D-WT sub-images obtained at each spectral channel [26]. These approaches also aim at linking the spectral and spatial domains, in some cases the spectral interpretation is not so straightforward [25] or of no concern [26], while in others the dimensionality of data matrix when passing from multispectral [14] to hyperspectral images become quite huge [24].

We recently proposed a novel approach to highlight the spatial-spectral interplay of the different components underlying a spectral imaging data set of a complex analytical system and published preliminary results [27] concerning a relatively simple Raman spectroscopy case study and a more complex one involving NIR spectral imaging datasets of an oil droplet in water and of biological fluids on cotton fabrics, respectively. However, we noticed that in systems of higher complexity, whose components show strong spectral and spatial overlap the analysis will become increasingly complex. To cope with this kind of situation, we here propose an extension and formalization leading to a novel method, called Image Decomposition, Encoding and Localization (*IDEL*). The method is meant to be unsupervised and exploratory.

IDEL relies on wavelet transform (WT) to resolve spatial features in distinct WT sub-images, then encodes this information in a set of descriptors (by using gray-level co-occurrence matrices [28]), and finally recovers specific spectral signatures for each spatial feature by multivariate data analysis. The encoded spatial information is fully exploited applying a semi-automatic procedure (that is data-driven) furnishing as a result a set of distinct spatial features linked to the specific spectral channels at which they are observable. In this way, clear and precise spatial features can be extracted, while chemical interpretability is maintained.

IDEL is challenged with a benchmark consisting of complex samples made of semen and lubricant stains on cotton fabrics analyzed with NIR imaging. There is significant spatial and spectral overlap between the stains and fabrics, and strong scattering effects are present. The localization of the fluid

on the substrate is of interest in forensic applications. As such, the segmentation of the biological fluid from the substrate as well as the removal of the significant scattering effects visible in the spectral imaging data is crucial. *IDEL* was able to isolate the stains from the fabrics, while preserving spectral information, as well as isolating a spatial structure previously unobserved. Moreover, the final obtained model is capable of isolating components also in new images, of similar type, once projected on it.

2. Materials & Methods

2.1 Methodology

The framework for WT decomposition and gray-level co-occurrence matrices is described in detail in reference [27] and briefly recalled in this section. The main novelty implemented in this work consists of a methodology where: i) only the most relevant spatial information is selected by applying PCA on the descriptors' matrix, which is based on picking the most significant scores (in terms of unique information) by means of convex peeling [29, 30] and ii) a semi-automatic procedure to link the spectral information to the relevant spatial information, establishing a correspondence among PCA scores and loadings. As a result, the most relevant wavelet sub-images are extracted. These sub-images form a new data cube that contains the most significant spatial information at specific spectral channels. In more general terms, the spatial structures are firstly resolved, exploiting the original data cube. Then, maintaining a direct link with the spectral signature, a reduced image data cube is retrieved in the WT domain. Subsequently, to interpret the corresponding information encoded in terms of individual spatial components, a PCA approach is proposed. In fig. 1, the three main steps of *IDEL* are schematically shown. These steps are explained in detail in the following sub-sections.

2.1.1 Spectral image decomposition and encoding

The first step consists of the decomposition of the individual images corresponding to each spectral channel by 2D-WT (see fig 1a). 2D-WT is a very powerful filtering method, highlighting the different

frequencies content of an image, while maintaining their localization with respect to the original domain. High- and low-pass filters are applied to decompose the signal into two disjoint subspaces holding the sets of details and approximation blocks (high and low frequencies, corresponding to sharp and smooth features, respectively). The decomposition is iterated on the approximation block, obtaining at each level a coarser representation of the image than in the previous approximation block and the filtered higher frequencies in the details. For image analysis, the same mono-dimensional wavelet filters are recursively applied along the two image directions. For each decomposition level, four blocks are obtained: 1) approximation (A): a low-pass filter is applied both row- and column-wise; 2) horizontal details (H): a low-pass filter is applied row-wise, then a high-pass filter, column-wise; 3) vertical details (V): a high-pass filter is applied row-wise, then a low-pass filter, column-wise; 4) diagonal details (D): a high-pass filter is applied both row- and column-wise. The specific direction along which the low- and high-pass filters are alternated, allows for specific textural patterns to be captured e.g., the H decomposition block highlights any pattern which would manifest horizontally, such as stripes (hence the name horizontal details). For the V and D blocks, vertical and diagonal textural patterns are highlighted, respectively, while the A block holds the original image with the details subtracted. We use the 2D stationary WT (2D-SWT) [31] which retains the size of the original image (see fig. S1, Supplementary Material), so that the decomposition blocks (from now on, referred to as sub-images A, H, V, and D), for each decomposition level, are equal in size to the raw image.

Wavelet filters are grouped in specific families, which differ in shape and symmetry, while amplitude is modulated in each family by the number of vanishing moments [32]. The choice of an appropriate wavelet filter is data dependent and providing an automatic tool to tackle this task is outside the scope of the paper. However, there are criteria detailed in literature [33, and references therein] to guide the choice of suitable wavelet filters. A general recommendation, that can be given is that the simplest *Haar* wavelet, which comes from the Daubechies-family (Daubechies-1) is usually a good starting point when, as in this case, the aim is exploratory. In fact, *Haar* can capture general

Journal Pre-proof

changes present in an image, not focusing on specific spatial features, and disentangle signals that range from sharp contrasting edges to broad structures.

In this work, the simplest *Haar* wavelet is applied, which comes from the Daubechies-family (Daubechies-1) and showed good performance (other wavelet filters, such as Daubechies-2, -5, -7, Symlet-2, -4 and Coiflet-3, -5 were tested, data not shown); the maximum decomposition level compatible with the image size was used. As is illustrated in fig. 1b, 2D-SWT is applied to the spectral imaging data.

To exploit the spatial information, the decomposition blocks are encoded into a set of descriptors that contain information on distinct local spatial features. This is done by calculating descriptors on the gray-level co-occurrence matrices (GLCM) derived from the A, H, V, and D sub-images. The GLCM method maps the spatial dependence of pixel-pairs in quantized gray-level images. The quantization was set at 128. As such, each image intensity is normalized and distributed across 128 gray-levels. A map is generated with size 128 by 128 elements, containing all possible quantized pixel-pairs. Selecting the appropriate number of gray-levels is similar to selecting the bin size in a mono-dimensional histogram and is always dependent on the size of, and information present in an image. A balance must be found between highlighting the relationships between neighboring pixels and not losing the details in the maps. Choosing a low number of gray-levels (large bin size) will result in a high number of counts over a small number of points, while choosing a high number of gray-levels (small bin size) will yield a low number of counts over a larger number of points (see fig. S2 in Supplementary Material for a visual representation).

On the pixel-pairs counting, two other parameters are of importance, namely the offset and angle. Both parameters must be attuned to the decomposition blocks and levels, due to the nature of WT. The offset determines the distance at which every neighboring pixel is observed with respect to the main pixel e.g., for a direct neighboring pixel, this distance is 1. This has been set to vary as $2^{level-1}$, with *level* being the WT decomposition level corresponding to the sub-image being codified. This

permits GLCM to account for the smoother patterns that are highlighted with increased WT decomposition levels, due to the removal of higher frequencies. The second parameter is the angle, or neighbors' location, which dictates the direction in which a neighboring pixel is located. We select the angle to maintain coherency between the directions of the WT decomposition details H, V, and D and the location of the investigated neighbor within the GLCM. Hence, the angle is set depending on the type of sub-image: H considers the top and bottom neighbors, V, the left and right neighbors, and D, the top-left, and bottom-right neighbors. These neighbors highlight the local differences within the sub-images. While for A, as there is no specific direction, the neighbors in all directions, are considered.

To encode the information carried by distinct patterns within an image, a set of eight descriptors was calculated from the GLCM, namely Energy, Contrast, Correlation, Variance, Inverse difference moment, Sum entropy, Information measure of correlation 1, and Maximal correlation coefficient. These are a subset of the descriptors proposed by Haralick et al. [28], which were selected as they are not much correlated with one another while describing all the relevant spatial features. We refer to [14] for a more in-depth survey of the selected descriptors.

As is illustrated in fig. 1c, a matrix of dimensions: *decomposition blocks x decomposition levels*, in the rows, and *spectral wavelengths* in the columns, is obtained for each descriptor, which is auto-scaled. Appending column-wise the matrices obtained for all descriptors, a so-called Descriptors' matrix (DM) is obtained (see fig. 1d).

2.1.2 Locating informative decomposed images

The DM contains descriptors on sub-images at every spectral channel, encoding spatial information in the rows and retaining spectral information in the columns. Applying PCA to DM, scores (fig 1e) and loadings (fig. 1f) thus relate to the spatial and spectral information, respectively. The number of PCs to consider is of course data dependent and here we used the scree-plot as a guideline. Each point in the scores plot corresponds to one descriptor of a sub-image (A, H, V or D) at a specific decomposition

Journal Pre-proof

level. Thus, looking at the scores, the most distinct spatial structures can be identified. The loadings plot, in conjunction with the scores plot, enables us to establish a link with the spectral channels. In fact, the loadings plot shows at which spectral wavelengths the largest variation of the descriptors within the different sub-images and decomposition levels is observed.

To this aim, we developed a semi-automatic procedure, which looks for relevant points in the scores plot while matching them to the loadings. It is a two-steps procedure.

The first step of the procedure consists of the selection of relevant sub-images from the scores plot. It is based on the estimation of the convex-hull of the score points (fig. 1e). Convex hull is applied, instead of e.g., a thresholding on scores values, as it depicts the minimum set of distinctive points enclosing all information captured in the scores plot. Here, we implemented a "peeling procedure" where the convex hull is applied twice. The first convex hull will remove the first "peel" of the data and the second will refine the selection. This accounts for situations where a few quite extreme points may skew the convex geometry too much [34, 35]. This procedure identifies the distinct sub-images that show the highest variation across spectral channels for the descriptors as, e.g., in fig. 1e, where the selected points (marked red) show significant variation on the first two PCs, meaning that they show high variation for a specific descriptor (within a certain sub-image at a given decomposition level across all spectral channels).

The second step is to match the salient spectral channels with the distinct spatial features (see fig. 1f). To do this, the scores and loadings must be reconducted in the same space by adequate scaling, as in a biplot [36]. The correspondence of the loading points with the selected score points is expressed in terms of angle, which evaluates the location of the scores and loadings with respect to the origin of the PC-space. To identify the loading points that have a correspondence with specific score points, a threshold is set around 20 degrees (zero degrees meaning perfect correspondence, and ninety degrees, no correspondence). The sign, of the scores and loadings, is not considered, meaning that a loading point that shows negative correlation (opposite location with respect to the

Journal Pre-proof

PC origin) to a score point is considered equal to a loading point that shows positive correlation. We assume that positive and negative correlations between the scores and loadings have equal importance.

A single sub-image can be selected multiple times if it showed significant variation across spectral channels in several different descriptors. In fact, there are eight different points in the scores plot corresponding to each sub-image at a specific decomposition level, one for each GLCM descriptor. To give a clear overview of the selected sub-images at distinct spectral channels and highlight the sub-images that show significant variation for several descriptors, a representation is generated. This representation, depicted as a Ω-map in fig. 1g, represents the decomposition blocks and levels of all sub-images *vs* the spectral channels. The Ω symbol indicates the sub-images selected by the procedure. The color coding on the color bar depicts the number of descriptors that were selected. In the end, only the sub-images that are required to explain the spatial features that make up the different spatial components in the wavelet decomposition are kept.

The selected sub-images are then reconstructed by inverse SWT and reorganized in the so-called Ω-data cube (fig. 1h). Even if the decomposed sub-images are of congruent size, reconstruction avoids spatial distortion with respect to the original image, which may be introduced at the deepest level of decomposition and brings the decomposed images back to original intensity scale. The Ω-data cube contains the wavelet sub-images at specific spectral channels, that isolated the significant spatial structures determined from a set of chosen descriptors. Also, the values in each of these sub-images, when assembling the Ω-data cube, are auto-scaled, and multiplied by $\sqrt{f}$, with f being the number of descriptors that have been selected for each selected sub-image. In this way, more weight is given to sub-image which show significant variation for more than one descriptor, meaning that different and distinctive spatial features are enhanced/captured by them.

2.1.3 Image fusion

Notwithstanding that the Ω-data cube contains a subset of reconstructed wavelet sub-images exploited by 2D-SWT decomposition of spectral imaging data, it still includes some redundant spatial information (the same spatial features are visible at more than one single spectral channel). Thus, it can be desirable to further distill the captured information. We generically refer to this task as "image fusion" and different approaches may be used. The simplest approach is to decompose the data matrix obtained after pixel-wise unfolding of the Ω-data cube by applying PCA. The refolded scores will provide images (fig. 1i) that combine the spatial patterns that show a similar variation in the Ω-domain. The representation and interpretation of the loadings is slightly more complex, as they do not encompass all channels of the original spectral domain (see fig. 1j). The loadings are organized in such a way that they will have the same dimensions as the Ω-map to get a clear overview on their importance (by means of the color bar) at a specific decomposition and spectral channel. This results in a so-called loadings map. Beside it, for each PC, a plot of the mean spectrum of the original spectral image is reported, with only the significant loadings highlighted by using distinct symbols/colors to indicate the corresponding wavelet sub-image, i.e. A, H, V, and D (fig. 1k). This is done to visualize any correspondence of the selected sub-images with any spectral bands in the original data set. The path from here can branch out, as extracting the spatial structures of the Ω-data cube can be done by several different fusion or modelling techniques e.g., one can apply MCR or Independent Component Analysis instead of PCA.

2.1.4 Ω-projection

An advantage of *IDEL* is that the generated PCA model can be used to project new imaging data, requiring only the 2D-SWT decomposition step to assemble the Ω-data for the test images (see section 3.3). Sub-images at the specific spectral channels (the ones belonging to the Ω of the training image) need to be calculated. Then, having the Ω-data of the new image, we can unfold and project it onto the PCA model, obtaining the scores, which in turn give the scores' images by refolding.

2.2 Data and preprocessing

The increased use of spectral images in forensic applications makes this methodology particularly interesting for body fluid detection [2-3, 37–39]. In such a scenario, forensic experts are often searching for compounds (such as blood, semen, and saliva) with specific spectral signatures that can link a crime scene to a victim, an assaulter or even a witness. However, those fluids usually appear on many different substrates, whose composition and texture characteristics can hamper its localization, making it difficult for the analyst to identify its origin and, consequently, to submit them to further DNA analyses, for example.

IDEL has been applied on ten spectral images of stained cotton fabrics. There are five differently colored (yellow, white, red, green, and black) cotton fabrics, each with a stain of either lubricant or semen. All semen samples were obtained from the same donor [3, 40], and the lubricant called KY-Jelly, mostly consisting of glycerol and hydroxyethyl-cellulose, came from the Durex© brand. The NIR imaging data was acquired by a Short-Wave Infrared (SWIR) SisuCHEMA imaging system from Specim (Oulu, Finland). The spectral range was 900 to 2500 nm with a spectral resolution FWHM of 10 nm and a spectral step size of 6.3 nm (256 spectral channels). The imaging system used had a lens of 50 mm and a pixel size of 156x156 mm^2. The squared pieces of fabric were stained with a droplet of semen and lubricant, and left to dry for a week at room temperature. We refer to Silva et al. [3] for more details on the samples and data acquisition. The pre-processing of the data in this work is not the standard procedure for NIR imaging data, as the aim of standard procedures is to generate bi-linearity within the data by harmonizing the scattering and removing the variance of the path length (e.g. multiplicative scattering correction [41], MSC and standard normal variate [42], SNV). The angle of analysis for IDEL is image processing while maintaining spectral correlation. As such, the intended purpose of pre-processing is to contrast the spatial features within the images, while maintaining spectral correlation. To achieve this purpose weighted least squares baseline correction is applied, where a baseline is estimated for each pixel. However, firstly, the data was smoothed with Savitzky-Golay [43] (11-point window, 2^{nd} order polynomial) to account for any unwanted spikes in the spectra.

Secondly, the first 40 and last 15 spectral channels were removed, as these show only noisy images, containing no significant information. And lastly, weighted least squares (WLS) baseline correction (3rd order) [44] is applied. A summary of the example dataset used in this work is shown in fig. 2. In addition, the scores maps and loadings profiles resulting from its PCA analysis after preprocessing by means of WLS and two more standard pretreatment algorithms for near-infrared data (i.e., MSC, and SNV) are displayed in figure S3 of Supplementary Material.

3. Results and Discussion

The stained cotton fabrics data are of interest for forensic applications and were used for the purpose of presumptive identification of biological fluids on textile. These data provide a meaningful benchmark to assess the efficiency of *IDEL*. At least two chemical constituents exist for each image, the cotton and the stain (lubricant or semen), but there is no spatial region without cotton, and the location of the stain may be detectable at selected spectral channels, but it is not clearly observed in the raw data as it is mixed with the cotton [16]. Moreover, the fabric and stain show overlapped spectral bands. The results obtained from three different images will be discussed, namely the yellow cotton fabric with a lubricant stain (LY), the white cotton fabric with a semen stain (SW) and the red cotton fabric with a semen stain (SR) (see fig. S4). In addition, the results of all ten datasets are provided in fig. S5.

For the LY and SW data sets, parameters were set as detailed in section 2.1.1. The results will go over the PCA analysis of the Ω-data cube, investigating the scores' images, loadings maps and highlighted loadings on the mean spectrum, as shown in fig. 1i, j and k.

3.1 Lubricant on yellow cotton fabric

The Ω-data cube for the LY data set consists of 140 sub-images, extracted from the wavelet decomposition, and the results are reported in fig. 3. The resulting scores' images (fig. 3a) of the first 4 PCs (explaining 58.3 % variance of the Ω-data cube) are considered, where four distinctive spatial

features are clearly recognizable. One can clearly identify the stain spot and the cotton fiber pattern, as will be discussed below.

The PC1 scores' image (fig. 3a) mainly shows the presence of an intense spot (almost in the center) which can be identified as a stain. The corresponding loadings map (see fig. 3b) shows that all the wavelet sub-images from every decomposition block (A, H, V and D) are contributing to the model, however the highest loadings values are mainly associated to approximation sub-images (A), which retain low frequency contributions in the original data set, hence smooth patterns. Figure 3c represents the relevant spectral wavelengths on the mean spectrum of the spectral image with the notable points being: i) approximation sub-images at decomposition levels 2 and 6 (A-2 and A-6), which are linked to positive loadings within the spectral region 1990 to 2060 nm, ii) sub-images A-4 and 5, linked to negative loadings around 2400 nm, and iii) A-6 linked to negative loadings around 1300 nm. Although it is extremely difficult to consider band assignment in NIR for such complex matrices, the band around 2000 nm suggests contributions from glycerol [45], one of the main compounds of the lubricant. The negative contribution at 1300 nm is interesting as well, as neither cotton nor glycerol absorb at that wavelength. This contribution could be linked to a solely physical effect, due to the lack of absorbance of either cotton or lubricant, or it could be linked to a third unknown component.

The PC2 scores' image (fig. 3a) clearly shows the diagonal texture associated to the cotton fibers. The loadings' map (fig. 3b) shows that the most relevant contributions are from the H and D sub-images in the spectral range from 1400 to 1550 nm. When looking at the loadings (see fig. 3), all these contributions relate to the band centered at 1494 nm. This can be attributed to the first overtone of O-H in cotton [46]. The main contribution comes from the D sub-images, however some minor contributions come from the H-sub-images. This can be attributed to the large spacing that is seen between the diagonal fibers, which can be captured in the horizontal details.

The PC3 scores' image (fig. 3a) is not straightforward to interpret. It shows some very smooth patterns, which are usually captured at the deepest decomposition levels (low frequency contributions in the spectral images) and mainly by approximations. However, details may also capture this type of information when, as in this case, there could be low frequency directional spatial patterns present (the most intense loadings values are from the H and V sub-images). When looking at the highest values in the loadings map (fig. 3b) and at their location on the mean spectrum (fig. 3c), the contributing spectral regions are quite spread and mainly on shoulders or along the spectral baseline. These patterns are quite difficult to interpret, and can originate from various sources, e.g. non-homogeneous illumination of the surface. These points can introduce minor variations in an image that can be seen in the deepest levels of a wavelet decomposition.

Finally, PC4, as for PC2, shows the texture of the cotton textile, however now its pattern has mostly a vertical direction. The main contributions are for details sub-images (mainly V) and again the relevant spectral region includes the band centered at 1494 nm.

Added to this is a contribution from the spectral band at about 2000 nm, which was not captured by PC2. This spectral channel is slightly shifted with respect to the contribution discussed in PC1. This could be attributed to the first overtone of R-CO-R. Possible reasoning for PC4 to be separated from PC2 is that the spatial structure is significantly different and is isolated as a different component. Even though they both originate from cotton, the overlapping fiber structures show significant differences.

Summarizing the results for the LY data set, different spatial features could be isolated, segmenting the stain and recovering the cotton fiber patterns across the whole image in the scores' images. A possible link to the spectral domain has also been established, where the interplay of chemical and physical information is observed.

Journal Pre-proof

Journal Pre-proof

3.2 Semen on white cotton fabric

The Ω-data cube consists of 491 sub-images, extracted from the wavelet decomposition. The results of the PCA analysis of the Ω-data cube for the SW data set are shown in fig. 4. The first four PCs (explaining 52.1 % variance of the Ω-data), which capture the different spatial structures, are discussed below.

The PC1 scores' image only shows the semen stain without any pattern related to the texture of the fabric (see fig. 4a). The loadings map (fig. 4b) highlights several contributions but the highest (in absolute terms) are from the A sub-images across most of the decomposition levels. Reporting the correlated loadings on the mean spectrum, the corresponding spectral regions are located at: 1300 nm, 1700 nm and 2200 nm, showing positive loadings values, 1450 nm, 1850 nm and 1940 nm, showing negative loadings values.

The contributions at 1700, 1850 and 2200 nm could relate to semen, as they could be attributed to protein bands [47]. However, the band at 1300 nm is not attributable to a specific component: it could be that this is solely associated to physical scattering effects that come into play, as something similar was observed in the lubricant example. The 1450 and 1940 nm bands could be attributed to water bands [48], as the loadings show negative values and a faint negative circle is observable in the lower left part of the corresponding score image (fig. 4a , PC1). A similar contribution can be seen on the PC3 scores' image (fig. 4a) but with an inverted sign (positive values of scores and loadings). Even if the samples were dried, it cannot be excluded that water on the border is reabsorbed due to the environmental conditions, since the humidity of the room (where the samples were stored) was not controlled.

As in the lubricant data, the PC2 scores' image depicts the texture linked to the cotton fibers. However, here the fiber orientation spatially manifests in the horizontal direction. As such, the H sub-images are mostly selected (fig. 4b, PC2). The salient loadings highlighted on the mean spectrum are

associated to the absorption band at 1494 nm, which has already been referred to as the first overtone of O-H stretching in cotton.

The PC3 scores' image shows, like for the lubricant data, smooth spatial patterns. However, a small intense circle is also visible in the bottom left part of the image. Looking at the salient loadings, both in the loadings map and reported on the mean spectrum, we see that mostly A and V sub-images have the highest absolute loadings; the relevant spectral channels are in large part the same as for PC1, e.g. 1300, 1450, 1830 and 1940 nm. In fact, the simultaneous absorbance around 1450 and 1940 nm could be linked to water, which could mean that what is observed is due a to a drying effect at the border of the semen stain. Analogously, similar, but negative spectral contributions were observed in PC1. In the image, the semen stain border has an elongated form in the vertical direction. As such, it is being captured by the vertical details (V sub-image). On the other hand, the A sub-images capture the small spot, which is linked to semen.

The PC4 scores' image shows the border of the semen stain, captured by H and V sub-images, contributing the most to the loadings map. However, the contribution from the texture of cotton is observable. Around the border of the stain, the spectral contributions from the cotton fabric and the semen stain are strongly overlapping. The salient spectral regions include the 1700 nm band, already discussed for PC1 as connected to semen, and the 1500 nm band connected to the cotton fibers, mentioned with regards to PC2.

The compression (or "fusion") step operated by PCA was extremely efficient to extract information, separating spatially not only the semen stain from the texture (which consists of the scattering effects of cotton), but also distinguishing the scattering around the border of the semen stain together with a possible drying effect of the semen.

3.3. Projection: Semen on red cotton fabric

We have seen that the proposed approach is very efficient to retrieve spatial information and interpret it in terms of spectral contributions. In particular, the scores' images obtained from the

analysis of the Ω-data help in discerning the various spatial components, which are not observable separately at any single spectral channel in the original data. The loadings highlight the spectral channels at which those components mostly manifest. A clear next step can be foreseen, which is evaluating if new (test) images projected on the loadings of a reference image can extract the same kind of specific spatial information in the scores' images.

The SR data set is investigated. The system is sufficiently similar to SW, but the shape of the stain, the scattering effects and the color of the cotton are different. The resulting scores' images are shown in fig. 5. In the projected scores' images, similar spatial features can be observed: the semen stain is isolated in PC1, the texture of the cotton fibers with some bordering effects is seen in PC2, in PC3 a bordering effect linked to the semen stain is visible, and finally, the joint border and scattering effects are highlighted along PC4. Although the texture of the cotton fibers is not completely isolated from the border effects in PC2, the semen stain has been isolated and correctly identified. These minor differences may be due to spatial and spectral differences between the data sets. The texture of the cotton fibers is not the same, i.e., it is oriented in a different direction with respect to the modelled image. In addition, the color of the fabric is different: red fabrics might exhibit a different absorption with respect to white. Also, the amount of deposited semen may not be the same, nor its position or shape. Very similar results were obtained by projecting, as test image, the green cotton fabric with a semen stain (SG), as shown in fig. 6a. It is worth noticing that, for both SR and SG, the squared prediction residuals (SPE) are in the same range of the calibration image (i.e. SW) as shown in fig. 5b and 6b, respectively.

Overall, these results seem very promising. Nonetheless, the projection (figure not shown for the sake of brevity) of the black fabric image with a semen stain (SB) and, to a minor extent, of the yellow fabric, while showing similar spatial features on scores images, resulted in high SPE signaling that when, the spatial structures and/or the spectral background (as it is the case of SB) of the test images

are very different from the calibration image much care should be taken in interpreting the scores maps, even if interesting spatial structure are unveiled.

4. Concluding remarks

IDEL utilizes WT, image encoding and PCA to extract decomposed sub-images that show significant variation across the spectral domain for spatial features related to distinct descriptors. Not only can it extract the distinct spatial-scattering effects present in a NIR spectral image, but also other components that show significant spatial differences between each other, while simultaneously having the capability to retain the spectral information that is linked to such captured spatial components. Thus, *IDEL* seems a very useful and powerful spectral imaging exploratory tool. However, some care must be paid when interpreting the highlighted spectral channels, as the previously discussed physico-chemical effects are difficult to separate from one another.

Once the model is built for components that have distinct spatial-spectral features, test images can be projected onto its space for their direct assessment. Also, the application of PCA to the Ω-data cube showed very promising results for spectral image interpretation. Some future work will be to utilize image fusion techniques to better extract and isolate spatial components.

The results obtained in this work can be generalized to any application field employing spectral imaging for the visualization of materials characterized by high morphological content, such as biological tissues [48], wooden materials [49–51], or remote sensing [52]. The integration of the proposed approach with other data analysis techniques, like multivariate curve resolution (MCR), will also be investigated.

5. Acknowledgements

Dr. C.S. Silva acknowledges financial support from: FACEPE (BFP-0800-1.06/17 and APQ-0576-1.06/17)

References

[1] I. Tahmasbian, N.K. Morgan, S. Hosseini Bai, M.W. Dunlop, A.F. Moss, Comparison of Hyperspectral Imaging and Near-Infrared Spectroscopy to Determine Nitrogen and Carbon Concentrations in Wheat, Remote Sensing 13 (2021) 1128. https://doi.org/10.3390/rs13061128.

[2] C. Malegori, E. Alladio, P. Oliveri, C. Manis, M. Vincenti, P. Garofano, F. Barni, A. Berti, Identification of invisible biological traces in forensic evidences by hyperspectral NIR imaging combined with chemometrics, Talanta 215 (2020) 120911. https://doi.org/10.1016/j.talanta.2020.120911.

[3] C.S. Silva, M.F. Pimentel, J.M. Amigo, R.S. Honorato, C. Pasquini, Detecting semen stains on fabrics using near infrared hyperspectral images and multivariate models, TrAC Trends in Analytical Chemistry 95 (2017) 23–35. https://doi.org/10.1016/j.trac.2017.07.026.

[4] H. Huang, L. Liu, M.O. Ngadi, Recent developments in hyperspectral imaging for assessment of food quality and safety, Sensors (Basel, Switzerland) 14 (2014) 7248–7276. https://doi.org/10.3390/s140407248.

[5] M. Manley, Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials, Chemical Society reviews 43 (2014) 8200–8214. https://doi.org/10.1039/C4CS00062E.

[6] J. Zhou, L. Yu, Q. Ding, R. Wang, Textile Fiber Identification Using Near-Infrared Spectroscopy and Pattern Recognition, Autex Research Journal 19 (2019) 201–209. https://doi.org/10.1515/aut-2018-0055.

[7] L.E. Agelet, C.R. Hurburgh, A Tutorial on Near Infrared Spectroscopy and Its Calibration, Critical Reviews in Analytical Chemistry 40 (2010) 246–260. https://doi.org/10.1080/10408347.2010.515468.

[8] B. Debus, R. Vitale, S. Sasaki, T. Asahi, M. Sliwa, C. Ruckebusch, A multivariate curve resolution approach to separate UV–vis scattering and absorption contributions for organic

nanoparticles, Chemometrics and Intelligent Laboratory Systems 160 (2017) 72–76. https://doi.org/10.1016/j.chemolab.2016.11.011.

[9] E.A. Magnussen, J.H. Solheim, U. Blazhko, V. Tafintseva, K. Tøndel, K.H. Liland, S. Dzurendova, V. Shapaval, C. Sandt, F. Borondics, A. Kohler, Deep convolutional neural network recovers pure absorbance spectra from highly scatter-distorted spectra of cells, Journal of biophotonics 13 (2020) e202000204. https://doi.org/10.1002/jbio.202000204.

[10] A. Kohler, J.H. Solheim, V. Tafintseva, B. Zimmermann, V. Shapaval, Model-Based Pre-Processing in Vibrational Spectroscopy, in: Comprehensive Chemometrics, Elsevier, 2020, pp. 83–100.

[11] F. Jamme, L. Duponchel, Neighbouring pixel data augmentation: a simple way to fuse spectral and spatial information for hyperspectral imaging data analysis, Journal of Chemometrics 31 (2017) e2882. https://doi.org/10.1002/cem.2882.

[12] M.H. Bharati, J. Liu, J.F. MacGregor, Image texture analysis: methods and comparisons, Chemometrics and Intelligent Laboratory Systems 72 (2004) 57–71. https://doi.org/10.1016/j.chemolab.2004.02.005.

[13] J.M. Prats-Montalbán, A. de Juan, A. Ferrer, Multivariate image analysis: A review with applications, Chemometrics and Intelligent Laboratory Systems 107 (2011) 1–23. https://doi.org/10.1016/j.chemolab.2011.03.002.

[14] M. Li Vigni, J.M. Prats-Montalban, A. Ferrer, M. Cocchi, Coupling 2D-wavelet decomposition and multivariate image analysis (2D WT-MIA), Journal of Chemometrics 32 (2018) e2970. https://doi.org/10.1002/cem.2970.

[15] R. Vitale, S. Hugelier, D. Cevoli, C. Ruckebusch, A spatial constraint to model and extract texture components in Multivariate Curve Resolution of near-infrared hyperspectral images, Analytica chimica acta 1095 (2020) 30–37. https://doi.org/10.1016/j.aca.2019.10.028.

[16] Z. Wang, P. Xu, B. Liu, Y. Cao, Z. Liu, Z. Liu, Hyperspectral imaging for underwater object detection, SR 41 (2021) 176–191. https://doi.org/10.1108/SR-07-2020-0165.

[17] G. Maragatham, S. Mansoor Roomi, A Review of Image Contrast Enhancement Methods and Techniques, RJASET 9 (2015) 309–326. https://doi.org/10.19026/rjaset.9.1409.

[18] Annual IEEE Computer Conference, IEEE International Conference on Image Processing, ICIP, IEEE International Conference on Image Processing (ICIP), 2014: 27-30 Oct. 2014, Paris, France, IEEE, Piscataway, NJ, 2014.

[19] J.-L. Xu, A.A. Gowen, Spatial‐spectral analysis method using texture features combined with PCA for information extraction in hyperspectral images, Journal of Chemometrics 34 (2020). https://doi.org/10.1002/cem.3132.

[20] N. Gorretta, J.M. Roger, G. Rabatel, V. Bellon-Maurel, C. Fiorio, C. Lelong, Hyperspectral image segmentation: The butterfly approach, in: Whispers 2009, IEEE, [Piscataway, NJ], 2009, pp. 1–4.

[21] J. Liu, J. MacGregor, On the extraction of spectral and spatial information from images, Chemom. Intel Lab Syst. 85 (2007), 119-130. https://doi.org/10.1016/j.chemolab.2006.05.011

[22] M.S. Reis, An integrated multiscale and multivariate image analysis framework for process monitoring of colour random textures: MSMIA, Chemom. Intel. Lab. Syst. 142 (2015), 36-48. http://dx.doi.org/10.1016/j.chemolab.2015.01.008

[23] P. Juneau, A. Garnier, C. Duchesne, The undecimated wavelet transform—multivariate image analysis (UWT-MIA) for simultaneous extraction of spectral and spatial information. Chemom. Intel. Lab Syst. 142 (2015), 304-318. http://dx.doi.org/10.1016/j.chemolab.2014.09.007

[24] A. Nardecchia, R. Vitale, L. Duponchel, Fusing spectral and spatial information with 2-D stationary wavelet transform (SWT 2-D) for a deeper exploration of spectroscopic images, Talanta 224 (2021) 121835. https://doi.org/10.1016/j.talanta.2020.121835.

[25] X. Guo, X. Huang, L. Zhang, Three-Dimensional Wavelet Texture Feature Extraction and Classification for Multi/Hyperspectral Imagery, IEEE Geosci. Remote Sensing Lett. 11 (2014) 2183–2187. https://doi.org/10.1109/LGRS.2014.2323963.

[26] Beauchemin M. Spatial pattern discovery for hyperspectral images based on multiresolution analysis, Int J Image Data Fusion. 3 (2012) 93-110.

[27] M. Ahmad, R. Vitale, C.S. Silva, C. Ruckebusch, M. Cocchi, Exploring local spatial features in hyperspectral images, Journal of Chemometrics 34 (2020). https://doi.org/10.1002/cem.3295.

[28] R.M. Haralick, K. Shanmugam, I. Dinstein, Textural Features for Image Classification, IEEE Trans. Syst., Man, Cybern. SMC-3 (1973) 610–621. https://doi.org/10.1109/TSMC.1973.4309314.

[29] H. Caussinus, P. Ettinger, R. Tomassone, Proceedings in computational statistics, Physica-Verl., Heidelberg, Wien, 1982.

[30] A. Cutler, L. Breiman, Archetypal Analysis, Technometrics 36 (1994) 338. https://doi.org/10.2307/1269949.

[31] G.P. Nason, B.W. Silverman, The Stationary Wavelet Transform and some Statistical Applications, in: A. Antoniadis (Ed.), Wavelets and statistics, Springer-Verlag, New York, 1995, pp. 281–299.

[32] Cohen A, Daubechies I., Jawerth B., Vial P., Multiresolution analysis, wavelets and fast wavelet transform on an interval. CRAS Paris, Ser. A, 1993; 316: 417-421.

[33] Prats-Montalbán J. M., Cocchi M. and Ferrer A., N-way modeling for wavelet filter determination in multivariate image analysis J. Chemometrics 2015; 29: 379–388

[34] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (2014) 2812–2831. https://doi.org/10.1039/C3AY41907J.

[35] J.A. Fernández Pierna, L. Jin, M. Daszykowski, F. Wahl, D.L. Massart, A methodology to detect outliers/inliers in prediction with PLS, Chemometrics and Intelligent Laboratory Systems 68 (2003) 17–28. https://doi.org/10.1016/S0169-7439(03)00084-4.

[36] I.T. Jolliffe, Principal Component Analysis, Springer International Publishing, Cham, 20.

[37] A. Majda, R. Wietecha-Posłuszny, A. Mendys, A. Wójtowicz, B. Łydżba-Kopczyńska, Hyperspectral imaging and multivariate analysis in the dried blood spots investigations, Appl. Phys. A 124 (2018). https://doi.org/10.1007/s00339-018-1739-6.

Journal Pre-proof

[38] M. Romaszewski, P. Głomb, A. Sochan, M. Cholewa, A dataset for evaluating blood detection in hyperspectral images, Forensic science international 320 (2021) 110701. https://doi.org/10.1016/j.forsciint.2021.110701.

[39] F. Zapata, F.E. Ortega-Ojeda, C. García-Ruiz, Revealing the location of semen, vaginal fluid and urine in stained evidence through near infrared chemical imaging, Talanta 166 (2017) 292–299. https://doi.org/10.1016/j.talanta.2017.01.086.

[40] D.H. Owen, D.F. Katz, A review of the physical and chemical properties of human semen and the formulation of a semen simulant, Journal of andrology 26 (2005) 459–469. https://doi.org/10.2164/jandrol.04104.

[41] Geladi, P., MacDougall, D., Martens, H., Linearization and scatter-correction for near-infrared reflectance spectra of meat, Appl. Spectrosc., 39 (1985) 491-500.

[42] Barnes, R., Dhanoa, M., Lister, S., Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc., 43 (1989), 772-777.

[43] A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, Anal. Chem. 36 (1964) 1627–1639. https://doi.org/10.1021/ac60214a047.

[44] Eigenvector Research, Inc., PLS_Toolbox function reference manual, available at http://wiki.eigenvector.com/index.php?title=Wlsbaseline, Eigenvector Research, Inc., 2021.

[45] K. Izutsu, Y. Hiyama, C. Yomota, T. Kawanishi, Near-infrared analysis of hydrogen-bonding in glass- and rubber-state amorphous saccharide solids, AAPS PharmSciTech 10 (2009) 524–529. https://doi.org/10.1208/s12249-009-9243-0.

[46] H.W. Siesler, Y. Ozaki, S. Kawata, H.M. Heise (Eds.), Near-Infrared Spectroscopy, Wiley-VCH Verlag GmbH, Weinheim, Germany, 2001.

[47] E.W. Ciurczak, D.A. Burns, Handbook of near-infrared analysis, 2nd ed., Marcel Dekker, New York, 2001.

[48] Y. Ozaki, Applications in Chemistry, in: H.W. Siesler, Y. Ozaki, S. Kawata, H.M. Heise (Eds.), Near-Infrared Spectroscopy, Wiley-VCH Verlag GmbH, Weinheim, Germany, 2001, pp. 179–211.

[49] M. Halicek, H. Fabelo, S. Ortega, G.M. Callico, B. Fei, In-Vivo and Ex-Vivo Tissue Analysis through Hyperspectral Imaging Techniques: Revealing the Invisible Features of Cancer, Cancers 11 (2019). https://doi.org/10.3390/cancers11060756.

[50] J. Sandak, A. Sandak, L. Legan, K. Retko, M. Kavčič, J. Kosel, F. Poohphajai, R.H. Diaz, V. Ponnuchamy, N. Sajinčič, O. Gordobil, Č. Tavzes, P. Ropret, Nondestructive Evaluation of Heritage Object Coatings with Four Hyperspectral Imaging Systems, Coatings 11 (2021) 244. https://doi.org/10.3390/coatings11020244.

[51] R. Vitale, P. Stefansson, F. Marini, C. Ruckebusch, I. Burud, H. Martens, Fast Analysis, Processing and Modeling of Hyperspectral Videos: Challenges and Possible Solutions, in: Comprehensive Chemometrics, Elsevier, 2020, pp. 395–409.

[52] P. Stefansson, J. Fortuna, H. Rahmati, I. Burud, T. Konevskikh, H. Martens, Chapter 2.12 - Hyperspectral time series analysis: hyperspectral image data streams interpreted by modeling known and unknown variations, in: J.M. Amigo (Ed.), Data Handling in Science and Technology Hyperspectral Imaging, Elsevier, 2020, pp. 305–331.

Legend of figures

Figure 1: *Illustration of DIEL. The methodology consists of three main steps. Firstly, "Spectral image decomposition and encoding", encompassing; a) a NIR spectral image, which is decomposed by means of wavelet transform into; b) blocks containing horizontal (H), vertical (V) and diagonal (D) details, and approximations (A) at different decomposition levels; c) that are then encoded into distinct descriptors and organized into a Descriptor's matrix. Secondly, "Locating the informative decomposed images": where d) the Descriptors' matrix is unfolded descriptor wise, retaining the spectral dimension, and e)-*

f) decomposed by principal component analysis. The convex hull of the resulting scores is highlighted in red and labelled in the scores plot, while the corresponding salient loadings are highlighted by a black point, inside the colored point in the loadings plot. g) The scores (on the convex hull) and their respective (salient) loadings are mapped in the Ω-map. The map reports on the "x-axis" the spectral channels and on the "y-axis" the decomposition block, to which each sub-image belongs, as well as the decomposition levels ordered from first to last (going down). Lastly, "Image fusion", where the sub-images that are localized in the Ω-map are extracted from the reconstructed wavelet decomposition and assembled into a Ω-data cube (h). Principal component analysis is applied on the unfolded Ω-data cube and the resulting (refolded) scores' images for the first two principal components are shown in i_{1-2}. The loadings are mapped and visualized in a so-called loadings' map (j_{1-2}), where the color coding is set according to the loadings values. The mean spectrum is shown in k_{1-2}, which highlights only loadings with absolute values > 0.075 (to declutter the figure, where negative values are denoted by a ✷ and positive ones by a O), colored according to the decomposition block: A (blue), H (red), V (black) and D (green) sub-images. The purple to red color coding relates to the spectral dimension throughout the figure.

Figure 2: *An illustrative data set is shown: (a) the spectral data cube, (b) the corresponding mean image and (c) 1 % of the spectra.*

Figure 3: *Results for the LY data set are shown. Scores' images (a_{1-4}), loadings' maps (b_{1-4}) and salient spectral channels on the mean spectrum (c_{1-4}) are shown for the first four principal components extracted from the analysis of the Ω-data cube.*

Figure 4: *Results for the SW data set are shown. Scores' images (a_{1-4}), loadings' maps (b_{1-4}) and salient spectral channels on the mean spectrum (c_{1-4}) are shown for the first four principal components extracted from the analysis of the Ω-data cube.*

Journal Pre-proof

Figure 5: *(a) Results of the projection of the SR data set onto the SW PCA model. Scores' images of the first four principal components are shown; (b) Plot of squared prediction residuals (SPE). SPE for calibration set (SW image) are shown in blue color.*

Figure 6: *(a) Results of the projection of the SG (semen stain on green fabric) data set onto the SW PCA model. Scores' images of the first four principal components are shown; (b) Plot of squared prediction errors (SPE). SPE for calibration set (SW image) are shown in blue color.*

Supplementary figures

Figure S1: *Framework of two-dimensional stationary wavelet transform. A low- (F_j) and high- (G_j) pass filter is applied row- and column-wise, in particular sequences, to retrieve distinct sub-images: Horizontal (H), Vertical (V) and Diagonal (D) details, and Approximations (A). ↑2 denotes an up-sampling that is applied to the resulting wavelet coefficients to retain the original image size.*

Figure S2*: The effect of the number of gray-levels in a GLCM to show the clear balance between a low and high number of bins / gray-levels.*

Figure S3: *PCA analysis of the semen stain on white cotton fabric data (SW), with different pre-processing techniques (WLS, top; MSC, bottom-left; SNV, bottom-middle; 2nd derivatives, bottom-right).*

Figure S4: *Mean images (across the spectral dimension) of the ten colored fabrics with either lubricant (left) or semen (right) stains.*

Figure S5: *Outcomes of the analysis of the ten images of colored cotton fabrics with either lubricant (left) or semen (right) stains. The first four scores' images and salient spectral channels on the mean spectrum are shown.*

Journal Pre-proof

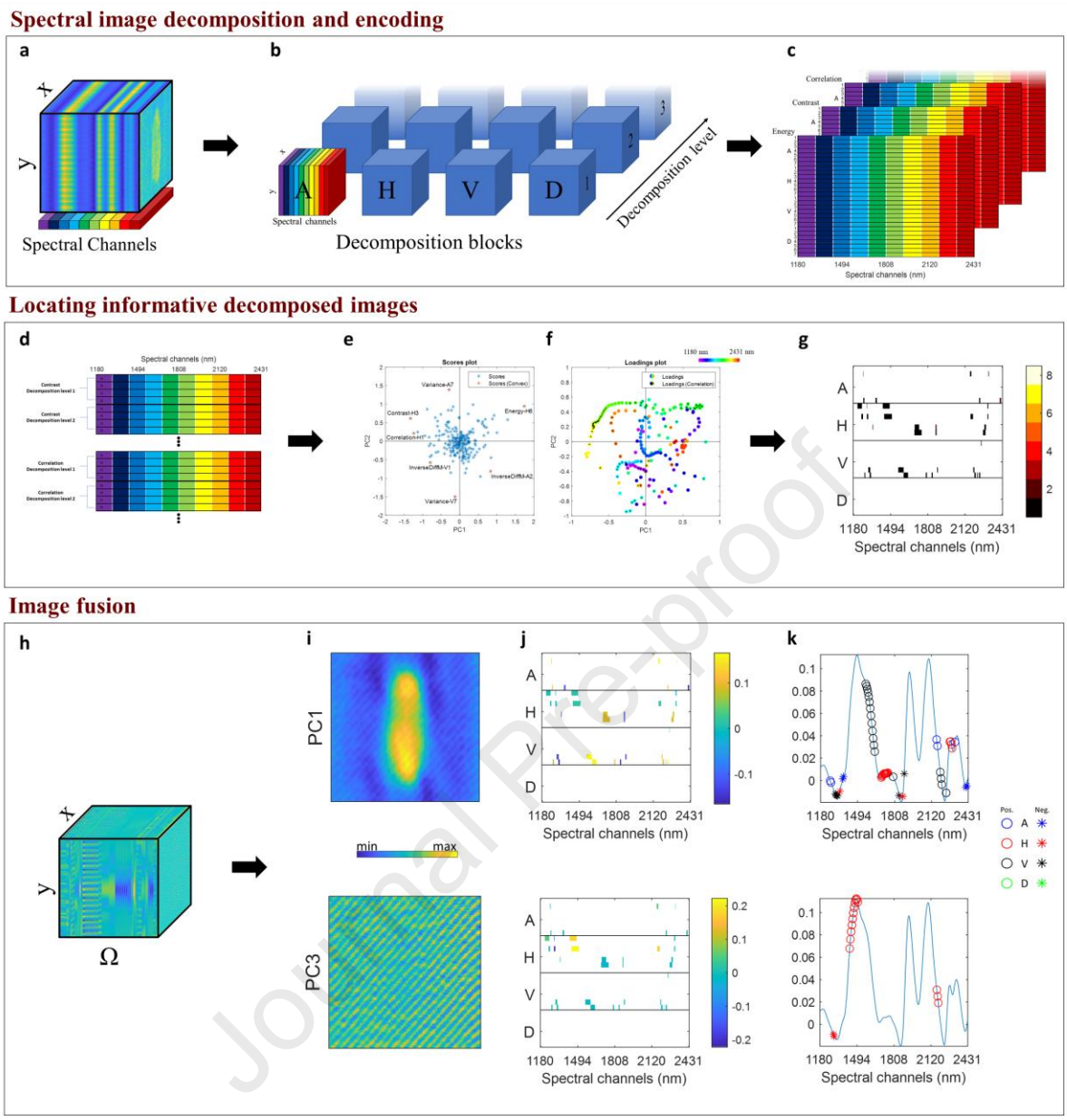

Spectral image decomposition and encoding
a
b
c
Spectral Channels
Spectral channels
A
H
V
D
Decomposition blocks
Decomposition level
Correlation
Contrast
Energy
Spectral channels (nm)
Locating informative decomposed images
d
e
f
g
Spectral channels (nm)
Scores plot
Loadings plot
1180 nm
2431 nm
PC1
PC2
Image fusion
h
i
j
k
PC1
PC3
min
max
Ω
Pos.
Neg.
1180 1494 1808 2120 2431
Spectral channels (nm)

Figure 1

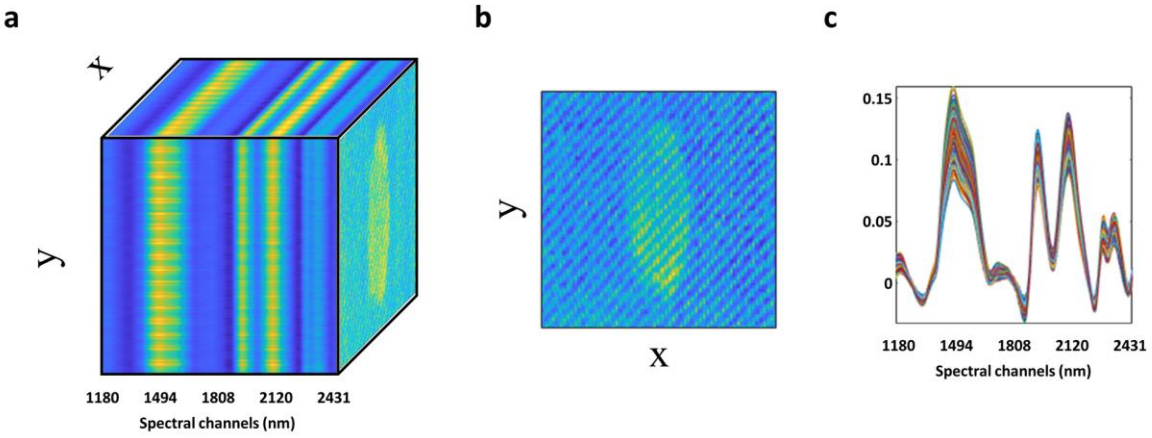


Figure 2

Journal Pre-proof

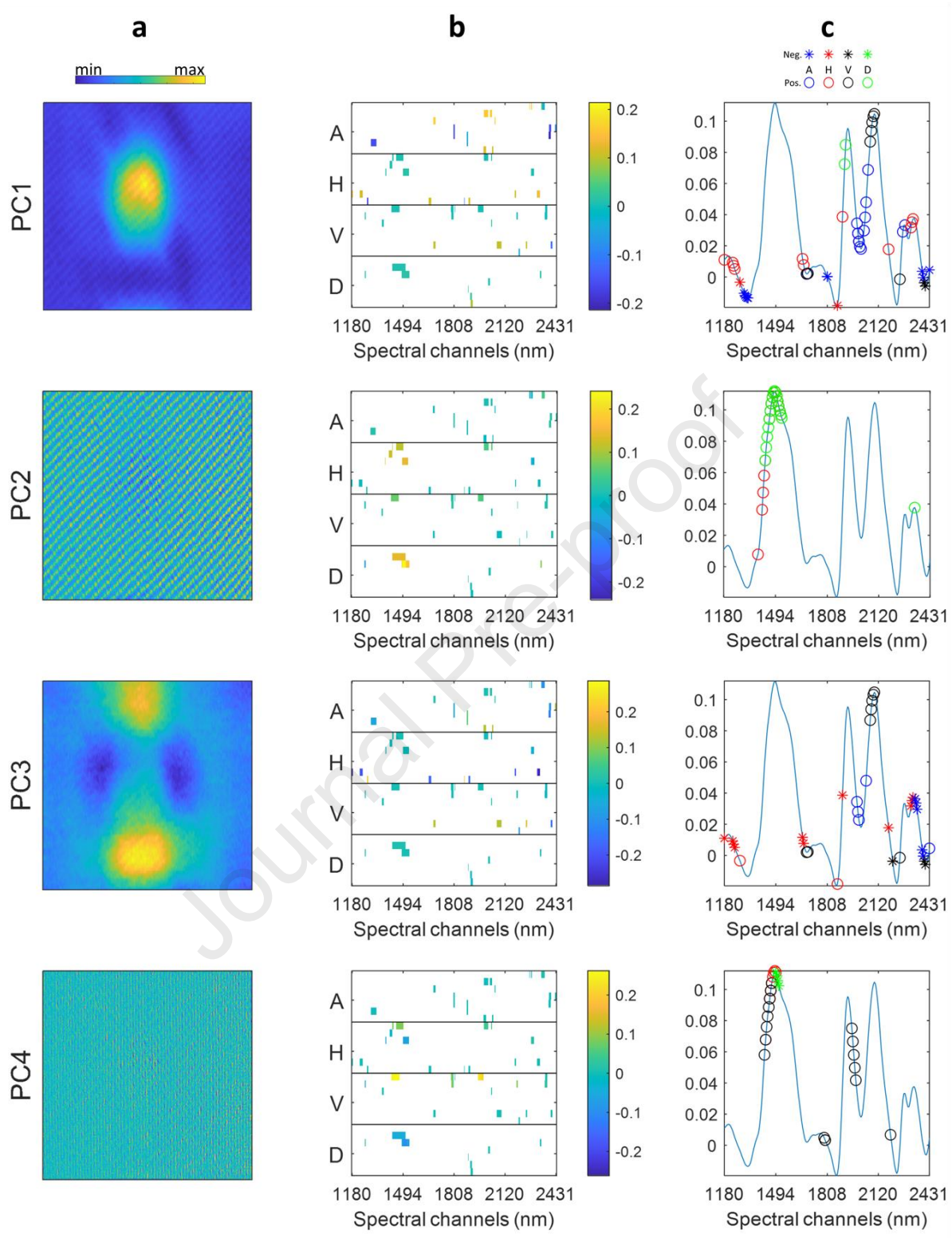


Figure 3

Journal Pre-proof

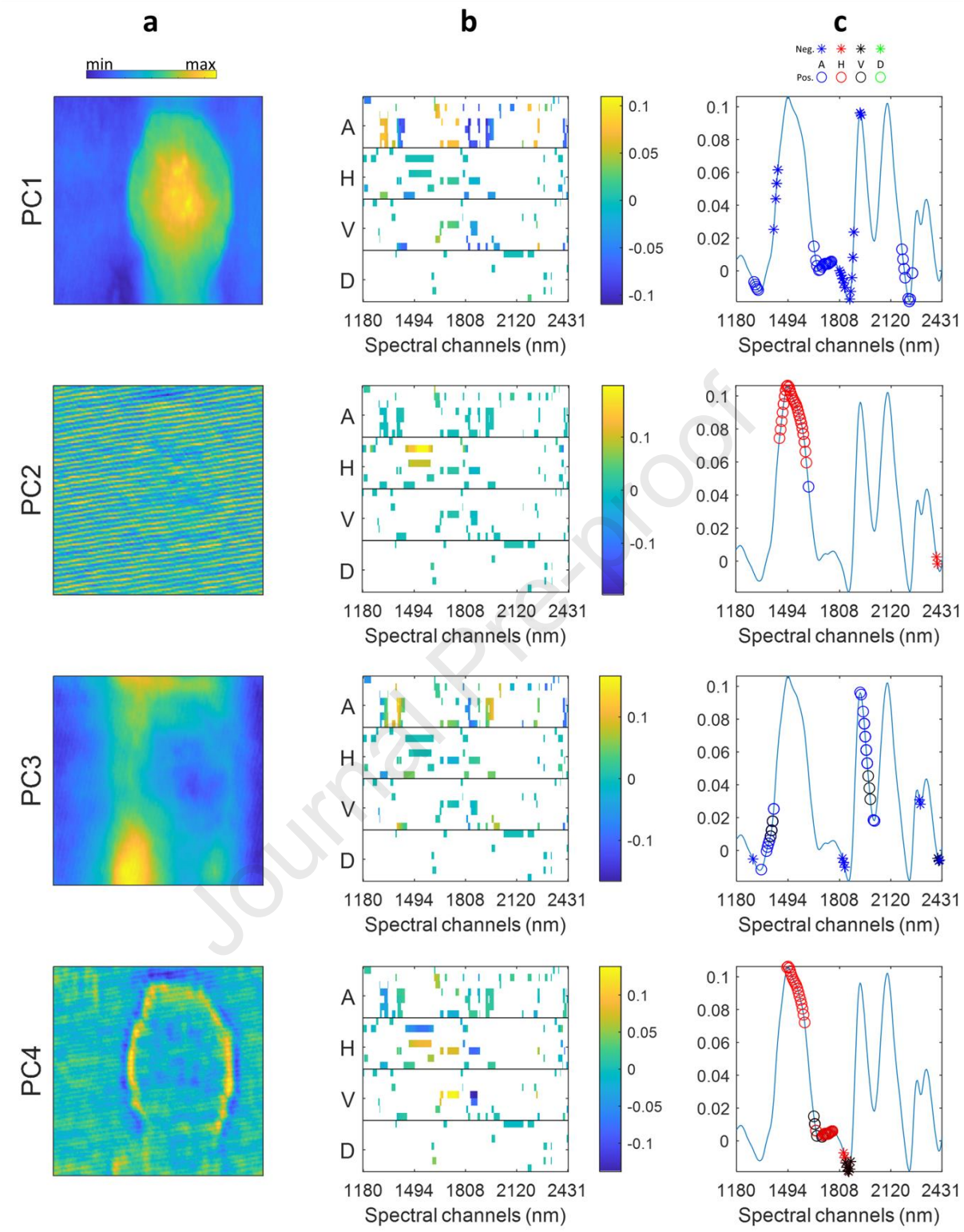


Figure 4

Journal Pre-proof

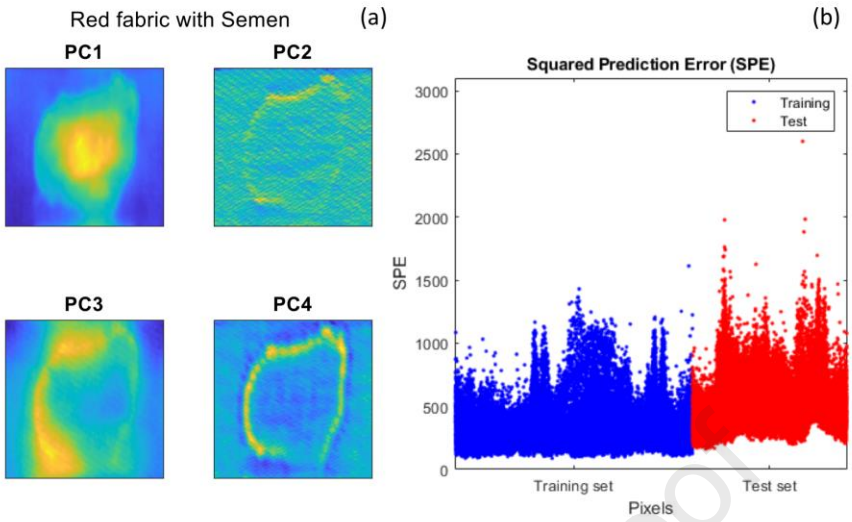


Figure 5

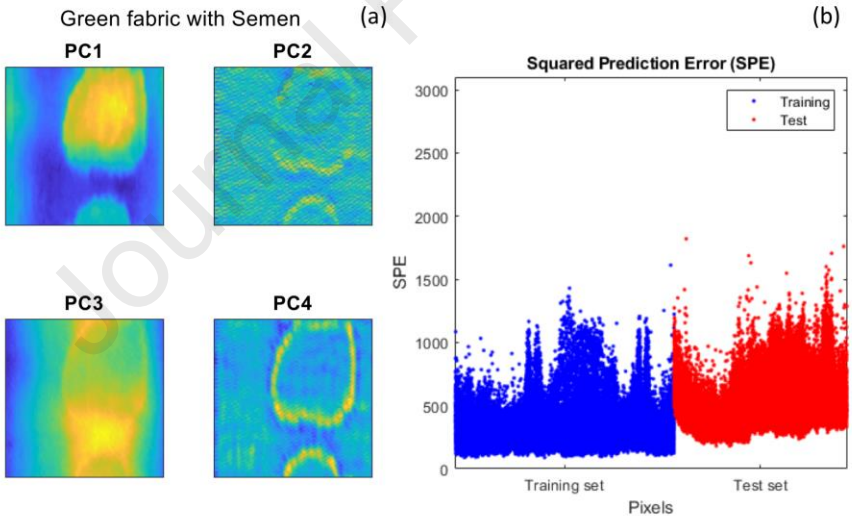


Figure 6

Journal Pre-proof

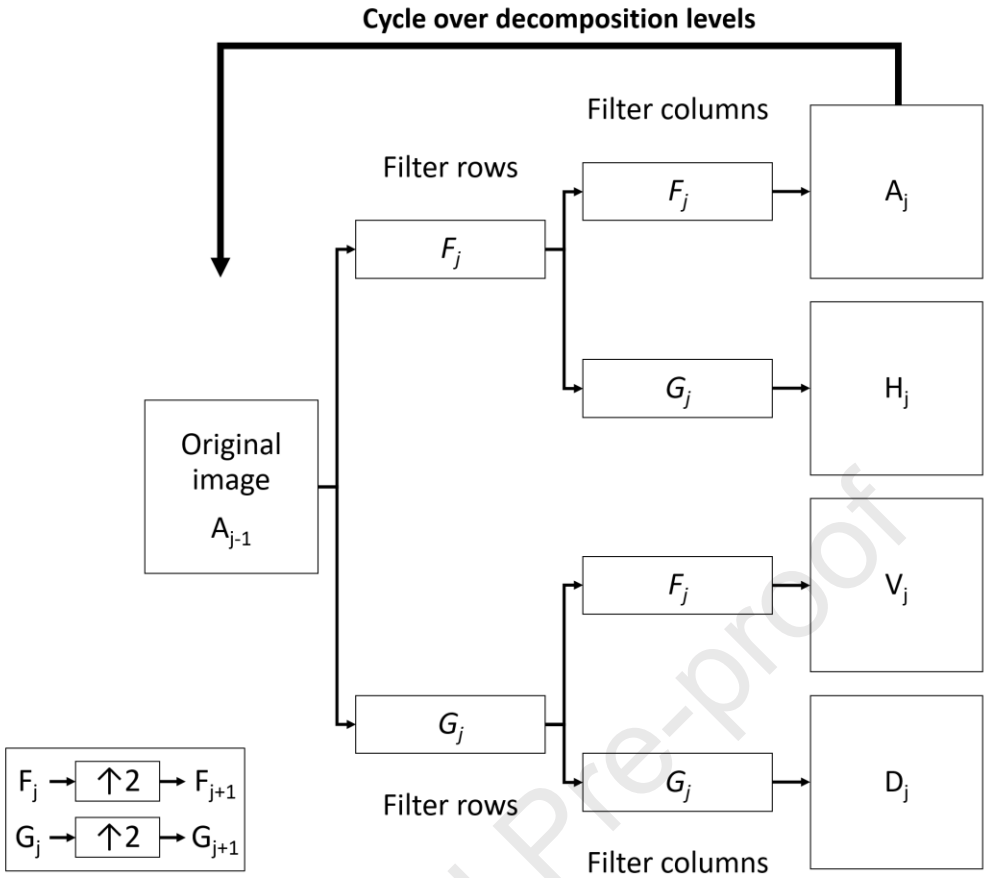


Figure S1 (Supplementary Material)

Journal Pre-proof

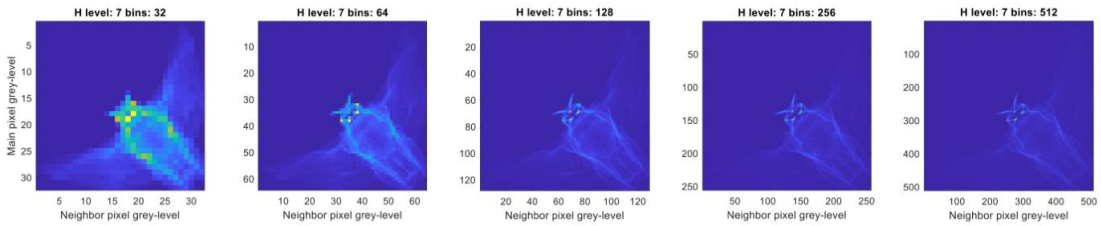


Figure S2 (Supplementary Material)

Journal Pre-proof

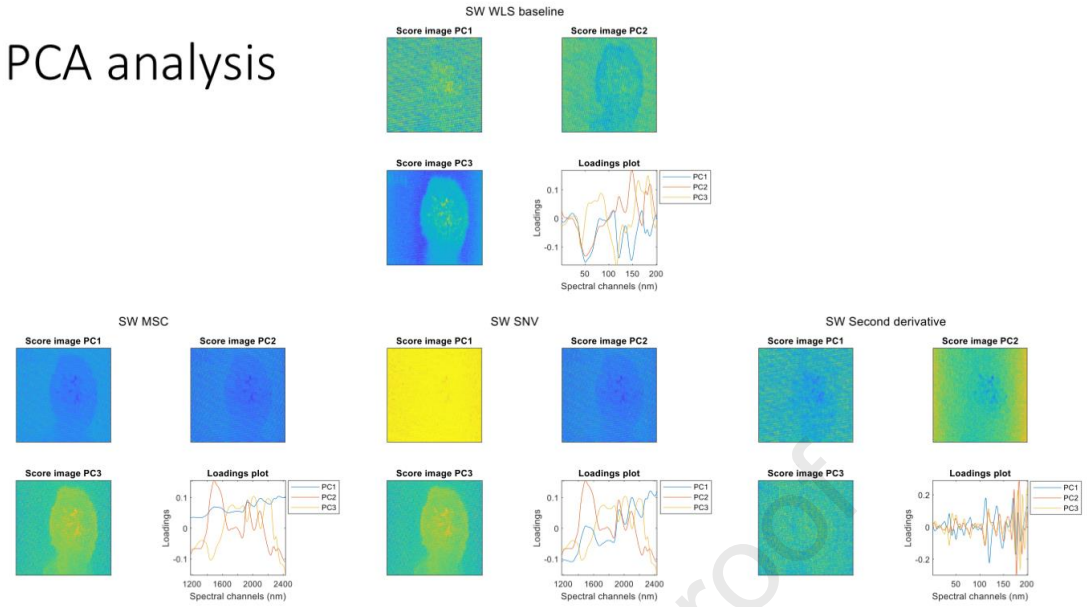


Figure S3 (Supplementary Material)

Journal Pre-proof

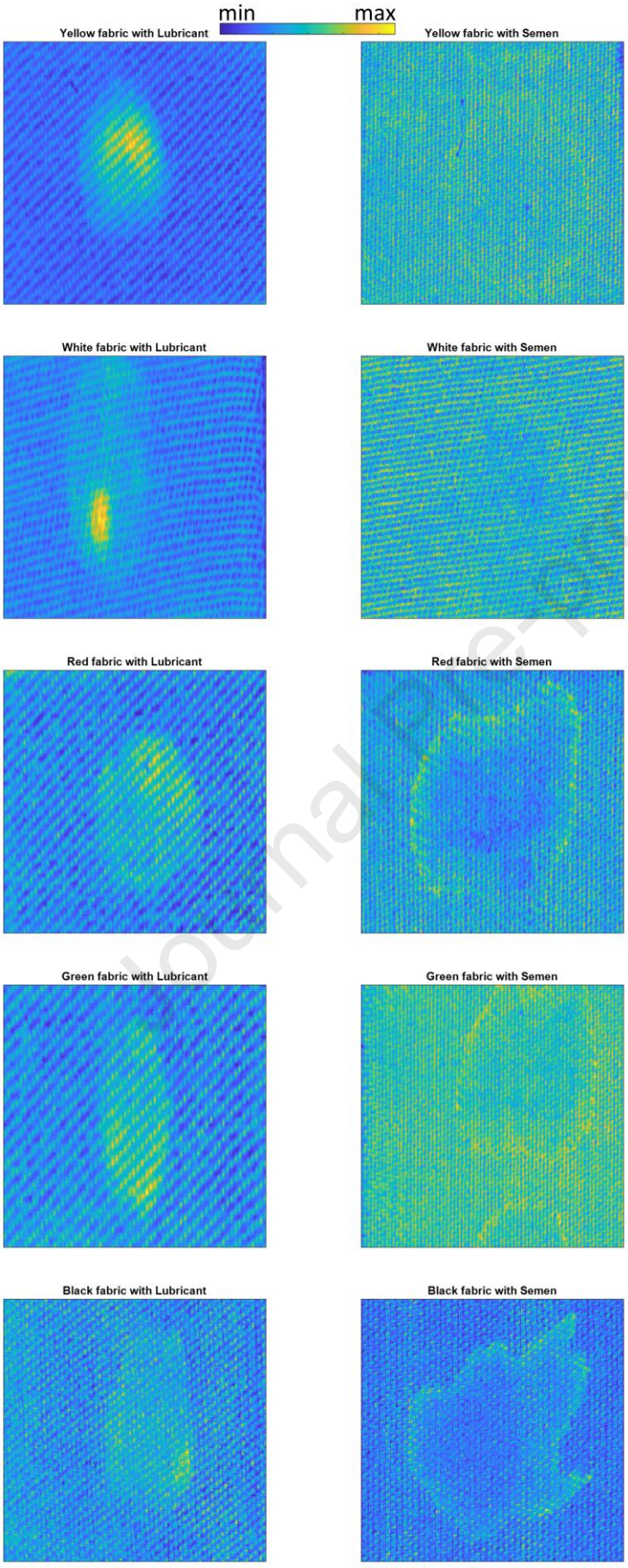

min
max
Yellow fabric with Lubricant
Yellow fabric with Semen
White fabric with Lubricant
White fabric with Semen
Red fabric with Lubricant
Red fabric with Semen
Green fabric with Lubricant
Green fabric with Semen
Black fabric with Lubricant
Black fabric with Semen

Figure S4

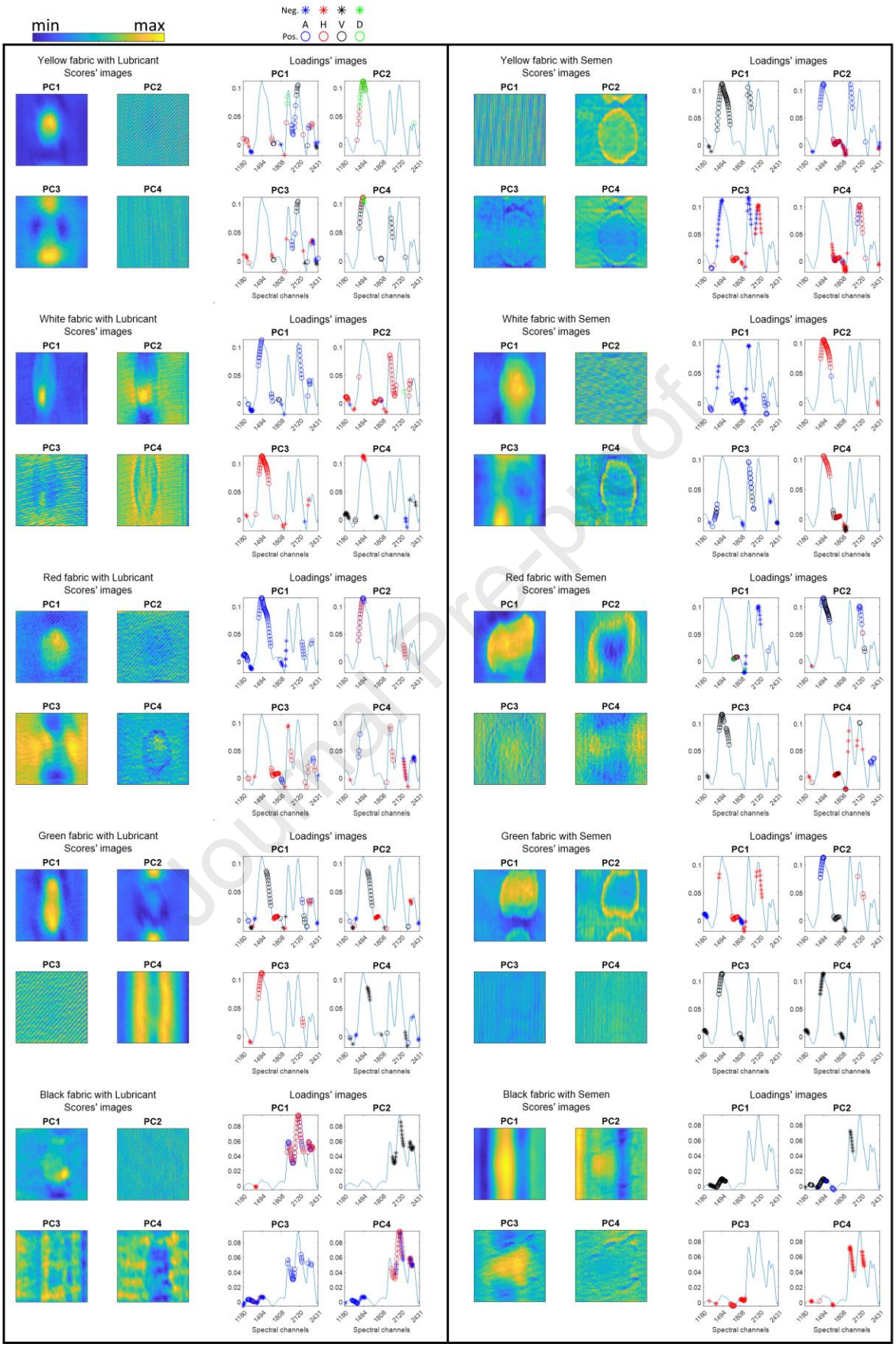


Figure S5 (Supplementary Material)

Journal Pre-proof

1. A novel method for unsupervised exploration of hyperspectral imaging data is presented.
2. The method is based on Image Decomposition, Encoding and Localization steps.
3. It retrieves distinct spatial features while linking them to specific spectral channels.
4. The method is tested on data sets of forensic interest.

Journal Pre-proof

Journal Pre-proof

Declaration of interests

☒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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proof