Unsupervised hyperspectral image segmentation of films: a hierarchical clustering-based approach

Lingxi Liu, Giovanni Delnevo[*] and Silvia Mirri

*Correspondence:
giovanni.delnevo2@unibo.it

Department of Computer
Science and Engineering,
University of Bologna, Bologna,
Italy

Abstract

Hyperspectral imaging (HSI) has been drastically applied in recent years to cultural heritage (CH) analysis, conservation, and also digital restoration. However, the efficient processing of the large datasets registered remains challenging and still in development. In this paper, we propose to use the hierarchical clustering algorithm (HCA) as an alternative machine learning approach to the most common practices, such as principal component analysis(PCA). HCA has shown its potential in the past decades for spectral data classification and segmentation in many other fields, maximizing the information to be extracted from the high-dimensional spectral dataset via the formation of the agglomerative hierarchical tree. However, to date, there has been very limited implementation of HCA in the field of cultural heritage. Data used in this experiment were acquired on real historic film samples with various degradation degrees, using a custom-made push-broom VNIR hyperspectral camera (380–780nm). With the proposed HCA workflow, multiple samples in the entire dataset were processed simultaneously and the degradation areas with distinctive characteristics were successfully segmented into clusters with various hierarchies. A range of algorithmic parameters was tested, including the grid sizes, metrics, and agglomeration methods, and the best combinations were proposed at the end. This novel application of the semi-automating and unsupervised HCA could provide a basis for future digital unfading, and show the potential to solve other CH problems such as pigment mapping.

Keywords: Hierarchical Clustering, Machine Learning, Cultural Heritage, Hyperspectral Imaging, Image Segmentation

Introduction

The application of imaging spectroscopy in the field of cultural heritage (CH) has drastically expanded over the last decade, supporting material diagnostics, chemical analysis, conservation practice, and also digital restoration [1–4]. Hyperspectral imaging (HSI) is one of the non-invasive analysis tools that register a spatial map with spectrum collected at each pixel position. It characterises detailed signatures of the material, attributed to the fine spectral resolution down to nanometer scale and wide sampling range varying from visible to infrared [5]. Due to its superior capacity to document and extract rich information from the inherently complex and delicate art materials[3], HSI has found

a fruitful application in the CH field. Numerous studies have demonstrated the effectiveness of employing hyperspectral imaging to identify and characterize colourants in a variety of media, including paintings, illuminated manuscripts, murals, and prints [6–8].

However, the implementation of HSI to study deteriorated film materials is yet to be fully explored. The colour film most typically has a trichromatic structure, in which the yellow, magenta, and cyan dyes locate in separate layers in the emulsion and combine in different proportions at each pixel upon exposure to the imaging objects [9]. However, those large synthetic organic molecules are intrinsically unstable and deterioration is inevitable, making the preservation of the original colour a challenging problem. Moreover, the degradation is non-linear among the dyes, which means certain types of dyes degrade at faster rates over time, and is often spatially inhomogeneous that certain areas are more damaged than the rest. Unlike in a painting, where areas painted with distinctive pigments are different in chemical compositions and the degradation products are unique, all pixels in a film have similar chemical constituents. Thus, the magnitude of colour loss needs to be accessed and it is affected by both factors: (1) image content that determines the original dye densities at each pixel, and (2) the degree of degradation that may vary across the images. This means, the points with similar original dye densities may end up in diverse degradation degrees (factor 2), and for pixels that are subjected to the same degree of degradation, the magnitudes to bring them back to the original states are still different if the original dye densities vary (factor 1). The complexity of the non-linear and inhomogeneous deterioration makes the spectroscopic measurement preferential to the simple RGB values, by providing extended information on the colour loss based on high-resolution spectral signatures [10]. Infrared spectroscopy has been applied to quantify the degradation degree of the transparent supports (cellulose-based materials) [11, 12], but it is a single point based analysis and not includes the study on emulsion layers. Regarding the colour studies of film using HSI, very few works have been reported. Trumpy et al. have proposed a "dye purification" digital restoration approach that uses HSI data acquired on the cross-section as references of the spectral densities [13]. However, their method is developed on synthetically faded images and does not consider the inhomogeneous degradation effect. Another work devoted to the digital restoration of films used a "vector quantization" algorithm to look up the most suitable unfaded spectrum from the spectral codebooks [14]. The faded and unfaded pair of references in a codebook is merely established via spatial relationship. It would further improve the performance of digital restoration and extend the applicability of the method to first segment the image into clusters of similar spectral features, considering both the dye density and degradation degree.

In the meantime, it is still challenging to efficiently elaborate and exploit the vast data (millions of recorded spectra) of high complexity. In this context, machine learning (ML) algorithms are highly promising to provide innovative solutions to heavy spectral data analysis. From simple to complicated applications, the ML algorithms are suited for large data sets, performed with minimal human interference. Moreover, it establishes a direct correspondence between the input and output data, reducing the necessity of in-depth knowledge of all the chemical compounds during the degradation process [15]. Principal components analysis (PCA) and minimum noise fraction transform (MNF) are the data analysis tools most widely used by conservation

scientists for their efficiency and simplicity [15–18]. These multivariate data analysis techniques could be used as the pre-processing step for the pigment classification or mapping, while still requiring manual selection or labeling of the representative channels [19, 20]. K-means clustering has also been applied as more sophisticated algorithms for feature detection or image segmentation [16, 21]. However, finding the proper k value can be challenging, and applying it to high-dimensional hyperspectral data may result in instability and convergence failures.

To efficiently extract information from the large quantity of high-dimensional HSI data, the hierarchical clustering algorithm (HCA) is proposed to use as an alternative approach to classify and segment the spectral images [22–24]. In the medical, chemical, and remote sensing field, the implementation of HCA on spectral data treatment is widely explored [25–28]. However, compared to other multivariates analysis tools such as PCA, very rare applications of HCA on hyperspectral data have been reported in the field of CH. Even though, some simple applications have been described in the stratigraphical analysis of painting materials. The research tasks are mainly to determine the chemical distribution of certain painting materials on the cross-section or to separate the artist's original painting layers and the restorer's retouches [16, 29, 30]. Without prior knowledge of the exact chemical composition of the complex mixture, the unsupervised HCA algorithm allowed for successful mapping of the cross-section taking advantage of the spectral data. This opens the door for a more in-depth investigation of applying HCA to image segmentation, in addition to stratigraphical analysis.

The goal of this project is to investigate the potential of clustering algorithms in assistance for further digital colour restoration. The restoration of cinematic films is a complex process that involves several stages, including digitization, pre-processing, color correction, and de-noising. The colour correction will be performed via Singular Spectrum Analysis (1D-SSA) based transformation. In the foreseen method, the first SSA component is used to compute the distance of the faded spectrum to the unfaded state, obtaining a transformation vector that could be applied to all the similar spectra. This process is repeated for each fading group since the degradation is inhomogeneous in the historical films. Thus, the effective segmentation of various degradation features is required for the correct reconstruction.

In this paper, we focus on the clustering aspect of the overall approach to digital film restoration by proposing and investigating an HCA-based data processing pipeline. Data sets used in this experiment were acquired with a custom-made push-broom VNIR hyperspectral camera (380–780 nm) on a set of historic cinematic film samples with various degrees of degradation. The data set was pre-processed and divided into sub-images by certain grid sizes (200 × 200, 100 × 100, 50 × 50) to learn the input clusters to the algorithm. Calculating the distances between every pair of clusters, the initial clusters were repeatedly merged into larger clusters according to defined agglomeration functions, until reached only one cluster. All the mergers were recorded, and the segmentation was retrieved from the final clusters obtained. The preliminary outcomes have already been published [31]. Here a more thorough analysis of all the relative algorithmic parameters is presented, including the grid sizes, metrics, and agglomeration methods. The qualitative analyses of the obtained results were presented and the best combinations of parameters were proposed at the end.

Liu *et al. Journal of Big Data*   (2023) 10:31

Page 4 of 16

The remainder of the paper is structured as follows. Section Materials and Methods presents the data used in this study and the overall methodology. Then, Section Results and Discussions details and discusses the results obtained and the limitation of the proposed method. Finally, Section Conclusions closes the paper, presenting some final remarks and future works.

Materials and methods

This Section describes the materials, the acquired hyperspectral images, and the pre-processing activities. Then, it presents the hierarchical clustering methodology with the relative parameters.

Materials

Five positive frames (Fig. 1) are selected to perform the analysis in this research. Provided by "L'immagine Ritrovata Film Restoration Laboratory" (Cinematique of Bologna), they are authentic historical films from the same scenario that report different fading effects. Physical measurements reveal that each frame is 34 mm × 19 mm (± 0.5 mm) in size. The film is FUJI N4 cellulose acetate film, produced between April and June 1986, according to edge marks printed on the perforation. As could be observed in Fig. 1, the degradation is at diverse stages among the image set and is also inhomogeneous within a single frame. The less severely degraded film, sample 23, appears purplish in colour and has higher dye density, which is supported by the reflectance spectra (Fig. 1b). The extensively deteriorated films (samples 22, 25, and 27) present a general pinkish hue due to the loss of cyan dyes, with certain yellowish areas more severely degraded that also partially lost the magenta dyes, preserving even lower dye density. Even though the ground truth is missing in this experiment, because of the inevitable degradation process regardless of the preservation condition, the relatively best-preserved sample 21 is considered as the reference frame.

Hyperspectral imaging

The spectral data of the sample frames were scanned by hyperspectral imaging. The HSI device in this experiment uses the push-broom system with 2448 across-track pixels and 2,048 wavelengths, composed of an imaging spectrograph (ImSpector V10E, Specim, Finland) and a monochromatic camera (Blackfly S, BFS-U3-51S5M-C, 5 MP, FLIR


Fig. 1 Optical images of the dataset selected for this experiment: frames number 21, 22, 23, 25, and 27

Integrated Imaging Solutions, Canada) [32]. Spectral images of the film samples were acquired in a reflectance geometry with broadband LED light source covering the spectral range from 380–1000nm. The spectral sampling rate is 0.3nm in the spectral range of 322.9–1025.3nm, where the considered effective range is from 380nm to 780nm. The effective spectral and spatial resolutions of the system were 2.9nm and 100μm, respectively, as evaluated by a gas discharge tube and spatial grids used for system calibration. Simple calibration and normalization of the spectra are performed at the time of registration.

Pre-processing

Each collected datacube has a size of 2448 pixels x 1400 pixels x 2048 wavelengths. The initial datacubes are over-abundant and contain information that is both spatially and spectrally out of the region of the interests. The uninformative range was thus removed during pre-processing of the data set. The spectra were binned in increments of 5 channels and the wavelength range less than 380nm and over 780nm were cropped for being excessively noisy. Furthermore, the uninformative border and perforation areas were removed. The image regions were cropped to multiples of hundreds of pixels for the convenience of further division of sub-images. After the initial processing, each individual reflectance datacube has a size of $1200 \times 1000 \times 240$ (Fig. 2a).

For a more thorough analysis of the data, selected spectra from various pixel positions on frame 22 were recovered, as shown in Fig. 2b. The varied degradation phenomena are reflected in the spectral patterns. As examples of the influences of factor 2 (introduced in the introduction), both spectra 1 and 3 were derived from the image's upper light region. The more extensively degraded pixel (spectrum 1) has lost
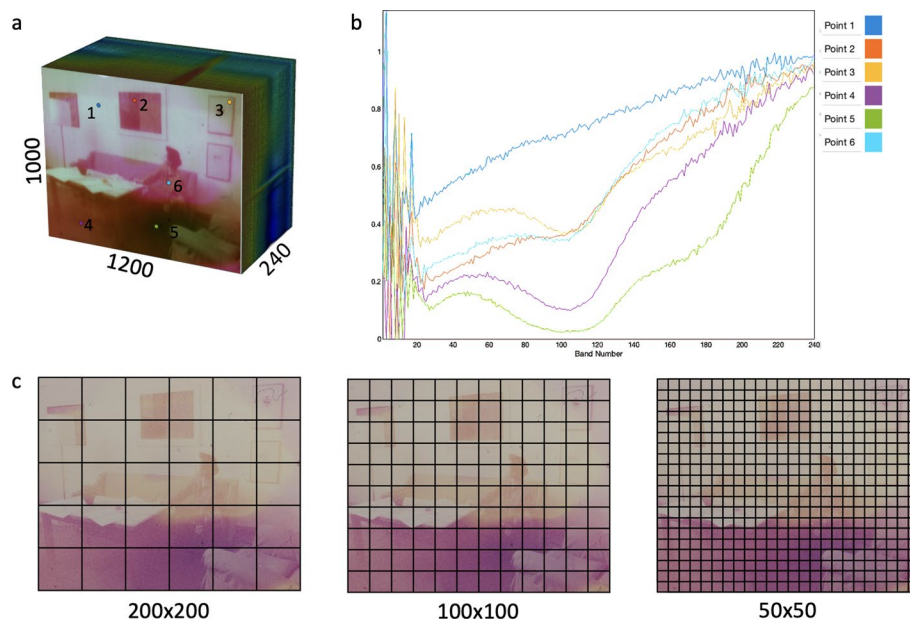
Fig. 2 Pre-processing phases of the hypercubes: **a** Illustration of the hypercube in region of interest, after cropping and binning. **b** comparison of spectra extracted from the hypercube, on which indicated the locations of the selected points. **c** Schematic representation of the grids evaluated: 200×200, 100×100, and 50×50

almost all the absorption features and has the highest reflectance value, whereas the less severely degraded corner (spectrum 3) still preserves the absorption around band 110. Spectra 4 and 5 come from the image's darker area of high original dye density, where point 4 is more colour-degraded than point 5. The relative lowest reflectance values (highest absorption by the dyes) are then found in spectra 5, with spectrum 4 partially losing the absorption around band 180. For the influence of factor 1, spectra 1 and 6 were both from the most degraded yellowish area. Spectra 6, extracted from the human figure, still preserved more dye density than spectra 1, which is originally very light and contains few dyes. On the other hand, spectra of similar degradation degree and original dye density present very similar spectral signatures, even if they belong to different image content. Spectra 2 and 6 were both extracted from the yellowish area, one from the wall painting and one from the human figure, and are almost overlapping in spectral shapes, suggesting similar requirements for digital restoration treatment. As compared to the simple RGB triple values, the distinctive spectral characteristics serve as the foundation for further clustering.

To further reduce the computational load, the datacubes were decomposed into small sub-images and only the centroids were used as the input for the algorithm. In terms of pixels, three grid sizes were chosen: 200×200, 100×100, and 50x50. According to each of these three groupings, every datacube in the set was divided respectively (Figure 2c). In this step, we obtained:

- For grid size 200×200: 150 sub-images (30 per frame).
- For grid size 100×100: 600 sub-images (120 per frame).
- For grid size 50×50: 2440 sub-images (480 per frame).

For each sub-image, the centroid spectrum was calculated by averaging all pixels to represent the entire sub-image. In the calculations that follow, each centroid spectrum will be treated as one initial cluster. This step was repeated for every sub-image in the whole dataset and then all the centroids were grouped together to form the initial input clusters.

Hierarchical clustering

Each group of resulting initial clusters was subjected to the agglomerative hierarchical clustering analysis. The closest two of the N centroids were repeatedly merged into larger clusters, until all of the spectra were in one cluster. All the mergers were recorded, and this information formed a hierarchical tree. All the data treatments were performed in Matlab R2021b with embedded functions. The processing pipeline is summarized as a flowchart (Fig. 3) and described step-by-step in the following paragraphs.

Grids and Distance metrics First of all, three groups of initial clusters were obtained by the method described in the pre-processing section. The same process was performed and repeated on each group of grid sizes (200×200, 100×100, and 50×50). Using those centroids as input clusters, five different metrics were tested for measuring the distances between every pair of initial clusters: Chebyshev, Cosine,
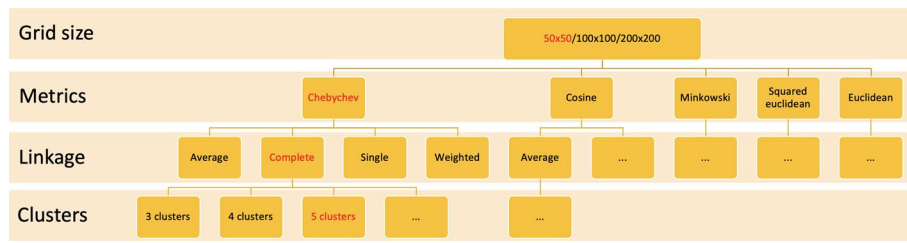
Fig. 3 The flow chart of the pipeline for determining the best combination of parameters for the hierarchical clustering. Parameters marked in red indicate one of the possible combinations
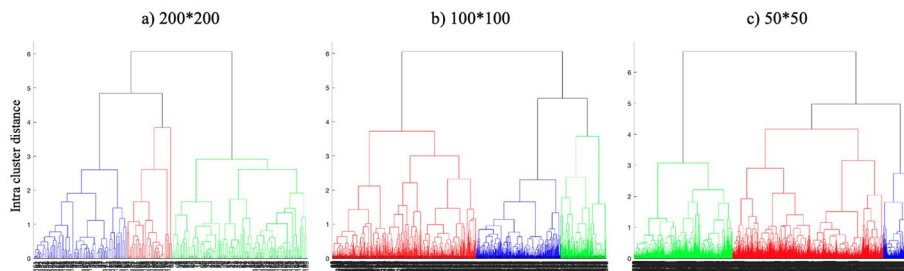
Fig. 4 Dendograms obtained using the Euclidean distance and the weighted method, varying different grid sizes: **a** 200x200, **b** 100 × 100, and **c** 50 × 50

Minkowski, Squared Euclidean, and Euclidean distance, according to different principles respectively. For N initial clusters (N = 150, 600, or 2400), the distance matrix d has a size of $1 \times M$, $M = N \times (N-1)/2$.

Linkage Then, the distance information generated was used to determine the similarity and proximity of every spectrum to each other. Pairs of clusters are merged into a new cluster at each step according to different agglomeration criteria, and the distance data between those linked leaves is recorded as the intra-cluster distance of the newly formed large cluster. In the ith step, the new cluster created is indexed as $M + 1$, while the final total nodes in the tree are $M-1$. The newly formed cluster would be regarded as one large cluster in the next step, and the intra-cluster distance was repeatedly calculated based on the current level of clustering. This linkage process was repeated until there was just one cluster remaining in the dataset that contained all of the initial clusters. Four different linkage functions were tested: Average, Complete, Single, and Weighted. They calculated respectively the unweighted average distance, farthest distance, shortest distance, and weighted average distance.

Hierarchical tree The grouping index and intra-cluster distance information were recorded during each step in the process. As a result, a hierarchical cluster tree is formed, with clusters at one level joining to generate larger clusters at the following levels. The dendrogram figures that plot the outcomes of the hierarchical tree, using the Euclidean metric and Weighted linkage function as parameters, are provided in Fig. 4 to better explain the procedure. The tree-shaped graphics show clearly how the initial clusters are developed in a multilevel hierarchy. The X-axis indicates the original cluster index, and the Y-axis represents the dissimilarity of clusters at each node. The tree consists of many nodes that document every cluster merge, from initial data points (the centroids of the sub-images) until all the points are linked in only one cluster. The height

Liu *et al. Journal of Big Data*   (2023) 10:31

Page 8 of 16

of each node is proportional to the intra-cluster distance of the two leaves that are connected. The hierarchical tree can assist in selecting the clustering level or scale that is most suitable for subsequent application.

In each grid size group, the intra-cluster distances of the hierarchical tree are generally at the same level, as shown in Fig. 4. Although the linking process varies, the formation of the trees is in similar structures: all data points are grouped in three main branches (coloured in red, green, and blue), and the proportion of data points covered in separated branches is similar. This division might correspond to the regions with different degradation features or areas with different dye intensities. The main rises of dissimilarity are contained in the last dozens of nodes that will be examined in detail in the following sections.

Clusters Finally, by specifying a threshold, the clustering results at a certain level could be retrieved from the tree. The entire dataset is classified into two clusters at the level of node 1, three clusters at the level of node 2, and so forth. Each cluster is a collection of multiple initial clusters; in other words, every initial cluster is assigned to a specific cluster group. By looking at the index of the initial clusters that indicate their original location on the dataset, the images could be segmented according to this labeling information. In this research, the level of 3, 4, and 5 clusters were selected to demonstrate the segmentation results, balancing both the accuracy and computational load.

The flowchart (Fig. 3) summarizes all the parameters that have been examined. For each step, tests of the parameters are repeated under every category of the previous results. There are a total of 180 possible parameter combinations: 3 grids × 5 metrics × 4 linkage functions × 3 cluster numbers. One example of a possible combination is highlighted with the chosen parameters marked in black. To better organize the paper, only selected significant results are demonstrated in the following results and discussions section.

Results and discussions

The evaluation of the results is performed in three main subsections: (i) Grid sizes, (ii) Metrics and (iii) Linkage methods. The results obtained with exhaustive combinations of algorithmic parameters are fully investigated, while only selected significant results are presented and compared in this paper due to the high number of images generated. Some parameter combinations with the best performance are suggested at the end of this section, and limitations and future works are discussed.

Grid size

The effect of different grid sizes was first evaluated. In this experiment, the same parameters employed in the preliminary work [31] are used. They consists of the Euclidean distance as the metric and the Weighted as the agglomeration method. They were applied to all three grid groups.

The results of the hierarchical clustering are reported in Fig. 5. For each grid size, the final three, four, and five clusters are depicted for sample 22. As shown, grid sizes of 200 × 200 and 100 × 100 obtained comparable results, while the 50 × 50 grid produced significantly finer segmentation than the other two. With the total number of clusters equal to three and four, the results for the grid sizes 200 × 200 and 100 × 100 are roughly
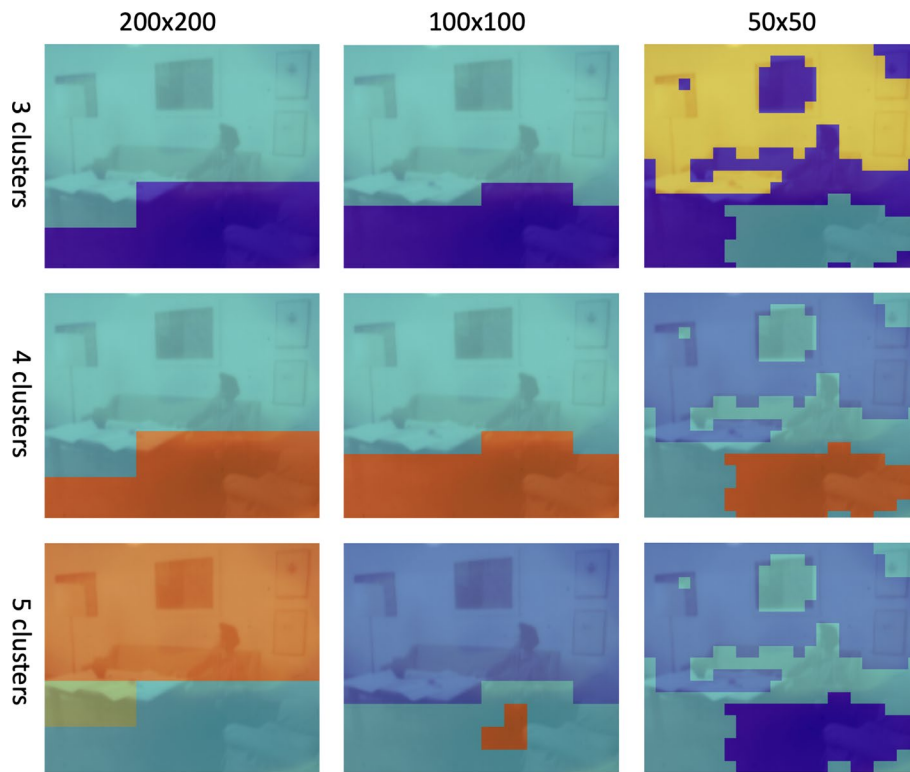
Fig. 5 Comparison of segmentation results obtained with different grid sizes, employing the Euclidean distance and the weighted agglomeration method

divided into two sections, the upper brighter part, and the lower darker part, without capturing more details. As a comparison, the 50×50 grid segmented the image into three sections with all cluster numbers and enabled the correct separation of the wall painting, the less deteriorated upper-right corner, and the stained area in the lower part of the image. By choosing five total clusters, the segmentation results in the 200×200 and 100×100 groups are slightly enhanced. However, they are still not comparable to the 50×50 group. The same tendency was observed in all the results. Hence, in the following experiments, the results obtained using the 50×50 grid size are reported.
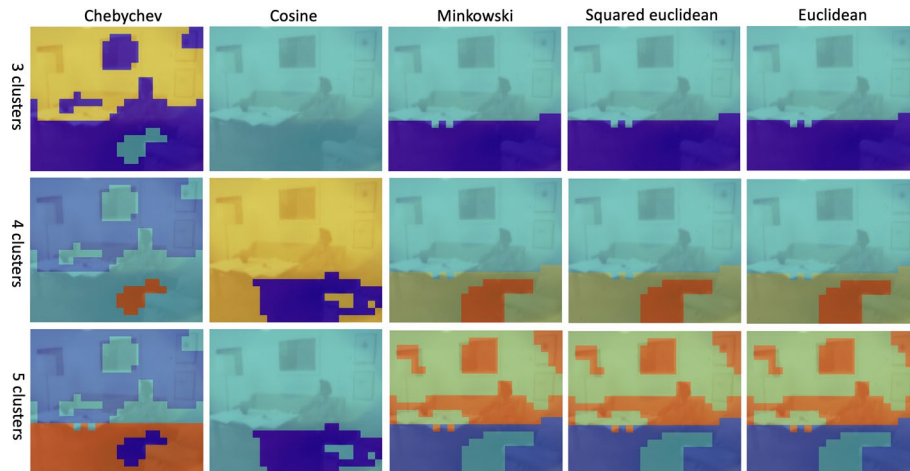
Distance metrics

Observing all the results, the weighted and complete agglomeration methods were found to be the most efficient among all linkage functions. To have a closer investigation of the metrics function, a comparison of all the metrics is presented, considering the 50×50 grid size, and evaluating both the weighted and complete agglomeration method. The metrics tested in this research are Chebyshev, Cosine, Minkowski, Squared Euclidean, and Euclidean distance. The results are presented for three, four, and five clusters.

The results using the weighted method are reported in Fig. 6. As shown, the Cosine metric has the worst performance, obtaining only a rough separation of the lighter and darker parts. Minkowski, Squared Euclidean, and Euclidean metrics obtained comparable results, segmenting images into three parts with details highlighted. However, the segmentation is not improved with a larger number of clusters. The Chebyshev metric
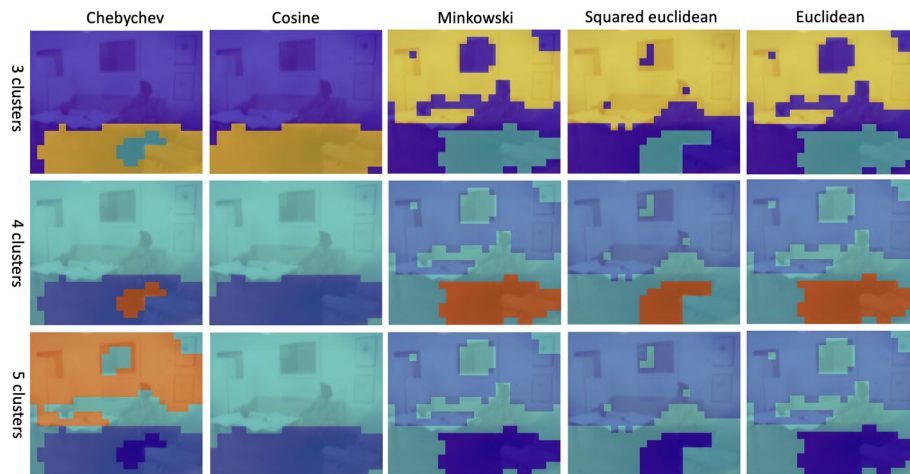
Fig. 6 Segmentation results with different distance metrics, using a grid size of 50 × 50 and the weighted agglomeration method

Fig. 7 Segmentation results with different distance metrics, using a grid size of 50 × 50 and the complete agglomeration method

has the best performance among all. For three and four total clusters, the images are segmented into three parts, detecting the boundary of the stain. With five total clusters, the lighter part is further separated into the extensively degraded area and the less deteriorated middle tone. The wall painting and the upper-right corner are successfully segmented, obtaining in total four major clusters on the image. In this case, it could be concluded that the Chebyshev combined with weighted method has the most preferable results.

Then, also the complete method is evaluated, varying the various distance metrics. As shown in Fig. 7, the Chebyshev metric also has outstanding performance with the maximum segmentation of four clusters on the image, highlighting the wall painting, corner area, and staining area. The Minkowski, Squared Euclidean, and Euclidean metrics also obtained better results as compared to the weighted method. The segmentation is gradually improved with increasing cluster numbers. For three clusters,

the images are only separated into two parts. And as the clusters increase, the stain areas, wall painting, and corner areas are successfully segmented, obtaining four major parts as well. In this case, all results obtained with the complete method are better than using the weighted method, while the Chebyshev, Minkowski, Squared Euclidean, and Euclidean metrics are comparable in the segmentation efficiency.

Linkage

Given the results presented in the above section, the Chebyshev and Euclidean metrics are selected to present a closer analysis of all the agglomeration functions for the hierarchical clustering. Figures 8 and 9 report results obtained with four linkage functions, (i.e., Average, Complete, Single, and Weighted) using respectively the Chebyshev and Euclidean metrics. In both cases, the Single agglomeration function is the least effective among all, failing to segment the image and classifying the entire image into one cluster.

  Using the Chebyshev distance, the Complete and Weighted methods outperformed other methods, successfully segmenting the image into four parts, with fine defining of the dark stain, most deteriorated light part, and the corner and middle tone. The Average function, in this case, has failed to separate the dark staining area in the lower part.

  With regard to the Euclidean metric, the Weighted and Complete functions have succeeded in separating all important features with four major clusters defined. Anyway, the best results are achieved by the Complete method. In fact, it is the only one that has segmented the lamp and the staining area, that neither the Weighted and Average methods were able to detected. The Weighted method in this case is not so satisfactory that obtained three rough parts with 3 clusters and is not improved as the cluster numbers increase.
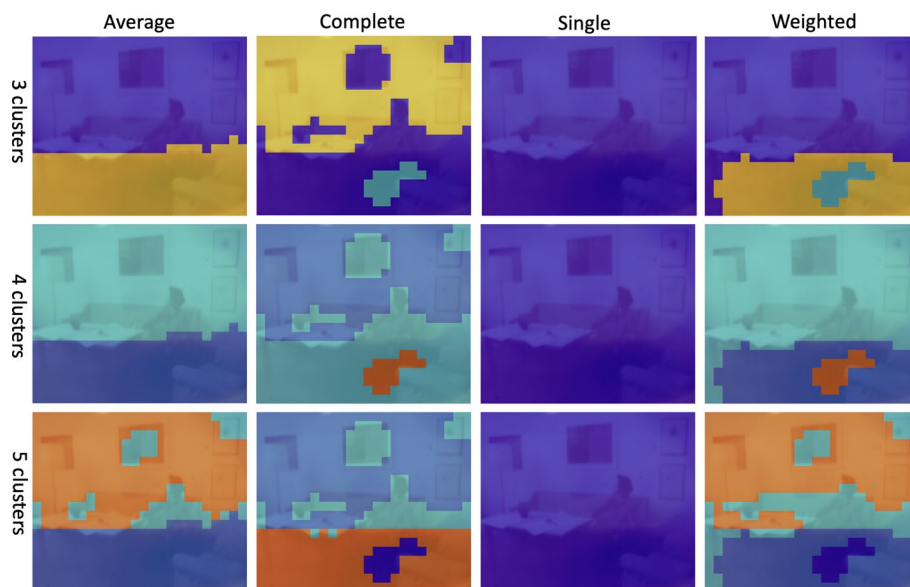
Fig. 8 Segmentation results varying the agglomeration method, using a grid size of 50 × 50 and the Chebyshev distance

Liu *et al. Journal of Big Data*   (2023) 10:31

Page 12 of 16
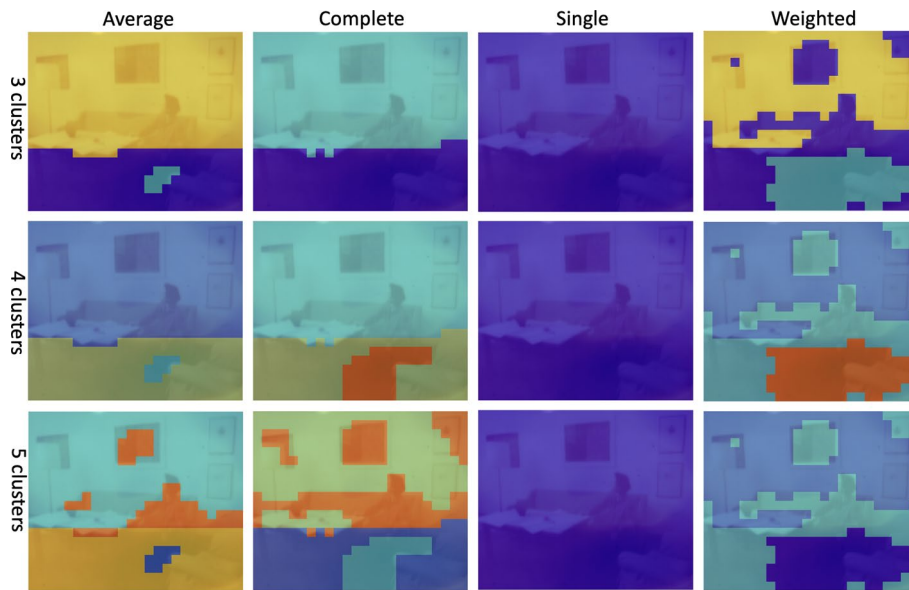
Fig. 9 Segmentation results varying the agglomeration method, using a grid size of 50 × 50 and the Euclidean distance
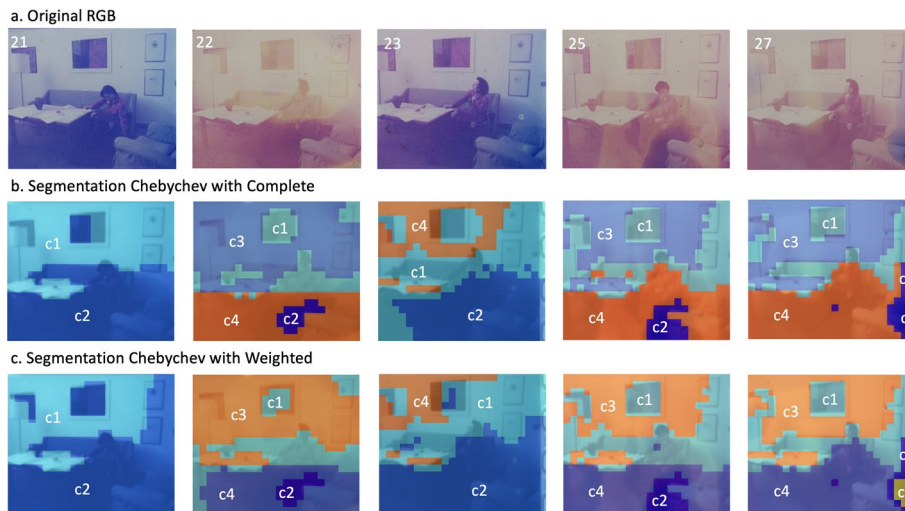
Fig. 10 Segmentation results on the full dataset (images 21, 22, 23, 25, and 27): **a** the original RGB images. **b** results obtained using the Chebyshev metric and the complete method. **c** results obtained using the Chebyshev metric and the weighted method

Considering the results obtained in all the experiments, the best results were obtained combining the Chebyshev metric with the Complete and the Weighted methods and are suggested for performing a similar task of segmenting hyperspectral images.

Segmentation results

In this Section, the segmentation results on the entire data set, including frames 21, 22, 23, 25, and 27, are reported. The results were obtained by selecting five clusters in total, for the algorithmic parameter combinations of Chebyshev metrics with complete

linkage function (Fig. 10b), and Chebyshev metrics with weighted method (Fig. 10c). As a comparison, the original RGB images of the frames are also reported in Fig. 10a. With both combinations, frames with similar deterioration effects have similar segmentation clusters. The reference frame 21 is segmented into two clusters, the bright top region with high reflectance (c1) and the dark ground portion (c2). Frame 23, which had only minor degradation, obtained a similar clustering structure to frame 21, while the more degraded pinkish upper left area is classified into a separate cluster c4. Frames 22, 25, and 27 have the most extensive, yet comparable, degradation. They have a similar clustering structure yet are different from the reference. They are all segmented into four fundamental clusters: (1) c3, the regions that have been most severely degraded and have lost the majority of the dyes; (2) c1, middle tone sections, including the wall painting and peripheral areas, where colours are still remained and are comparable to those in frame 21; (3) c2, which is the small portion on the ground part that still preserves a large amount of dyes; (4) c4, the rest of the dark ground area which is severely deteriorated. Frame 27 has one more separate cluster (c5), that contains the blue strips on the bottom-right side of the image and that does not appear in other cases. Throughout the entire data set, it could be observed that areas with similar degradation features on diverse frames are classified in the same cluster.

On the other hand, within the same frame, the inhomogenous degradation areas are successfully segmented into different clusters. For example, on frames 22, 25, and 27, the most degraded regions (c3), middle tone areas (c1), the degraded darker parts (c4), and the special features (c2 and c5) are segmented with very fine contours, successfully separating the wall painting hanging on the wall, the lamp on the left side, and the right upper corner with less degraded shades. The two combinations of parameters have both obtained satisfactory results with slight differences. Overall, comparing qualitatively the segmentation results, it can be conclude that the two proposed combinations of algorithmic parameters is effective for the segmentation task performed on the full data set, with the 50*50 grid size.

Discussions and limitations

As anticipated in the Introduction section, the objective of the overall project is to digitally restore the colours of the film, performing 1D-SSA. In order to apply it, there is the need for the effective segmentation of various degradation features in the films. In this paper, we investigated the use of clustering algorithms to segment the HSI spectra acquired on films. From a qualitative point of view, HCA was able to separate degradation features among the different frames of film. Unfortunately, the lack of ground-truth values does not allow us to conduct also quantitative analysis. Since in the literature no other works are focusing on HCA applied to hyperspectral images, we considered the several parameters of the Hierarchical Clustering (i.e., the distance metric and the linkage approach) as hyperparameters and we conducted several tests to get the best combination of them. Considering the final results, the distance metric that performed worse was the cosine distance, probably due to the nature of the data. In fact, it is usually employed for text similarities [33] or for recommendation systems [34]. The other distance metrics, being all of them special cases of the same function, varying just one parameter (i.e., the value of p that was 1 in our case for the Minkowski, 2

for the Euclidean, and tends to infinity for the Chebychev one), perform pretty similar. Concerning the linkage method, instead, the poor performance of the Single method, which uses the smallest distance between objects in two clusters, was expected. Such an agglomeration method has the tendency of producing long and thin clusters in which elements at opposite ends of a cluster can be really distant from each other, favouring the separation of outliers first. For this reason, it is commonly used for the clustering of galaxies in astronomy [35]. For the opposite reason, better performance was expected for the Complete method, as it works in the opposite principle compared to the single method, generating highly compact clusters. Since in the average method the distance among clusters is computed as the average among the distances between all the elements of the two clusters, this allows a segmentation based on major variances however less sensitive to small variations as compared to Complete method. Finally, the weighted method has good performances, similar to the ones of the Complete.

Though through this first stage demonstration of applying HCA to film, the main degradation features are separated, the precision and accuracy of the segmentation should still be improved. This is mainly limited by the division of sub-images, which applies fixed grids. On the one hand, the noise of every single spectrum inside the grid is smoothed and the computational load is largely decreased. On the other hand, the unique spectral features are flattened and the adjacent pixels of different degradation degrees or image content are all represented by the centroid spectrum. Even if we evaluated several grid sizes, it is clear that such an approach is a major limitation of our approach. Given that the restoration requires highly accurate treatment, especially in CH, the spatial relationship of the pixels must be considered in addition to the spectral dimension. Future works are then expected to combine the HCA with the superpixel technique [36]. The sub-images could be formed by SLIC (Simple Linear Iterative Clustering) algorithm [37] that generates a number of compact, nearly uniform superpixels, maintaining spatial relationships and color features. Then it can be fused with the HSI data and an HCA segmentation with higher precision can be performed on the superpixels obtained. Finally, the colour correction is expected based on the segmentation.

Conclusions

The hierarchical clustering-based approach proposed in this study offers new possibilities in the spectral image segmentation of films. Applied to the hyperspectral dataset acquired on historical film frames, the unsupervised machine learning algorithm has successfully differentiated and segmented the fading areas considering the full range of degradation features. A simple data pre-processing strategy is tested and the smaller grid size is proven to be more promising. Various algorithmic parameters and their combinations were evaluated, including the grid sizes, metrics, and linkage functions. The Chebyshev metrics combined with the Complete or Weighted linkage function were tested to have the best performance. The leveled clustering structure documented in the hierarchical tree allows an efficient and more flexible classification of complex data with unknown chemical nature or mixture. With the segmented cluster groups, the digital re-colourisation could be applied separately to each group of clusters. From future perspectives, the HCA-based approach has the potential to be applied to spectral data obtained on other art objects, such as paintings, to fulfill tasks like pigment mapping.

Liu *et al. Journal of Big Data*   (2023) 10:31

Page 15 of 16

Abbreviations

HCA   Hierarchical clustering algorithm
HSI   Hyperspectral imaging
CH   Cultural heritage
PCA   Principal component analysis
MNF   Minimum noise fraction transform
ML   Machine learning
SSA   Singular spectrum analysis

Author contributions

LL performed the data processing and interpretation, provided the graphs, and drafted the manuscript. GD and SM conceptualized the work, supervised the project, and revised the final manuscript. All authors read and approved the final manuscript.

Availability of data and materials

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable (no humans or animal involved).

Consent for publication

Not applicable.

Competing Interests

The authors declare that they have no competing interests.

References

1. Liang H. Advances in multispectral and hyperspectral imaging for archaeology and art conservation. Appl Phys A. 2012;106(2):309–23.
2. Dooley KA, Lomax S, Zeibel JG, Miliani C, Ricciardi P, Hoenigswald A, Loew M, Delaney JK. Mapping of egg yolk and animal skin glue paint binders in Early Renaissance paintings using near infrared reflectance imaging spectroscopy. Analyst. 2013;11:111.
3. Fischer C, Kakoulli I. Multispectral and hyperspectral imaging technologies in conservation: current research and potential applications. Stud Conserv. 2006;51:3–16.
4. Alfeld M, de Viguerie L. Recent developments in spectroscopic imaging techniques for historical paintings - A review. Spectrochimica Acta Part Atomic Spect. 2017;136:81–105.
5. Daveri A, Paziani S, Marmion M, Harju H, Vidman A, Azzarelli M, Vagnini M. New perspectives in the non-invasive, in situ identification of painting materials: The advanced MWIR hyperspectral imaging. TrAC Trends Analy Chem. 2018;98:143–8.
6. Catelli E, Randeberg LL, Alsberg BK, Gebremariam KF, Bracci S. An explorative chemometric approach applied to hyperspectral images for the study of illuminated manuscripts. Spectrochimica Acta Part Mol Biomol Spect. 2017;177:69–78.
7. Biron C, Mounier A, Le Bourdon G, Servant L, Chapoulie R, Daniel F. Revealing the colours of ukiyo-e prints by short wave infrared range hyperspectral imaging (SWIR). Microchem J. 2020;155:111.
8. Daniel F, Mounier A, Pérez-Arantegui J, Pardos C, Prieto-Taboada N, Fdez-Ortiz de Vallejuelo S, Castro K. Hyperspectral imaging applied to the analysis of Goya paintings in the Museum of Zaragoza (Spain). Microchem J. 2016;126:113–20.
9. Wilhelm HG, Brower C. The Permanence and Care of Color Photographs: Traditional and Digital Color Prints, Color Negatives, Slides, and Motion Pictures, 1st. ed. Grinnell: Preservation Pub; 1993.
10. Picollo M, Cucci C, Casini A, Stefani L. Hyper-spectral imaging technique in the cultural heritage field: New possible scenarios. Sensors. 2020;20(10):2843. https://doi.org/10.3390/s20102843.

Liu *et al. Journal of Big Data*   (2023) 10:31

Page 16 of 16

11. Nunes S, Ramacciotti F, Neves A, Angelin EM, Ramos AM, Roldão É, Wallaszkovits N, Armijo AA, Melo MJ. A diagnostic tool for assessing the conservation condition of cellulose nitrate and acetate in heritage collections: quantifying the degree of substitution by infrared spectroscopy. Herit Sci. 2020;8:1–14.

12. Lozano MVC, Catelli E, Sciutto G, Prati S, Genorini E, Mazzeo R. A non-invasive diagnostic tool for cellulose acetate films using a portable miniaturized near infrared spectrometer. Talanta. 2023;255: 124223.

13. Trumpy G, Flueckiger B, Dye purification: an image-processing technique for the digital restoration of chromogenic film. In,. Colour and Visual Computing Symposium (CVCS). IEEE. 2018;2018(1):5.

14. Liu L, Catelli E, Katsaggelos A, Sciutto G, Mazzeo R, Milanic M, Stergar J, Prati S, Walton M. Digital restoration of colour cinematic films using imaging spectroscopy and machine learning. Sci Rep. 2022;12(1):21982.

15. Fiorucci M, Khoroshiltseva M, Pontil M, Traviglia A, Del Bue A, James S. Machine Learning for Cultural Heritage: A Survey. Pattern Rec Lett. 2020;133:102–8.

16. Capobianco G, Bracciale MP, Sali D, Sbardella F, Belloni P, Bonifazi G, Serranti S, Santarelli ML, Cestelli Guidi M. Chemometrics approach to FT-IR hyperspectral imaging analysis of degradation products in artwork cross-section. Microchem J. 2017;132:69–76.

17. Pan N, Hou M, Lv S, Hu Y, Zhao X, Ma Q, Li S, Shaker A. Extracting faded mural patterns based on the combination of spatial-spectral feature of hyperspectral image. J Cult Herit. 2017;27:80–7.

18. Hou M, Zhou P, Lv S, Hu Y, Zhao X, Wu W, He H, Li S, Tan L. Virtual restoration of stains on ancient paintings with maximum noise fraction transformation based on the hyperspectral imaging. J Cult Herit. 2018;34:136–44.

19. Sun M, Zhang D, Wang Z, Ren J, Chai B, Sun J. What's wrong with the murals at the mogao grottoes: a near-infrared hyperspectral imaging method. Sci Rep. 2015;5(1):14371.

20. Lin Y, Xu C, Lyu S. Disease regions recognition on mural hyperspectral images combined by mnf and bp neural network. J Phys Conf Ser. 2019;1325: 012095.

21. Cornelis B, Ružić T, Gezels E, Dooms A, Pižurica A, Platiša L, Cornelis J, Martens M, De Mey M, Daubechies I. Crack detection and inpainting for virtual restoration of paintings: The case of the Ghent Altarpiece. Signal Proc. 2013;93(3):605–19.

22. Sanghoon Lee, Crawford MM. Hierarchical clustering approach for unsupervised image classification of hyperspectral data. In: IEEE International IEEE International IEEE International Geoscience and Remote Sensing Symposium, 2004. IGARSS '04. Proceedings. 2004;2. pp. 941–944. https://doi.org/10.1109/IGARSS.2004.1368563.

23. Nunez-Iglesias J, Kennedy R, Parag T, Shi J, Chklovskii DB. Machine Learning of Hierarchical Clustering to Segment 2D and 3D Images. PLOS ONE. 2013;8(8):71715.

24. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview wiley interdisciplinary reviews. Data Mining Knowledge Discovery. 2012;2(1):86–97.

25. Kumar N, Uppala P, Duddu K, Sreedhar H, Varma V, Guzman G, Walsh M, Sethi A. Hyperspectral Tissue Image Segmentation Using Semi-Supervised NMF and Hierarchical Clustering. IEEE Transact Med Imag. 2019;38(5):1304–13.

26. Yu P. Applications of Hierarchical Cluster Analysis (CLA) and Principal Component Analysis (PCA) in Feed Structure and Feed Molecular Chemistry Research, Using Synchrotron-Based Fourier Transform Infrared (FTIR) Microspectroscopy. J Agric Food Chem. 2005;53(18):7115–27.

27. Gilbert N, Mewis RE, Sutcliffe OB. Classification of fentanyl analogues through principal component analysis (PCA) and hierarchical clustering of GC-MS data. Forensic Chem. 2020;21: 100287.

28. Hirsch E, Agassi E. Detection of gaseous plumes in hyperspectral images using hierarchical clustering. Applied optics. 2007;46(25):6368–74.

29. Wilson N, MacRae C. An automated hybrid clustering technique applied to spectral data sets. Micro Microanaly. 2005;11(S02):434–5.

30. Disser A, Dillmann P, Leroy M, l'Héritier M, Bauvais S, Fluzin P. Iron supply for the building of metz cathedral: new methodological development for provenance studies and historical considerations. Archaeometry. 2017;59(3):493–510.

31. Liu L, Delnevo G, Mirri S. Hierarchical clustering as an unsurpervised machine learning algorithm for hyperspectral image segmentation of films In on Information Technology for Social Good. GoodIT. 2022;22:397–402.

32. Dolenec R, Rogelj L, Stergar J, Milanic M. Modular multi-wavelength led based light source for hyperspectral imaging. Eur Conf Biomed Optics. 2019;11075:56.

33. Lahitani AR, Permanasari AE, Setiawan NA. Cosine similarity to determine similarity measure: Study case in online essay assessment. In: 2016 4th International Conference on Cyber and IT Service Management. 2016,1–6 IEEE. https://doi.org/10.1109/CITSM.2016.7577578.

34. Mansur F, Patel V, Patel M. A review on recommender systems 2017 International on Innovations in Information. Embed Commun Sys (ICIIECS). 2017;1:6.

35. Feigelson E. Classification in astronomy: Past and present. Adv Machine Lear Data Mining Astronomy. 2012;1:3–10.

36. Fan Y-R, Huang T-Z. Hyperspectral image restoration via superpixel segmentation of smooth band. Neurocomputing. 2021;455:340–52.

37. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans Pattern Analy Machine Intelligence. 2012;34(11):2274–82.

Publisher's Note