UNSUPERVISED CLUSTERING FOR WORKS OF ART USING HYPERSPECTRAL IMAGING: A CASE STUDY ON EDVARD MUNCH'S *SELF-PORTRAIT (1905)*

Dipendra J. Mandal, Hilda Deborah, Sony George, and Jon Y. Hardeberg
Department of Computer Science, Norwegian University of Science and Technology (NTNU), Norway

ABSTRACT

The study of pigments in historical works of art is of significant value for conservators and art historians, providing insight into artistic techniques and the preservation of cultural heritage. Hyperspectral imaging aids in the identification and classification of pigments, facilitating conservation efforts. However, the challenge lies in identifying these pigments in artworks where ground-truth data are unavailable, necessitating unsupervised clustering techniques. In this paper, the performance of dimensionality reduction and cluster estimator techniques are evaluated, further proposing a workflow for unsupervised clustering for paintings without known pigments. A case study is conducted on Edvard Munch's Self-Portrait (1905), providing valuable insights into a relatively unexplored artwork within the cultural heritage domain.

Index Terms— Hyperspectral imaging, cultural heritage, painting, unsupervised clustering

1. INTRODUCTION

Knowledge about pigments used in historical artworks helps conservators and art historians gain valuable insights into the materials and techniques employed by artists from different periods and regions. When dealing with an artwork for which we have no prior knowledge of its details, one of the crucial steps is determining the distinct number of pigments or endmembers present in the artwork. Using multi- or hyperspectral imaging (HSI) techniques, one can analyze the spectral signatures of different areas of the artwork. This helps to identify and classify the pigments used, which is vital information for the preservation and restoration of the artwork within the realm of cultural heritage (CH).

HSI, also known as imaging spectroscopy, provides detailed, non-invasive, and material-specific information about artworks and artifacts. Its ability to discriminate between materials, analyze subsurface features, and improve visualization sets it apart from traditional imaging methods and contributes significantly to preserving and understanding our CH. It has been widely used as a complementary tool for conservation studies, especially for paintings [1]. One of its most popular uses is to map pigments or colorants over the spatial extent of a painting [2], allowing for extrapolation of the knowledge obtained by means of, usually, point-based analytical methods. In a typical scenario, an identification of which pigments are available in a painting would have already been performed. For an HSI-based pigment mapping, this means that a spectral library of known pigment can be constructed, and any supervised approach can be used to perform the mapping. However, in most cases in CH, we do not have such information and, therefore, we need to analyze the data using unsupervised techniques.

This paper proposes a workflow to estimate and map the clusters in an artwork without available ground truth information. We explore three dimensionality reduction techniques, i.e., principal component analysis (PCA) [3], t-distributed stochastic neighbor embedding (t-SNE) [4], and uniform manifold approximation and projection (UMAP) [5]. They are further evaluated by several k-estimators, with a particular focus on spectral datasets associated with pigments commonly used in paintings from the historical periods. The objective is to identify the most suitable dimensionality reduction technique and K-estimator for clustering analysis on a pre-established dataset with a known number of clusters (pigment mockup) and apply this optimized workflow to the case study painting, i.e., Edvard Munch's Self-Portrait (1905). The case study will provide valuable information since, to the best of our knowledge, there are not many details available for this painting.

2. ON UNSUPERVISED CLUSTERING

Clustering is a statistical data analysis technique, mostly used as an unsupervised machine learning task that organizes entities into groups based on their shared features [6]. Clustering algorithms aim to maintain the distribution characteristics of the input data by grouping similar entities within the same cluster. When applied to hyperspectral images, the data points to be clustered usually represent the individual pixels in an image, with the features corresponding to the pixel values across different spectral bands. A proficient clustering algorithm should effectively group pixels with comparable spectral signatures, essentially identifying pixels that likely correspond to the same pigments.

Hierarchical and partitional clustering are two fundamental approaches to clustering data [7]. Hierarchical clustering does not require specifying the number of clusters beforehand, and it organizes data points into a tree-like structure (dendrogram) by recursively merging or splitting clusters based on similarity or distance. This can be further categorized into Agglomerative clustering (bottom-up approach) and Divisive clustering (top-down approach). Hierarchical clustering is computationally expensive and, thus, not suitable for large datasets. In addition, it does not separate overlapping clusters [6]. On the other hand, partial clustering is an approach that divides data points into non-overlapping groups or partitions based on the initially specified number of clusters. Each data point belongs to one and only one cluster (disjoint clusters). There are also some other clustering algorithms based on a variety of theories and techniques that can be used as hierarchical and partitional clustering; some of these include graph theory-based [8], fuzzy theory-based [9], mixture densities-based [10], neural network-based [11], kernel-based clustering [12], etc. Details about various clustering algorithms can be found in [7, 12].

The choice of clustering method and strategy usually depends on the specific dataset, computational resources, and the trade-offs between accuracy and computational complexity. Hierarchical clustering is computationally complex, especially for large datasets, because it operates pairwise to build the hierarchical structure of clusters. On the other hand, partitional clustering methods are often computationally efficient; however, they often require prior knowledge or an estimate of the cluster count. The key challenge lies in selecting the optimal number of clusters k, a critical decision that significantly influences cluster quality. Several techniques (k-estimators), such as the elbow method [13], the silhouette score [14], Davies-Bouldin index [15], and gap statistics [16], can help determine the ideal cluster count.

Elbow method determines the optimal k by iteratively calculating the within-group sum of squares (WCSS) for different values of k and selecting the point where the WCSS graph starts to level off as adding the additional clusters does not add sufficient information. The **silhouette score** summarizes the variation within-cluster and between-clusters. It is calculated by measuring the distances between data points within the same cluster and to the nearest neighboring cluster. The optimal k is determined by finding the value of the peak silhouette coefficient. The **Davies-Bouldin index** (DBI) is another metric to measure the separation and compactness of clusters. It computes the ratio of within-cluster distance to between-cluster distance for all clusters and their nearest neighbors, recording the maximum ratio for each cluster. The final index value is the average of these maximum ratios. The lower the DBI, the better the clustering quality. The **gap statistic** is used to compare the quality to a reference distribution. Calculate the difference between the performance of the clustering algorithm on actual data and its performance on random data. A larger gap suggests better clustering quality, indicating a suitable number of clusters. It can be calculated with and without taking the logarithm of the observed WCSS values. Using logarithms is helpful for datasets with widely varying WCSS values.

High-dimensional data, e.g., hyperspectral data, contain numerous features that represent variables or attributes. While offering valuable information, high dimensionality increases computational complexity, requiring more memory and resources. Dimensionality reduction techniques, by mainly linear projection (e.g., PCA) and nonlinear embedding (e.g., t-SNE, UMAP), aim to preserve essential information while reducing dimensionality. **PCA** transforms data into a lower-dimensional space and focuses on preserving the global structure and variance [3]. **t-SNE** is a nonlinear approach that can capture complex, nonlinear relationships between data points, making it useful for preserving local structures in the data. It is often used for clustering and visualization tasks when local relationships are more important than global structure [4]. **UMAP** is another non-linear technique designed to preserve local and global structures in data [5]. It employs a nearest-neighbor approach with efficient algorithms, making it better suited for handling large datasets. On the contrary, t-SNE quickly becomes computationally expensive as the size of the dataset increases.

3. MATERIALS AND METHODS

3.1. Object

Figure 1 depicts *"Self-Portrait with Moustache and Starched Collar"* (*Selvportrett med bart og høy snipp*), a captivating work by the renowned Norwegian artist Edvard Munch, created in 1905. This is a relatively small painting, measur-

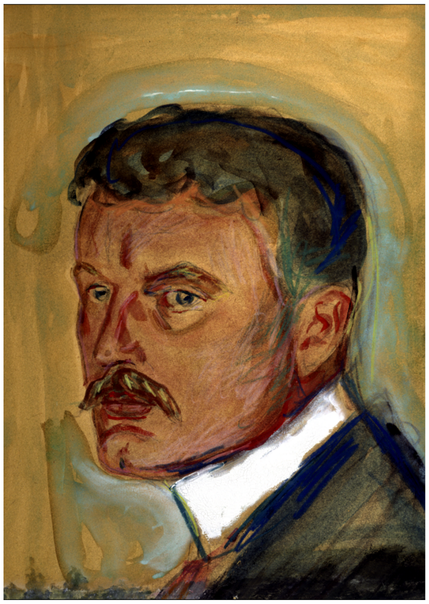

Fig. 1: *Self-Portrait with Mustache and Starched Collar (1905)* by Edvard Munch, National Museum, Norway. The color image was generated from the HSI data using bands at 614.52, 563.63, and 458.24 nm.

ing approximately 46.5×35.5 centimeters (18.3×14 inches). This modest size invites viewers to closely examine the details of Munch's self-representation and the nuances of his expression. In the context of this painting, the specific pigments employed, or the number of distinct end members remain unidentified. However, several studies delved into the details of Edvard Munch's artworks [17], providing valuable insights into the artistic materials employed by Munch, spanning the period from 1885 to 1927.

3.2. Image acquisition

Hyperspectral images were acquired at the National Museum, Oslo, Norway, where the painting was located. HySpex VNIR-1600, a line scanner camera developed by Norsk Elektro Optikk, was used. The painting was placed on an easel and the camera was on an X-Y translation stage. We used a 1-meter cylindrical lens for the acquisition that captures 1600 spatial pixels across a line with a field of view of approximately 30 cm. A quartz-enveloped halogen-tungsten broadband light source covers the broad spectrum of 400-2500 nm. A computer equipped with HySpex GROUND software provided by the manufacturing company controlled the hyperspectral acquisition system. This software automatically synchronizes the scanning speed for the user-defined integration time. Reference targets where also captured along with the painting to use for reflectance calculations.

3.3. Radiance to reflectance factor

Hyperspectral image obtained from the acquisition comes in terms of spectral radiance. To perform analysis on materials or surfaces, spectral reflectance is needed since it is the inherent property of materials and is independent of the illumination condition. Given the use of a calibration target G and its known reflectance R_1^G, we can estimate the illumination factor I from its radiance image R_0^G by:

$$I = \frac{R_0^G}{R_1^G}. \tag{1}$$

The reflectance factor R_1^T of the target radiance image R_0^T will therefore be computed as:

$$R_1^T = \frac{R_0^T}{I}. \tag{2}$$

Specifically for the case study image we use, the spectral reflectance image is derived by assuming that the whitest point in the image is supposed to have a relatively flat reflectance signal close to unity $G = \{1, \forall\lambda\}$, where λ is the wavelengths.

4. RESULTS AND DISCUSSION

4.1. Validation experiment results

We conducted dimensionality reduction on datasets with varying known numbers of clusters within the datasets. Different k-estimation methods were also applied to these lower-dimensional data to determine the optimal value of k. The results obtained were subsequently compared to the actual number of groups in the data set (Figure 2 illustrates the workflow). The summary of the results is presented in Table 1. Here, n represents the number of actual clusters in the dataset. By increasing n, we introduce greater complexity to the dataset. As illustrated in Figure 3, P1 to P11 represent distinct clusters. As more spectra overlap or become similar to others with slight differences in magnitude and shape, they can be considered representations of overlapping clusters.

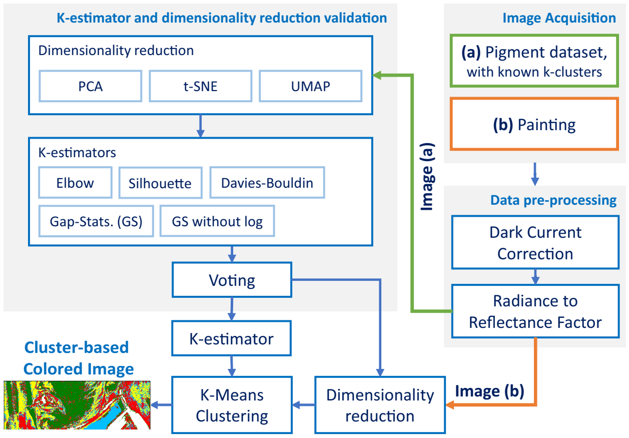

Fig. 2: Workflow for unsupervised clustering for works of art.

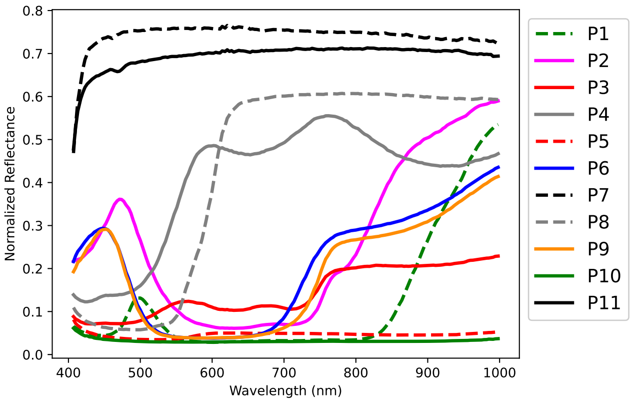

Fig. 3: Spectra of various pigments used within the known dataset. They represent the most commonly used pigments within the works of art across various historical periods [18]

.

Table 1: Summary of results obtained from three different data reduction techniques and a range of k estimators to ascertain optimal cluster values for datasets with varying known cluster counts n.

	Elbow Method			Silhouette Score			Davies-Bouldin Index			Gap Statistics (GS)			GS without Log		
n	PCA	t-SNE	UMAP	PCA	t-SNE	UMAP	PCA	t-SNE	UMAP	PCA	t-SNE	UMAP	PCA	t-SNE	UMAP
2	**2**	4	3	**2**	13	**2**	**2**	10	**2**	11	15	14	**2**	**2**	3
4	**4**	6	**4**	2	13	**4**	2	15	**4**	15	15	14	**4**	2	**4**
6	5	**6**	**6**	**6**	15	**6**	**6**	14	**6**	14	14	12	**6**	2	**6**
7	5	6	**7**	**7**	14	**7**	2	12	8	15	15	13	4	2	**7**
9	4	6	8	7	15	8	2	13	8	11	14	15	5	2	12
11	4	6	8	8	15	10	8	14	10	14	15	15	4	2	14

The results presented in Table 1 indicate that, paired with the Elbow and silhouette estimators, UMAP excels in estimating k. This success can be attributed to the ability of UMAP to capture local and global data structures, making it a valuable tool in scenarios where accurate determination of the cluster number is crucial. Although PCA performs adequately for a lower number of clusters, its performance deteriorates as the complexity of the clustering structure increases. This limitation arises from PCA's linear nature, which may struggle to accurately represent intricate non-linear relationships within the data. It should be noted that t-SNE consistently struggles to effectively estimate the number of clusters using various estimator methods for the dataset used. This observation suggests that t-SNE's primary focus on preserving pairwise similarities may not be well suited for these specific datasets due to its emphasis on local relationships.

4.2. The case study

Building upon the experiments and analyses conducted in the previous section, which involved a dataset with a known number of clusters, we observed that the use of silhouette methods in conjunction with UMAP consistently delivered the most optimal results in determining the appropriate value of k. Therefore, we employed this workflow to establish the value k for Munch's Self-Portrait (1905). As depicted in Figure 4, a specific section of the portrait was selected to estimate k, carefully chosen to encapsulate the overall characteristics. Having acquired the value of k through this workflow, we applied the K-Means clustering algorithm to the entire painting. The resultant clustering is presented in Figure 6. We identified the optimal number of clusters, i.e., $k = 5$ (See silhouette coefficient plot in Figure 5), through UMAP and the silhouette method. This clustering analysis unveils distinct groupings within the dataset, each corresponding to different pigments or mixtures of pigments. This preliminary analysis offers valuable insights into the dataset's diversity, which can serve as a useful starting point for curators and conservators.

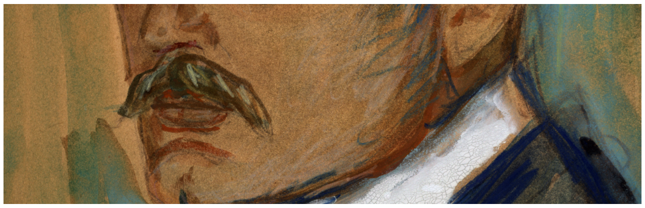

Fig. 4: Segment of Edvard Munch's Self-Portrait (1905) thoroughly reproduces the comprehensive features (pigments) employed in the composition. Thus, this segment is used to estimate k for the case study image.

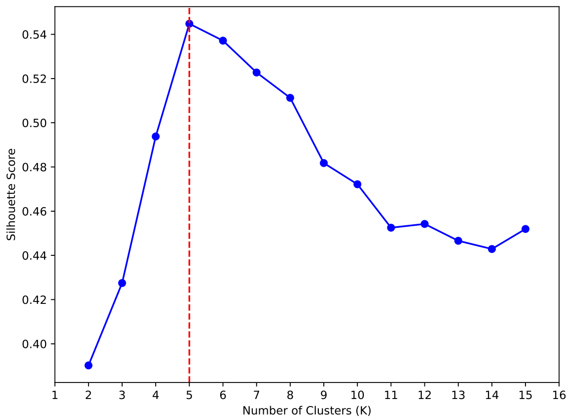

Fig. 5: Silhouette coefficients plotted against different numbers of clusters; the red dashed line indicates the number of clusters (k) where the data points are well-separated and appropriately assigned to the correct clusters.

5. CONCLUSION

In conclusion, this paper has successfully demonstrated the application of K-means clustering algorithms to determine the number of pigment clusters for unlabeled data, demonstrating a robust workflow. This study highlights the effectiveness of

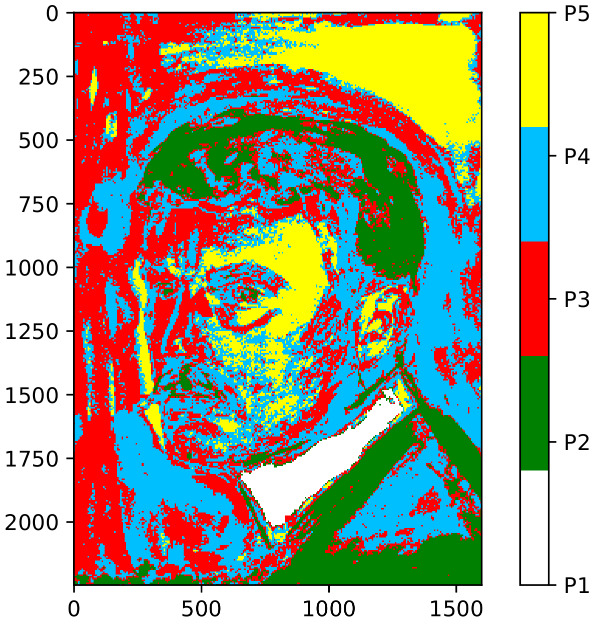

Fig. 6: K-Means clustering output applied to Edvard Munch's self-portrait, highlighting distinct clusters corresponding to different pigments.

combining UMAP with the silhouette method for precisely estimating cluster numbers in datasets that resemble works of art. These findings emphasize the crucial role of selecting an appropriate dimensionality reduction technique consistent with the unique characteristics and the specific challenges posed by the clustering task. By applying this workflow to Edvard Munch's self-portrait, we have shed light on a relatively unexplored artwork within the cultural heritage domain, further illustrating the potential of computational methods in art analysis and preservation.

6. REFERENCES

[1] C. Cucci, J. K. Delaney, and M. Picollo, "Reflectance hyperspectral imaging for investigation of works of art: Old master paintings and illuminated manuscripts," *Acc. Chem. Res.*, vol. 49, no. 10, pp. 2070–2079, 2016.

[2] H. Deborah, S. George, and J. Y. Hardeberg, "Pigment mapping of the Scream (1893) based on hyperspectral imaging," in *Image and Signal Processing*, 2014, pp. 247–256.

[3] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, no. 6, pp. 417, 1933.

[4] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008.

[5] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[6] T. S. Madhulatha, "An overview on clustering methods," *IOSR J. Comput. Eng.*, vol. 2, 2012.

[7] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, 2005.

[8] F. Harary, *Graph Theory*, Addison-Wesley, 1971.

[9] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy cluster analysis: methods for classification, data analysis and image recognition*, John Wiley & Sons, 1999.

[10] A. Cuevas, M. Febrero, and R. Fraiman, "Cluster analysis: A further approach based on density estimation," *Comput. Stat. Data Anal.*, vol. 36, no. 4, pp. 441–459, 2001.

[11] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[12] K.-R. Müller, S. Mika, K. Tsuda, and K. Schölkopf, "An introduction to kernel-based learning algorithms," in *Handbook of Neural Network Signal Processing*. CRC Press, 2018.

[13] T. M. Kodinariya and P. R. Makwana, "Review on determining number of cluster in K-means clustering," *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, no. 6, pp. 90–95, 2013.

[14] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[15] J. Xiao, J. Lu, and X. Li, "Davies Bouldin Index based hierarchical initialization K-means," *Intell. Data Anal.*, vol. 21, no. 6, pp. 1327–1338, 2017.

[16] M. Mohajer, K.-H. Englmeier, and V. J. Schmid, "A comparison of gap statistic definitions with and without logarithm function," *arXiv preprint arXiv:1103.4767*, 2011.

[17] B. Singer, T. E. Aslaksby, B. Topalova-Casadiego, and E. S. Tveit, "Investigation of materials used by Edvard Munch," *Stud. Conserv.*, vol. 55, no. 4, pp. 274–292, 2010.

[18] D. J. Mandal, M. Pedersen, S. George, H. Deborah, and C. Boust, "An experiment-based comparative analysis of pigment classification algorithms using hyperspectral imaging," *J. Imaging Sci. Technol.*, vol. 67, no. 3, pp. 030403–1–030403–1, 2023.