Analysis of 5 Paintings of Shot Marilyns by Andy Warhol

Yuchen Liu, Tianhong Lu, Jiayuan Wu
STA 160
June 2023
University of California, Davis
yvcliu@ucdavis.edu; thllu@ucdavis.edu; njywu@ucdavis.edu

Abstract

In this paper, our main focus is on the data analysis and image segmentation of Andy Warhol's "Shot Marilyns" artworks. We aim to explore various methods such as EM clustering, k-means, and HC trees to gain insights into these artworks. By utilizing color information, we seek to uncover and analyze the differences between the various images. Additionally, we aim to understand the specific embellishments that Warhol has applied in her paintings by examining the actual images. Through this research, we aim to provide a deeper understanding of Warhol's artistic choices and techniques employed in the "Shot Marilyns" series.

1 Introduction

The "Shot Marilyns" series by Andy Warhol is highly significant and worth studying. These paintings, created in 1964, feature Marilyn Monroe's portrait with different colored backgrounds. The series gained immense fame due to a peculiar incident where four of the five paintings were accidentally shot, leaving bullet holes through Marilyn's forehead. This incident added an extraordinary backstory to the artworks. The "Shot Marilyns" hold historical and cultural value, reflecting Warhol's iconic style and the enigmatic atmosphere of his famous studio, The Factory. The series showcases Warhol's ability to captivate the art world, with one of the shot canvases selling for a record-breaking 4 million in 1989. The paintings symbolize the larger-than-life environment of Warhol's creative scene and remain a testament to his enduring legacy.

In this analysis, we will delve into five distinct "Shot Marilyns" paintings by Andy Warhol, examining the colors and outlines across the five paintings which have different colored backgrounds, such as red, orange, light blue, sage blue, and turquoise. For these portraits, we will be using machine learning techniques to extract major colors of each image by separating layers of colors and analyzing the distribution of colors.

In this project, our focus is on unsupervised learning and clustering methods, specifically EM (Expectation-Maximization), MCMC (Markov Chain Monte Carlo), K-means, and HC trees (Hierarchical Clustering). These techniques enable us to analyze data without labeled information or predefined categories.

EM algorithm iteratively estimates parameters for probabilistic models and is valuable when dealing with missing or incomplete data. MCMC methods allow efficient sampling from complex probability distributions, making them useful for exploring high-dimensional data. K-means clustering partitions data into K distinct clusters by minimizing the sum of squared distances within each cluster. HC trees

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

create a hierarchical structure of clusters, offering a flexible approach for analyzing data at various levels of granularity.

By leveraging these unsupervised learning and clustering techniques, we aim to gain insights into patterns, relationships, and groupings within the data. This project enables us to explore the characteristics and structures of the analyzed dataset in a data-driven and exploratory manner.

2 Data Exploration

In this analysis, red, orange, blue, sage blue, and turquoise "Shot Marilyns" paintings by Andy Warhol will be used. The size of each image is 960 x 960 pixels matrix. Given the complexity of working with a 960x960x5 matrix where each pixel contains both RGB color and x-y coordinate information, directly clustering on the entire dataset would be challenging. The inclusion of both types of information complicates the clustering process due to the high dimensionality and the potential noise introduced by the x-y coordinate data.

To address this challenge, we can adopt a two-step approach. Firstly, we can focus on the RGB color information alone for clustering. By disregarding the x-y coordinates at this stage, we can reduce the dimensionality and potentially uncover meaningful patterns and relationships within the color space.

Once we have clustered the RGB color data and identified groups of similar colors, we can then introduce the x-y coordinates as a secondary step. This can help reduce the noise and refine the clustering results by considering spatial information. By incorporating the x-y coordinates, we can better capture the spatial structure and organization of the image, leading to more accurate and meaningful clusters.

By sequentially considering the RGB color information and then incorporating the x-y coordinates, we can effectively tackle the challenges posed by the high-dimensional and complex nature of the dataset. This approach allows us to extract valuable insights from the data while mitigating the impact of noise and improving the clustering results.

(a) Red (b) Orange (c) Blue

(d) Sage Blue (e) Turquoise

Figure 1: Five Shot Marilyns Paintings

Each individual images are converted to RGB mode. Here are the part of RGB information about each image (Figure 2):

	R	G	B
0	176	26	6
1	178	26	7
2	177	24	7
3	175	22	6
4	175	20	5
...	...	...	...
921595	172	37	3
921596	170	36	2
921597	169	36	2
921598	168	36	1
921599	166	34	1

921600 rows × 3 columns

(a) Red

	R	G	B
0	208	118	111
1	222	114	105
2	223	106	94
3	219	113	96
4	206	122	98
...	...	...	...
921595	214	121	71
921596	216	121	67
921597	216	117	61
921598	224	122	60
921599	228	120	53

921600 rows × 3 columns

(b) Orange

	R	G	B
0	86	140	200
1	86	140	200
2	86	140	200
3	86	140	200
4	86	140	200
...	...	...	...
921595	86	140	200
921596	86	140	200
921597	86	140	200
921598	86	140	200
921599	86	140	200

921600 rows × 3 columns

(c) Blue

	R	G	B
0	148	181	169
1	151	184	173
2	151	185	173
3	151	187	175
4	149	184	173
...	...	...	...
921595	155	188	177
921596	158	191	180
921597	152	189	177
921598	153	189	177
921599	159	195	183

921600 rows × 3 columns

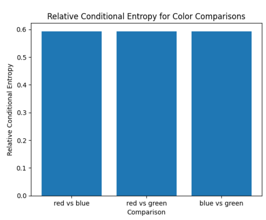
(d) Sage Blue

	R	G	B
0	148	181	169
1	151	184	173
2	151	185	173
3	151	187	175
4	149	184	173
...	...	...	...
921595	155	188	177
921596	158	191	180
921597	152	189	177
921598	153	189	177
921599	159	195	183

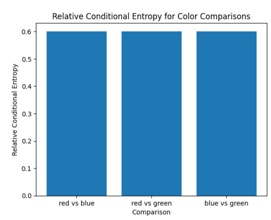
921600 rows × 3 columns

(e) Turquoise

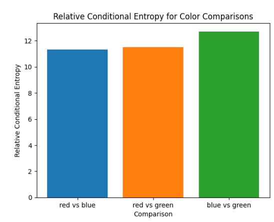
Figure 2: RGB information of each Image



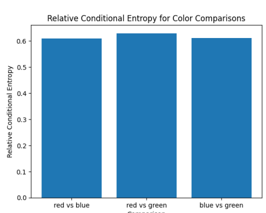
(a) blue



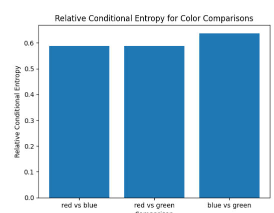
(b) orange



(c) red



(d) sage



(e) turquoise

Figure 3: Relative Conditional Entropy for each image

In Figure 3, we examined the association between pairs of fundamental colors present in all five images. To understand this association, we utilized relative conditional entropy. This measure allows us to quantify the amount of information shared between two colors. The relative conditional entropy (HR) that we employed is defined within the interval [0, 1]. A value of zero for HR indicates that the two colors in question are completely determined by each other. In other words, if one color is known, it provides all the information needed to uniquely determine the other color, and vice versa. On the other hand, a value of one for HR suggests that the two colors are completely independent of each other. In this case, knowing one color provides no information about the other color. By calculating the relative conditional entropy, we gain insights into the degree of interdependence or independence between pairs of colors. This analysis helps us understand the nature of the relationships between different colors within the images and provides valuable information about their underlying composition and visual dynamics.

3 Color Layer Separation

To achieve our objective, we begin by clustering the RGB color values. We observe noticeable color variations among different elements such as hair, face, and eyeshadow. By leveraging these color differences, we anticipate achieving a desirable separation by partitioning the image into regions with similar colors. We refer to these regions with the same color as color layers. By isolating these color layers, we can identify and extract distinct areas within the image that exhibit similar color characteristics.

Due to the presence of multiple color layers in each of the five paintings in the series, manually processing each layer would be time-consuming. First is by unsupervised automatic clustering, Second approach utilizes region of interest (ROI) and color cutoff techniques for clustering.

3.1 unsupervised clustering

After plotting the pixel colors in the RGB space, we discovered that the color distribution is highly unbalanced and does not conform to any recognizable distribution pattern. Moreover, clustering methods such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Hierarchical Clustering require substantial memory resources, making it impractical to cluster the entire painting.

To overcome these challenges, we propose utilizing the EM (Expectation-Maximization) and K-means clustering algorithms. By representing all pixels in the RGB space, clustering algorithms enable us to group pixels with similar color features together. Our objective is to obtain a concise and meaningful color segmentation through clustering.
By employing these clustering algorithms, we aim to partition the pixels based on their color

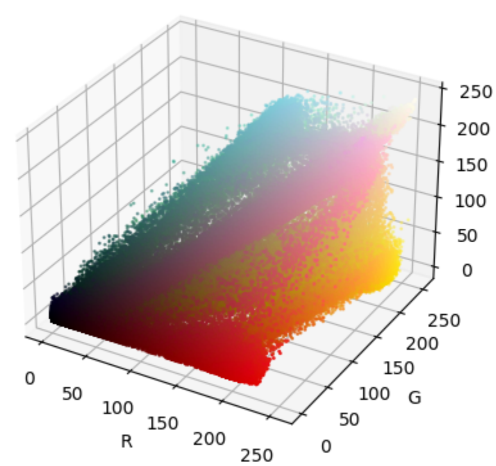


Figure 4: RGB space

similarities. This allows us to extract distinct color clusters that can provide a brief representation of the color distribution within the painting. The clustering results will help us identify significant color boundaries and understand how colors are organized within the artwork.

To achieve our objective, we begin by clustering the RGB color values to observe noticeable variations among different elements such as hair, face, and eyeshadow. By partitioning the image into regions with similar colors, known as color layers, we can extract distinct areas that exhibit similar color characteristics. However, manually processing each layer in the five paintings would be time-consuming. To address this, we propose using unsupervised clustering techniques such as EM (Expectation-Maximization) and K-means algorithms. By representing pixels in the RGB space and grouping them based on color similarities, we aim to obtain a concise and meaningful color segmentation. This approach allows us to identify significant color boundaries and understand the organization of colors within the artwork.

3.1.1 k-means

The k-means clustering algorithm makes certain assumptions to perform the clustering task effectively.K-means assumes that the data points within each cluster are closer to the centroid of that cluster compared to the centroids of other clusters. It assumes that the clusters are spherical and have approximately equal sizes.

K-means also assumes that the data points within each cluster are closer to the centroid of that cluster compared to the centroids of other clusters. It assumes that the clusters are spherical and have approximately equal sizes.

The K-means algorithm works in the following steps:

1. **Initialization:** Depending on the choice of users, K initial centroids will be selected from the dataset. Forty Method, Random partition and K-Means++ are some popular methods that might be used for this step.
2. **Assignment:** The next step is to add each data point to the cluster with the closest centroid. The similarity between points is often measured using distance measures such as Euclidean distance or Manhattan distance. For instance, if d_{ij} denotes the Euclidean distance between the i^{th} datapoint and the j^{th} centroid, the data point i will be assigned to the cluster j that minimizes d_{ij}.
3. **Update:** After all points get an assigned cluster, the centroids need to be relocated. Instead of using data points as centroids, the mean of all the data points assigned to the cluster will be the new centroid.

Repeat step 2 and step 3 until the algorithm converges.

Figure 5: K-means cluster

In addition to the information from our previous conversation, the k-means algorithm has a recognizable cutoff for each color, but it becomes less sensitive when dealing with dim colors. This is because k-means clusters points based on the total distance from each point to its centroid within the same cluster. As colors approach the point (0, 0, 0), they tend to converge, and k-means exhibits a weaker recognition of distinct clusters in these dim regions.

However, the assumptions of k-means, which assume spherical clusters and equal variances,might be severely violated in the RGB space. When examining the RGB space, we observe stick-like distributions emanating from (0, 0, 0) in the XYZ space. K-means assigns all points in the space to their nearest centroids, assuming that colors from the same event cloud represent similar features. However, this clustering approach may not provide informative results since points from different event cloud might have even closer distances to other color clouds. Furthermore, each color within a event cloud may represent diverse characteristics or variations.

Therefore, it is important to consider these limitations of k-means when using it for color clustering. Other approaches, such as EM clustering or alternative methods that can capture the underlying structure and distributions of the data, may be more suitable for obtaining meaningful insights from the color information in the RGB space.
 In the k-means clustering approach, we also replace the original colors of each color layer in the

Figure 6: 6-color by K-means

image with the RGB value of its centroid. This process condenses the image to a simplified version with only a few representative colors, reducing the complexity and making it visually appealing. By assigning each pixel to a cluster and using the centroid color as a replacement, we emphasize the dominant colors and minimize color variations within each layer.

3.1.2 Expectation-Maximization

The Expectation-Maximization (EM) algorithm is a powerful probabilistic method used for clustering and parameter estimation. Unlike the k-means algorithm, which assigns points to clusters based on distances, EM treats the clustering problem as a statistical inference task.

The EM algorithm iteratively estimates the parameters of a probability distribution model by alternating between two steps: the E-step and the M-step. In the E-step, it computes the probability (or likelihood) of each data point belonging to each cluster based on the current parameter estimates. Then, in the M-step, it updates the parameters by maximizing the expected log-likelihood based on the assigned probabilities from the E-step.

The EM algorithm makes certain assumptions about the underlying data distribution in order to perform effective clustering.

The EM algorithm assumes that the data is generated from a mixture model, where each cluster follows a specific probability distribution. In the case of color clustering, a Gaussian mixture model is often used, where each color cluster follows a multivariate Gaussian distribution.

The algorithm also assumes that the dimensions of the data, such as the RGB values, are statistically independent of each other within each cluster. This assumption allows the algorithm to estimate the parameters of the probability distributions for each cluster more accurately.

The EM algorithm requires prior knowledge or estimation of the number of clusters in the data. This information is necessary for initializing the algorithm and determining the number of components in the Gaussian mixture model.

If the stick-like shape or cloud observed in the RGB space is due to the performance of the paint on the painting, it is reasonable to expect that the distribution of colors within each cloud would roughly follow a normal distribution. This assumption aligns with the underlying assumption of the EM algorithm, which assumes that each cluster in the Gaussian mixture model follows a multivariate Gaussian distribution.

In the context of color clustering, the EM algorithm can be used to estimate the parameters of a mixture model, such as a Gaussian Mixture Model (GMM), where each cluster represents a color distribution. By iteratively updating the means, variances, and mixing proportions of the color distributions, EM can effectively capture the underlying structure of the color space and identify distinct color clusters.

The EM algorithm is particularly useful when dealing with complex and overlapping color distributions, as it incorporates the probabilistic nature of the data. It can also handle missing or incomplete data through the estimation of latent variables.

The EM algorithm works in the following steps:

1. Initialize θ^2 and σ^2 with random guess.
2. For i in T epoch:
   (a) Step 1: Compute u_i, Σ_i based on
   $$u_i = (\theta_t'\theta_t + \delta_t^2 I_r)^{-1}\theta_t' x_i$$
   $$\Sigma_i = \delta_t^2(\theta_t'\theta_t + \delta_t^2 I_r)^{-1}$$
   (b) Step 2: Use u_i, Σ_i to update δ_t^2, we can get δ_{t+1}^2 as follows:
   $$\delta_{t+1}^2 = \sum_{i=1}^{n}[(x_i - \theta_t u_i)'(x_i - \theta_t u_i) + \theta_t'\theta_t\Sigma_i]$$
   (c) Step 3: for $k = 1$ to n: Use u_i, Σ_i, and x_{ik} to update θ_{tk} to get $\theta_{(t+1)k}$:
   $$\theta_{(t+1)k} = \left[\sum_{i=1}^{n}(\Sigma_i + u_i^2)\right]^{-1}\sum_{i=1}^{n} u_i x_{ik}$$
   (d) Step 4: Repeat the above steps until convergence.

The Expectation-Maximization (EM) algorithm is generally more effective than k-means in separating close colors, such as the eyeshadow and clothing, resulting in less noise in the picture. This effectiveness is due to the assumption made by EM, which is that the underlying data distribution can

Figure 7: Cluster EM

be modeled by a Gaussian mixture. By estimating the parameters of these Gaussian components, EM is able to capture the subtle variations in color and separate them into distinct clusters.

However, the EM algorithm still encounters challenges in separating certain regions, such as the eyeshadow and teeth, and there are also issues with color boundaries. These limitations arise from the scanning process and the limitations of the color range (255, 255, 255) in the scanning copy. Some of the data points may reach the boundaries, causing deviations from the expected Gaussian mixture distribution. These deviations can result in misclassifications or incomplete separation of certain color regions.

Figure 8: 6-color by EM

To mitigate these issues, further preprocessing steps or adjustments to the EM algorithm may be required.

The combination of these two methods is the most efficient way to extract major colors of images. Through clustering, we would mainly focus on analyzing the background, face, hair, and eyeshadow, which are the most distinct parts of the portraits. We would use EM clustering to extract the layer of face and use K-means clustering to extract the layer of hair, background, and eyeshadow.

3.1.3 Conclusion

We utilize the K-means clustering algorithm to identify clusters based on the color characteristics of the paintings. The clustering algorithm enables the identification of similarities and differences in the color compositions, potentially highlighting specific color themes or variations across the artworks. The 3D visualization facilitates a comprehensive understanding of the clusters' positions in the RGB color space, aiding the interpretation of color relationships and their significance in the context of the artwork series. Overall, this analysis provides a quantitative and visual exploration of the color characteristics of the "Shot Marilyns" paintings, enhancing our understanding of the artistic choices made by Andy Warhol.

3.2 Cluster by ROI

Clustering on ROI (Region of Interest) refers to the process of applying clustering algorithms specifically to selected regions or subsets of data, rather than the entire dataset. In the context of image analysis, ROI typically refers to a specific area or region within an image that contains relevant or significant information for analysis.

When performing clustering on ROI, the goal is to focus on specific regions of interest that are expected to exhibit distinct patterns or characteristics. By isolating and clustering these regions separately, we can gain more detailed insights and potentially improve the accuracy of clustering results.

The process of clustering on ROI involves several steps. First, the ROI needs to be defined or selected within the image. This can be done manually by specifying the coordinates or boundaries of the ROI,

3.2.1 ROI clustering

In this case, clustering on the non-specific layer refers to applying clustering algorithms to group pixels based on their coordinates and RGB cutoff values. Instead of relying solely on the output of the EM algorithm, which may not provide a clear separation of desired regions, we incorporate additional criteria to define clusters.
The process involves setting cutoff values for both the coordinates (x, y) and RGB values to identify distinct regions within the image. By defining specific ranges or thresholds for these parameters, we can group pixels that fall within the defined ranges into separate clusters.

For example, we can set cutoff values for the x and y coordinates to define rectangular or circular regions of interest within the image. Pixels falling within these regions will be assigned to their respective clusters.

Similarly, we can set cutoff values for the RGB values to group pixels with similar colors into clusters. By specifying ranges or thresholds for the RGB components (R, G, B), we can separate pixels based on their color similarity.

Figure 9: Marilyn Monroe Colored

3.2.2 Color substitution

After splitting the colors and obtaining clusters based on the clustering algorithm, the next step is to replace the original colors with the corresponding colors from the color image of Marilyn Monroe. This process involves mapping the colors from the original image to the clusters identified in the segmentation step.

3.2.3 Conclusion

For each cluster, we can assign a representative color from the color image of Marilyn Monroe that closely matches the dominant color of that cluster. This can be achieved by finding the average or most frequent color within the cluster and replacing all the pixels in that cluster with the corresponding color from the Marilyn Monroe image.

By replacing the original colors with the colors from the Marilyn Monroe image, we can achieve a more realistic and visually appealing representation of the segmented image. This process helps in preserving the overall appearance and color characteristics of Marilyn Monroe in the final result.

Clustering based on coordinate and RGB cutoff values allows for a more targeted and controlled grouping of pixels, enabling the identification of specific regions or objects within the image. It helps overcome limitations or uncertainties in the initial clustering results from the EM algorithm, providing a more refined and interpretable segmentation of the image based on both spatial and color information.

(a) Replace eye shadow (b) Replace hair color

(c) Replace face color (d) Replace cloth color

Figure 10: Color Transferred Images

3.3 Conclusion

In conclusion, the utilization of clustering techniques such as EM and K-means has facilitated the examination and categorization of color patterns within Andy Warhol's "Shot Marilyns" series. Employing unsupervised learning algorithms, we successfully identified distinct clusters of colors present in each painting. The EM algorithm demonstrated its efficacy in effectively separating closely related colors and minimizing noise, while the K-means algorithm provided a simplified representation by focusing on dominant colors. By substituting the original colors with the centroid values of each cluster, we created images with a reduced color palette that emphasized the unique characteristics of each artwork. However, it is important to acknowledge the limitations observed, particularly in accurately distinguishing certain color boundaries and dealing with extreme color values. Nonetheless, these clustering techniques have yielded valuable insights into the color composition of the "Shot Marilyns" series, underscoring Andy Warhol's exceptional technical and aesthetic prowess that transcends reality and statistical models.

We generate a frequency histogram to examine the distribution of clusters in order to gain insights into the occurrence and frequency of different color patterns within the dataset. By incorporating the RGB values of the centroids into the bar colors, the plot provides information about the associated colors of each cluster. The bars in the histogram are colored based on the RGB values of the corresponding centroids, which are scaled to the range [0, 1] by dividing them by 255. By observing the plot, we can easily identify the dominant and less frequent clusters within the dataset. It provides us with valuable visual insights into the frequency distribution of clusters obtained from our previous analysis. The resulting histogram allows us to identify dominant and less frequent color patterns within the dataset. It aids in identifying the dominant color patterns, understanding their frequencies, and gaining insights into the color composition of the paintings.

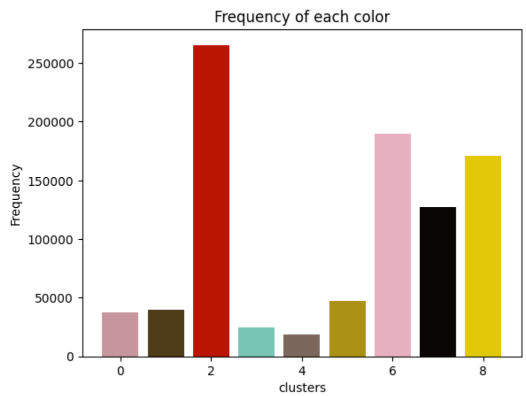
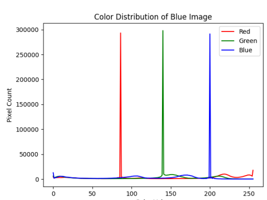
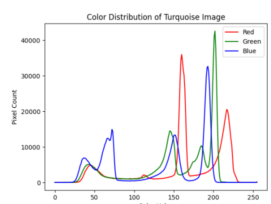


Figure 11: histogram of freqency of each image

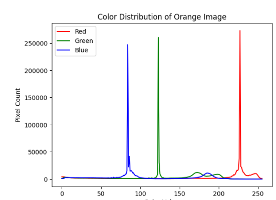
The color distribution (Figure 12) also helps to visually see the color composition of each portraits. These color distributions are very different than each other.



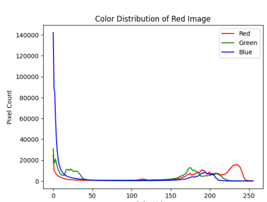
(a) blue



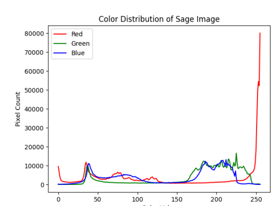
(b) turquoise



(c) orange



(d) red



(e) sage

Figure 12: Color distribution

4 Color Analysis

4.1 Color Trend on each Face Feature

Each images are separated into RGB information. The distributions of red, green, and blue differ between each image.

(a) blue-back (b) blue-eyeshadow (c) blue-face (d) blue-hair

(e) orange-back (f) orange-eyeshadow (g) orange-face (h) orange-hair

(i) red-back (j) red-eyeshadow (k) red-face (l) red-hair

(m) sage-back (n) sage-eyeshadow (o) sage-face (p) sage-hair

(q) Turquoise-back (r) Turquoise-eyeshadow (s) Turquoise-face (t) Turquoise-hair

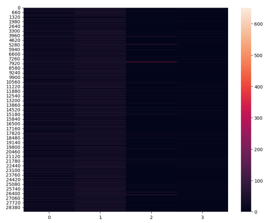
Figure 13: RGB space: Ordered by images(x) and face features(y)

In the generated plots in Figure of RGB space, each dot corresponds to a single pixel from the images. The color of each dot indicates the intensity of its red, green, or blue coordinate. The range of colors spans from low intensity (represented as white) to high intensity (represented as black). This visual representation allows us to gain a better understanding of how colors are utilized and distributed throughout the images. In each panel, progressing from left to right, we present the comparison of red versus green (with blue represented in grayscale), red versus blue (with green represented in grayscale), and green versus blue (with red represented in grayscale). This arrangement allows us to visually analyze the relationships and interactions between the different color channels in the image.

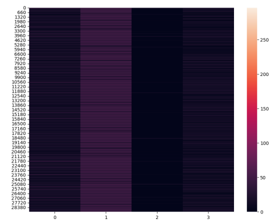
Upon observing the left column of the plots, it becomes evident that the red version of the image showcases the most intense background color among the variations, suggesting a wider and more extensive range of colors. Conversely, the sage version exhibits the least intense background color, signifying a narrower and more limited range of colors. Turning our attention to the eyeshadow region, we observe that the red version displays the most intense eyeshadow color. When comparing the face regions across the five images, the Turquoise version exhibits the least intense face color, while the blue version demonstrates relatively higher intensity. Analyzing the hair regions, the red version display relatively more intense colors compared to the other variations.

In summary, these visual analyses provide insights into the color distribution and intensity within different image elements. The observations reveal variations in background, eyeshadow, face, and hair colors across the different versions of the "Shot Marilyns" paintings by Andy Warhol. The red version is the most intense use of color among the five versions, which might indicate that in a strong background, Andy Warhol more likely tends to use more intense color in drawing hair, face, and eyeshadow.

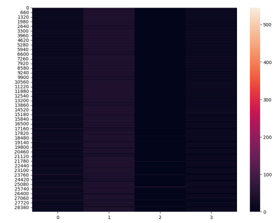
4.2 Heatmap



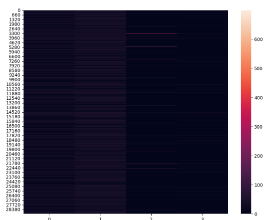
(a) red



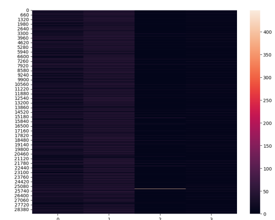
(b) orange



(c) blue



(d) turquoise



(e) sage

Figure 14: Color distribution presented by Heatmap

The heatmap represents the similarity or dissimilarity between different blocks of the image based on the RGB values. Lighter colors indicate higher similarity or clustering within those blocks. From Figure 11, we can observe red and orange images have higher value in red (column 1) in RGB than other images. Blue image have higher value in blue (column 3) in RGB than other images. All images have higher value in red (column 1) than other colors (column 2 and 3) in RGB since all images have red lip. In summary, the heatmap patterns of the five images in the analysis exhibit similarities, indicating common underlying structures or patterns. However, variations are observed in terms of the RGB colors used in each image. These differences in color contribute to the distinct visual characteristics of each artwork.

5 Conclusion

In conclusion, this study focused on analyzing and segmenting Andy Warhol's "Shot Marilyns" artworks using various clustering techniques. By exploring color information and employing clustering algorithms such as EM, k-means, and HC trees, we aimed to uncover the artistic choices and unique embellishments in the series.

We examined the "Shot Marilyns" paintings by Andy Warhol, considering their diverse colors and pixel information. Due to the dataset's complexity, directly clustering the entire set was challenging. To overcome this, we adopted a two-step approach, first clustering the RGB color information to reveal meaningful patterns, followed by incorporating spatial coordinates for refined results. By leveraging unsupervised clustering algorithms like EM and K-means, we obtained concise color segmentation, identified significant boundaries, and gained insights into color organization within the artworks. This analysis contributes to our understanding of Warhol's artistic choices and highlights the benefits of combining clustering techniques with image segmentation.

We used relative conditional entropy to examine the association between fundamental colors in the five images. This measure quantifies the shared information between colors, ranging from complete dependence to complete independence. By calculating relative conditional entropy, we gain insights into the interdependence or independence of color pairs, helping us understand the relationships and composition of colors within the images.

The application of clustering techniques, specifically EM and K-means algorithms, has enabled us to analyze and categorize color patterns in Andy Warhol's "Shot Marilyns" series. These unsupervised learning methods successfully identified distinct color clusters, with the EM algorithm effectively handling closely related colors and minimizing noise. By substituting original colors with cluster centroid values, we created simplified images that emphasized the unique characteristics of each artwork. However, limitations were observed in accurately distinguishing certain color boundaries and extreme color values. Nevertheless, these clustering techniques provided valuable insights into the color composition of the series, highlighting Warhol's technical and aesthetic prowess. Additionally, the generated frequency histogram offered visual insights into the distribution and dominance of color clusters, aiding in the identification of frequent and less common color patterns in the dataset.

The visual analysis of the generated plots in Figure of RGB space allows us to gain insights into the color distribution and intensity within different elements of Andy Warhol's "Shot Marilyns" paintings. By comparing the color channels, we observe variations in background, eyeshadow, face, and hair colors across the different versions. The red version stands out with the most intense background color and eyeshadow, while the sage version has the least intense background color. The Turquoise version exhibits the least intense face color, while the blue version shows relatively higher intensity. These findings suggest that Andy Warhol tends to use more intense colors in hair, face, and eyeshadow when there is a strong background. These visual analyses provide a valuable understanding of the color utilization and distribution in the artworks.

The heatmap analysis based on RGB values shows similarities in the patterns of the five images, suggesting common underlying structures. However, variations are observed in terms of the specific RGB colors used in each image, contributing to their distinct visual characteristics.

6 References

[1] Liao, S., Koehl, P., Schultens, J. et al. The geometry of colors in van Gogh's Sunflowers. Herit Sci 9, 136 (2021). https://doi.org/10.1186/s40494-021-00608-y