A NEW LENS ON ART HISTORY: USING COMPLEX NETWORK ANALYSIS AND UNSUPERVISED MACHINE LEARNING

by

Khloe N. Downie

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Master of Science (Applied Physics).

Golden, Colorado

Date ________________________

Signed: ________________________
Khloe N. Downie

Signed: ________________________
Dr. Lincoln D. Carr
Thesis Advisor

Signed: ________________________
Dr. Dorit Hammerling
Thesis Advisor

Golden, Colorado

Date ________________________

Signed: ________________________
Dr. Frederic Sarazin
Professor and Head
Department of Physics

ABSTRACT

Visual art is a wholly complex and inherently human creation not easily analyzed and interpreted by digital technology. The subjectivity of art makes its interpretative understanding elusive to machines while virtually instantaneous to the human viewer. This project attempts to demonstrate that there is a relationship between the measurable visual characteristics and the communicative characteristics of art. In doing so, we hope to offer a machine-based software tool that supplements the traditional critical approach to historical art found in art history and art theory. This AI-generated perspective will offer innovative insights to the inherent interpretative information found in art.

This project's methods are seated in the creation of a python-based feature extraction software. The software is an analog to the pluralistic critical approach of art theory. It abstracts images of historical paintings into complex network representations that contain the digital equivalent of formalist elements and principles of design present therein. By measuring the images' network representations, we obtain quantitative descriptions of their innate visual features. We, then, reduce the dimensionality of the measurement data set and find a clustering of the images. From those clusters, we can draw mathematical conclusions about the interpretative characteristics of the images held within. We postulate that the evaluative conclusions enabled by our method's AI-generated art movements will reach beyond those present in traditional art theory.

We measure the interpretative precision of the clustering we obtain using the precision and recall performance measures. We compare the resulting performance from our software to that of a random clustering of images. In doing so, we prove that our software indeed performs better and is statistically distinct from a random grouping of paintings in terms of critical and formalist evaluation. Beyond that, we show that our resulting clustering has greater success in terms of the performance metrics than the critically accepted historical

art movements.

These results show that, using complex networks to embody formalist elements and principles of design, measuring those networks, and clustering the paintings based on that data, we are not only able to create distinct groupings of images with common formalist components but common critical interpretations as well. Because of this, we can confirm the existence of an untapped empirical relationship between machine-measurable visual characteristics of images and the communicative concepts held within those images. That we obtained performance better than the historical movements shows that our methods offer the first steps to discerning and building on this correlation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

Information Content . . . . I_X

Shannon Entropy . . . . H_X

Number of Bins . . . . B

Number of Nodes/Superpixels . . . . N

Descriptors of type j in cluster i . . . . $d_{i,j}$

Count of descriptors of type j in cluster i . . . . $n_{i,j}$

Standardized count of descriptors of type j in cluster i . . . . $\sigma_{i,j}$

Significant descriptors of type j in cluster i . . . . $d_{ij,S}$

Insignificant descriptors j in cluster i . . . . $d_{ij,I}$

Number of clusters in an iteration . . . . K

Size frequency of cluster i . . . . c_i

Precision of cluster i . . . . P_i

Recall of cluster i . . . . R_i

Image Sample Size . . . . M

Empirical cumulative function for distribution 1 of size n . . . . $F_{1,n}$

Kolmogorov-Smirnov Statistic for two distributions of size n and m . . . . $D_{n,m}$

KS Null Rejection Level . . . . α

Two-Point Correlation for images i and j . . . . $P_C(i,j)$

LIST OF ABBREVIATIONS

Artificial Intelligence	AI
Natural Language Processing	NLP
Machine Learning	ML
Complex Network Analysis	CNA
Simple Linear Iterative Clustering	SLIC
Edge Creation Technique	ECT
Hue-Saturation-Light Color Space	HSL
Principal Component Analysis	PCA
AI-Generated Movements	AIGM
True Positive	TP
False Positive	FP
False Negative	FN
Kolmogorov-Smirnov	KS

ACKNOWLEDGMENTS

I would like to thank my mentor, Dr. Lincoln Carr, for allowing me to pursue a non-traditional thesis. Without his ever-present support, I would've never been able to work on a project that combines two of my greatest personal interests into one academic inquiry. I would like to send a huge thank you to my committee for always showing great interest in my research. I'd like to express gratitude to the faculty in the Colorado School of Mines Physics Department that I had the pleasure of learning from. The full support that I received from the physics community is what got me through both my grad and undergrad experience. This project would never have succeeded without Simon Reisig, who always instantly and willingly helped with any programming questions I ever had. I'd like to thank Joel Been for his constant presence in the QTL, allowing me to vent my ideas and struggles. Of course, I wouldn't be even close to obtaining the great education that I am fortunate enough to have if not for the support of my family. Specifically, I'd like to thank my grandmother, Barbara, for taking (sometimes forcing) me to a constant stream of museums, cultural sites, and art and science camps, sparking my passion for art, history, and science. I'd like to show appreciation to my dad for, without fail, asking me what I learned that day when picking me up from school, instilling my excitement for education. A major thanks goes to Booker T. Washington High School for preparing me for all that is the Colorado School of Mines. Finally, I'd like to thank my childhood art teacher, Mrs. Duvall, for showing me the creative space that I hold within.

For those who find comfort in the subjectivity of art and inspiration from the objectivity of science.

CHAPTER 1

INTRODUCING COMPLEXITY SCIENCE TO VISUAL ART ANALYSES

Visual art—the free expression of emotion, events, concepts, and ideas—is a wholly complex and inherently human creation, one not easily transferable to or understood by digital technology. The subjectivity of art, in both the creation and interpretation aspects of the field, makes the foundational understanding of visual art elusive to machines while virtually instantaneous to humans. How can this dichotomy be bridged? Is there a way for a computer to understand or, at the very least, discriminate between the different messages and concepts within visual art? Can we mathematically describe the artistic lexicon? This project attempts to construct this bridge—to form a pathway for machines to detect the visual subtleties within art that hint at an underlying emotional, contextual, or historical meaning. The hope is for this project, and subsequent projects on the matter, to eventually lead to a greater visual sentiment analysis technique for computer vision problems.

To generate a meaningful automated understanding of visual art, this project builds on complex network analysis, a compilation of machine learning tools, art theory, and information theory to create a python-based feature extraction software that organizes given images into groupings that are internally similar and externally dissimilar. In this, the art to be "understood" by the machine are digitized paintings whose historical art movement is known. Ultimately, our project is a proof of concept for a computer vision categorization problem. We require the machine to create groupings of the input images based on what it can understand—color space, spatial composition, and so on. In performing this analysis, we hope to confirm that there is a relationship between visual characteristics and human interpretations of art.

Our objective is to outline an algorithmic blueprint that introduces an artificial intelligence perspective to art history that supplements the traditional art theory perspective. We want to probe the inherent information of visual art and use that to investigate whether critical art theory excludes the consideration of certain features fundamental to a comprehensive description of art. We choose to do this investigation using *unsupervised* machine learning. We use unsupervised techniques at every available opportunity. This is done in order to obtain that artificial intelligence perspective that we seek. We want to work strictly within the machine's self-contained capabilities, avoiding a human or art theory bias–a bias that would naturally come about with supervised learning. By doing so, we theorize that our software could offer interpretative insights uninfluenced, and potentially undiscovered, by traditional, qualitative art analyses.

1.1 Research Questions

We would like to answer the following questions:

1. Is there a way for a computer to catalog the interpretive concepts exemplified in visual art strictly using the inherent visual information contained within? In other words, is there a relationship between the measurable visual characteristics and the qualitative, communicative, and meaningful characteristics of art?

2. Does art history neglect features necessary for a complete description and interpretation of art? Can we create a new, machine-based interpretive tool to supplement those that already exist?

3. How can we measure the inherent information in art?

4. How do we bridge the subjectivity of art and the objectivity of science?

1.2 Motivation: Computer Science And Linguistics

The work in this project was inspired by other works combining art and science, particularly in the fields of computer science and linguistics.

1.2.1 Semantic Networks

A *semantic network* is a complex network that represents a body of text, that uses nodes to represent the words in the text and edges to represent the relationships between those words. Our project can attribute a significant part of its inspiration to the work done in [1]. This project uses semantic networks to identify literary movements. Diego *et al* use complex networks to represent historical prose, and though a multi-step process involving network measurement, dimensionality reduction, and clustering, they ultimately obtain results that accurately cluster the texts into their respective literary movements. Our project loosely builds upon the algorithmic structure provided by this paper.

1.2.2 Natural Language Processing

Natural language processing (NLP) is a type of computational research used to understand human language. NLP is commonly used for sentiment analysis–an approach to linguistic science that uses ML models to investigate evaluative statements (opinions and emotions) of the public for predictive or feedback purposes. Semantic sentiment analysis corroborates our belief that visual sentiment analysis is possible. Semantic sentiment analysis uses lexical (vocabulary) and syntactic (spatial and contextual relationships between words) information found in text or speech for computation [2]. Our theory is that the translation of semantic to artistic/visual sentiment analysis would involve using the artistic lexicon, i.e., the elements of design, and the relationships between those techniques, i.e., the principles of design, for computation.

It is important to note that, because evaluative statements are subjective in nature and conditional to individual words, it is inefficient to use simple NLP text classification methods, like those used for objective statements, for semantic sentiment analysis. Instead, complex networks have offered a more precise method for sentiment analysis, due to their ability to maintain a rich informational structure, by utilizing their local and global properties [3]. Since visual art is highly subjective, this leads us to believe that employing complex networks for computation could be advantageous for our purposes, as well.

The use of probabilistic models to examine language, like with NLP, has led to an improved linguistic science in general. NLP has allowed researchers to find important new applications in understanding human language processing and in modelling linguistic semantics and pragmatics. Because of this, linguistic science has since turned to more empirical methods of investigation [2], indicating the potential for computational approaches to improve upon and offer new analyses in art theory and art history.

CHAPTER 2

BACKGROUND: COMPLEXITY, ART, AND INFORMATION THEORIES

This project works heavily with complex network analysis, art theory, and information theory. To fully understand the steps taken and calculations done in our project, we provide here a sufficient overview those fields.

2.1 Complex Network Analysis

Complex network analysis (CNA) is a type of modelling used to algorithmically examine many-body systems containing some form of interaction. Due to the versatility of its methods, one can apply CNA's paradigm to virtually any desired system, if that system has discrete elements that can interact or are connected in some way. CNA abstracts a system into a relational representation [4]. The elements in the system are represented as *nodes*, and the connections between the elements are represented as *edges*. These networks (otherwise known as graphs) can contain a high amount of specificity; it is possible for the graph to contain specific properties for each individual node and edge. That is, it is not necessary for all elements in a system to have the same characteristics or for all the interactions to be homogeneous. A key postulate to CNA is there is generally some emergence associated with the representational systems; behaviors and characteristics non-obvious with traditional modelling become apparent with the evolution and/or measurement of their complex network representation [5].

2.2 Art Theory and Art History

Our project builds upon the idea of a *pluralistic critical interpretation* of art. Throughout art history, there have been many noteworthy critical approaches to art theory and history, each focusing on how the art community should think about a work of art. Generally, these theories focused exclusively on one of two predominant outlooks: to interpret art strictly in terms of visual form, or to interpret art contextually (in terms of historical, social, or psychological contexts). *Formalism* was the leading theory that championed the former, coming into popularity in the early 20th century and offering *elements and principals of design* as the lexicon with which a critic should evaluate art [6]. As time progressed, art evolved and, with it, so did its interpretative theories. In the mid-20th century, *critical theory* came into focus, emphasizing the need to follow a *semiotic* (symbol-based) and contextual evaluative method for art [7]. The formalist and critical theories lived in conflict with one another until the late 20th century, when pluralistic interpretation–an evaluative method that synthesizes both visual form and semiotics–came into popularity. The critics that supported pluralistic theory argued that products of culture can have many valid interpretations and that one cannot have a complete understanding of a work of art without considering both concepts [7], [8].

The method in this project works to combine formalism and critical art theories to interpret art, much like the pluralistic approach does. However, unlike the traditional methods, we will evaluate art using the theories in serial. As introduced in Section 1.1, one of our objectives is to investigate whether there is an innate and consistent connection between the visual forms in an image and the emotional, contextual, social, etc. concepts exemplified in the art. Instead of synthesizing the two theories in parallel, we instead look at an image with regard to formalist terms–elements and principles–to have our machine (theoretically) sort images in critical terms, into groupings that are internally symbolically and/or contextually similar and externally dissimilar. It is important to note that, though we use formalist ideas to build our software, we are not having the machine define or assign meaning to those terms, as a formalist critic would do. In other words, we have our machine look at the art as surfaces, measuring the inborn color, spatial distribution, and relative angular composition, and return what it sees as significant groupings, from which we will attempt to capture the relationship between face-value characteristics and human meaning. By doing this, we are doing as [6] suggests; we are defining a new empirical and consistent evaluative method with potential usefulness for modern artistic analyses.

Another theory important to our project is the Gestalt theory of visual perception and psychology. Essentially, this theory states that human psychology makes it so, when viewing an image, the sum of its parts is greater than its whole, called a *gestalt*. This is a common occurrence in our everyday lives. Examples of this would be simultaneous contrast, when a color looks different when placed on different colored backgrounds, or similarity/proximity grouping, when we associate elements that look alike/are close together as belonging together. This tendency is innate in humans and can affect the way we evaluate art [9]. Machines, on the other hand, do not experience this phenomenon since they do not have a human psychology. They view images at face-value, allowing our technique to go beyond the natural human bias that comes with art evaluation and strengthening our argument of our method's consistency.

Because our software is loosely based on the formalist lexicon, we will outline these concepts. A *composition* is governed and formed by the elements and principles of design. Elements of design are the basic components found in art, such as color or line. Principals of design are the use of elements in a specific way to create a certain visual effect, such as balance or movement. Any element can be used to create any principal; there is no elemental exclusivity in how principals are formed. We capitalize on the relationship between the elements and principles when we create our complex network representations of the images, discussed in detail in Section 3.4. A list of the different elements and principles of design and their descriptions can be found in Table 2.1.

Table 2.1: Elements and Principals of Design

Elements of Design	
Element	Description
Point	Specific position, no extension
Line	1D; Can have length and direction, but no depth
Shape	2D; A closed contour
Form	3D; A combination of line, point, shape; Describes a volume
Color	Hue and saturation
Value	Relative lightness and darkness
Texture	Surface quality (physical or simulated)
Space	Area between and around objects
Principals of Design	
Principle	Description
Balance	Visual equilibrium of similar, opposing or contrasting elements that make a unified whole; Symmetrical or asymmetrical
Emphasis	Marks location in a composition which most strongly draws the viewer's attention
Movement	Visual flow through the composition; Forces movement of viewer's eye
Pattern	Repeated object, symbol, or element
Repetition	Reuse of the same, similar, or different objects throughout the design
Proportion	Comparative relationship between 2 or more elements; Creates a sense of order between elements used
Rhythm	Alternation of elements with defined intervals; Establishes pattern, texture, and/or movement
Variety	Complex relationship between elements; Creates visual interest in specific area
Unity	Uniform relationship between elements and the composition; Creates a sense of completeness
Contrast	Difference within an element or between elements; Makes element stand out

The analysis of our results requires some familiarity with critical theory. Ultimately, critical theory looks at abstraction level, genre, and message in unison. The abstraction level of any given painting can be one of four options, listed below.

- Wholly Abstract: The painting contains content with no resemblance to natural shapes.
- Organically Abstract: The painting contains content with some resemblance to natural organic forms.
- Semi-abstract: The figures and other objects within the painting are discernible to an extent.
- Naturalistic: The figurative and other content within the painting is instantly recognizable.

For the organically abstract, semi-abstract, and naturalistic abstraction levels, one can determine genre. The options for genre are as follows: historical (the portrayal of an actual event), portraiture, genre (the portrayal of an everyday event, with no recognizable characters), landscape, and still life.

Historical movements are groupings of art in which the ideas, concepts, emotions portrayed therein are similar. Historical movements have a general message or purpose, such as a reaction to contemporary society or a rejection of tradition, that are decided and agreed on by prominent critics or schools of art. We use these movements as a basis for comparison for the groupings our machine creates, or *AI-generated movements* as we will call them. Our project examines work from the following historical movements, listed in chronological order: Renaissance, Baroque, Rococo, Neoclassicism, Romanticism, Realism, Impressionism, Post-Impressionism, Fauvism, Expressionism, Cubism, and Surrealism. These movements span from the 1400s to the 1970s. (Note: We only look at paintings that are qualified as fine-art.) We also examine randomly pixelated images, images of randomly chosen shapes, and uniform color images in serial with the actual art as a sort of sanity check of our results. For the sake of clarity in reading this report, we will refer to these images as *simple movements* and the actual art as *complex* or *historical movements*. Art history characterizations of the complex movements as well as example images from each can be found in Appendix A.

2.3 Information Theory

Information theory is used extensively in our project, specifically in results analysis. Information theory is applied probability, mathematically examining the inherent information in communicative processes. Since art is essentially a form of communication, information theory is perfect for our project's purposes. For the main body of our thesis, we only use information theory in idea. When doing our arrow of time analysis (located in Appendix E), we use information theory in practice, building upon the following metrics: information content and entropy.

Information content, sometimes called the surprisal or Shannon information, is a relatively simple measure involving the probability of a random variable. It can be thought of as the "surprise" of or information gained from an outcome. Given an event x with probability P_x, the information content is defined in Equation 2.1.

$$I_X = -\log_2(P_x) \tag{2.1}$$

The base of 2 means that the information content is in units of bits. A probability of 100% would return a surprisal of 0, meaning we have gained no information from that measurement. The lower the probability of an event, the greater its surprisal.

Entropy builds on information content. It is the expected value of the the information content of a random variable, quantifying how surprising–how much information we gain–measuring an event is on average. The lower the entropy, the more information is inherent in the system, and the surprise is low upon measurement [5]. Given a set of events X with probabilities $p(x)$, the entropy of the system is given by Equation 2.2.

$$H_X = -\sum_{x \in X} p(x) \log_2 p(x) \tag{2.2}$$

Again, base 2 means the entropy is in terms of bits. A set of events with equal probability would return the maximal entropy.

CHAPTER 3

METHOD: PROBING FORMALIST ART THEORY USING COMPLEX NETWORKS

3.1 Method Outline

To examine the connections between the visual information and contextual information of art, we built a python-based feature extraction software, using a compilation of programming and modeling techniques. Our software takes the following steps:

1. Decompose an image with image segmentation to obtain the nodes in its representative complex network.
2. Create edges between the nodes, based on a user-chosen image characteristic.
3. Take and record measurements of the network. Repeat steps 1-3 for all sampled images.
4. Perform a dimensionality reduction on the measurement data set.
5. Cluster the images in the reduced measurement space.

3.2 Data Preparation

Our work requires a large data set of paintings of which the movement and general interpretation is defined and known. We collected 49 digital image samples from each movement (both simple and complex) listed in Section 2.2. The images of the complex movements were not normalized in their resolution nor color gamut, which introduces potential noise. To reduce some of this inconsistency, before creating their complex network representations, we normalize all images to the same resolution, 500x500 pixels, using the *Pixel Area Relation* interpolation method. Though there may be some informational benefit to maintaining the original shape of an image (not all paintings are square originally), we chose to normalize size to simplify our results' interpretations.

3.3 Network Creation: Nodes

After normalization, our software proceeds by creating a representative complex network for an image. Section 2.1 states that, for CNA modelling to be feasible, we require discrete elements to act as nodes in the network. The most obvious choice for the network's nodes would be the image's individual pixels. Though this may be efficient for small images, we work with relatively large images–containing thousands of pixels–so this choice would be computationally inefficient. It would also be informationally inefficient, since formalist visual forms (the elements listed in Table 2.1), what we are attempting to capture, live in a space larger than a single pixel. Hence, we must transform the image into a representational object that is easier to manage. To do this, we use *image segmentation*. In short, image segmentation is a machine learning tool that divides an image into clusters of pixels, known as *superpixels*, based on certain characteristics of the local pixel neighborhoods. Instead of creating thousands of nodes, image segmentation allows us to create only tens to hundreds of nodes while capturing and maintaining the key visual information of the image.

There is a variety of segmentation techniques to use, each creating the superpixels based on different pixel characteristics. There are supervised and unsupervised algorithms available, as well. This project focuses on three unsupervised techniques: Felzenszwalb, Quickshift, and Simple Linear Iterative Clustering (SLIC). We chose to use unsupervised techniques to avoid introducing human bias (remember that we want an AI-perspective on art). Since we are working in a novel space and are unsure of what visual forms have the most significant relationship to human interpretation, we chose to examine all three segmentation techniques as hyperparameters in our software. Later, we will determine which, if any, techniques return the most relevant results. See [10], [11], and [12] for more information on the segmentation techniques we use. Figure 3.1 shows the resulting segmentation (and re-scaling) for each of the before-mentioned segmentation techniques for an example image. The yellow lines indicate the borders of each superpixel, and the red dots mark their centers.

3.4 Network Creation: Edges

Our next step is to determine the edges between the nodes. This is where we capitalize on the relationship between the elements and principles of design. As stated, principles are different combinations of elements to create a particular visual effect. To implement this idea in our complex network representations, we think of the visual/image characteristics of the superpixels–represented as nodes in our network–as elements, and we think of those characteristics' similarities or dissimilarities between nodes as principles. Theoretically, if we view the superpixel visual characteristics as elements, then the relationships between those characteristics should embody the principles of design. Then, when we measure a network, we should actually be measuring the principles in use in that work of art. Thus, we look at the similarities between a chosen visual characteristic of the superpixels to create the edges between their representative nodes.

As one can see in Table 2.1, principles are not determined by the presence of similar/dissimilar elements alone but by their presence *and* their spatial proximity. Hence, in addition to using visual similarities to create edges, we introduce a spatial embedding to the network. We use a radial distance threshold (with a lower and upper bound) for this. Simply put, if the centers of any two superpixels are found to be within this radial distance interval and their characteristic of focus is determined to be 'similar enough' (based on chosen binning methods), those nodes in the network will have an edge between them. Not only does this allow us to more accurately represent the principles, it also grants us the ability to look at both short and long range visual relationships of the image.

We look at three superpixel characteristics for edge creation: pixel size, color, and angular orientation. We call these the *edge creation techniques* (ECTs), and they are treated as hyperparameters when the software is run.

(a) Original Image

(b) Felzenszwalb

(c) Quickshift

(d) SLIC

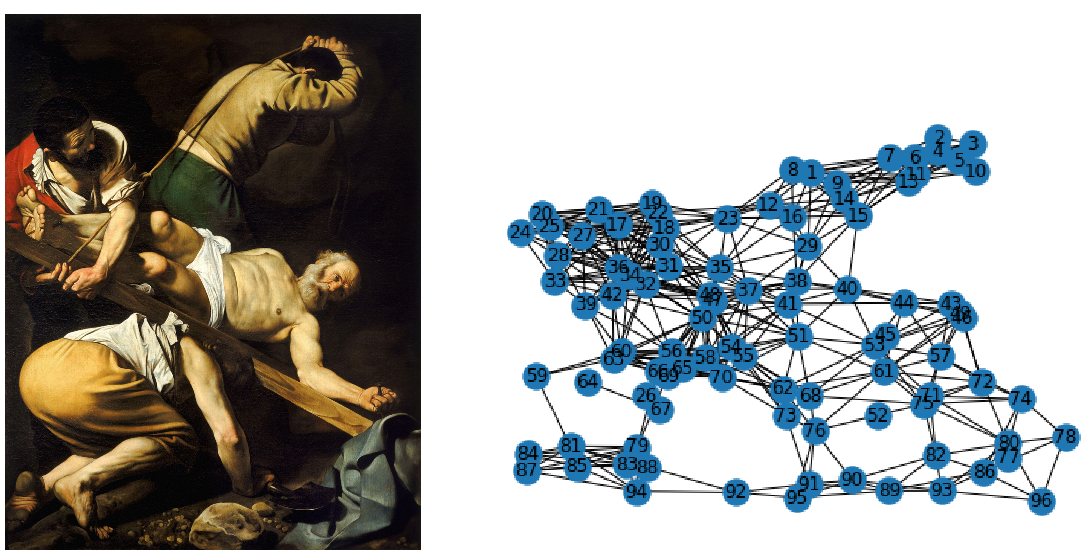
Figure 3.1: Segmentation Example for The Crucifixion of St. Peter by Caravaggio

3.4.1 Pixel Size

The simplest of the ECTs is the pixel size ECT. When this technique is called, the software creates a histogram from the distribution of superpixel pixel sizes (total count of pixels within). The number of bins for the histogram is chosen using Equation 3.1, where B is the number of bins and N is the number of superpixels within the image.

$$B = \text{round}(\sqrt{N}) \tag{3.1}$$

From there, the bin edges determine the similarity threshold in which nodes will have an edge between them. In other words, if and only if any two nodes are found to be within the same bin in the pixel size histogram (and are within the chosen radial distance interval), they will have an edge between them. Our theory is that this ECT will be best at detecting principles created from the shape, space, and form elements. One can find an example network (and original image) for the Pixel Size ECT in Figure 3.2. The nodes in this and the following ECT example figures are generated by the Felzenszwalb segmentation technique.



(a) Original Image

(b) Pixel Size Complex Network

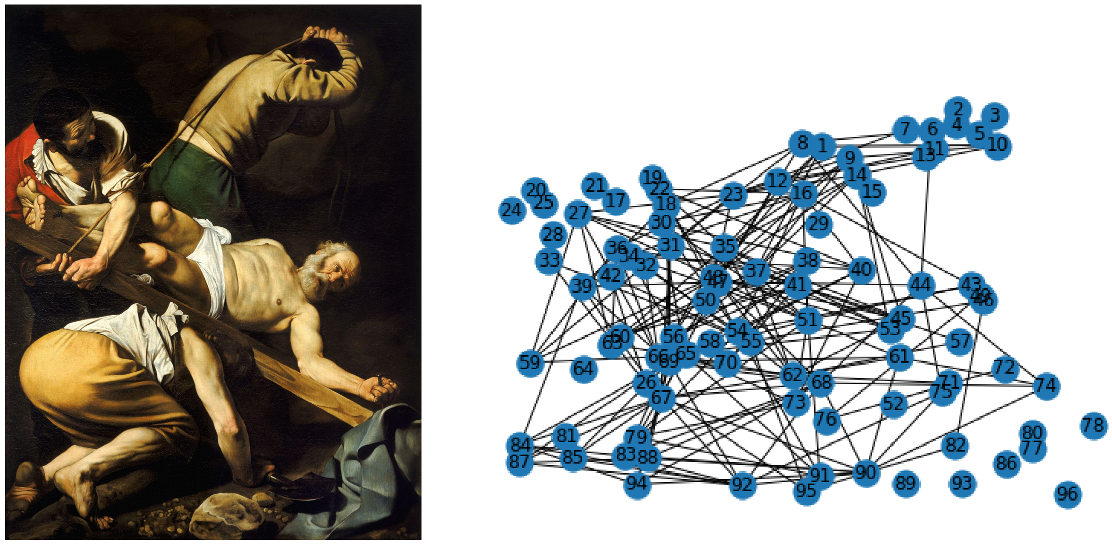
Figure 3.2: Example Image and Pixel Size Complex Network Representation

3.4.2 Color

When called, the color ECT uses the superpixels' average color *identity* to create edges. Inherent to a pixel is its color, given by a three-integer vector. Generally, digital technologies use the RGB color space to describe a pixel, but this space is not intuitive to how a human, and especially not an artist, thinks about color. Hence, after finding the average red, green, and blue value for each superpixel from the individual pixels it contains, we convert to a more intuitive color space, hue-saturation-light (HSL). Then, the software uses a predetermined binning to identify that average HSL color vector. We use three different binning metrics for this ECT, each associated with an axis in the HSL space. The hue binning reflects the 12 primary, secondary, and tertiary colors found on the traditional color wheel. The saturation axis is binned to correlate to high, intermediate, and low saturation, and the lightness axis' binning correlates to dark, precise, and light value[1]. If any two superpixels are found to be within the same bin *for all three HSL axes*, there is an edge placed between their representative nodes. Of course, the radial distance interval applies in this ECT, as well. We predict that the color ECT will excel at detecting principles generated by the color and value elements.

It is important to note that these bins were created by eye in the HSL space, hence introducing a high amount of personal bias into the software. (Recall from Section 2.2 the concepts of the gestalt and simultaneous contrast.) This is the difficulty of combining art, inherently subjective, with science, which requires consistency and reproducibility. Because color is so foundational to the interpretations of art, we found it necessary to bin according to traditional color divisions (at least to the best of our ability).

[1]The exact binning metrics for the color ECT can be found in Appendix B.



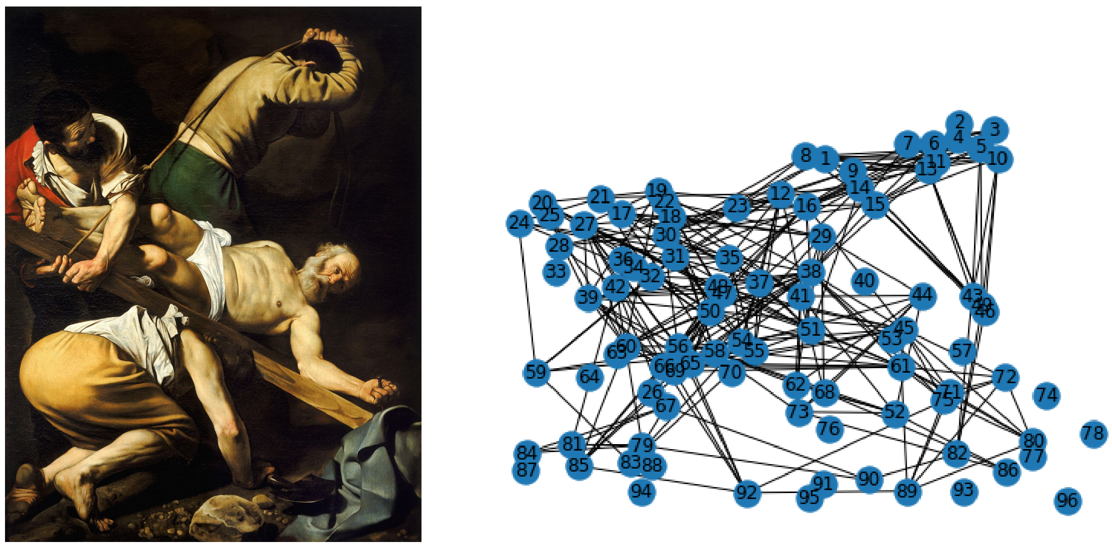
(a) Original Image (b) Color Complex Network

Figure 3.3: Example Image and Color Complex Network Representation

3.4.3 Orientation

The final ECT looks at the relative angular orientation of the superpixels in the image. Specifically, the orientation ECT examines the angle between the rows of the pixels and the major axis of the superpixel's equivalent ellipse. An *equivalent ellipse* is one that has the same second image moments as the superpixel [13]. The angle can be between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$ (other half is ignored due to angular symmetry). As usual, we use a binning metric to create edges between the representative nodes. We divide the angular space evenly into four sections, best described as points on a compass: north-northeast, east-northeast, east-southeast, south-southeast.

We exclude the consideration of angles within $\frac{\pi}{24}$ radians from perfectly horizontal or vertical. We do this to create a differentiation between images within more angular features and those with more horizontal or vertical features. Because the movement principle, a key concept in formalist interpretation, is generally associated with angular features, we wanted an ECT that can detect whether the image contains more or less diagonal features. The exclusion of those angles ensures that the software will not create edges for those superpixels with perfectly or nearly perfect horizontal/vertical angles; thus, images with more movement will have more edges and vice versa. We anticipate that the orientation ECT will be most strongly correlated with the line and shape elements. As usual, an example of this complex network representation can be found in Figure 3.4.



(a) Original Image

(b) Orientation Complex Network

Figure 3.4: Example Image and Orientation Complex Network Representation

3.5 Measurement

Once we create a complete complex network representation of the image, we can take measurements. We have little sense of what (if any) visual forms differentiate between human interpretations; because of this, we must extract those features from what we have–the complex networks. The first step of this *feature extraction* process is to collect a highly descriptive, and even redundant, initial set of measured data. Thus, we instruct the software to take several measurements to ensure we obtain an exhaustive description of the networks. The size of the data set will be quite large because of this, but we will mediate this issue with a dimensionality reduction (the second step of feature extraction), discussed later. Table 3.1 describes each of the measurements taken.

Table 3.1: Network Measurements

Measurement	Type
Degree, k	Distribution
Clustering Coefficient, C	Distribution
Cliques, q	Count, Clique Number
Components, p	Count, Component Number
Assortativity, r	Spectrum [-1, 1]
Degree Centrality, c^d	Distribution
Betweenness Centrality, c^b	Distribution
Closeness Centrality, c^c	Distribution
Bridges, b	Count
Edge Density, σ	Spectrum [0, 1]
Number of Nodes, N	Count

Each of the network measurements in Table 3.1 are commonly found in complex network analysis literature. If the measurement type is distribution, that means that the measurement is microscopic (specific to each element) and requires us to take a distribution over the nodes or edges to get an overall description of the network. For all distribution measurement types, we take the first four moments of the distribution: mean, variance, skew, and kurtosis. A count measurement type means we find the number of the focus concept in the network. A spectrum measurement type means the concept of interest can describe the overall network with one value, with the endpoints of that scale given by the tuple next to it. See [5] for detailed descriptions of each.

3.6 Post-Measurement Processing

After creating and measuring the representative network for each image, using a chosen segmentation and edge creation technique, we are left with a large array of measurements. Eventually, we want the machine to find emergent groupings of the networks, based on that measurement data we found. Before such clustering, though, we need to complete the second step of our feature extraction process by reducing the dimensionality of the data set (while maintaining its core statistical information). To do this, we use *principal component analysis* (PCA). PCA is a linear dimensionality reduction technique. In short, PCA works to find new set of variables from a series of linear combinations of those that already exist, that maximize the variance of the data and are uncorrelated with each other [14]. These new variables–the principal components–are not defined by a priori, but by the data set itself, which makes it quite adaptive and perfect for our use.

Before applying PCA, we must, first, rescale the data within each measurement dimension. Since this tool finds the principal components that maximize variance, having one measurement set with more variance than another could lead the technique to find a false direction of maximal variance, weighted more toward the variable with greater variance. We use a standard scalar to do this, shifting the data to have a mean of 0 and a standard deviation of 1. After rescaling, we perform PCA on the dataset. Instead of choosing an arbitrary number of principal components to reduce to, we allow *scikit-learn* (Python package) to choose the minimum number of components such that a chosen fraction of the original variance (in our case, 0.95) between data is retained.

As stated, reducing the dimensionality of the data sets using PCA essentially creates axes on which the data is maximally interpretable. We capitalize on that variance to take our final processing step, by clustering the data. There are several clustering techniques to use, and each has their own benefits and drawbacks. The algorithm we chose was agglomerative hierarchical clustering. It is important to note that this algorithm does not require number of clusters as a parameter, meaning the algorithm itself determines the optimal number of clusters for the data based on other metrics; we chose a clustering technique with this capability since instructing the algorithm to look for a specific number of clusters would introduce bias and arbitrariness. As mentioned before, we want the software to return an AI perspective on art history, meaning we must allow the model to determine important parameters, such as optimal number of clusters, through and through. After clustering, we construct a table that shows cluster assignments, or what images belong to what clusters. The cluster tables are, then, used for our results analysis.

CHAPTER 4

EVALUATIVE PERFORMANCE AND STATISTICAL ANALYSIS

We now have a clustering–an AI-generated grouping–of images for each radial distance interval, segmentation choice, and ECT combination. We call these groupings *AI-generated movements* (AIGM). These AIGM were generated from examining the digital equivalent of formalist elements and principles only. Recall that our objective is to prove that there is an empirical relationship between the innate and measurable visual features of the images and the critical evaluation of those images (see Section 1.1); to satisfy this objective, we must now find a way to determine if the clusters are interpretatively similar internally and dissimilar externally. To do this, we must collect external data of the painting's critical interpretations, on what the paintings are displaying and/or trying to communicate; we will call this evaluative data *descriptors*. Once we have the decided descriptors of each painting, we can translate over to find the statistically significant descriptors in each cluster and, hence, use those descriptors as a tool to measure the software's ability to categorize critical theory evaluations. In addition, we collect external information for the formalist evaluations of the images for the sake of comparing the divisional power of the software in the two artistic theories.

To determine the descriptors for each painting, we decided to follow an interpretative decision tree in each of the evaluative paradigms (critical and formalist). Visual representations of these decision trees can be found in Figure 4.1 and Figure 4.2. The formalist descriptor assignment process looks at four basic groups of elements of design: shape, space, and perspective; brushstrokes and texture; line, form, and value; and color. For each element group, we ask basic formalist questions. For example, in the shape, space, and perspective element group, we evaluate and record whether an image is balanced in space, the level of movement of the eye through the image, etc. Because these concepts generally lie on a continuous spectrum, i.e., perfectly balanced to perfectly imbalanced, we binned these options to maintain consistency throughout all descriptor assignment. A more detailed explanation of this binning process and descriptor definitions can be found in Appendix C.

Our critical descriptor assignment process breaks the critical evaluation process into discrete steps. For each image, we first ask the level of abstraction. If the painting is given a wholly abstract abstraction level, no other descriptors are given since the critical interpretation is generally too ambiguous and subjective. Then, for all other abstraction levels, we move on to determine the genre of the image. (See Section 2.2 for a list of all possible abstraction levels and genres.) From there, we go into assigning genre-specific descriptors. These are generally broken down into content–what is physically there or happening–and message–what the artist is attempting to say or portray with the painting. Out of all the descriptors, both formalist and critical, we care most about the genre-specific descriptors in our analysis, since we want to prove that there is a machine-measurable relationship between visual form and critical interpretation.

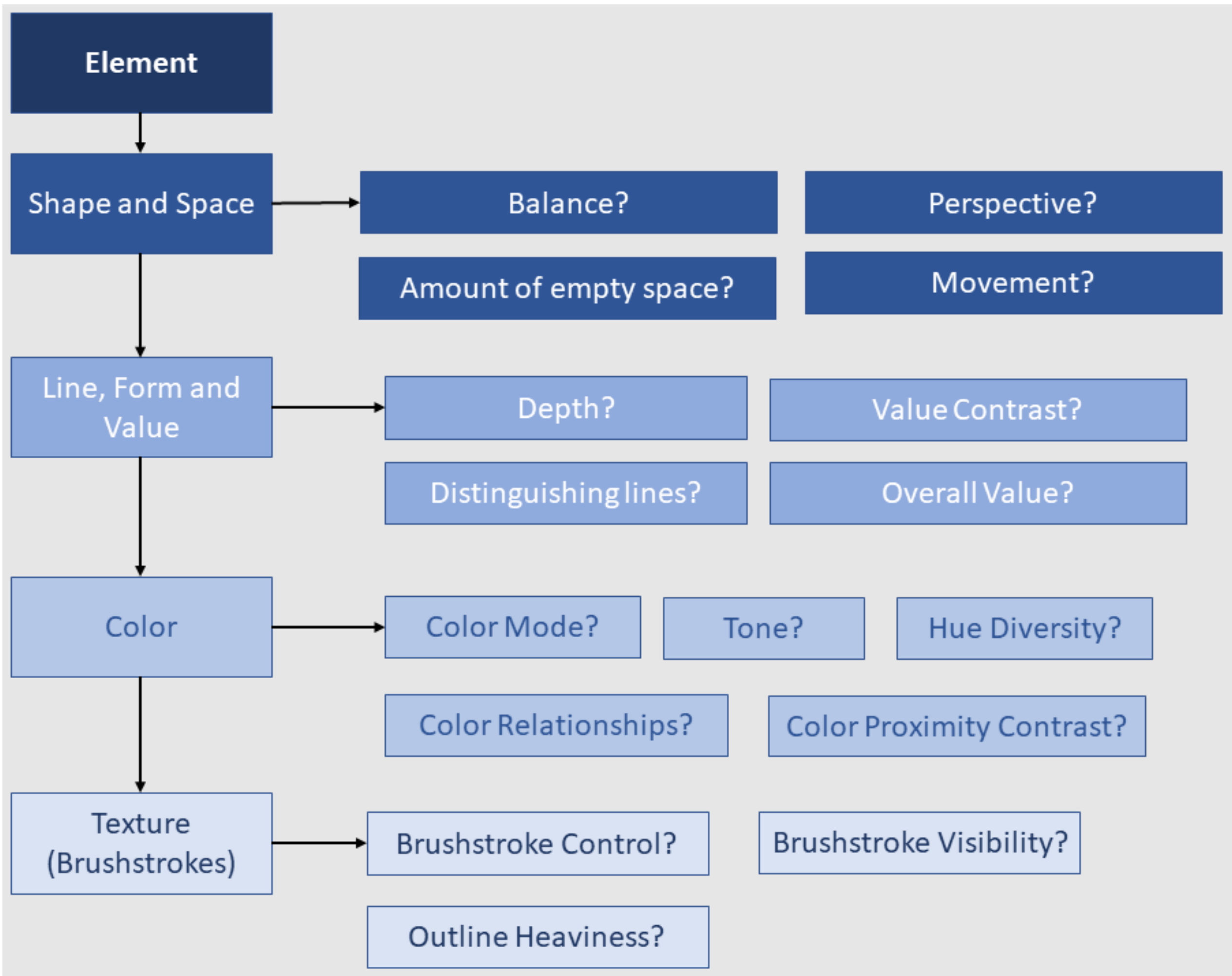


Figure 4.1: Formalist Theory Descriptors Evaluation Tree: This evaluation tree is used to determine the formalist descriptors for each image. We evaluate each painting element by element. We ask a set of questions for each of the element groups listed. The possible answers to each question listed can be found in Appendix C.

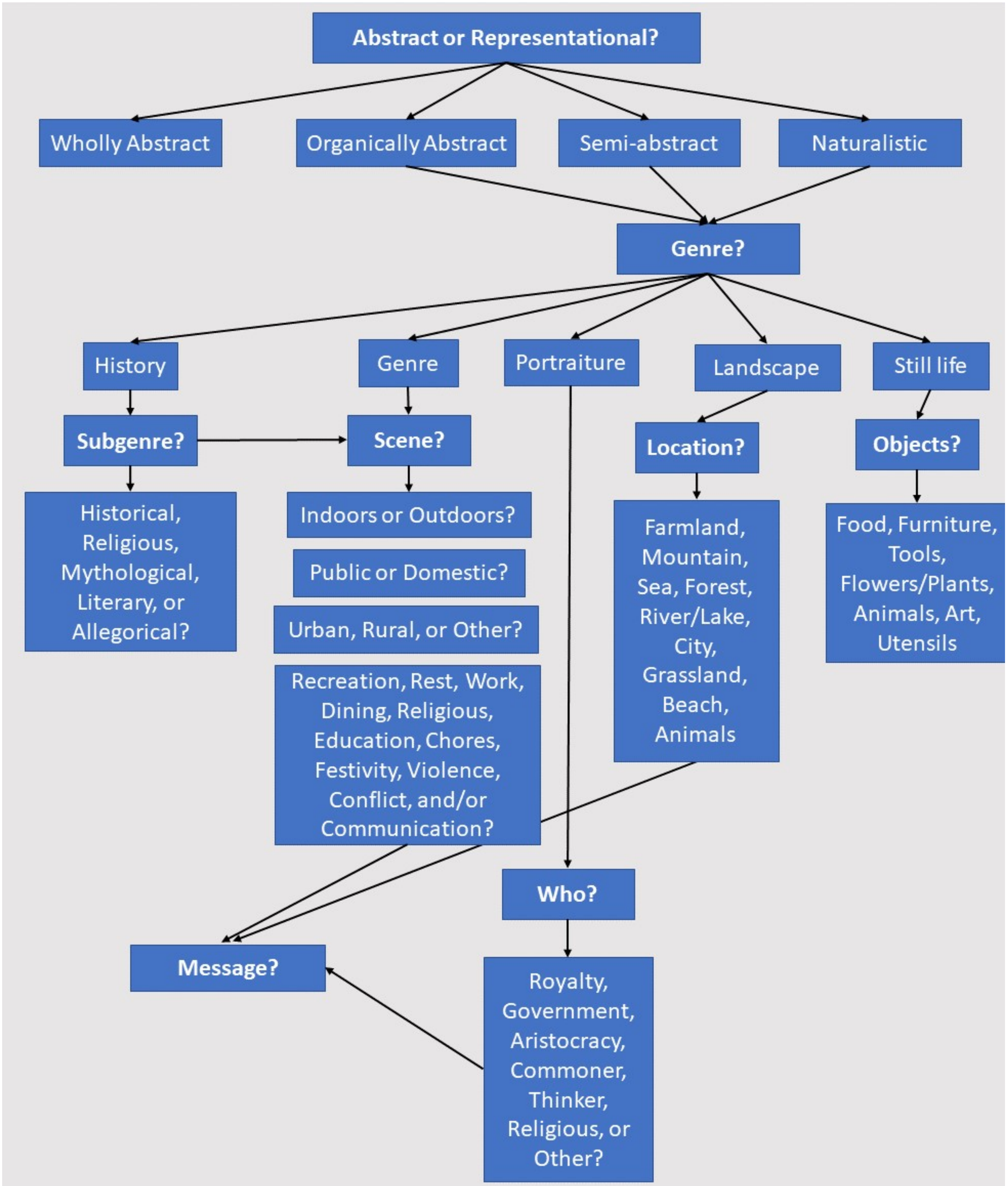


Figure 4.2: Critical Theory Descriptors Evaluation Tree: We follow this evaluation tree to assign appropriate critical descriptors for each painting. We, first, determine the abstraction level of the image. If not wholly abstract, we move onto determine the genre of the painting. Based on that genre, we assign genre-specific descriptors. For all genres, with the exception of still life, we eventually assign message descriptors–the descriptors that we care most about. (Message descriptors are not listed here due to the size of the set. All descriptors are listed in Appendix C.)

It is important to note that the descriptors were decided upon and assigned to the paintings by the researcher, with help from various, publicly available art history resources (particularly [15], [16], and [17]). The researcher is not an art history expert. Hence, there is a significant opportunity for error introduced here. This is discussed more in Chapter 4.5. More detailed explanations of the descriptors as well as an analysis of the overall distribution (over all images) of the descriptors can be found in Appendix C.

Recall that we include randomly pixelated, random shapes, and uniform images in our analysis. For these images, we chose the following descriptive assignments. For the critical descriptors, we simply assigned those images' abstraction levels as wholly abstract. (Wholly abstract images are given no other assignments.) For the formalist descriptors, the images are given their corresponding image type (random, random shapes, or uniform) as that element group's descriptor.

To determine the significant descriptors for each cluster in an iteration[2], we, first, divide the descriptors by type, found in Table 4.1.

Table 4.1: Painting Descriptor Types and Their Corresponding Artistic Theories

Descriptor Type	Artistic Theory
All Formalist Descriptors	Formalist
All Critical Descriptors	Critical
Abstraction Level	Critical
Genre	Critical
Genre-Specific	Critical
Shape, Space, and Perspective	Formalist
Texture and Brushstrokes	Formalist
Line, Form, and Value	Formalist
Color	Formalist

For each type, we find the frequency of occurrence (count) of each descriptor within the cluster. We find the statistical significance of each descriptor by standardizing the count

[2]By iteration, we mean a specific hyperparameter combination, i.e. a specific edge creation technique, segmentation choice, and radial distance interval combination. In other words, an iteration means one run of the software over a set of images.

data: we subtract the mean and divide by the standard deviation of the count distribution for each data point. This standardized data can be read as standard deviations from the mean. We define a significant descriptor as one whose standardized count is 1.5 standard deviations greater than the mean. A visual representation of the significant descriptor process can be found in Figure 4.3.

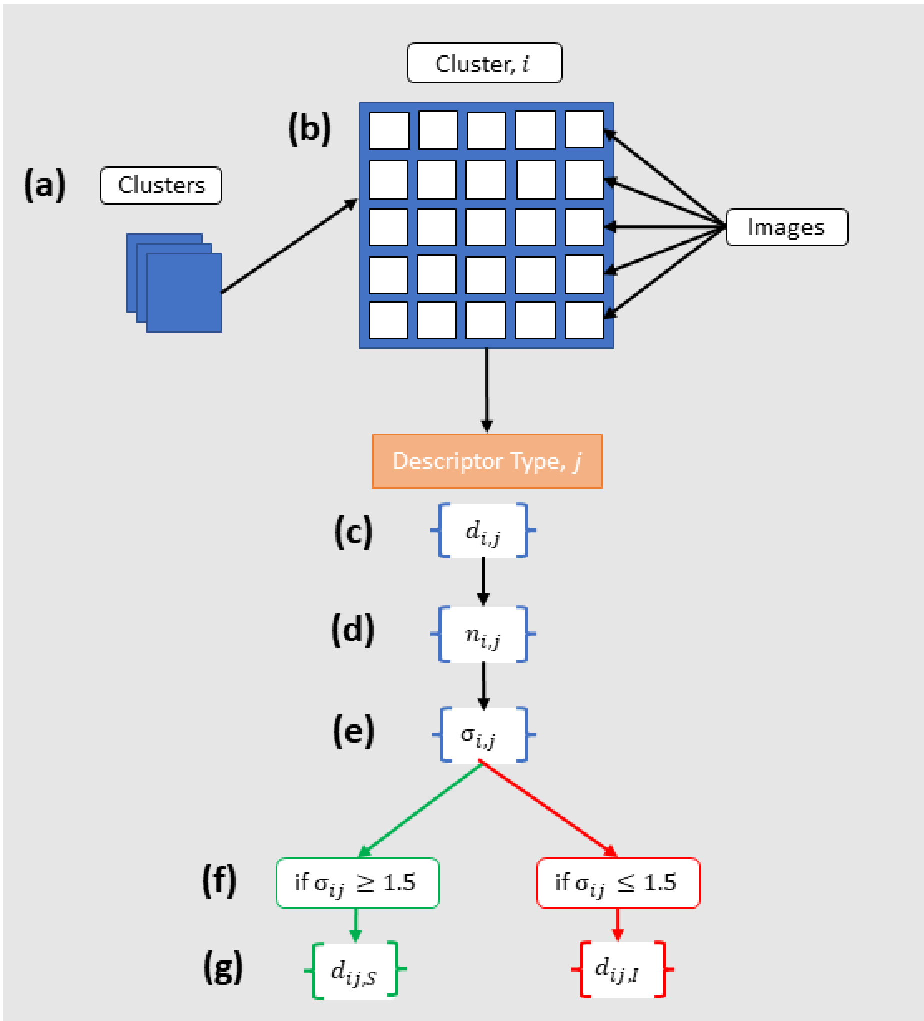


Figure 4.3: Visual Representation of Significant Descriptors Calculation: **(a)** This represents all the clusters for a specific hyperparameter iteration. **(b)** We work with one cluster at a time, finding all descriptors for all images in the cluster. **(c)** First, we split the descriptors in the cluster into their specific types. **(d)** Then, we count all the unique descriptors in that type. **(e)** Next, we standardize the data. **(f)** If the standardized data for a descriptor of any type is greater than 1.5, it is marked as significant (green). Otherwise, the descriptor is insignificant (red). **(g)** Finally, we export the significant and insignificant descriptors of each type for that cluster and move onto the next cluster in the iteration.

Notice that we do not consider the overall (over all images in the dataset) frequency of the descriptor when determining its significance. This could have some effect on the representational accuracy of a significant descriptor. This is because those descriptors with a greater overall representation in the entire data set will more likely have a greater representation in any one cluster, leading the more-frequent descriptors to overshadow those less-frequent descriptors in the significance calculation. We chose to ignore this point for the sake of simplicity and because, when deciding the dictionary of descriptors for assignment, we eliminated any descriptors with an overall count of less than 10.

Table 4.2 displays the decided significant descriptors (for all descriptor types except overall) for the historical movements that we study. After an examination of the historically accepted interpretations of the movements, the significant descriptors decided by our methods tend to be quite accurate to reality, though not perfect. This confirms that both our descriptor assignment process and significance calculations are somewhat reliable, reducing some perceived error. See Appendix A to compare the descriptors in the table to those accepted historical movement evaluations decided upon by traditional schools of art.

Table 4.2: Historical Movement Significant Descriptors

Movement	Significant Descriptors
Renaissance	Naturalistic, Harmonious, Religious, Outdoors, Spirituality, Greatness, Private, History, Distinct Lines, Direct Perspective, Small Empty Space
Baroque	Naturalistic, Blended Color, Indoors, Private, Public, Morality, Religious, Conflict, Drama, History, Distinct Lines
Rococo	Naturalistic, Harmonious, Private, Recreation, Joy, Abundance, Outdoors, Rural, Genre, Distinct Lines
Neoclassicism	Naturalistic, Blended Color, Outdoors, Conflict, Communication, Drama, Private, Antiquity, Morality, History, Distinct Lines, Intermediate Perspective
Romanticism	Naturalistic, Outdoors, Rural, Conflict, Nature, Public, Violence, Drama, Government, Distinct Lines, Highly Dynamic
Realism	Naturalistic, Genre, Distinct Lines, No Outlines
Impressionism	Capture of Conditions, Harmonious, Outdoors, Public, Recreation, Occupation of Time, Urban, Intermediate Perspective, Small Empty Space, No Outlines
Post Impressionism	Harmonious Color, Expressionistic, Outdoors, Public, Rural, Occupation of Time, Nature, Private, Genre, Distinct Lines, Highly Dynamic
Fauvism	Deconstruction, Lively, Color, Texture, Beauty, Landscape, Visible Brushstrokes
Expressionism	Semi-Abstract, Expressionistic, Blended Color, Outdoors, Public, Color, Genre, Direct Perspective, Mix Controlled Brushstrokes
Cubism	Expressionistic, Color, Deconstruction, Flat Depth, Direct Perspective, Mostly Controlled Brushstrokes
Surrealism	Harmonious, Blended Color, Outdoors, Private, the Mind, Deconstruction, Genre, Distinct Lines, Highly Dynamic, Intermediate Perspective, Blended Brushstrokes

Once we have a list of all significant and insignificant descriptors per cluster for each type, we are ready to evaluate the performance of our software.

4.1 Precision and Recall

We now have a list of significant and insignificant descriptors for each cluster. To quantify our software's ability to find pronounced and compelling groupings of paintings with a common critical interpretation, we need to contrast the significant descriptors to all those that are present. More specifically, we would like to look at the number of images within the cluster that are defined by its significant descriptor(s) and compare that to both those outside of that cluster, and to the number of images held within that are without the significant descriptor(s). To do this, we use the precision and recall metrics. These measures are commonly used to gauge performance in classification problems.

Precision and recall generally calculate the precision and sensitivity of machine-assigned labels. They look at the ratios of relevant (assigned for a group) and retrieved (all in a group) labels. Specifically, precision asks how many of the *retrieved* instances are *relevant*, and recall calculates how many *relevant* items are *retrieved.* Both measures are calculated over all groups in the set (clusters in an iteration) through averaging and weighted averaging. To do this calculation, the metrics require us to mark the data as *true positives* (TP), *false positives* (FP), or *false negatives* (FN). The TP are the number of relevant instances retrieved per group, FP are the number of retrieved instances in a group that aren't relevant, and FN are the number of relevant instances not retrieved per group. For our purposes, the TP will be those images that are defined by the significant descriptor of the cluster they're contained within. The FP will be the images within the cluster that aren't defined by its significant descriptor. The FN will be those images defined by the significant descriptor of one cluster but are contained in another cluster (unless that other cluster has the same significant descriptor). The expressions used to find precision and recall can be found in Equations 4.1 and 4.2. The measures can range from 0 to 1.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.2}$$

For a particular iteration, we calculate the average and the weighted average of the precision and recall over all the clusters present (for a particular descriptor type). For the weighted average, we use the cluster size frequency (cluster size divided by the total number of images) as the weight. The expressions for these can be found in Equations 4.3

and 4.4, where K is the number of clusters present, c_i is the size frequency of cluster i, and $(P/R)_i$ is the precision or recall of cluster i.

$$\text{Average} = \frac{\sum_{i=1}^{K} (P/R)_i}{K} \tag{4.3}$$

$$\text{Weighted} = \sum_{i=1}^{K} c_i (P/R)_i \tag{4.4}$$

4.2 Null Hypothesis and Bootstrap Testing

We use precision and recall to determine how good our software is at creating distinct critical theoretic divisions in the paintings. But what does "good" mean in our case? We have no means of determining how well the software performs in terms of our objective without a basis of comparison. To give the returned precision and recalls from our software relevance, we must compare to the performance measures of an arbitrary grouping of paintings. Put another way, we need a statistical authentication that our software creates groupings of paintings (AIGMs) that are more critically distinct than a random grouping of the same paintings. Hence, for this project's statistical analysis, we decided to perform a null hypothesis test through bootstrapping and comparing real (from our software) and random precision and recall distributions.

Our null hypothesis states that our software's precision and recall (for either average or weighted) generally performs worse or the same as a random clustering of the sampled images and that the real and random precision and recall distributions come from the same underlying distribution. Our alternative hypothesis states that the two distribution sets do not come from the same underlying distribution, *and* the real data consistently performs better than the random data. By rejecting our null hypothesis, we would obtain a statistical confirmation that our software has created groupings of the paintings that have a non-trivial common contextual/critical interpretation, beyond what would be possible with a random clustering of the same images. Because our software creates clusters based strictly on inherent and measurable visual characteristics, this would essentially mean that there does exist an empirical relationship between visual form and critical interpretations and that our software extracts that relationship, at least to some extent. All in all, a rejection of our null hypothesis would confirm our theory outlined in research question one in Section 1.1.

To reject or sustain the null hypothesis, we require large amounts of data. This is where we implement the bootstrapping technique. For a particular sample size M, we sample M images with replacement from the total set of 734 images (including the randomly pixelated, random shapes, and uniform images). We create the complex network representations, take measurements, and cluster the data for all hyperparameter iterations. Then, for all those hyperparameter iterations, we calculate the significant descriptors and the average and weighted precision and recall *for each descriptor type* listed in the first section of this chapter. This data is recorded, and we repeat this process 99 more times. (Note that with each trial, we resample M new images with replacement.)

After collecting all performance data for the real case, we do the same for a random condition. For each trial, we use the same set of sampled images as in the real case. For each hyperparameter iteration in that trial, we find the average cluster size, the cluster size standard deviation, and the number of clusters of the real data cluster assignments. We create a normal distribution from the mean and standard deviation and decide cluster sizes from that (same number of clusters from the real data). Then, we shuffle the sampled images, and assign each image to a random cluster, maintaining the cluster sizes obtained from the normal distribution. From this random cluster assignment, we follow the same analysis steps as with the real data: we find significant descriptors and average and weighted performance measures for each descriptor type, record the data, and repeat the process 99 more times.

From the different trials, we acquire distributions of the performance measures that we can use to compare our software's performance to the random performance, i.e., to conduct our null hypothesis test. To do this comparison, we split the data–both real and random–by hyperparameter focus and descriptor type. We split by hyperparameter for two reasons. First, if we don't split like this, we risk washing the data. The hyperparameters treat the image data differently, and if we were to look at the distributions over all hyperparameters, any possible distinction from random may be lost in the data. Second, we split by hyperparameter because we want to discover if there is a way for a machine to catalog interpretative concepts in art strictly using visual information. Treating all hyperparameters as one would eliminate our ability to determine which visual factors (i.e., which hyperparameters) play a key role in this. We split by descriptor type, again, because we want to see what evaluations our methods perform best in.

After splitting the entire 100-trial data set by hyperparameter focus and descriptor type, we examine the real and random distributions together. First, we look at the histograms and density curves of both data sets for each performance measure[3]. We do this to get a visual idea of performance and distribution shape, as well as to obtain the samples' probability distributions for the subsequent calculation. Then, we perform a Two-sample Kolmogorov-Smirnov (KS) Test on the two probability distributions just obtained. This test is a nonparametric statistical test that answers the question of whether two samples could've been drawn from the same underlying probability distribution; this is perfect for our use, since it answers one of the questions of our own null hypothesis. The KS Test uses the empirical cumulative distributions, $F_{1,n}$ and $F_{2,m}$, of the samples to determine rejection of the null hypothesis: whether the two distributions come from the same (but unknown) distribution. The KS statistic is given in Equation 4.5, where n and m are the sizes of the first and second samples, respectively, and $\sup_x$ is the supremum function.

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| \tag{4.5}$$

The null hypothesis for the KS Test is rejected at level α if Equation 4.6 is satisfied.

$$D_{n,m} > \sqrt{-\ln\left(\frac{\alpha}{2}\right)\frac{1+\frac{m}{n}}{2m}} \tag{4.6}$$

[3]Keep in mind that the performance measures we look at average recall, average precision, weighted recall, and weighted precision.

We also determine the p value with this calculation, interpreted as how likely it is to have found the real distribution if the null hypothesis were true. If found to be less than 0.05, it is generally accepted that the two distributions are not from the same underlying distribution, and the null hypothesis is rejected.

We performed our statistical analysis–from data collection to p value–for a sample size of 100 images. Summary plots for the KS statistics for all performance measures for the 100 samples sample size can be found on the following pages. Note that the α level we chose for our KS test is 0.05. The descriptor type index key for the plots can be found in Table 4.3.

Table 4.3: Descriptor Type Index Key for KS Test Summary Plots

Descriptor Index	Descriptor Type
0	All Formalist Descriptors
1	All Critical Descriptors
2	Abstraction Level
3	Genre
4	Genre-Specific
5	Color
6	Line, Form, and Value
7	Shape, Space, and Perspective
8	Texture and Brushstrokes

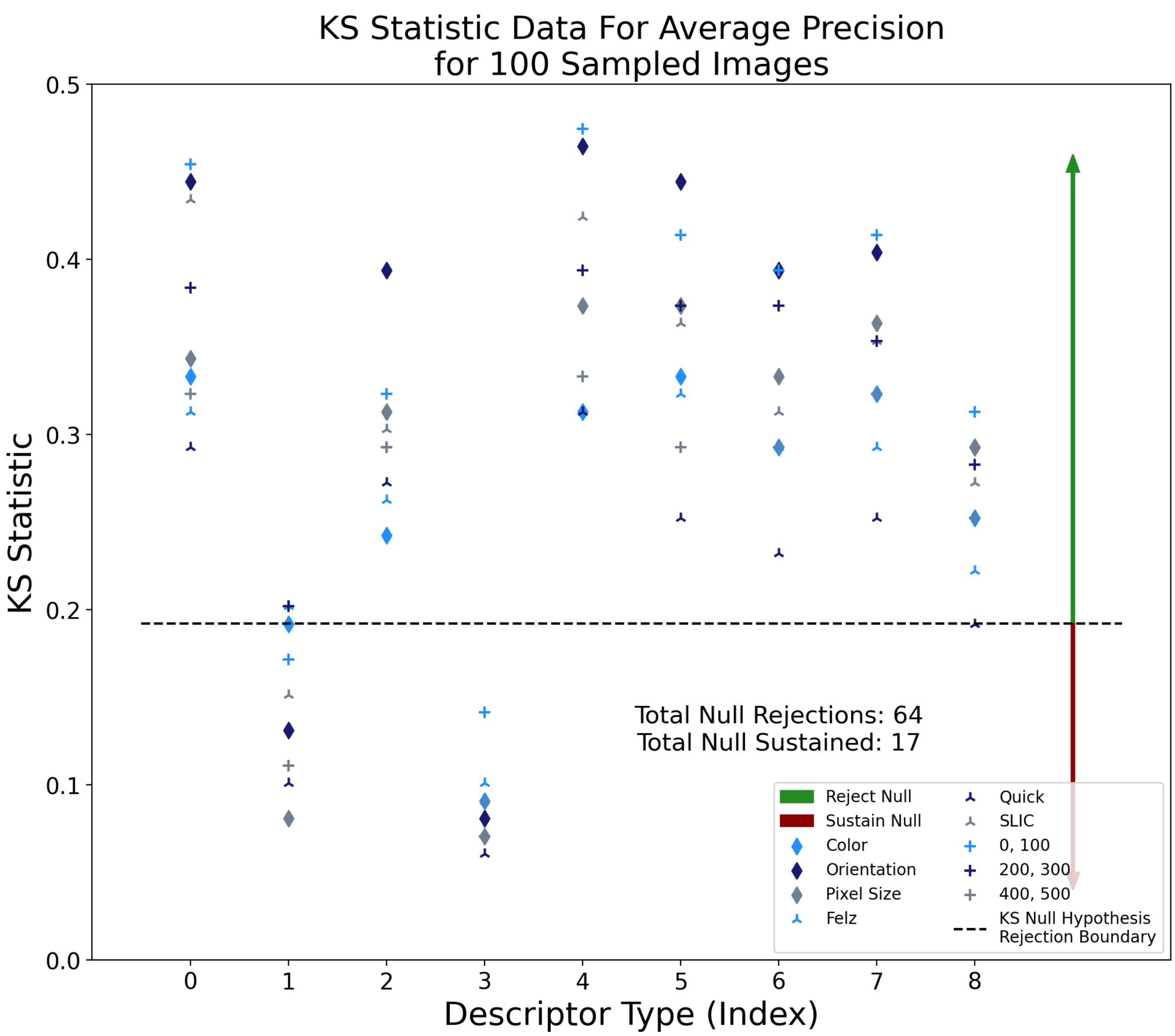


Figure 4.4: Kolomogorov-Smirnov Two-Sample Test Summary for Average Precision

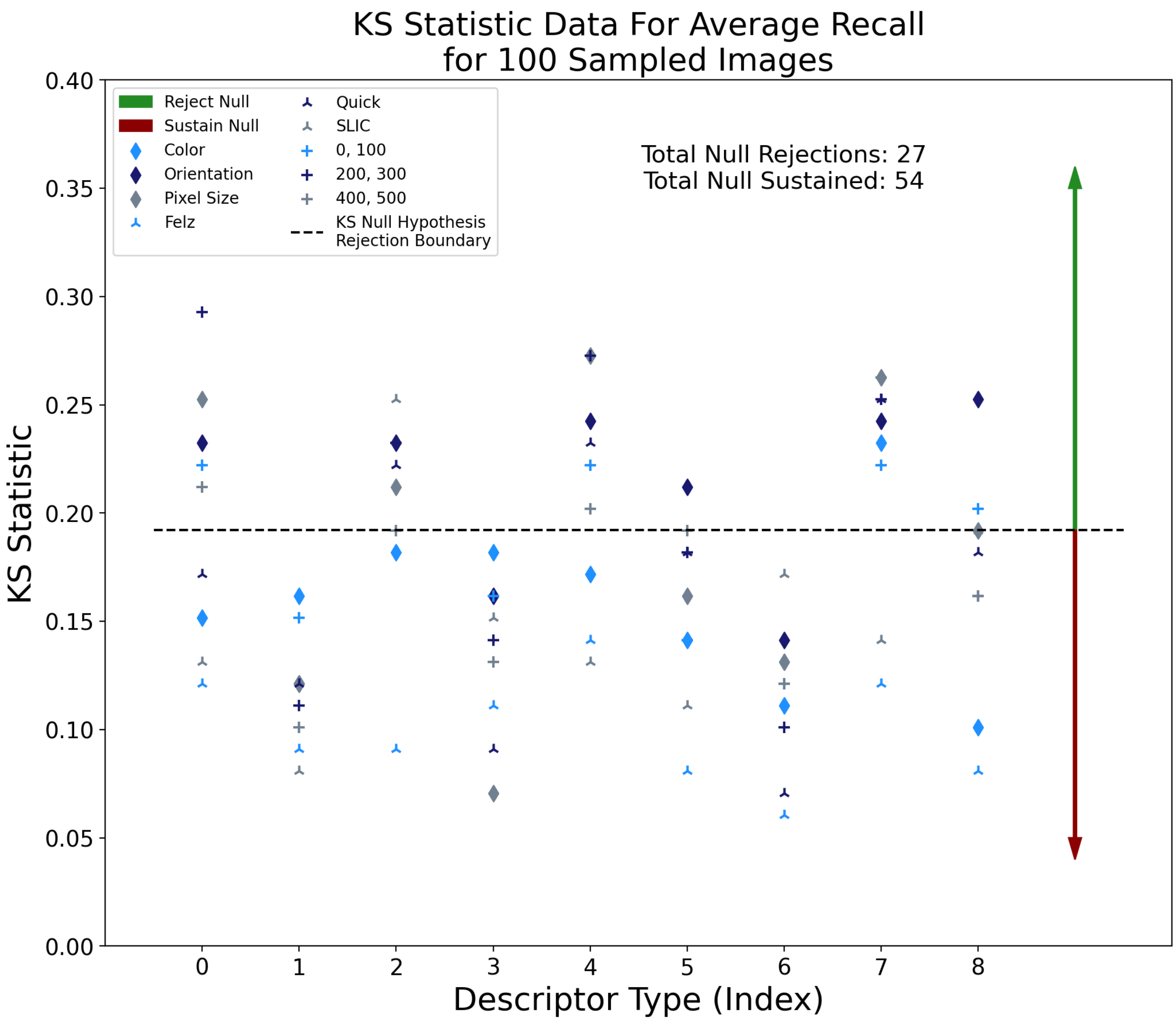


Figure 4.5: Kolomogorov-Smirnov Two-Sample Test Summary for Average Recall

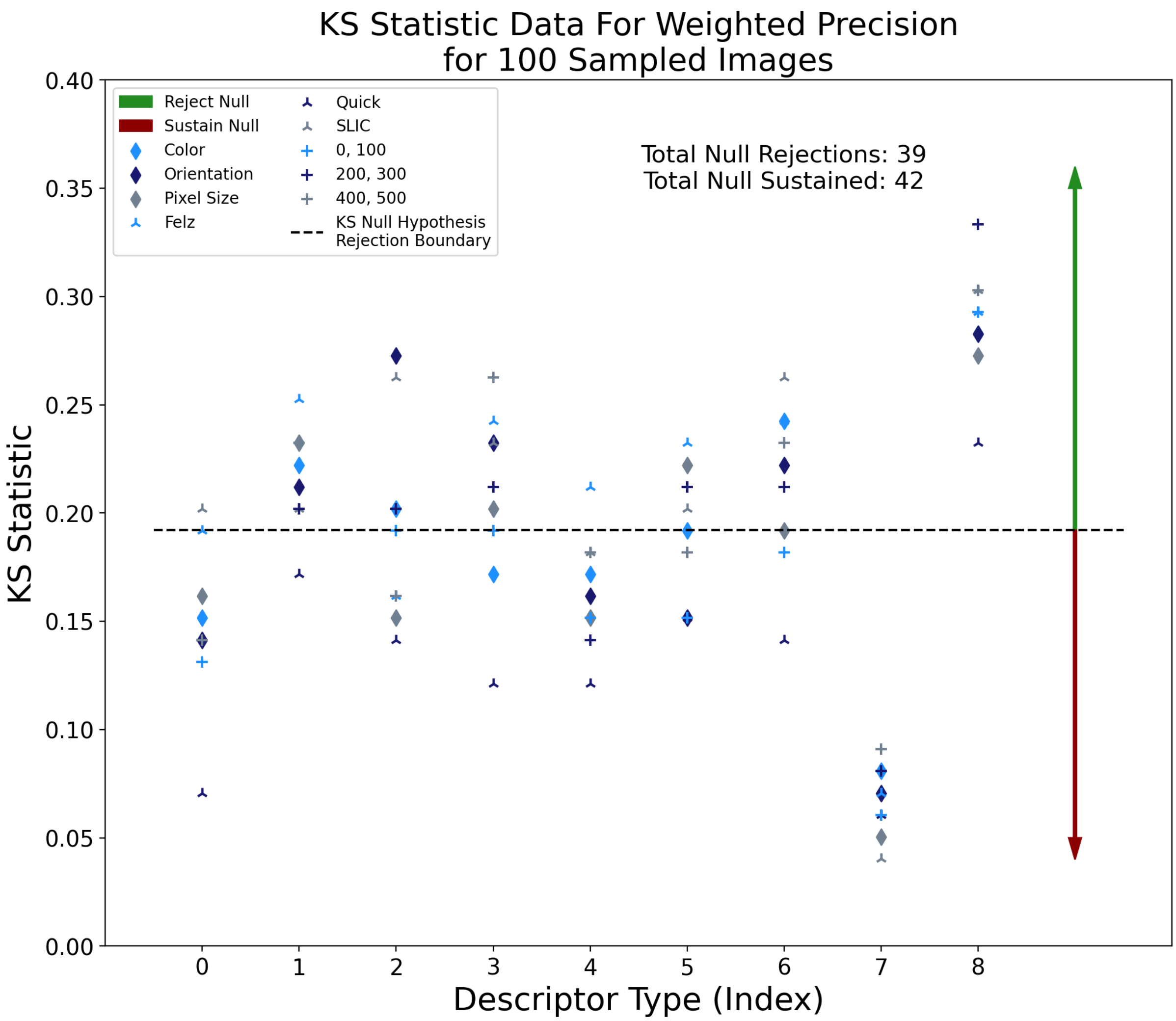


Figure 4.6: Kolomogorov-Smirnov Two-Sample Test Summary for Weighted Precision

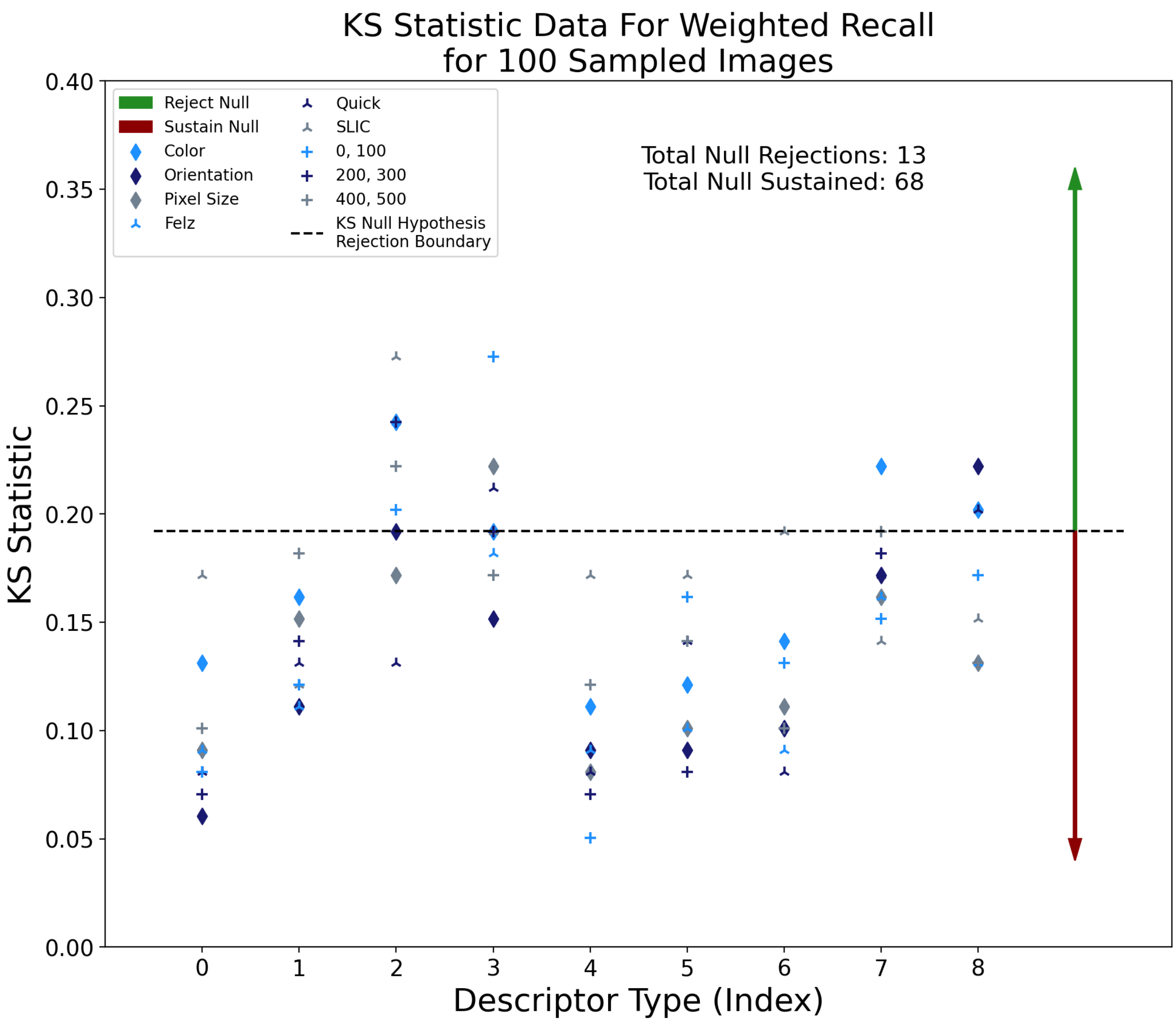


Figure 4.7: Kolomogorov-Smirnov Two-Sample Test Summary for Weighted Recall

As mentioned previously, the KS test only answers one part of our null hypothesis–whether the random and real data come from the same underlying distribution. We want to find those hyperparameter focus/descriptor type combinations that perform better than random *in addition* to having statistically distinct distributions. We define better performance as those combinations whose mean performance measure is greater than that of the random data. All combinations that pass the KS test and whose real data mean is greater than that of the random are considered to reject our null hypothesis.

A comprehensive table for those combinations that reject the null for each sample size can be found in Appendix D. Table D.1 in said Appendix shows that a significant amount of our combinations indeed reject our null hypothesis, showing that our methods are successful in distinguishing critically distinct and non-trivial clusters of images, strictly using visual form. Out of all descriptor types, abstraction level, genre-specific, and, in particular, shape and space elements tend to perform best (i.e., have the most hyperparameter counterparts that reject the null).

In addition to looking at specific hyperparameter/descriptor type combinations, we would like to examine the performance of the critical descriptor types in comparison to the formalist descriptor types over all hyperparameters. Table 4.4 displays the fractions of null rejecting cluster systems for each performance metric. We have split the descriptor types by their corresponding theory; the types and their corresponding art theories can be found in Table 4.1. We combined the hyperparameter focuses in the fraction calculation.

Table 4.4: Fraction of Null-rejecting Iterations Split by Artistic Theory

Performance Metric:	Average Precision	Average Recall	Weighted Precision	Weighted Recall
Formalist Descriptors	0.98	0.34	0.08	0.02
Critical Descriptors	0.53	0.34	0.14	0.25

Table 4.4 shows that the average precision performance metric tends to find the most AIGM that reject our null hypothesis. In other words, on average, we find more interpretative distinction, in both formalist and critical paradigms, between clusters when looking through the lens of average precision. This table also shows that the formalist descriptor types tend to have better performance than the critical descriptor types. We discuss this more in Chapter 5. Ultimately, though, non-trivial fractions of all the descriptor types analyzed do end up rejecting our null hypothesis, confirming this project's success in terms of our objective.

As a final analysis, we would like to inspect the absolute performance metrics of each hyperparameter/descriptor type combination that rejects the null hypothesis, rather than their performances relative to their random counterparts. Table 4.5 displays the average performance metrics for the different null-rejected combinations. Due to the large volume of passing combinations, we only display those performance metrics greater than 0.6. Recall that all performance metrics range from 0 to 1, where 1 would be perfect performance. Thus, Table 4.5 shows the most successful hyperparameter/descriptor type combinations. Appendix D contains the probability distributions of the mean for the (null-rejecting) performance measures.

Table 4.5: Average Performance Metrics for Most Successful Hyperparameter/Descriptor Combinations

Hyperparameter Focus	Descriptor Type	Performance Metric	Average Value
Orientation ECT	Color	Average Recall	0.605
Pixel Size ECT	Genre-Specific	Average Recall	0.621
0 to 100 Radial Interval	Texture and Brushstrokes	Average Recall	0.628
400 to 500 Radial Interval	Genre-Specific	Average Recall	0.631
200 to 300 Radial Interval	All Formalist Descriptors	Average Recall	0.632
Orientation ECT	All Formalist Descriptors	Average Recall	0.643
200 to 300 Radial Interval	Genre-Specific	Average Recall	0.657
0 to 100 Radial Interval	All Formalist Descriptors	Average Recall	0.666
Orientation ECT	Genre-Specific	Average Recall	0.672
0 to 100 Radial Interval	Genre-Specific	Average Recall	0.686

Notice that genre-specific descriptor type, the type that contains the content and message descriptors for the critical theory paradigm, appears frequently in the table of top performers. On top of that, the corresponding average values for the descriptor fall in the upper half of possible performance, indicating successful performance in terms of evaluative distinguishability. All this together suggests that, beyond the fact that the empirical relationship between measurable visual characteristics and the quantitative, communicative, and meaningful characteristics of art does exist (which was confirmed by a rejection of our null hypothesis), our proposed methods allow a machine to pick up on and capitalize on that relationship to some level of success[4].

For the sake of comparison, Table 4.6 displays the top performance metrics for the historical movements that we analyze in this project. For the sake of space, we only display those metrics with a value greater than 0.4. Note that these metrics were calculated in the same way as with the clustering data.

Table 4.6: Most Successful Descriptor Types for the Historical Movements

Descriptor Type	Performance Metric	Value
Line, Form, and Value	Average Recall	0.416
Line, Form, and Value	Weighted Recall	0.417
Color	Average Recall	0.417
Color	Weighted Recall	0.418
Color	Average Precision	0.422
Color	Weighted Precision	0.422
All Formalist Descriptors	Average Recall	0.432
All Formalist Descriptors	Weighted Recall	0.432
Genre-Specific	Average Precision	0.440
Genre-Specific	Weighted Precision	0.441
Genre-Specific	Average Recall	0.498
Genre-Specific	Weighted Recall	0.499

Table 4.6 shows that, though the genre-specific descriptor does consistently appear in the top performances for the historical movements, the actual performance does not reach quite to the level of that of the our software's returned data, displayed in Table 4.5. Firstly, the fact that our genre-specific descriptor type appears consistently in the top performance for the historical movements confirms that our descriptor assignment method (described in the first half of this chapter) is sufficient. In critical theory, the definition of historical movements is the collections of paintings with common philosophies or messages. The genre-specific descriptors hold the philosophical ideas of the paintings. Hence, having that descriptor type as a top performer–meaning that it is best differentiated in the

[4]Though Table 4.5 only shows average recall metrics, this is only because average recall happens to be the only metric that obtains values greater than 0.6. We only showed those metrics greater than 0.6 for the sake of space. There exist examples of other performance metrics that perform better than 0.5. See Appendix D.

groupings of images–confirms that our descriptor assignment is in some measure accurate.

Secondly, the fact that our (non-random) clustering data returns performance metrics that consistently perform *better* than the historical movements, especially in terms of the genre-specific descriptor types, shows that our software creates AIGM *more* interpretatively distinct than the critically accepted historical movements. This has great implications in terms of our second research question. We will discuss this in more detail in Chapter 5.

Recall that in Chapter 3 we predicted that certain ECTs would perform best with certain formalist elements. We postulated that the Pixel Size ECT would perform best with shape, space, and form, the Color ECT with color and value, and the Orientation ECT with Line and Shape. Table 4.7 shows the mean precision metrics for those (null-rejecting) combinations for our sample size of 100 images.

Table 4.7: Null Rejecting Formalist Descriptor Types For the Edge Creation Techniques

ECT	Descriptor Type	Performance Metric	Mean
Color	Shape, Space, and Perspective	Average Precision	0.333
		Weighted Recall	0.508
		Average Recall	0.525
	All Formalist Descriptors	Average Precision	0.361
	Color	Average Precision	0.368
	Line, Form and Value	Average Precision	0.385
	Texture	Average Precision	0.394
Orientation	Shape, Space, and Perspective	Average Precision	0.317
		Average Recall	0.500
	Color	Average Precision	0.364
		Average Recall	0.605
	All Formalist Descriptors	Average Precision	0.365
		Average Recall	0.643
	Line, Form, and Value	Average Precision	0.373
	Texture	Average Precision	0.394
		Average Recall	0.593
Pixel Size	Shape, Space, and Perspective	Average Precision	0.320
	Color	Average Precision	0.364
	All Formalist Descriptors	Average Precision	0.372
		Average Recall	0.599
	Line, Form, and Value	Average Precision	0.377
	Texture	Average Precision	0.391

Table 4.7 rejects the predictions we made. None of the ECTs display best performance for the formalist descriptor types we predicted. This shows us that, though we based our methods on the ideas of formalist elements, the mapping we created in our software does in edge creation not totally embody those ideas.

4.3 A Case Study: New Insights to Expressionism

To show the evaluative benefit of our software, this section presents a brief case study of a cluster system returned from our methods. The set of AIGM that we will examine was created (from a sample of 100 images) with the Color ECT, Quickshift segmentation, and the 0 to 100 radial distance interval. Table 4.8 shows the performance metrics for all descriptor types for this AIGM set.

Table 4.8: Case Study Performance Metrics

Descriptor Type	Average Precision	Average Recall	Weighted Precision	Weighted Recall
All Formalist Descriptors	0.707	0.919	0.405	0.919
All Critical Descriptors	0.104	0.319	0.212	0.656
Abstraction Level	0.723	0.632	0.470	0.458
Genres	0.0	0.0	0.0	0.0
Genre-Specific	0.719	0.907	0.419	0.899
Color	0.660	0.669	0.322	0.451
Line, Form, and Value	0.729	0.950	0.448	0.993
Shape, Space, and Perspective	0.653	0.768	0.288	0.625
Texture and Brushstrokes	0.700	0.950	0.399	0.993

This example was chosen because of its relatively high performance across all performance metrics. Note that, when compared to Table 4.6, this iteration performs significantly better than the historical movements with regard to the genre-specific descriptor type. As stated in the previous section, a higher performance in this type means that the AIGM are more interpretatively distinct than the critically accepted historical movements. Essentially, a higher performance indicates an informational gain in terms of the critical evaluation of the new groupings of images. This leads us to ask: what information are we gaining from this new representation of the paintings? We must look at the clustering assignments to answer this.

Table 4.9 shows how the sampled images' corresponding historical movements are split into the clusters. Note that, in Table 4.9, we are only displaying the movement count for clusters 0 and 2; we do this because clusters 1, 3, and 4[5] do not contain any of the historical paintings, only those from the simple movements[6]. In fact, the simple movements' images are perfectly divided into clusters that only contain that type. That is, cluster 1 only contains random shape images, cluster 3 only contains uniform images, and cluster 4 only contains randomly pixelated images. This tells us that our software can not only distinguish between different orders of random information (random images), but between that and ordered (uniform images) and/or complex (historical paintings) information, as well, at least to some extent[7].

[5] There are five clusters in total in this iteration.

[6] Recall that the simple movements are those containing randomly pixelated, random shapes, or uniform color images.

[7] We can only say this to some extent because we did not perform a statistical confirmation of this statement. This is a good sanity check for us, though.

Table 4.9: Case Study Movement Divisions

Movement	Cluster 0	Cluster 2
Romanticism	3	2
Cubism	3	3
Neoclassicism	8	2
Impressionism	6	1
Expressionism	3	5
Rococo	3	3
Realism	6	1
Baroque	5	2
Surrealism	4	3
Renaissance	3	4
Post Impressionism	2	1
Fauvism	0	8

As Table 4.9 shows us, the historical movements are spread between the two non-simple clusters, with the exception of the Fauvism movement. The similar distributions between the two clusters shows us that our methods are not finding critical distinction along the same lines as the historical movements. The fact that (almost) all movements are spread among both clusters tells us that our software has found an alternative way of diving the images, of making critically distinct groupings of the paintings, to the traditional, qualitative analyses done in art theory. To give tangible evidence to this claim, we would like to compare the paintings in either cluster that come from a single historical movement. Doing so will allow us to inspect the new qualitative division–i.e., the new lens through which we can examine art–that our software has made. Table 4.10 gives a brief summary of the accepted critical evaluations of those paintings from the Expressionism movement in either cluster, found in Figure 4.8[8].

[8]Because these images are used strictly for non-profit and academic purposes and because they are used sparingly, they fall under Fair Use copyright policy.

(a) Cluster 0

(b) Cluster 2

Figure 4.8: Paintings from the Expressionism Movement in Clusters 0 and 2

Table 4.10: Evaluations for Expressionism Paintings in Clusters 0 and 1

<table>
<tr><th>Painting</th><th>Interpretation</th></tr>
<tr><td colspan="2">Cluster 0</td></tr>
<tr><td>The Red Studio - Matisse - 1911</td><td>A play with the spatial composition of an image; Deconstruction of spatial illusion</td></tr>
<tr><td>Harmony in Red - Matisse - 1908</td><td>Deconstruction of linear space; flatness</td></tr>
<tr><td>Street, Dresden - Kirchner - 1908</td><td>Observation of urban bustle; A deconstruction of the jarring experience of modern urban life</td></tr>
<tr><td colspan="2">Cluster 2</td></tr>
<tr><td>Grazing Horses IV - Marc - 1911</td><td rowspan="3">Animals represent human spirituality; the eternality of nature and our human relationships with it</td></tr>
<tr><td>The Little Blue Horse - Marc - 1911</td></tr>
<tr><td>Dog Lying in the Snow - Marc - 1911</td></tr>
<tr><td>Dance of Life - Munch - 1900</td><td>The cycle of life; ”life’s dance”</td></tr>
<tr><td>Death and Fire - Klee - 1940</td><td>A painting of his own death; abstraction of sad mortality</td></tr>
</table>

A quick look at Table 4.10 gives insight to the new interpretative lens that our software offers. Though both groups belong to the Expressionism art periods, our methods have found key communicative differences between the two groups of paintings. Though the main ideas may be different, all the paintings in cluster 0 have a common sub-theme of deconstruction, whether that be of linear space, like with the Matisse paintings, or of a concept, like with the Kirchner painting. Cluster 2, on the other hand, tends to deal more with spirituality and mortality. Though not obvious upon first inspection, each painting in cluster 2 has a core element that deals with the human relation with life and death. In short, there is a key interpretative difference between the Expressionist paintings in cluster 0 and cluster 2. All in all, this newfound interpretative difference demonstrates that our software has taken paintings that historically belong to the same group and has found finer critical differentiation within. In other words, our methods have allowed us to gain a non-obvious critical perspective to art.

4.4 Two-Point Correlations

As part of our analysis, we thought to examine the two-point correlations of images within the clusters. In other words, we want to ask how often any two ”similar” images are found in the same cluster. To do this, we need to determine what ”similar” means. Looking only at the overall descriptor types (all formalist and all critical descriptors), we decided that the similar images are those with the most descriptors in common in either overall descriptor type. We set the cut off for similarity as the 50 most common pairs of images in the type. Note that the total set of similar images for either type does not necessarily mean all images are equally similar with every other image; just the image pairs we collected are highly similar (as we define it). We do this to have some sort of image differentiation for clustering purposes.

We calculate two-point correlation in the following way. We work with one overall descriptor type at a time. We use the set of similar images from the chosen descriptor type to create a cluster system, using the same methods described in Chapter 3. For any two images i and j within that system, we find the empirical probability of those two images belonging to the same cluster using Equation 4.7, where $n_{c,i/j}$ is the count of image i or j in cluster c, $N_{i/j}$ is the total count of image i or j in the entire sampled data set, and S is the set of clusters that contain both images.

$$P_C(i,j) = \frac{\sum_{c=1}^{S} n_{c,i} + n_{c,j}}{N_i + N_j} \tag{4.7}$$

Equation 4.7 is the two-point correlation that we seek. We want to know how often our software will place any two similar images in the same cluster, and hence, we again apply the bootstrapping method to obtain a distribution of the two-point correlation. Focusing on one overall descriptor type at a time, we run our software 100 times, obtaining 100 trials of each hyperparameter iteration (ECT/Segmentation/Radial Distance combination). We will analyze the statistics of co-occurrence for similar images in both the formalist and critical realms. This will be completed after the presentation of this thesis.

4.5 Error Considerations

Our software has a high opportunity for uncertainty since not all factors necessary to this type of investigation are considered. In this section, we will outline the factors not considered in practice, listing them in order of what we think of having the most to the least effect on the results.

1. *Art Descriptors:* As discussed in Chapter 4, the researcher decided upon the set of descriptor types and descriptors therein and defined the paintings according to those descriptors independently, with help from art history resources. The researcher is not an expert in art history, and thus, the choice to do this independently does introduce some personal bias into the results analysis. However, when we look at Table 4.2 in Chapter 4 and compare those descriptive distributions to critically decided definitions of the movements themselves (found in Appendix A), there is a notable equivalence.

   The exception to the match would be the Realism and Expressionism movements. For each of these movements, I would add in some descriptors to obtain an evaluation more accurate to the critical. The additions can be found in the list below.

   - Realism: pain/strife, nature, ugliness, uneasiness/anxiety
   - Expressionism: drama, uneasiness

   These exceptions are likely an issue with the sampling we did for the movements. I predict that we did not collect enough paintings to accurately represent the movements. With that being said, the movements not mentioned in the list have more exact critical evaluations to the ones we found with our assigned descriptors, allowing for more confidence in our results. Nevertheless, we must still keep this in mind. To counter this issue, one would need to consult an expert in the art history field to determine the best, most representative images to use per historical movement, the proper set of descriptors to use, and the best assignments per image.

Even then, due to the subjectivity of art, no one definition will be strictly reproducible with every observer. Such is the incongruity of art and science.

2. *Original Images, Image Capture, and Resolution:* Because no single comprehensive resource for historical artwork exists, we were not able to obtain images with normalized characteristics like resolution or color gamut. Generally, none of the images used were captured by the same devices nor in the same conditions, meaning the spatial and color resolutions are variable [18]. This will affect the image segmentation and edge creation in our software, which could have some effect on results.

3. *Binning Metrics:* : In our color ECT, we identify the average color of a superpixel and create edges from that. The color identification is done by binning all possible colors in HSL space to correlate to the primary, secondary, and tertiary colors found on the traditional color wheel. The binning was done by eye; there was no rigorous nor strictly repeatable process we used for this. Because our project is a meeting point between art theory and scientific computation, there exists little precedent for a meticulous binning of colors for our use. Our variable binning leaves room for inaccurate and/or imprecise results.

4. *Linear Dimensionality Reduction:* We use a linear dimensionality reduction in this project. It is possible that it would be best to use a non-linear technique, such as diffusion maps.

5. *Image Re-scaling and Interpolation:* We use Pixel Area Relation to interpolate our images to a standard size before analysis. It is possible that the initial re-scaling can affect results since image segmentation and our ECTs are so dependent on the spatial composition of the image. A different interpolation method could be best for this analysis. There could be some informational benefit to maintaining the original shape, as well. Answering this question would require more extensive research into

interpolation methods and their effect on image processing. See [19] for more information.

6. *Western culture Focus:* This project focuses on paintings from strictly Western cultures. We may obtain more relevant results if we were to obtain paintings from various cultures, rather than from our Euro-centric data set.

CHAPTER 5

DISCUSSION, CONCLUSIONS, AND OUTLOOK

The analysis done in Chapter 4 demonstrates that our software has potential as a powerful supplemental tool for both the formalist and critical art theories. The statistical testing we have completed enables us to answer the research questions set forth in Section 1.1. Our principal objective was to complete a proof of concept, answering the following questions: Is there a relationship between the measurable visual characteristics and the qualitative, communicative, and meaningful characteristics of art, and can we train a machine to discern and utilize that relationship? To answer this question, we conducted a null hypothesis test. Our null hypothesis stated that our software would return a distribution of performance measures that were of the same underlying distribution as a random clustering and that those measures would perform worse or at the same level as the random case. We found that a significant fraction of our hyperparameter/descriptor type combinations indeed rejected this null hypothesis, meaning they behave statistically different from random. The hyperparameters of our software determine how we identify the images' visual forms for computation (segmentation), what visual elements to test as having some relation to critical interpretation (edge creation techniques), and what range of interaction within the image to investigate (radial interval). Put another way, the hyperparameters are our independent variables. The descriptor types are the different artistic lexicons that we are testing for significant correlation with visual form. A rejection of our null hypothesis for a particular hyperparameter/descriptor type combination means that our software has created groupings of images, using those specific visual characteristics and in that specific artistic lexicon, that are of notable critical differentiation, beyond what would be possible with a random clustering assignment. Fundamentally, this indicates that not only does a non-trivial relationship exist between those specific independent variables and that artistic lexicon (whether it be of the formalist or critical paradigms), but it also proves that our software has detected and implemented that relationship to some extent. In summary, we have shown that those specific visual characteristics and artistic lexicons in the hyperparameter/descriptor type combinations that reject our null hypothesis do have an empirical relationship that can be capitalized on for visual sentiment analysis.

In addition to this discovery, we found that our genre-specific descriptors, those that contain the philosophical information about each painting, consistently return performance measures that are not only distinct from random but achieved the greatest success in terms of creating interpretatively distinct AIGM. Many of the performance metrics for this descriptor type return average values greater than 0.5. For these reasons, we can confirm our hypothesis that there exists a relationship between visual form and, specifically, *critical* evaluation. It is important to note, however, that the formalist descriptors did return a better performance in terms of rejecting our null hypothesis. This makes sense since we based our complex network creation strictly on those formalist ideas. The important consequence of our conclusions, though, is that we did find success with the critical evaluations, as well, without using that critical evaluation data in our modelling system.

As a disclaimer, we'd like to highlight the fact that our methodology does not have the capability to detect any three-dimensional aspects of the paintings. Often, artists purposely apply thick layers of paint to their work, creating a physical texture on the canvas. This additional dimension can be (and generally is) a pertinent part of the overall evaluation of the work. Because we work strictly with digitized scans of the paintings, it is not possible to obtain this information, which could have some effect on the efficacy of our results. Three-dimensional scans containing additional information on paintings would provide source data for a future extension of our work to take texture into account.

Our second research question asks whether we can supplement the traditional, qualitative critical approach to art analysis. Based on the comparison between the historical movements and the AIGM done in Chapter 4, we would argue that the methods we have presented do offer new insights to art–particularly, new critical/interpretative divisions of paintings–that could add to the traditional analyses of works of art. Historical movements are groupings of paintings with a common philosophical message or goal. Hence, the genre-specific descriptors should and do perform best in terms of evaluating the interpretative distinguishability of the critically decided and accepted groups of paintings. The AIGM generated by our software, however, tend to outperform the historical movements in terms of this descriptor. This means that our software creates AIGM that are *more* philosophically distinct than the historical movements, ultimately validating the idea that traditional art theory neglects certain factors that would, when added to those evaluations that already exist, provide a new lens through which to understand art. This also verifies that our methods offer the first step into tapping into that undiscovered information.

With the next research question, we ask how to measure the inherent information in art. The methods proposed in this project offer a first insight into this pursuit–a confirmation that information theory is applicable to artistic pursuits. Though we do have some success in the implementation of the software, there are ambiguities and ample room for error. In particular, the descriptor system that we use was created by the researcher, who is not an art history nor art theory expert. Though there is some evidence that leads us to believe that the descriptor assignment process was sufficient for our current purposes, we believe the analyses done can be improved with the involvement of art historians or art theory experts. The same can be said with the binning metrics used in the edge creation techniques. In addition to that, our analysis would be improved if the original images were all collected with the same methods, at the same time and in the same conditions, rather than collected from online resources. Finally, the use of a different programming tool could be beneficial. We believe that an even better and more conclusive analysis can be done if one were to experiment with different clustering and dimensionality reduction techniques, or even with a neural network analysis.

Our final research question addresses a question that we maintain throughout this report–how to bridge the subjectivity of art and the objectivity of science. Unfortunately, there is no clear answer to this yet. Many of the choices made in this report were arbitrary. Because of the utter volume of choices one can make–color bins, descriptors, original images, etc.–and the ever-present human subjectivity that comes with visual perception and artistic interpretation, it's not clear that there can ever be a strictly reproducible analysis of art. With that being said, we argue that reproducibility is not the most important concept in this project. The pillar we build this project on is the creative exploration of art through technological means. Art will always be subjective. Hence, we argue that it does not matter how subjective our means are, as long as the ends are explainable and offer insights previously unseen in the art world.

One condition we'd like to address in more detail is that the linear dimensionality reduction technique works. Due to the novelty of our objective, it was unclear what features of the complex network representations would return the greatest correlation with critical interpretation. Hence, we implemented a feature extraction process, with the second step being to reduce the dimensionality of our measurement data set, creating a smaller, non-redundant data space. This reduced space allowed for an optimized clustering of the images into similar critical interpretation. The axes in the new space—the principal components—were created from a linear combination of the original 28 measurement dimensions. The question is why this worked for us. Linear dimensionality reduction has limitations. The principal component analysis (PCA) technique will not work if the joint distribution of the measurement data does not follow a multivariate normal distribution. In other words, if the data has any nonlinear correlations, PCA does not have the capabilities to handle that behavior.

Why did this relatively simple transformation of the measurement data allow for a confirmation of our hypothesis i.e., for a critically distinct clustering of the images? We argue that this is because the measurement data already had a regularized structure before dimensionality reduction. The art that we analyze obviously contains some nonlinear spatial relationships in the visual features that we investigate. (Otherwise, the artworks wouldn't be very interesting to look at.) When we convert an image into its complex network representation, however, we are regularizing those correlations. Those nonlinearities are detected with image segmentation, deconstructed with our superpixel measurement[9], and reconstructed into a more regular relational form with our binning/edge creation methods. This is the power of the complex network analysis modelling system; any nonlinear spatial features present in the images are collapsed into associative features that, when measured, can be handled by the linear transformation that PCA performs.

We would like to note that our analysis only looks at visual features on a relatively large scale. The edge creation techniques (ECTs) outlined in Section 3.4 look at low order visual characteristics of the images. We do this on purpose; as stated, our work is unprecedented, and because of that, we wanted to look at the simplest artistic/visual features available as a first step. However, when we do this—when we strictly focus on low order visual characteristics—we lose information about the small-scale attributes of the images. Take the color ECT for example. To create the complex network representation for this ECT, we found and used the average hue, saturation, and light values of the superpixels. However, when we do this averaging, we eliminate access to the finer variations in color which could hold valuable interpretative information.

Part of this loss in descriptive resolution has to do with the image segmentation step of our methods[10]. Though the image segmentation does divide the image into smaller, meaningful components, it does not divide to the scale, of say, brushstrokes. The divisional scale of the segmentation algorithms (for the way we configured them) is more on the order of object detection—not brushstroke detection. This does not mean, however, that it is not possible to detect small-scale features using our methods. All it would take is to modify

[9] By measurement, we mean of the pixel size, average color, or angular orientation of each individual pixel.

[10] Another part could be attributed to the resolution of the images we work with. We predict that 500 by 500 pixels is too small to accurately detect brushstrokes.

different parts of the software. As said when this project was introduced, our working objective was to outline an algorithmic blueprint with which one can conduct a creative investigation of art. In other words, this project's methods are a framework that can be adjusted at any step and for any particular use. For instance, to detect brushstrokes in an image, one could, first, modify the image segmentation algorithms' hyperparameters to detect smaller features of brushstroke scale. Next, one could create a new ECT that embodies whatever desired visual feature of those small-scale components, such as color, color gradient, shape, size, path variation, etc. Beyond that, it is not necessary to use a binning method to create edges. Any similarity/dissimilarity logic could be used for edge creation. In summary, with the work presented, we have offered a versatile framework with which one can pursue any visual investigation of information in art. We predict that, based on the preliminary success of our methods, one could apply the modifications just outlined for fraud detection in or a materials analysis of works of art.

REFERENCES CITED

[1] D. R. Amancia *et al.* Identification of literary movements using complex networks to represent texts. *New Journal of Physics*, **14**(4), 2012. doi: https://doi.org/10.1088/1367-2630/14/4/043029.

[2] J. Hirschberg and C. D. Manning. Advances in natural language processing. *Science*, **349**(6245):261–266, 2015. doi: https://doi.org/10.1126/science.aaa8685.

[3] A. Ali *et al* M. Khan, M. Durrani. Sentiment analysis and the complex natural language. *Complex Adaptive Systems Modeling*, **4**(2), 2016. doi: https://doi.org/10.1186/s40294-016-0016-9.

[4] D. Zinoviev. *Complex Network Analysis in Python*. The Pragmatic Bookshelf, Raleigh, NC, 2018.

[5] S. Thurner, R. A. Hanel, and P. Klimek. *Introduction to the Theory of Complex Systems*. Oxford University Press, Oxford, 2019.

[6] M. Johnson. The elements and principles of design: Written in finger jello? *Art Education*, **48**(1):57–61, 1995.

[7] D. Nadaner. Painting in an era of critical theory. *Studies in Art Education*, **39**(2): 168–182, 1998.

[8] D. Tekiner. Formalist art criticism and the politics of meaning. *Social Justice*, **33**(2): 31–41, 2006.

[9] R. R. Behrens. Art, design and gestalt theory. *Leonardo*, **31**(4):299–303, 1998.

[10] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, **59**:167–181, 2004. doi: https://doi.org/10.1023/B:VISI.0000022288.19776.77.

[11] A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. *Proceedings: Computer Vision ECCV 2008*, pages 705–718, 2008.

[12] R. Achanta *et al.* Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intellegence*, **34**(11):2274–2281, 2012.

[13] L. Rocha, L. Velho, and P. C. P. Carvalho. Image moments-based structuring and tracking of objects. *Proceedings. XV Brazilian Symposium on Computer Graphics and Image Processing*, pages 99–105, 2002. doi: 10.1109/SIBGRA.2002.1167130.

[14] I. T. Jolliffe and J. Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **374**(2065), 2016.

[15] Europe 1300-1800. `https://smarthistory.org/europe-1300-1800/`, .

[16] Europe: 19th century. `https://smarthistory.org/europe-19th-century/`, .

[17] Modernisms 1900-1980. `https://smarthistory.org/modernisms-1900-1980/`, .

[18] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Pearson, New York, NY, 4 edition, 2018.

[19] J. P. Allebach. *Handbook of Image and Video Processing*. Elsevier Academic Press, 2005.

APPENDIX A

HISTORICAL MOVEMENT CHARACTERIZATION

This appendix outlines the critical and formalist theory evaluations of the historical art movements that we use in our analyses. We also list representative paintings from each movement.

A.1 Renaissance

- Timeframe: 1400 - 1600 A.D.
- Thematic Characteristics:
  - Idealized beauty
  - Religious depictions with a realistic perspective
  - Humanism: heightened influence of classical (Greek and Roman) antiquity
  - Naturalism: accurate depiction of the observable world
  - Depiction of proper/preferred gender roles
- Artistic Characteristics:
  - Linear Perspective
  - Perfected anatomical depiction of humans; complex poses and postures of humans
  - Humans in complex group formats
- Representative Artists:
  - Sandro Botticelli
  - Leonardo da Vinci

- Michelangelo

- Representative Paintings:
  - The Baptism of Christ by Verrocchio c. 1475
  - Mona Lisa by da Vinci c. 1503
  - La Primavera by Bottecelli c. 1482
  - Danae by Correggio c. 1530

A.2 Baroque

- Timeframe: 1585 - 1730 A.D.
- Thematic Characteristics:
  - Powerful, dramatic realism
  - Excuberant Ornamentation
  - Replication of observed reality
  - Emotional intensity
  - Instability, intensity, vividness
  - Directed involvement of viewer
- Artistic Characteristics:
  - Bold contrasts of light and dark
  - Tightly cropped compositions to enhance physical and emotional immediacy of narrative
  - Vibrant palette
  - Balanced compositions
  - Dynamic movement: Diagonals

  - Idealized forms

- Representative Artists:
  - Rembrandt
  - Caravaggio
  - Rubens

- Representative Paintings:
  - The Crucifixion of St. Peter by Caravaggio c. 1600
  - The Calling of Saint Matthew by Caravaggio c. 1660
  - The Proposition by Leyster c. 1631
  - Judith and her Maidservant with the Head of Holofernes by Gentileschi c. 1623

A.3 Rococo

- Timeframe: 1700 - 1775 A.D.

- Thematic Characteristics:
  - Lack of seriousness
  - Fashionable ideal
  - Perpetual youth
  - Pleasure, indulgence, and sexual gratification

- Artistic Characteristics:
  - Pastel color palette
  - Overflowing presence of nature
  - Loose brushstrokes

- Asymmetrical curves

- Representative Artists:
  - Fragonard
  - Boucher
  - Le Brun

- Representative Paintings:
  - The Lock by Fragonard c. 1779
  - Diana After the Hunt by Boucher c. 1745
  - The Blue Boy by Gainsborough c. 1770
  - The Bathers by Fragonard c. 1765

A.4 Neoclassicism

- Timeframe: 1700 - 1850 A.D.

- Thematic Characteristics:
  - Enlightenment: clear-headed thinking, rationality, seriousness, ethics, austerity
  - Patriotism, civic virtue, ethics, purpose, reason, discipline
  - Criticism of corruption of monarchy and aristocracy
  - Appreciation of Greek and Roman political systems

- Artistic Characteristics:
  - No evidence of brushstrokes
  - Clarity of form
  - Sober colors

    - Shallow space
    - Strong horizontals and verticals
- Representative Artists:
    - David
    - Ingres
    - West
- Representative Paintings:
    - Achilles Receiving the Ambassadors of Agamemnon by Ingres c. 1801
    - Burial of Atala by Trioson c. 1808
    - Napoleon Crossing the Alps by David c. 1805
    - The Oath of Horatii by David c. 1784

A.5 Romanticism

- Timeframe: 1800 - 1848 A.D.
- Thematic Characteristics:
    - Emotionally expressive
    - Physically direct; involving the viewer
    - Focus on nature; fear of the unknown; nature's dominance over man
    - Dramatic scenes: horrific images, revolution
    - Passion, sensitivity, imagination
    - A reaction against industrialism
- Artistic Characteristics:

  - Vivid colors
  - Visible, unrestrained brushstrokes
  - Unrefined outlines
  - Emphasis on color over form
  - Emphasis on the sky

- Representative Artists:
  - Courbet
  - Delacroix
  - Turner

- Representative Paintings:
  - The Shipwreck by Turner c. 1805
  - The Desperate Man by Courbet c. 1845
  - Lady Liberty Leading the People by Delocroix c. 1868

A.6 Realism

- Timeframe: 1840 - 1900 A.D.

- Thematic Characteristics:
  - Portrayal of artists' contemporary environment
  - The accurate, detailed, un-embellished depiction of nature or contemporary life
  - Humble, serious tone
  - Raw and natural (like a photograph)
  - Portrayal of the ugliness, grittiness of the world

- Artistic Characteristics:

  - Dark, warm, Earthy color palettes
  - Emphasis on correct reflection of shadow, light, depth, perspective
- Representative Artists:
  - Repin
  - Manet
  - Courbet
- Representative Paintings:
  - Barge Haulers on the Volga by Repin c. 1873
  - Ivan the Terrible and His Son Ivan by Repin c. 1885
  - Portrait of an Unknown Woman by Kromskoi c. 1883

A.7 Impressionism

- Timeframe: 1860 - 1886 A.D.
- Thematic Characteristics:
  - Landscape and genre scenes
  - Attempt to capture a particular moment in time by pinpointing specific atmospheric conditions
  - Capturing the rapid pace of contemporary life and the fleeting conditions of light
  - Color over form
- Artistic Characteristics:
  - Optical blending: vibrant colors
  - Quickly shifting light on the surface of forms
  - Highly visible (patchy) brushstrokes

- Light, pastel color palette
- Unbalenced compositions

- Representative Artists:
  - Degas
  - Monet
  - Renoir
- Representative Paintings:
  - Women Walking on the Beach by Sorolla c. 1909
  - The Ballet Class by Degas c. 1874
  - The Water Lilies by Monet c. 1880
  - At the Races in the Countryside by Degas c. 1869

A.8 Post Impressionism

- Timeframe: 1886 - 1905 A.D.
- Thematic Characteristics:
  - The extension and rejection of Impressionism: brought "scientific rigor" to the period
- Artistic Characteristics:
  - Pointilism and optical mixture: chromatic intensity
  - Avoided mixing complementary colors
  - Clearly defined, intensified color
  - Flattened (shallow) space
  - Asymmetry; imbalanced compositions

- Representative Artists:
  - van Gogh
  - Gauguin
  - Cezanne
- Representative Paintings:
  - A Wheatfield with Cypresses by van Gogh c. 1889
  - Breton Girls Dancing by Gauguin c. 1888
  - The Card Players by Cezanne c. 1895
  - The Starry Night by van Gogh c. 1889

A.9 Fauvism

- Timeframe: 1904 - 1915 A.D.
- Thematic Characteristics:
  - The rejection of naturalism
  - Making art for art
- Artistic Characteristics:
  - Bold colors; limited color choices, emphasizing bright shades
  - Flat plane
  - Monochromism
  - Simplified representation/abstraction
  - Inconsistent brushwork
- Representative Artists:

  - Derain
  - Matisse
  - Vlaminck
- Representative Paintings:
  - Landscape Near Chatou by Derain c. 1904
  - Portrait of Matisse by Derain c. 1905
  - Dance by Matisse c. 1909
  - The River Seine at Chatou by Vlaminck c. 1906

A.10 Expressionism

- Timeframe: 1905 - 1920 A.D.
- Thematic Characteristics:
  - Portraying emotion to the fullest intensity
  - Expression of what artists' felt, not saw
  - Reaction to the surrounding world
- Artistic Characteristics:
  - Flattened forms
  - Reduced detail
  - Heavy outlines
  - Solid geometry
  - Intense color
  - Rough brushwork
- Representative Artists:

    - Kandinsky
    - Picasso
    - Munch
    - Kirchner
- Representative Paintings:
    - Blue Horse I by Marc c. 1911
    - Street Berlin by Kirchner c. 1913
    - The Old Guitarist by Picasso c. 1903
    - The Scream by Munch c. 1893

A.11 Cubism

- Timeframe: 1907 - 1937
- Thematic Characteristics:
    - Disassembling of an observation/form
    - The depiction of forms over time, from different perspectives
    - Reduced content, color, emotion
- Artistic Characteristics:
    - Pale tones and darker shadows, much like with a relief
    - Monochromatic color
    - Focus on line, form, structure
    - Shallow space
- Representative Artists:

- Picasso
- Gris
- Braque

- Representative Paintings:
  - Bottle and Glass on a Table by Gris c. 1914
  - The Round Table by Braque c. 1929
  - Three Women by Picasso c. 1908
  - Nature Morte au Compotier by Picasso c. 1914

A.12 Surrealism

- Timeframe: 1922 - 1970
- Thematic Characteristics:
  - Looks to the mind as a source of liberation
  - Subconscious thought and identity: influenced by Freud
  - Unexpected/illogical juxtaposition of objects
  - 'Surrendering' control to the art-making process
  - Primitive, child-like symbolism
- Artistic Characteristics:
  - Distorted figures and biomorphic shapes
  - Randomness
- Representative Artists:
  - Dali

  - Ernst
  - Magritte
  - Kahlo
- Representative Paintings:
  - Persistence of Memory by Dali c. 1931
  - The Barbarians by Ernst c. 1937
  - The Broken Column by Kahlo c. 1944
  - The Son of Man by Magritte c. 1964

APPENDIX B

COLOR ECT BINNING METRICS

Table B.1, Table B.2, and Table B.3 show the exact binning metrics for the different HSL axes. As mentioned, we work in HSL since the color space is more intuitive to how we think about color (in terms of the visible spectrum). Keep in mind that these bins were differentiated by eye. As explained by the Gestalt theory of color perception, we view colors differently based on the background they are placed upon, so these bin edges can be variable depending on the person.

Table B.1: Hue Axis Bin Labels and Edges

Hue Bin	Hue Axis Interval [0 - 360)
Red-orange	11 - 26
Orange	26 - 36
Yellow-orange	36 - 46
Yellow	46 - 61
Yellow-green	61 - 91
Green	91 - 151
Blue-green	151 - 201
Blue-violet	251 - 271
Violet	271 - 291
Red-violet	291 - 341
Red	341 - 11

Table B.2: Saturation Axis Bin Labels and Edges

Saturation Bin	Saturation Axis Interval [0 - 100]
True grey	0 - 6
Low Saturation	6 - 41
Medium Saturation	41 - 71
High Saturation	71 - 100

Table B.3: Lightness Axis Bin Labels and Edges

Lightness Bin	Lightness Axis Interval [0 - 100]
True Black	0 - 6
Dark	6 - 31
Precise	31 - 70
Light	70 - 95
True White	95 - 100

APPENDIX C
ART DESCRIPTORS

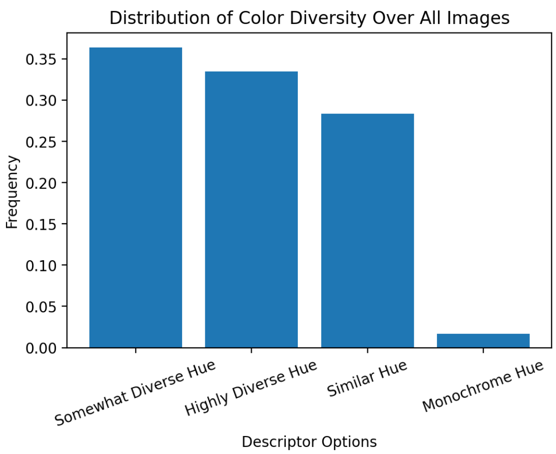
Table C.1 shows how we conduct the binning for the different formalist questions we ask in the descriptor assignment process. Note that if any question did not have an obvious answer for an image, that image was given no descriptor for that particular question. Table C.2 shows all possible critical descriptors for assignment. Figure C.1 through Figure C.4 display the count distribution for all formalist descriptor types (and the options therein) over all the images we collected. Figure C.5 displays the descriptor distributions for the critical descriptors over all images. Note that the genre-specific descriptors do not have labels, since there were too many to list.

Table C.1: All Formalist Descriptors

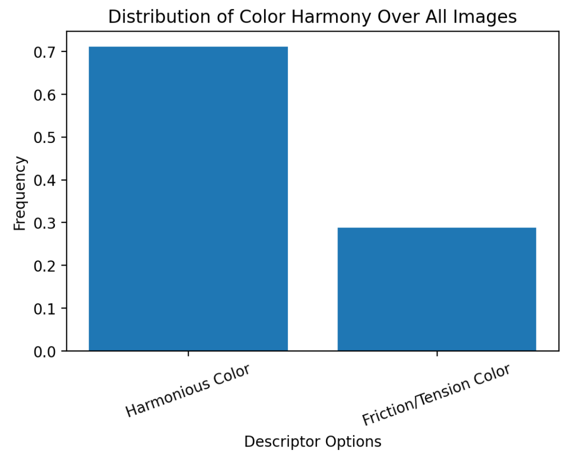
Question	Formalist Descriptors			
	Shape, Space, and Perspective			
Balance?	Highly Balanced Space	Mostly Balanced Space	Mostly Imbalanced Space	Highly Imbalanced Space
Movement?	Highly Dynamic Movement	Intermediate Movement	Highly Static Movement	-
Empty Space?	Small Empty Space	Intermediate Empty Space	Large Empty Space	-
Perspective?	Direct Perspective	Intermediate Perspective	Removed Perspective	-
	Brushstrokes and Texture			
Brushstroke Control?	Mostly Controlled Brushstrokes	Mixed Brushstrokes	Mostly Wild Brushstrokes	-
Outlines?	Light/No Outlines	Mix Outlines	Heavy Outlines	
Brushstroke Visibility?	Blended Brushstrokes	Mixed Visibility Brushstrokes	Visible Brushstrokes	-
	Line, Form, and Value			
Depth?	Flat Depth	Shallow Depth	Regular Depth	Deep Depth
Value Contrast?	Low Value Contrast	Intermediate Value Contrast	High Value Contrast	-
Average Value?	Light Overall Value	Balanced Overall Value	Dark Overall Value	
Distinct Lines Between Objects?	Indistinct Lines	Intermediate Lines	Indistinct Lines	-
	Color			
Color Mode?	Naturalistic Color	Capture Color	Expressionistic Color	-
Color Diversity?	Monochrome Hue	Similar Hue	Somewhat Diverse Hue	Highly Diverse Hue
Tone?	Vibrant Tone	True Tone	Muted/Tinted Tone	-
Color Relation-ships?	Harmonious Color	Friction in Color	-	-
Proximity Contrast?	Blended Color Contrast	Stark Color Contrast	-	-

Table C.2: All Critical Descriptors

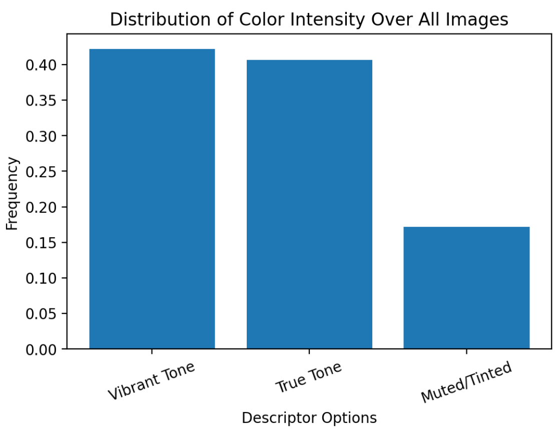
Critical Descriptors			
Abstraction Level			
Wholly Abstract	Organically Abstract	Semi-Abstract	Naturalistic
Genre			
Historical	Portraiture	Genre	Landscape
Still life	-	-	-
Action, Location, Person, or Object			
Indoors	Outdoors	Public Domain	Private Domain
Rural Realm	Urban Realm	Recreation	Rest/Liesure
Work	Dining	Religious Action	Education
Chores/Travel	Festivity	Violence	Conflict
Communication	Historical subgenre	Religious subgenre	Mythological Subgenre
Literary Subgenre	Allegorical Subgenre	Royalty	Government/Military
Aristocracy	Commoner	Thinker/Artist	Religious
Other/Self	Farm/grassland	Mountain	Sea/ocean
Forest	River/lake	City	Beach
Human(s)	Animals	Buildings	Food
Furniture	Tools	Flowers/Plants	-
Message			
Community/family	Spirituality	Cycle of life	Science
Government	Social Consciousness	the Body	the Mind
Power	Beauty/Ideal	Love	Morality
Occupation of Time	Peaceful	Mystery	Danger
Bountiful	Unforgiving	Genius	Strife/pain
Exhaustion	Greatness	Joy	Youth/Sensitivity
Serious	Dedication	Sadness	Mischievous
Thoughtfulness	Harmony	New Directions/Change	Primativism
Fast Pace	Lively/Energetic	Nature Interaction	Deconstruction
Sensuality	Sexuality	Minimalism	Casual
Drama	Elegance	Ugly	Antiquity
Color	Texture	Abundance/Luxury	Isolation
Detachment	Vacancy/Boredom	Antiquity	Revolution
Patriotism	Strength	Confidence	Cross Culturalism
Fear	Hope	Warm/Inviting	Peace
Loss	Uneasiness/Anxiety	Comfort	Gender
Discipline	Independence	Momentary	Intimacy



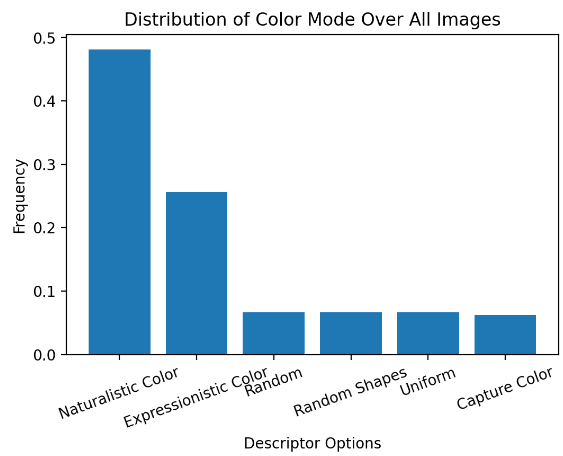
(a) Color Diversity Distribution



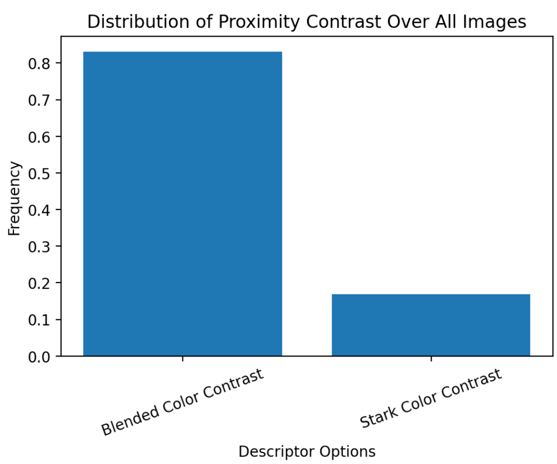
(b) Color Harmony Distribution



(c) Color Intensity Distribution

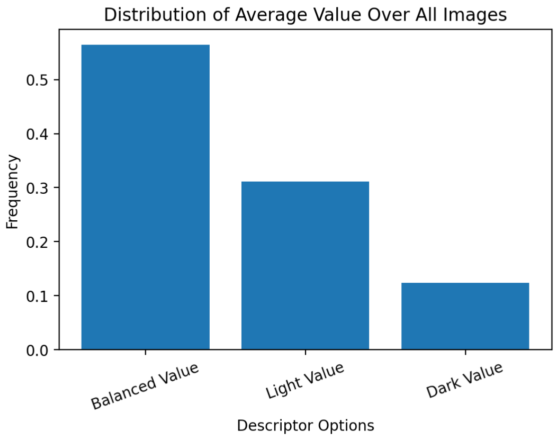


(d) Color Mode Distribution

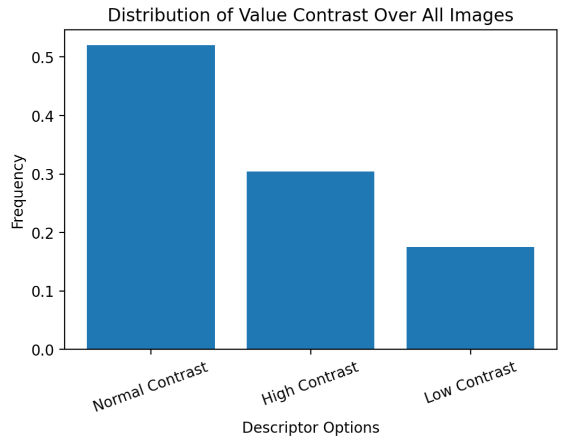


(e) Proximity Contrast Distribution

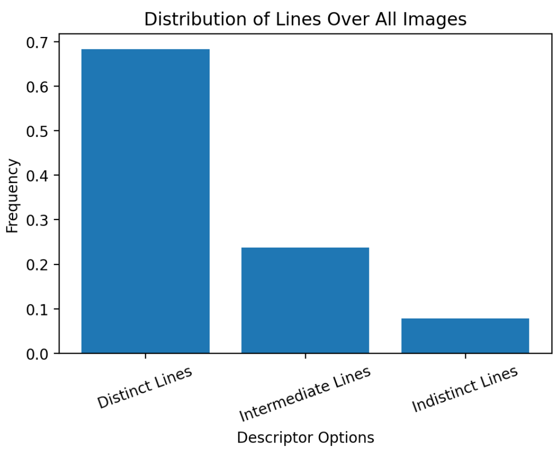
Figure C.1: Descriptor Distributions For the Color Element Group



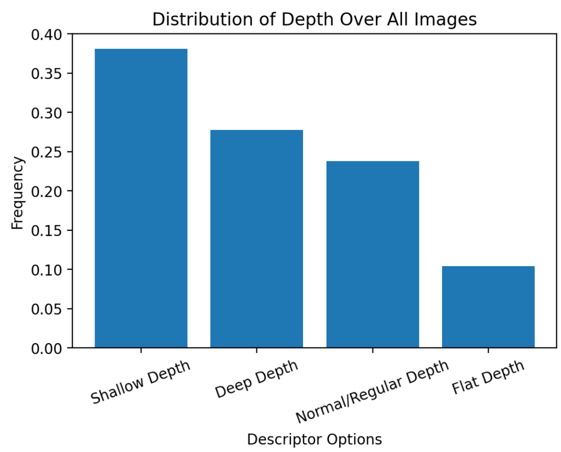
(a) Average Value Distribution



(b) Value Contrast Distribution

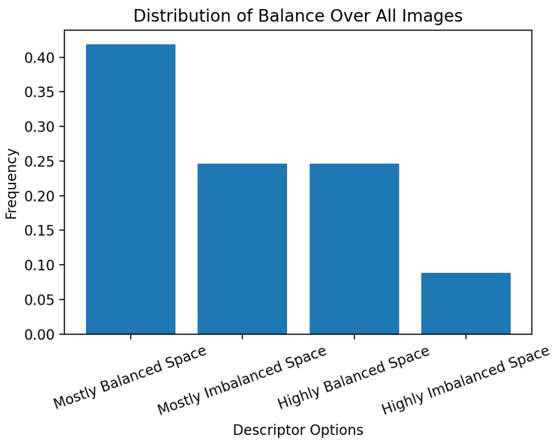


(c) Distinctive Lines Distribution

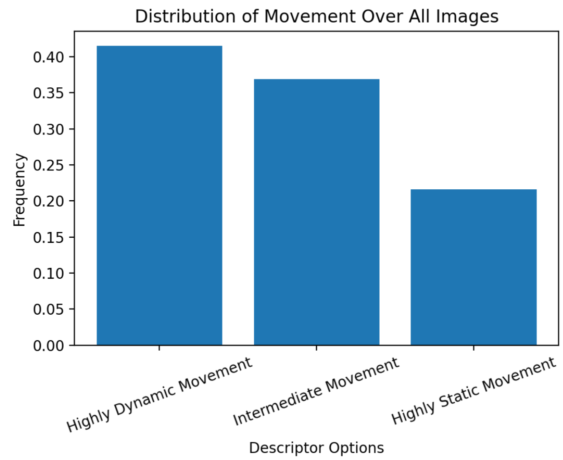


(d) Depth Distribution

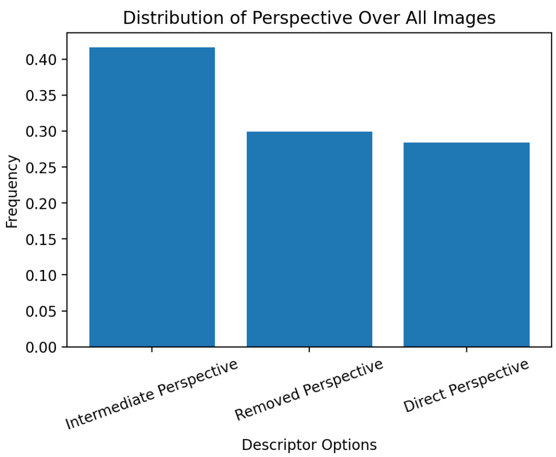
Figure C.2: Descriptor Distributions For the Line, Form, and Value Element Group



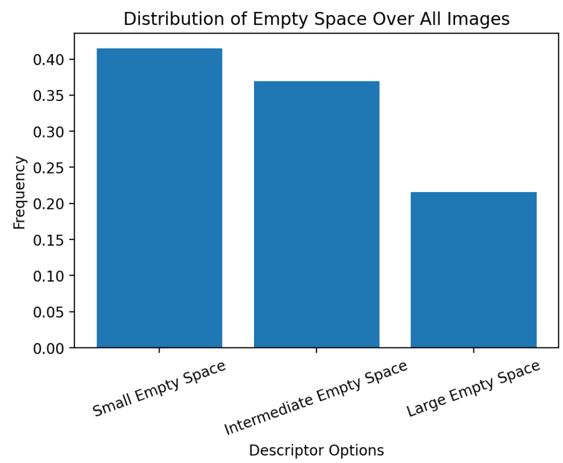
(a) Balance Distribution



(b) Movement Distribution

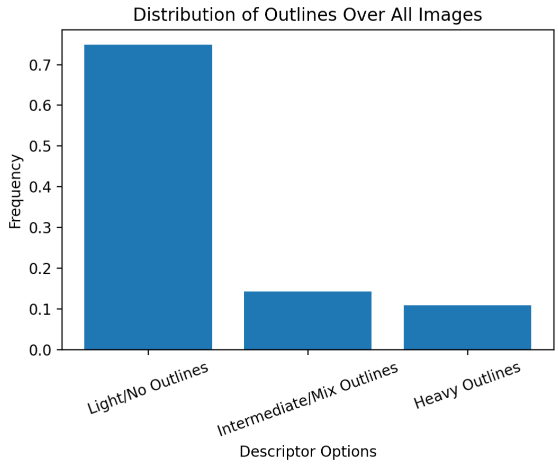


(c) Perspective Distribution

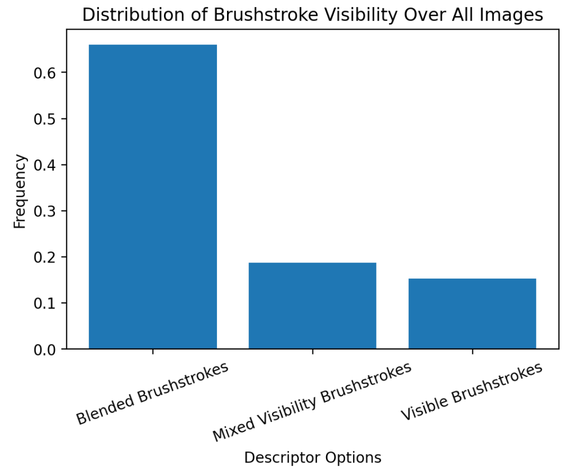


(d) Empty Space Distribution

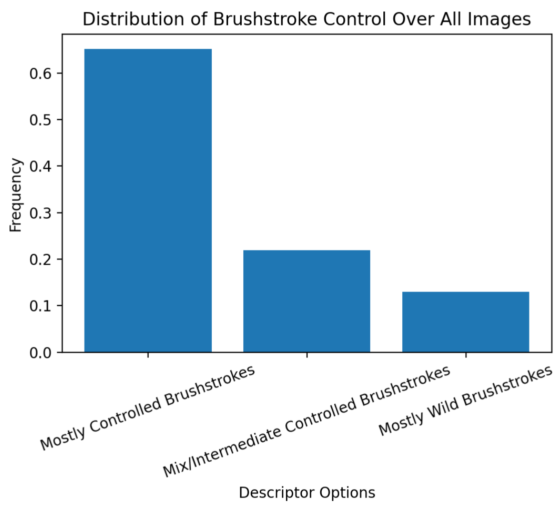
Figure C.3: Descriptor Distributions For the Shape, Space, and Perspective Element Group



(a) Outline Distribution

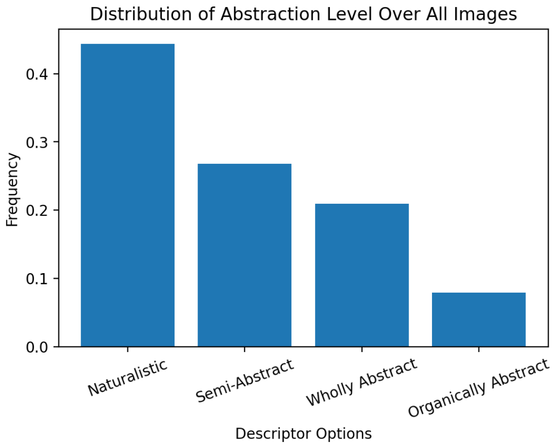


(b) Brushstroke Visibility Distribution



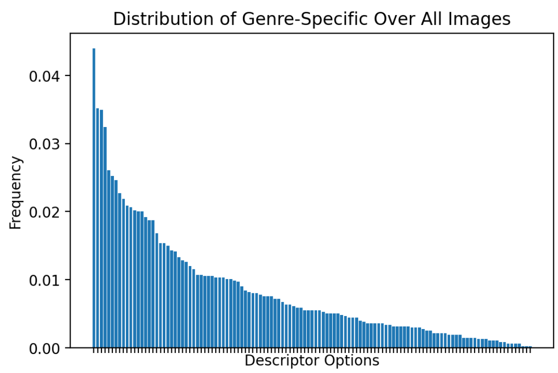
(c) Brushstroke Control Distribution

Figure C.4: Descriptor Distributions For the Texture and Brushstrokes Element Group



(a) Abstraction Level Distribution

(b) Genre Distribution



(c) Genre-Specific Distribution

Figure C.5: Descriptor Distributions for Critical Descriptors

APPENDIX D

SUPPLEMENTAL RESULTS

D.1 Null Rejected Performance Metric Distributions

Figure D.1 shows the distributions for the different null-rejected performance metrics for the 100 sample size. We do not specify descriptor type nor hyperparameter focus.

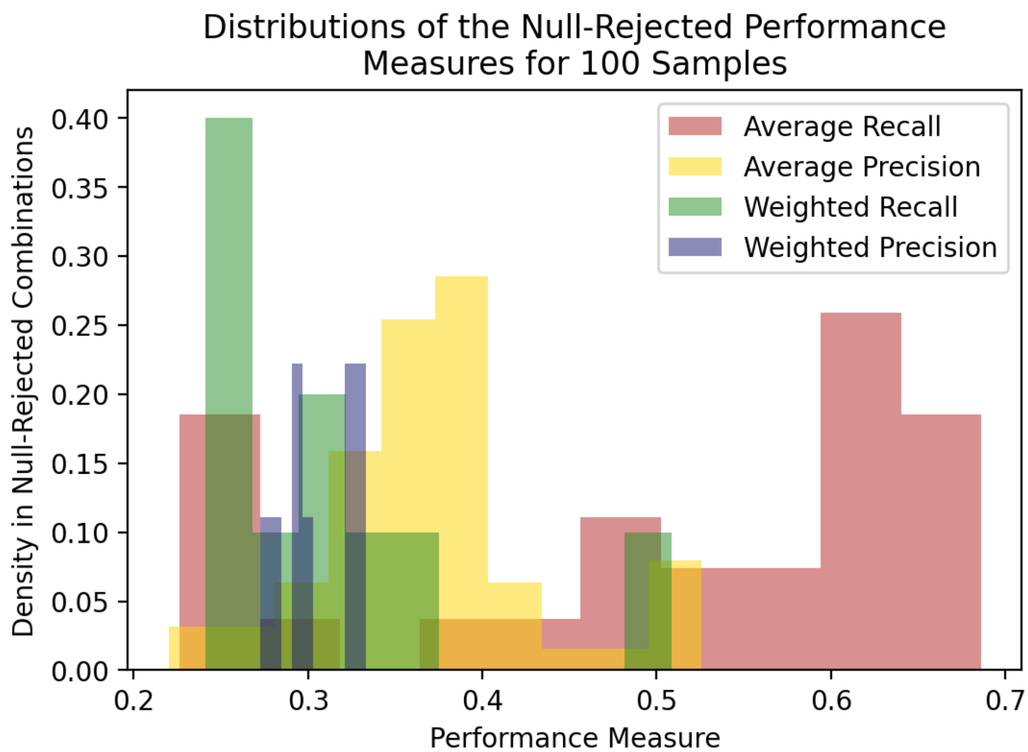


Figure D.1: Null Rejected Performance Metric Distributions

D.2 Null Rejected Hyperparameter-Descriptor Combinations

Table D.1 shows all combinations of the hyperparameter focus, descriptor type, and performance measures that end up rejecting our null hypothesis. Recall that a rejection of the null hypothesis means that, for a particular hyperparameter/descriptor/performance combination, the real data from our bootstrap data collection performs *better* than the corresponding random data and that we are statistically certain the two distributions were *not* drawn from the same underlying distribution.

Table D.1: Hyperparameter and Descriptor Combinations that Reject the Null Hypothesis

Hyperparameter Focus	Descriptor Type	Performance Measure
Color ECT	All Formalist Descriptors	Average Precision
	Abstraction Level	Average Precision Weighted Precision Weighted Recall
	Genre-specific	Average Precision
	Color	Average Precision
	Line, Form, and Value	Average Precision
	Shape, Space, and Perspective	Average Precision Average Recall Weighted Recall
	Texture and Brushstrokes	Average Precision
Orientation ECT	All Formalist Descriptors	Average Precision Average Recall
	Abstraction Level	Average Precision Average Recall Weighted Precision
	Genre-specific	Average Precision Average Recall
	Color	Average Precision Average Recall
	Line, Form, and Value	Average Precision
	Shape, Space, and Perspective	Average Precision Average Recall
	Texture and Brushstrokes	Average Precision Average Recall
Pixel Size ECT	All Formalist Descriptors	Average Precision Average Recall
	Abstraction Level	Average Precision Average Recall
	Genres	Weighted Recall
	Genre-specific	Average Precision Average Recall
	Color	Average Precision
	Line, Form, and Value	Average Precision
	Shape, Space, and Perspective	Average Precision Average Recall
	Texture and Brushstrokes	Average Precision

Table D.1: Continued.

Hyperparameter Focus	Descriptor Type	Performance Measure
Felzenswalb Segmentation	All Formalist Descriptors	Average Precision
	All Critical Descriptors	Average Precision
	Abstraction Level	Average Precision Weighted Recall
	Genre-specific	Average Precision Weighted Precision
	Color	Average Precision
	Line, Form, and Value	Average Precision
	Shape, Space, and Perspective	Average Precision
	Texture and Brushstrokes	Average Precision
Quickshift Segmentation	All Formalist Descriptors	Average Precision
	Abstraction Level	Average Precision Average Recall
	Genres	Weighted Recall
	Genre-specific	Average Precision Average Recall
	Color	Average Precision
	Line, Form, and Value	Average Precision
	Shape, Space, and Perspective	Average Precision Average Recall
SLIC Segmentation	All Formalist Descriptors	Average Precision Weighted Precision
	Abstraction Level	Average Precision Average Recall Weighted Precision Weighted Recall
	Genre-specific	Average Precision
	Color	Average Precision Weighted Precision
	Line, Form, and Value	Average Precision Weighted Precision
	Shape, Space, and Perspective	Average Precision
	Texture and Brushstrokes	Average Precision
0, 100 Radial Interval	All Formalist Descriptors	Average Precision Average Recall
	Abstraction Level	Average Precision Average Recall Weighted Recall
	Genres	Weighted Recall
	Genre-specific	Average Precision Average Recall
	Color	Average Precision

Table D.1: Continued.

Hyperparameter Focus	Descriptor Type	Performance Measure
0, 100 Radial Interval	Line, Form, and Value	Average Precision
	Shape, Space, and Perspective	Average Precision Average Recall
	Texture and Brushstrokes	Average Precision Average Recall
200, 300 Radial Interval	All Formalist Descriptors	Average Precision Average Recall
	Abstraction Level	Average Precision Average Recall Weighted Precision Weighted Recall
	Genre-specific	Average Precision Average Recall
	Color	Average Precision
	Line, Form, and Value	Average Precision Weighted Precision
	Shape, Space, and Perspective	Average Precision Average Recall
	Texture and Brushstrokes	Average Precision
400, 500 Radial Interval	All Formalist Descriptors	Average Precision Average Recall
	Abstraction Level	Average Precision Weighted Recall
	Genre-specific	Average Precision Average Recall
	Color	Average Precision
	Line, Form, and Value	Average Precision
	Shape, Space, and Perspective	Average Precision Average Recall
	Texture and Brushstrokes	Average Precision

APPENDIX E

ARROW OF TIME ANALYSIS

As a tangent to our project, we also conducted an arrow of time analysis, using entropical measures. Due to our null result for this investigation, it was decided to place this investigation in the appendix, rather than as apart of the main body of this report. This analysis is outlined below.

E.1 Theory

In our research, we looked extensively at Shannon entropy and its generalization to digital images. As stated in Section 2.3, with Shannon entropy, the lower the value, the more information is inherent in the system, and the "surprise" is low. This idea can be extended to images. Fundamentally, entropy should measure the uncertainty of an image. If totally one color, our generalized entropy would be zero, and if totally randomized, the entropy would be maximal. The paintings that we analyze would lie somewhere in between these extremes, where meaningful information lies. Classical paintings–those from the Renaissance, Baroque, Neoclassical, and Realism movements–tend to illustrate a specific event or situation. They tend to have a (relatively) well-communicated message, tone, or feeling that is easily understood by the viewer. The paintings of these movements also tend to be more visually ordered (though this varies, depending on context) [15], [16]. Thus, our prediction is that these movements will have a lower generalized entropy–a lower surprise factor.

To break away from tradition, the later movements–Surrealism, Expressionism, Cubism, and Impressionism–introduced disorder to art. This is not the classically defined disorder of displaying chaotic events but a totally stylistic disorder. These movements brought painting styles, color schemes, shapes, and ideas previously unseen in fine art; they were rejecting traditional forms of expression and escaping order [16], [17]. With this came less explicit information; the intended meaning became more opaque [8]. Recall that classical movements generally had easily interpretable subject matter and explicit tone. Hence, the movements had more inherent interpretative information. For the later movements, the interpretation is more subject to the viewer. The tone, feeling, and content is intended to be less unanimously interpretable. Thus, I predict that these movements will have a higher entropy–a higher uncertainty.

This theory, if true, coincidentally follows the second law of thermodynamics: the law that entropy tends to increase over time. What does this mean, philosophically? As society becomes more complex, as it has over the past half millennia, does our art mathematically reflect the new uncertainties and diversities of thought? This idea, if confirmed, would return powerful conclusions on society, philosophy, art, and their evolutions.

The generalized entropy we use for this analysis can be found in Equation E.1, where $P_{R,G,B}$ indicates the probability of a particular color vector within the image being examined.

$$H_{R,G,B} = -\sum P_{R,G,B}(r,g,b) \log_{256} P_{R,G,B}(r,g,b) \tag{E.1}$$

We use base 256 for the logarithm since there are 256 possible red, green, and blue states for a color vector. The above entropy is known as 1st order entropy. In this investigation, we also look at 0th and 2nd order entropy. The 0th order entropy is simply a count of all the different states (color vectors) present. The 2nd order entropy, known as Rényi entropy, can be found in Equation E.2.

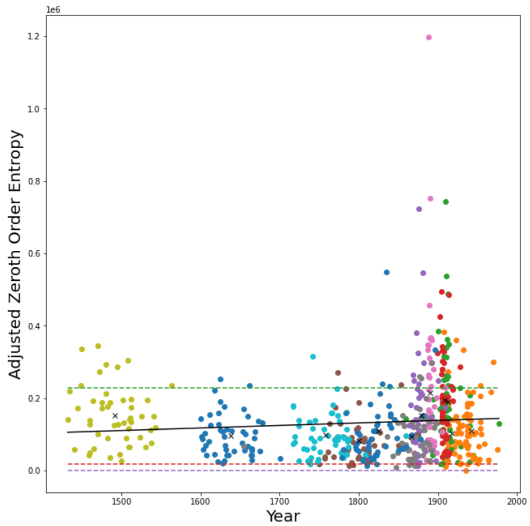
$$H_{2,(R,G,B)} = -\log_{256}\big(\sum P_{R,G,B}(r,g,b)^2\big) \tag{E.2}$$

The 2nd order entropy is known as the Rényi Entropy. This entropy is commonly thought of as an entropic measure of diversity, where the sum inside of the log is true diversity. Ultimately, we can use 2nd order entropy to look at the categorical diversity between color vectors.

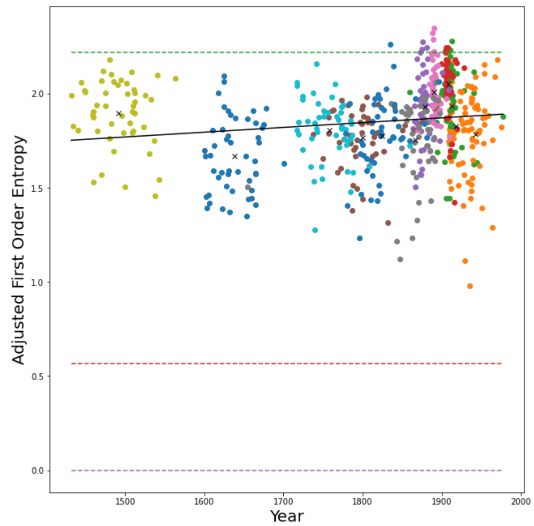
E.2 Results

We found the 0th, 1st, and 2nd order entropies as well as the true diversity for all the images that we used in our main analysis, including those from the simple movements. The results from this investigation can be found in Figure E.1. For each of the plots in Figure E.1, we also performed a linear regression on the data (black line). The regression metrics can be found in Table E.1.

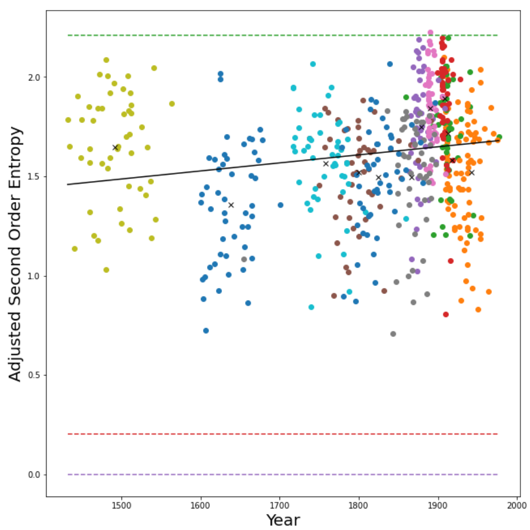
The R-value, p-value, and standard error found in Table E.1 tell us that we cannot confirm our hypothesis that the color-based generalized Shannon entropy of paintings increases over time, at least linearly. R-values act as a goodness-of-fit test, with possible values ranging from 0–the worst possible fit for the data–to 1–a perfect fit of the data. The R-values for all measures are less than 0.2, indicating a poor fit of the least-squares linear regression. The p-values act as a null hypothesis test, testing whether the data as a true slope of zero. The minuscule p-values that we obtained seem to indicate the null hypothesis can be rejected, but we would prefer to conduct more tests to confirm this.



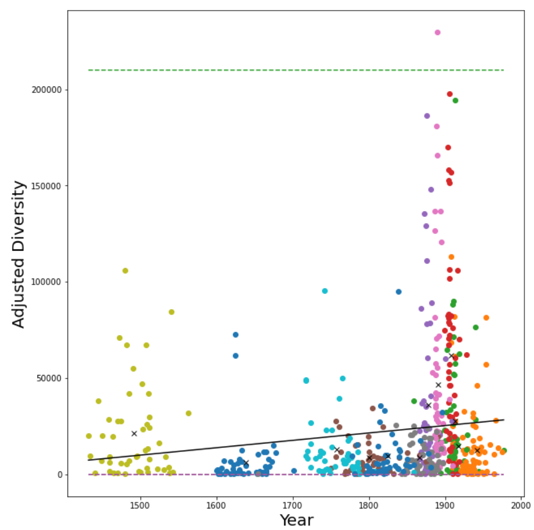
(a) Zeroth Order Entropy



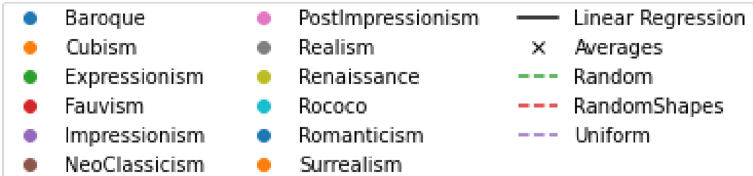
(b) First Order Entropy



(c) Second Order Entropy



(d) True Diversity



(e) Legend

Figure E.1: Arrow of Time Results

Table E.1: Arrow of Time Linear Regression Metrics

Measure	Slope	Intercept	R-value	p-value
Zeroth Order Entropy	70.019	5577.090	0.0811	0.0491
First Order Entropy	2.5E-4	1.389	0.1531	0.0001
Second Order Entropy	4.0E-4	0.878	0.1751	1.9E-5
Diversity	38.360	-47652.903	0.1485	3.0E-4

E.3 Remarks

Alas, we cannot confirm our theory that the generalized Shannon entropy of images increases over time, but hope is not lost. We will list some investigations one could do to continue the search into the theory.

- *Introduce color binning.* One may want to consider implementing the same binning metrics that we use in our color ECT (see Section 3.4) for this entropic investigation. Instead of computing the information measures with individual color vectors as states, one would use the color identities (the combined hue, saturation, and light bins) as the states. The number of states would decrease dramatically, and the correlated probabilities would likely be significantly larger. The motivation behind this is that, generally, artists don't think of color in terms of a discrete spectrum, like digital color spaces do, but in terms of a continuous spectrum that gradually changes from one color to another. For instance, an artist may create an monochromatic painting that contains various hues of red. Treating this image as we have, the entropy would likely be very similar to a non-monochromatic artwork, since all the different color vectors corresponding to the hues of red are treated as independent states. Introducing the color binning would make it so that monochromatic painting has fewer independent states than the non-monochromatic painting, potentially reducing the entropy.

- *Collect more data.* We work exclusively with Western fine art from the 1400s through the 1970s. How would this analysis change if we increased the time span from prehistoric periods through modern art? What if we explored the art from other cultures?

- *Look at the rise and fall of civilizations.* One of the postulates of our theory is that art reflects the increasing disorder of society. One way to explore this idea is to look at the art of a specific area of the world where civilizations would rise, collapse, and restart over centuries. Assuming we have access to the art from the different civilizations, we could investigate how the generalized entropy of those civilizations' art changes with the changes in their societal structure. Is there a lower artistic entropy when the civilization is budding? Does it increase with the growth of the society? Does it come to a climax and fall as the civilization collapses and restarts?

- *Use the interpretability of the art instead.* Instead of looking at the color for entropy, what if we looked at the relative ambiguity of the human interpretations of that art? Do the classical historic movements–those with clear messages and intent–have a lower evaluative entropy? Do the artworks with a more opaque message have a higher (semantic/evaluative) entropy?