Intersect, Vol 15, No 1 (2021)

Using Convolutional Neural Networks To Compare Paintings By Style

Moritz Alkofer
Technical University of Berlin

One of the main challenges in the field of Computer Vision and Art History is the extraction of a numerical representation of an artwork's style. Calculating such representation allows art historians to automatically analyze large digital collections of art. In this study we aim to transfer an approach of numerical style extraction originally developed for artistic style transfer to the task of comparing paintings by style. The approach uses a Convolutional Neural Network (CNN) trained on object detection to derive an image's style representation in the form of Gram matrices. We use it to compare paintings either by clustering a set of paintings or retrieving a paintings' most similar paintings from the set of paintings. We hypothesize that using a different CNN architecture trained on artistic style (instead of object) detection would lead to a significant increase in comparison quality. Using an object detection network we achieve a clustering accuracy of 22%. Using a network specifically trained on artistic style detection increases the clustering accuracy by 44%. Directly using the art detection networks output instead of Gram matrices yields an accuracy of 42%. Overall, we conclude that the approach is suitable to compare paintings by style. We significantly improved the approach's accuracy by changing the network architecture and training and show that for the improved network, Gram matrices provide little benefit.

Introduction

Recent improvements in Computer Vision and the availability of large public accessible fine art collections (Tan, Chan, Aguirre, & Tanaka, 2016) have led to new research challenges at the intersection of Computer Vision, Machine Learning and Art History.

An important challenge is the extraction of a vector representation capturing latent features of an artwork's style, for it allows automated painting analysis. It assists curators in organizing large digital collections, retrieving artworks, detecting forgery and many other tasks. Such automatic art analysis remains difficult for it involves understanding features inherent to human perception, such as content, composition, brushstroke, and overall form (Gardner, 1970). Many of those components originate from the formal elements of paintings such as lines, volumes, colours and textures.

Convolutional Neural Networks (CNNs) have shown great success in understanding such features. CNNs are a type of feed-forward machine learning algorithms that are capable of processing images. They consist of an input layer, an output layer, and multiple hidden layers in between, each of which applies a non-linear function to its input. Feed-forward means that each of these layers receives an output from former layers as input and passes its own output on to later layers. For training, a large dataset consisting of images and labels is used in a two-step process. During forward propagation the images are processed by the layers and a prediction for the input's label is calculated. During backpropagation the prediction's error or "loss" is calculated by comparing the prediction to the label. Then the network's layers are updated to minimize this loss. During this training process the filters learn to represent the input at different levels of abstraction, where early representations correspond to simple shapes such as lines and later layers to complex concepts such as content and composition. Popular examples of CNNs are ResNet architectures (He, Zsang, Ren & Sun, 2016), VGGs (Simonyan, Zisserman, 2014) or the inception architecture (Szegendy et al., 2015).

Gatys et al. have developed a way to use these representations for the translation of a painting's style features into a vector (Gatys et al., 2016). First, they use a Convolutional Neural Network trained on object detection to extract abstract patterns from the paintings. To obtain a style representation they compute the correlation of these abstract patterns, resulting in a so called Gram matrix (Gatys, Ecker, & Bethge, 2015). The approach has successfully been tested for style transfer, but not for painting comparison.

In this article, we explore if the approach used by Gatys et al. are applicable to the task of painting comparison and evaluate proposed improvements. Furthermore, we discover a novel way of representing style. The approach is used to compare paintings both by clustering a dataset of paintings by style and retrieving a paintings most similar images from a dataset. To compute the style representations three different methods are employed: First Gram matrices are calculated in the way proposed by (Gatys et al., 2016), using a network trained on a generic task. The thesis shows the feasibility of using Gram matrices for the stylistic comparison of paintings. Second, the network for calculating the Gram matrices is replaced by a modern architecture that is trained on a style recognition task. It is shown that the change of architecture and training data significantly improves the representation's quality in comparison to the first method. Third, a style representation taken directly from a layer of the modern network trained on recognizing style is used. It yields results only slightly worse than those achieved using the second method.

The results indicate that Gram matrices can be applied to the task of style comparison. We show that Gaty's approach can be significantly improved by changing the network architecture and training it on a style specific task. We furthermore discover that using an internal embedding of a network trained to detect style provides a representation of comparable quality.

Related Work

One of the first attempts of style abstraction was conducted by D. Keren (Keren, 2002) who derived features from discrete cosine transformation to train a naïve bias classifier. The methodology of extracting various low level features capturing shape, texture, edge and colour properties to train a classifier such as Support Vector Machines or K-Nearest Neighbors is shared amongst most of the earlier studies. A comprehensive overview of these earlier studies and other uses of computational methods in art history is given in (Brachmann & Redies, 2017).

Recent advancements in Computer Vision and the field of computational fine arts have allowed experimenters to overcome the necessity of hand-engineering features. This was made possible when the appearance of large, well labelled public datasets such as Wikiart (Karayev et al., 2014), and the increase in computational power allowed for the application of deep neural networks such as CNNs. Instead of computing hand-engineered features, deep neural networks are capable of automatically extracting relevant patterns and finding intricate relationships amongst those. The approach is especially interesting for it needs no prior knowledge about art and relevant features of style. The first application of deep learning in the field of digital art history was conducted by Karayev et al. who utilized a CNN to classify painting styles (Karayev et al., 2014). They used a CNN trained for object recognition to derive painting patterns and managed to outperform most of the hand-engineered features on the task of classifying painting style (Karayev et al., 2014). The efficiency of CNN-based feature extraction was further confirmed for style (Bar, Levy, & Wolf, 2014), artist (David & Netanyahu, 2016) and genre classification (Cetinic & Grgic, 2016).

Methods & Materials

1.1 Method

To evaluate different methods of painting comparison, we apply the following framework compromising to following major components:

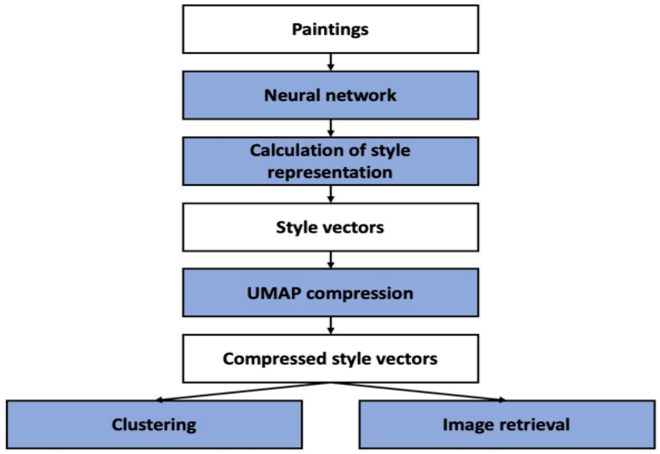


FIGURE 1. A framework for painting comparison. First, a Convolutional Neural Network is used to process a painting dataset. Using the network, style vectors are calculated to capture the painting's latent style features. Finally, the style vectors are compressed and used for painting comparison.

(1) First, A CNN is used to process the paintings. To investigate the applicability of the approach proposed by Gatys et al. a VGG-19 network (Simonyan, Zisserman, 2014) trained on object detection is employed. To study the extend of possible improvements the network is replaced by a Resnet-34 (He, Zsang, Ren & Sun, 2016) trained on style detection. (2) Then, for both networks numeric style representations are calculated using Gram matrices. For the Resnet-34, the direct network embedding is discovered to be an efficient representation, too. (3) In the next step the style vectors are compressed using the UMAP Algorithm. (4) Finally, the compressed vectors are used to compare the paintings by clustering a dataset and retrieving an input painting's most similar paintings from a dataset.

1.2 Datasets

For training the network on a style task, an unpublished dataset consisting of 300.000 paintings was used. It contains paintings from all epochs starting in the year 1300 labelled for artist and/or art-epoch.

For testing, the clustering algorithm the Wikiart dataset presented (Bar et al., 2014) was used. To the author's knowledge, it is the most common dataset to evaluate art-related classification tasks and was used for example by (Bar et al., 2014; Chu & Wu, 2018; David & Netanyahu, 2016; Girshick, Donahue, Darrell, & Malik, 2014; Hentschel, Wiradarma, & Sack, 2016; Seguin, Striolo, Kaplan, et al., 2016). The dataset consists of over 80,000 paintings, making it one of the biggest publicly available labelled datasets.

The dataset covers paintings with various styles from 27 art epochs such as Baroque, Ukiyo-E or Abstract Expressionism.

All paintings are downsampled to 226 x 226 pixels. This provides a reasonable trade-off between resolution and computation time necessary for processing. Since textures such as brush strokes are vital for a paintings style more emphasis was given to resolution.

1.3 Convolutional Neural Network Architectures

Two different CNN architectures were implemented. The first network was used to study the feasibility of using Gram matrices for comparing paintings by style. For this, VGG-19 pretrained on the ImageNet dataset for object detection was employed. It was used to create numeric style representations in the form of Gram matrices. Whilst training on object detection the network has developed filters that recognize a wide range of patterns. Despite being formed for object detection many of the filters are relevant for style recognition and applicable to the task of extracting style from an image.

The second architecture was selected to improve the approach. The ResNet-34 architecture was chosen because it achieves state of the art results for artistic style detection (Lecoutre, Negrevergne, & Yger, 2017). In order to train the network on recognizing style, it was trained both on artist- as well as art-epoch-classification at the same time. This yields filters explicitly optimized for recognizing patterns relevant to the style of a painting. Training on both the prediction of a painting's artist and epoch allows the use of paintings labelled for artist and/or art epoch, resulting in a significantly larger dataset.

To perform simultaneous training on two tasks, the ResNet-34 architecture was modified to have two classification heads. This means the output of the final convolutional layer was fed to two separate dense networks, creating two separate predictions for each painting's artist and art epoch. The network's loss function was calculated by summing the loss of both heads. When the label for artist or art epoch was not available, the corresponding loss was set to zero. The network was trained on the training. The images were augmented using random transformations such as rotation and random cropping. The optimisation was conducted using the Adam optimizer with a batch size of 50. Using a 80/20 train-test split the network achieved an artist prediction accuracy of 73% and an style prediction accuracy of 57%.

1.4 Style Representation

Three different methods for calculating style vectors were applied: The first method was to calculate the painting's Gram matrix on a VGG-19 network pretrained on object detection. For its easy implementation and success with style transfer, the implementation of Gatys et al. was followed exactly. With increasing depth, the network's filters recognize increasingly complex patterns. By calculating the Gram matrix for layers at various depths an abstract representation of image style at various levels of textural detail was obtained.

The second method improves the stylistic representation by replacing the VGG-19 object detection network with a ResNet-34 trained to detect style.

The third method uses the ResNet-34 network's internal representation of the artwork. This is done by using the network's last hidden layer's embedding as a style vector. As the network was trained on a style detection task the layer was optimized to calculate representations describing the style of the painting.

1.5 UMAP and the 'Curse of Dimensionality'

When trying to cluster the data without dimensionality reduction, the clusters show no homogeneity in style, due to the 'Curse of dimensionality' (Bellman, 1966). It describes the rapid increase in volume when adding more dimensions to a mathematical space. As a result, the available data quickly becomes sparse. Consequently, the amount of data needed to obtain statistically sound results often grows exponentially. To compare two painting's style vectors the 'curse' needed to be 'banished'. For doing so, the style vectors dimensionality was compressed using the Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, & Melville, 2020) algorithm to compress the vectors $\vec{y}$ to 128 dimensions.
UMAP essentially builds a neighbour graph in the original feature space and then tries to find a similar graph in lower dimensions. The detailed process is fairly complex and can be found in the original paper. The algorithm was initialized with n_neighbours = 15 and min distance = 0.1.

1.6 Painting Style Comparison

To test if the approaches are suitable for the comparison of paintings and to evaluate the quality of the style representation two tasks were performed: First, the clustering of a test dataset on style. Second, retrieving the test dataset's paintings that are the most similar to a given input painting.

1.6.1 Clustering

For clustering, the compressed style vectors were clustered using the k-Means algorithm (MacQueen et al., 1967). To account for the stylistic diversity within the test dataset's 27 art epochs k = 128 centroids were used.

1.6.2 Clustering Quality Score

Conducting a human study on the subjective impression of the clusters' homogeneity would have provided interesting insights but was prohibitively expensive. Therefore the art-epoch labels were used as a proxy for style to quantify the clustering quality. A clustering method's 'quality score' q was calculated as

$$q = \frac{|postive|}{|paintings|} \tag{1}$$

To define the positive group each cluster was assigned the art-epoch label most frequently occurring in the cluster. Positive then denoted the subgroup of paintings that are in a cluster whose art-epoch label is the same as theirs.

1.6.3 Retrieval of Similar Images

For the retrieval of an input painting's most similar images the cosine similarity *d* of the respective style vectors $\vec{s}$ was used as a measure of similarity.

$$d = \frac{\vec{s}_1 \cdot \vec{s}_2}{\|\vec{s}_1\| \, \|\vec{s}_2\|} \tag{2}$$

1.7 Visualizing Network Filters

The filters learned during the training process were visualized using the optimization approach introduced by (Erhan, Bengio, Courville, & Vincent, 2009). This allows us to get insight into what patterns are used by the network to discriminate styles. It was done by iteratively generating an image maximally activate a filter. This process separates the patterns causing a filter's activation from the patterns merely correlating with it. The method provides a useful tool for understanding what a filter is specifically looking for.

First, a random image was initialized. It was fed to the trained networks with fixed weights and the average activation of the targeted filter was extracted from the desired layer. Then, the gradients of the extracted filter's activation were computed with respect to the input image pixel values. These gradients are then utilized to update the pixel values in a way that maximizes the average activation of the chosen filter. The steps are repeated until the image results in a filter's maximal activation. If such an image was optimized without any constraints, an image consisting of noise and high-frequency patterns would emerge from the optimisation. Dealing with this high-frequency noise has been a major challenge in feature visualization research. It can be reduced by applying various transformations in each optimisation step. Transformations such as jittering and rotating increase the transformation robustness. Blurring the image in every step adds a frequency penalization.

Results

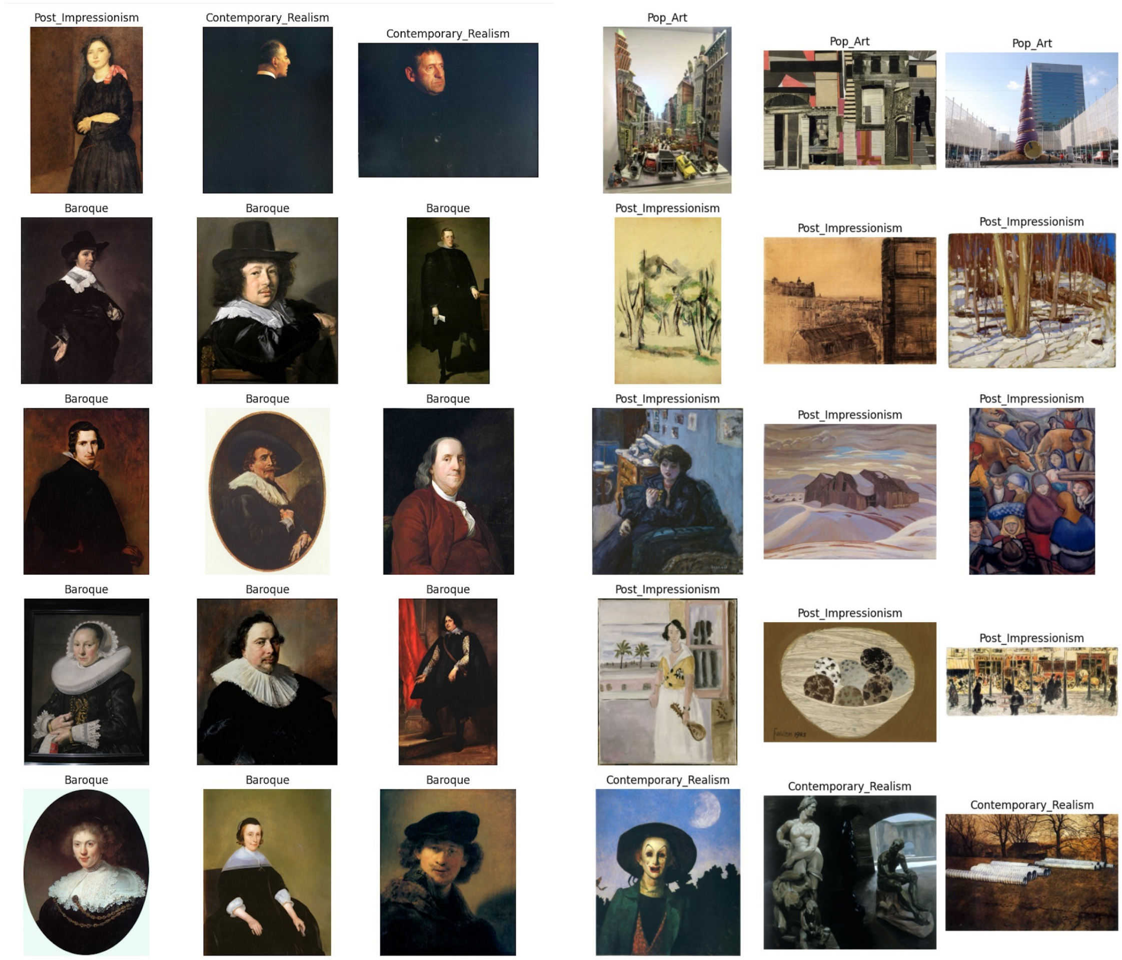
To evaluate the algorithms, two methods were applied. First, the algorithms are used to cluster the dataset and the clustering quality score is calculated. Second, the algorithms are used to find paintings most similar for a given painting. This process is evaluated subjectively. Lastly, the object detection networks activations were visualized and compared to those of the art-epoch/artist detection network.

1.8 Clustering Quality Score

Table 1 shows the clustering quality scores for clustering the test dataset using different approaches. Fig. 1 shows an excerpt from a cluster created by using the artist/art-epoch detection network with Gram Matrix calculation.

Approach	Quality Score
Object detection network - Gram	0.21
Artist/art-epoch detection network - Gram	0.40
Artist/art-epoch detection network - embedding	0.38

TABLE 1. Quality scores for clustering the Wikiart dataset using the k-means clustering algorithm. The style vectors are calculated using three different approaches. The quality score represents the percentage of paintings that are in a cluster whose most frequent art-epoch label is the same as theirs.



(a) a good cluster sample (b) a bad cluster sample

FIGURE 2. Two excerpts from clusters created by clustering artworks using Gram Style representations derived from ResNet-34 pretrained on image-style/artist classification.

1.9 Image Retrieval

To directly compare what the different algorithms consider to be similar for a specific image the most similar images for an input were retrieved (Fig. 2). The paintings considered to be most similar to the input are those whose style vectors have the largest cosine similarity to the input. The images retrieved by using the object detection network manage to capture the general colour tone and large patterns of the painting. However, they clearly differ from the input image's style.

Retrieving images using the network trained on art classification yields images of high visual similarity to the input image. The paintings retrieved in Figure 2a share the input's colour tone and fine, repetitive brushwork.

The paintings retrieved in Figure 2b share the input's medieval scenery consisting of multiple humans in front of the sky.

(a) Images retrieved for a abstract painting

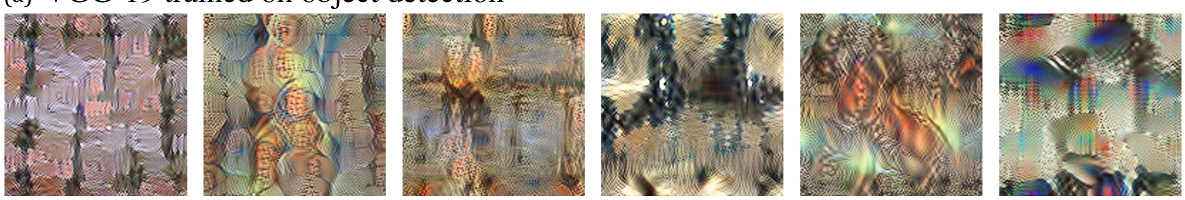
(b) Images retrieved for a scenic painting with humans

FIGURE 3. Paintings retrieved as most similar to the input image. Similarity is assessed by taking the cosine similarity of the painting's style vectors $\vec{s}$.

1.10 ResNet Activation Visualisation

To compare the filters used for the computation of the Gram matrices the last convolutional layer of the used networks was visualized. They were visualized by generating images that maximally activate a targeted filter as described in section 3.7.
The filters seen in Figure 3a were activated by images containing strong colour contrasts and clearly defined edges. The images show similar patterns in high frequency. They hint at objects such as eyes. The filters seen in Figure 3b are activated by more toned down colours and more homogeneous textures. The patterns generally are of lower frequency and varying size. Both small patterns such as brushstrokes visible in the first filter and large patterns such as a seascape visible in filter three emerge from the optimisation.

(a) VGG-19 trained on object detection



(b) ResNet-34 trained on art classification

FIGURE 4. Activations of the convolutional layer used to compute the gram matrices. The filters are activated by vastly different patterns.

Related Work

This work demonstrates the feasibility to cluster and compare images by style using features derived by calculating Gram matrices on a CNN pretrained for object detection. It further shows that the accuracy of stylistic representations could be significantly improved by using a network trained for art-epoch/style discrimination as a feature extractor. It is discovered that directly applying the convolutional embedding of the style specific network for image comparison yields comparable results to the Gram matrix representation.
It is difficult to evaluate the results against the related work. To the author's knowledge all similar papers aimed for the recreation of art-epoch clusters. An example is provided by (Gultepe, Edward. Conturo, & Makrehchi,

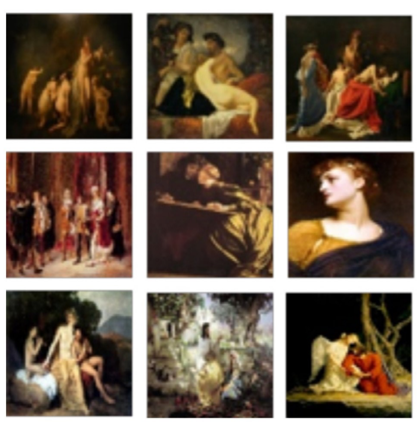
2018) who used deep learning feature extraction and the K-means algorithm to recreate 8 epoch clusters for 6,776 paintings. They achieved an F-Score of 0.469. Using handcrafted features and a classical SVM classifier to cluster eight painting epochs (Spehr et al. 2009) achieved an accuracy of 0.405.

Due to the human subjectiveness on which the definition of an epoch is based this approach was not followed by the author. Thus the cluster number was not determined by the semi-arbitrary borders set by art history but set large, allowing for much finer separation. However, recreating epoch clusters will be included in future studies for it allows for a better comparison of approaches. To further increase comparability more evaluation metrics will be employed.

One of the main limitations of this study is the difficulty to evaluate the results. This is due to the subjective nature of style. Also using the epoch label as a proxy for style inevitably leads to partially wrong judgements. As seen in Figure 4 the epochal label can be used as a proxy for style only so far. Images from the same epoch can look very different and images from different epochs can look very similar.

(a) Salvador Dali ’The persistance of memory’

(b) Rene Margritte ’Ceci n’est pas une pipe’



(c) Examples from Academism

(d) Examples from Realism

FIGURE 5. Two images from the same epoch can express very different style. Picture 4a and 4b both are classified as works of abstract expressionism, but

look very different. Images retrieved from (*Die Beständigkeit der Erinnerung und Salvador Dalis Beitrag zum Surrealismus*, n.d.; *The Treachery of Images (This is Not a Pipe) (La trahison des images [Ceci n'est pas une pipe])*, n.d.). 4c and 4d (both adopted from (Hentschel et al., 2016) are images that look very similar, but are labelled differently.

The method of style extraction by calculating Gram matrices was originally developed for style transfer. Thus, the improvements introduced in this work are likely to yield an increase in the quality of style transfer images if applied to the task. Furthermore, the development of methods for organizing large image datasets and retrieving similar images without manual annotation could provide a useful tool for art historians' provenance research.

For further research we suggest studying the impact of the improvements made to the algorithm in more detail. Thus, to conduct a further study including VGG-19 trained on a style specific task as well as ResNet-34 trained on object detection.

It is expected that fine-tuning the algorithm would allow a further increase in accuracy. For this, an investigation on the trade-off between the curse of high dimensionality and the loss of information inherent to the compression of data would be interesting. Finding the balance is likely to improve the meaningfulness of comparing two artwork's numerical style representation. Further investigation of the clustering algorithm is likely to increase performance, too. It would be interesting to find an optimal number of centroids k for the K-means algorithm minimizing the number of equivalent or heterogeneous clusters. Furthermore, spectral clustering (Ng, Jordan, & Weiss, 2001) could be used instead of K-means, which is recommended by (Gultepe, Edward. Conturo, & Makrehchi, 2018) for its unique ability to find non-spherical groupings. Another interesting study would be the investigation of the image resolution's impact on the comparison results. As a higher resolution retains many relevant patterns such as fine brushwork higher resolution is likely to significantly increase the quality of comparison.

References

Bar, Y., Levy, N., & Wolf, L. (2014). Classification of artistic styles using binarized features derived from a deep neural network. In *European conference on computer vision* (pp. 71–84).

Bellman, R. (1966). Dynamic programming. *Science*, *153*(3731), 34–37.

Brachmann, A., & Redies, C. (2017). Computational and experimental approaches to visual aesthetics. *Frontiers in computational neuroscience*, *11*, 102.

Cetinic, E., & Grgic, S. (2016). Genre classification of paintings. In *2016 international symposium elmar* (pp. 201–204).

Chu, W.-T., & Wu, Y.-L. (2018). Image style classification based on learnt deep correlation features. *IEEE Transactions on Multimedia*, *20*(9), 2491–2502.

David, O. E., & Netanyahu, N. S. (2016). Deeppainter: Painter classification using deep convolutional autoencoders. In *International conference on artificial neural networks* (pp. 20–28).

Die best¨andigkeit der erinnerung und salvador dal´ıs beitrag zum surrealismus.(n.d.).

https://blog.singulart.com/de/2019/09/19/die-bestaendigkeit-der-erinnerung-und -salvador-dalis-beitrag-zum-surrealismus/,.

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009, 01). Visualizing higher-layer features of a deep network. *Technical Report, Univerist´e de Montr´eal*.

Gardner, H. (1970). Children's sensitivity to painting styles. *Child Development*, 813–821.

Gatys, L., Ecker, A. S., & Bethge, M. (2015). Texture synthesis using convolutional neural networks. Advances in neural information processing systems, 28, 262-270.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016, 06). Image style transfer using convolutional neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)*.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 580–587).

Gultepe, E., Edward. Conturo, T., & Makrehchi, M. (2018). Predicting and grouping digitized paintings by style using unsupervised feature learning. *Journal of Cultural Heritage*, *31*, 13-23.

Retrieved from https://www.sciencedirect.com/science/article/pii/S1296207417301474 doi: https://doi.org/10.1016/j.culher.2017.11.008

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Hentschel, C., Wiradarma, T. P., & Sack, H. (2016). Fine tuning cnns with scarce training data—adapting imagenet to art epoch classification. In *2016 ieee international conference on image processing (icip)* (pp. 3693–3697).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Karayev, S., Trentacoste, M., Han, H., Agarwala, A., Darrell, T., Hertzmann, A., & Winnemoeller, H. (2014). *Recognizing image style*.

Keren, D. (2002). Painter identification using local features and naive bayes. In *Object recognition supported by user interaction for service robots* (Vol. 2, p. 474-477 vol.2). doi: 10.1109/ICPR .2002.1048341

Lecoutre, A., Negrevergne, B., & Yger, F. (2017, 11). Recognizing art style automatically in painting with deep learning. In M.-L. Zhang & Y.-K. Noh (Eds.), *Proceedings of the ninth asian conference on machine learning* (Vol. 77, pp. 327–342). PMLR. Retrieved from http://proceedings.mlr .press/v77/lecoutre17a.html

MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297).

McInnes, L., Healy, J., & Melville, J. (2020). *Umap: Uniform manifold approximation and projection for dimension reduction*.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th international conference on neural information processing systems: Natural and synthetic* (p. 849–856). Cambridge, MA, USA: MIT Press.

Seguin, B., Striolo, C., Kaplan, F., et al. (2016). Visual link retrieval in a database of paintings. In *European conference on computer vision* (pp. 753–767).

Spehr, M., Wallraven, C., & Fleming, R. W. (2009, May). Image statistics for clustering paintings according to their visual appearance. In Computational Aesthetics 2009: Eurographics Workshop on Computational Aesthetics in Graphics, Visualization and Imaging (pp. 57-64). Eurographics.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions.

In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

Tan, W. R., Chan, C. S., Aguirre, H. E., & Tanaka, K. (2016). Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. In *2016 ieee international conference on image processing (icip)* (p. 3703-3707). doi: 10.1109/ICIP.2016.7533051

The treachery of images (this is not a pipe) (la trahison des images [ceci n'est pas une pipe]). (n.d.). http://collections.lacma.org/node/239578, note = Accessed: 2021-06-05.