More Than Met the Eye: A Novel Deep Learning Embedding-Clustering Analysis of The Metropolitan Museum of Art's Photography Collection

James T. Wang
Carnegie Mellon University
Pittsburgh, PA 15213
jameswang@cmu.edu

Abstract

Is it possible to analyze a museum's collection by the aesthetic and semantic content of objects within, rather than historical metadata? In this paper, we propose and implement a novel methodology to analyze a collection of historical photographs. Each image is mapped to an embedding vector in a latent space by a Resnet-18 neural network model trained on ImageNet image classification. Then, the embeddings are clustered via a Ward hierarchical clustering model into "conceptual clusters." We apply this methodology to the Metropolitan Museum of Art's photograph collection, which primarily consists of 19th-century photos. Qualitatively, it performs surprisingly well in extracting clusters representing specific aesthetic ideas, semantic subject types, and photographic technologies, especially since neither the embedding network nor the clustering model has prior knowledge about the specific subject matter. With these clusters, we analyze their contents, measure how they change over time, and speculate on further applications of the methodology.

1 Introduction

How do concepts in photography change in popularity over time? Is it possible to analyze a museum's collection by the aesthetic content of objects within, rather than historical metadata? Is it possible to quantitatively measure the aesthetic-conceptual similarity between works of visual art?

While these questions have been previously impossible to answer, recent developments in deep machine learning offer a new possibility: image embedding. Image embedding is where a deep neural network, trained on millions of generic images, leverages what it has learned to place images onto a low-dimensional latent space. In other words, it creates a cloud of points floating in an abstract space, where each point (dubbed an "embedding") represents an image; similar images are close together, while dissimilar images are far apart. Therefore, we can quantitatively measure the aesthetic difference between artworks with the distance between their embeddings.

Furthermore, suppose we further group these embeddings into clusters. In that case, we can discover "genres" of artwork where many photographs follow the same aesthetic ideas, such as headshot portraits, natural landscapes, and mugshots.

In this paper, we apply this novel methodolgy to analyze the Metropolitan Museum of Art (the Met) photography department collection. By looking into these clusters' contents and their distribution across time, we can gain an intimate understanding of the Met's extensive collection of photography, its composition, and how it relates to photography history. Therefore, this analysis offers a new way to understand museum collections and how they evaluate and acquire historical objects.

2 Literature Review

There has been rising interest in quantitative analyses of visual artwork due to the convergence of two trends. First, museums and other institutions have increasingly focused on collection digitization and

Large-scale Social Phenomena (Fall 2021).

open access, where digitized records and object metadata are made publicly available with minimal copyright restrictions via the Internet. This movement has been spearheaded by the OpenGLAM (galleries, libraries, archives, and museums) initiative [3]. The movement toward the release of cultural information has led to open access of the collections of the Metropolitan Museum of Art [16], the Smithsonian [7], and the National Gallery of Art [2].

Additionally, the field of deep machine learning has made significant progress in analyzing visual data in a meaningful manner. For example, the ImageNet Large Scale Visual Recognition Challenge classification task, a competition to build a machine learning model to classify the contents of millions of images into a thousand categories, has seen an increase in accuracy from 71.8% to 96.4% within a few years, from 2010 to 2015 [13][9]. Additionally, the idea of embeddings, where a neural network maps words to points in a relatively-low-dimensional latent space, has been translated from its original use in linguistic applications [11] toward image data [8].

There has been a significant amount of research toward the application of machine vision to collections of artwork. However, previous research has rarely analyzed the collection as a whole. For example, work has been done to automatically create image descriptions [5] and identify forgeries [6]. Similarity metrics for artwork have been created to predict style or genre [14] and identify compositional references to earlier artwork [15]. Notably, a team of Carnegie Mellon University and University of Pittsburgh researchers utilized a similar method to create a high-level grid-based visualization of paintings held by the National Gallery of Art, as a means to compare the breadth of different collections [10].

3 Methods

The Metropolitan Museum of Art offers a public dataset and API for information about objects across its collection. Additionally, every public-domain work in its collection is available under Creative Commons Zero as part of its Open Access program, including 5,819 works in the Photographs department, primarily from the 19th century.

These photographs are mapped onto a 512-dimensional latent embedding space by a Resnet-18 Img2Vec neural network trained on the ImageNet dataset. Resnet ("residual network") is a performant but relatively simple neural network architecture, shown in Figure 1. It has been trained to classify images on ImageNet, a standard benchmark dataset of 1.3 million Flickr photographs spread across 1000 categories [13].

Even though the average ImageNet image is quite different-looking from the average Met collection image, the latent space produced by the embedding model is very well-behaved. This pre-training process has allowed the network to learn the textures, patterns, and objects common across all photography, such as faces, bodies, and buildings. For example, Figure 2 contains a sample of images, along with 3 of their nearest neighbors and 3 of their farthest neighbors, showing that the latent space matches well with qualitative aesthetic and semantic distance. Additionally, Figure 3 displays a cluster in a projected view of the latent space, and the clusters are dense & relatively well-separated.

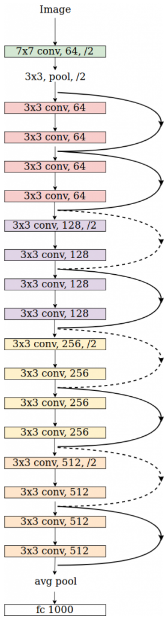

Figure 1: The Resnet-18 neural network architecture

Thousands of points scattered across 512 dimensions is hard to interpret, so the points were further grouped into clusters using a Ward agglomerative hierarchical clustering algorithm. This algorithm initially treats each point as its own cluster, then recursively merges neighboring clusters while minimizing within-cluster variance. The result is a large binary tree of clusters, dubbed a "dendrogram," as shown in Figure 4. Each leaf node represents an image-point, and each non-leaf node represents the merging of two clusters into a larger cluster. The height represents the distance between clusters. In other words, each cluster could be understood as a semantic or aesthetic idea; as we go up the dendrogram, similar concepts are grouped together into more and more general ideas.

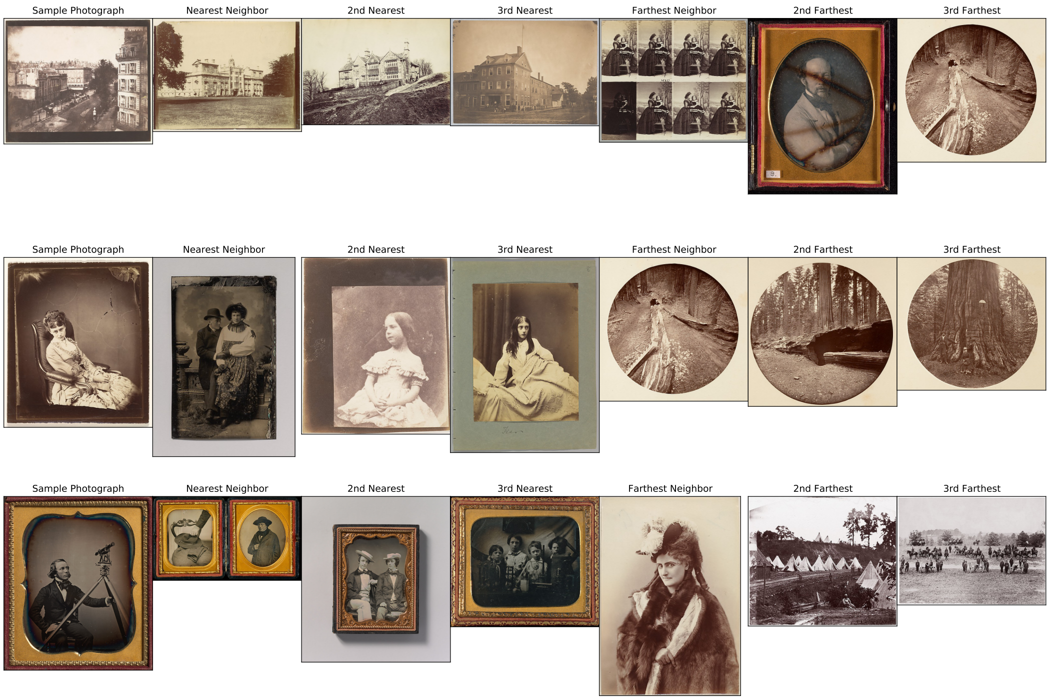

Figure 2: A sample of photographs, along with the photos that are nearest and farthest from it in the latent embedding space

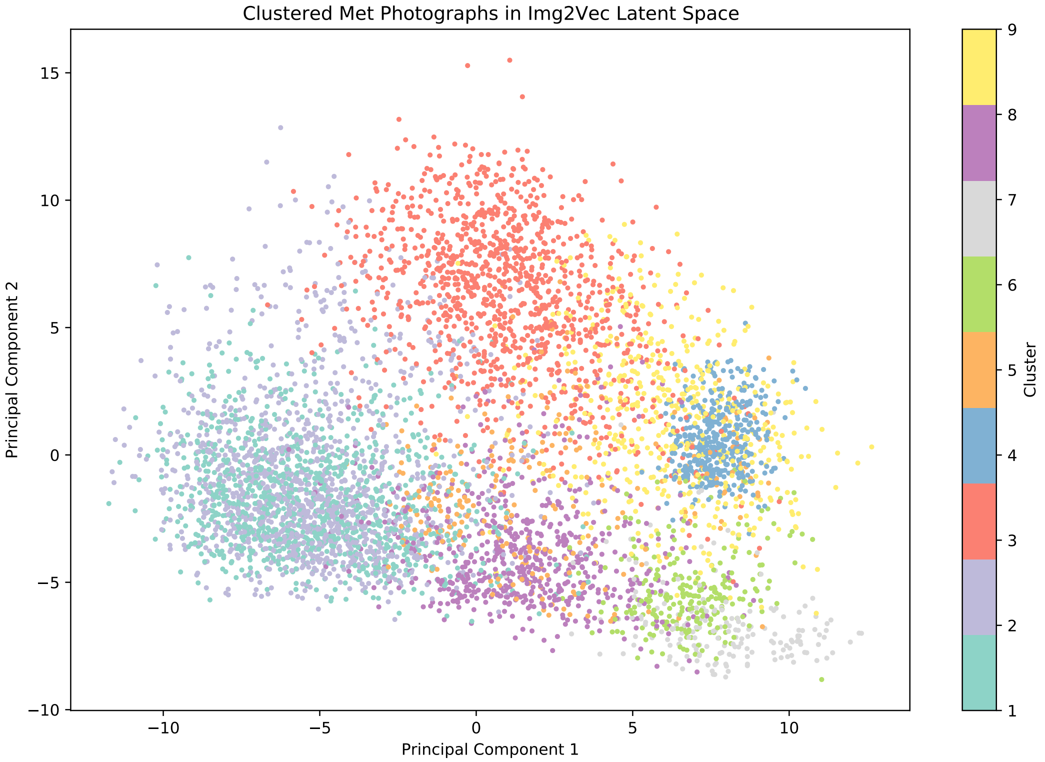

Figure 3: A PCA-projection of the latent space, and the clusters within. An animated, 3-dimensional version of this plot can be seen at `https://imgur.com/a/bpnxxEJ`.

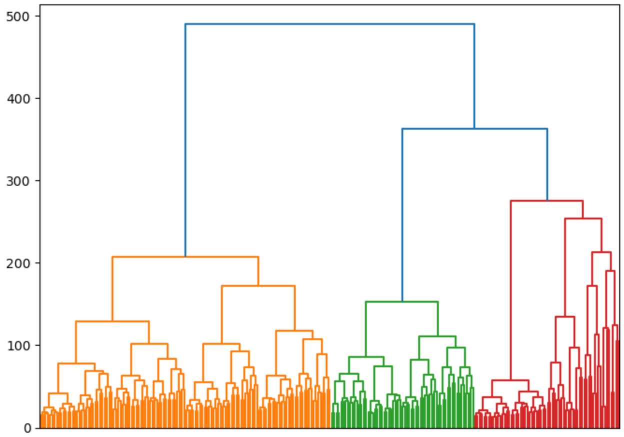

Figure 4: The Ward clustering dendrogram

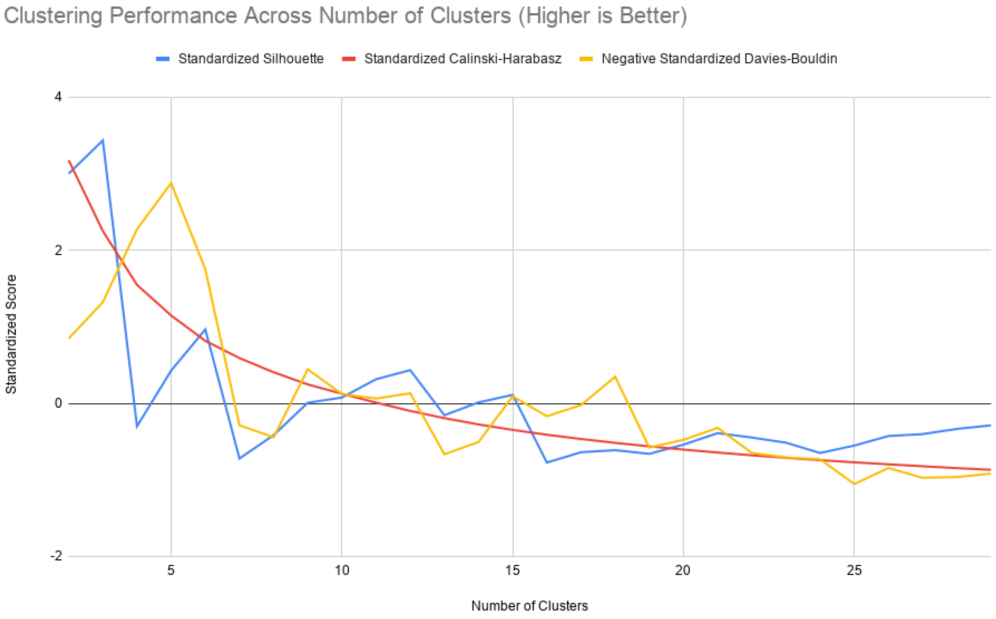

Figure 5: Clustering performance across number of clusters

We can select the "resolution" (so to speak) of our clustering method by choosing how many clusters we will ultimately use, by dictating when we will stop merging clusters. Figure 5 shows three clustering performance metrics across the number of clusters; we decided to use 9 clusters. This number is not too few as to fail to capture the diversity of concepts across 19th-century photography, yet not too many as to perform poorly and be difficult to interpret.

4 Results

With these 9 clusters, each cluster's contents were inspected to understand the ideas they represent and the reason they were grouped together. With these understandings in hand, we analyzed how cluster size varied with time and how it relates to photography history. Finally, a preliminary novelty-transience analysis was done as a potential future application of this embedding-clustering methodology.

4.1 Cluster Interpretation

Figure 6 shows a sampling of images across all 9 clusters; there are common threads between each cluster's images. The likely themes of each cluster are as following:

1. Cluster 1 primarily consists of landscapes of nature and other organic subjects. It includes work from U.S. Civil War photographers such as Mathew Brady and Western United States explorers such as Carleton Watkins & Timothy O'Sullivan. Additionally, it includes the more natural, non-architectural photos of archeological photographers like Félix Teynard and Auguste Salzmann.
2. Cluster 2 is architectural photography. There are many photographs from *Missions Héliographiques*, a project to photograph decaying French landmarks and monuments. However, the bulk of these photographs are from expeditions by Teynard, Salzmann, and other archeological photographers.
3. Cluster 3 are half- and full-body portraits. Three portraitists created over 40% of this cluster. The first is Pierre-Louis Pierson, who photographed the rich and famous in Paris, France. The other two are David Octavius Hill and Robert Adamson of the profoundly influential Hill & Adamson photography studio in Edinburgh, Scotland.
4. Cluster 4 are mugshots. Every photograph in the cluster except one was taken by Alphonse Bertillon, French police officer and inventor of the standardized mugshot.
5. Cluster 5 consists of a hodgepodge of photos, but most of them have a round border. This cluster is interesting because it contains over 91% of the work of Guillaume-Benjamin-Amand Duchenne de Boulogne. He was a French neurologist who researched electrophysiology by electrocuting the face muscles of an "old toothless man" and photographing the results.
6. Cluster 6 are daguerreotypes and tintypes. These types of photographs were sensitive to exposure to air and stored in protective cases. 53% of the cluster are from unknown artists; these photographs were relatively affordable and sold to everyday people by a large industry of obscure portraitists.
7. Cluster 7 contains work from two books: *Photographs of British Algae: Cyanotype Impressions* by Anna Atkins, and *Photographic Views of Sherman's Campaign* by George Bernard. Photographs from both books were digitized by the Met in a similar fashion, where the entire book page is captured, with the rest of the book and a grey table visible behind it. *British Algae* consists of botanical impressions using an early form of photography dubbed the cyanotype, giving each print a bright blue color. Meanwhile, *Sherman's Campaign* contains Bernard's U.S. Civil War photographs taken while under General Sherman's command.
8. Cluster 8 contains another hodgepodge of photos from across time and style. However, it appears that most of the images have been digitized by the Met in such a fashion where the backing paper or table is also captured around the photo. It appears that several stereographs are also included due to the similar framing, which are prints containing two offset images to act as an early form of 3D imagery.
9. Cluster 9 contains many *cartes-de-visite*. These were cheap, small "visiting cards" to be traded among friends, which later transformed into mass-produced photos of the rich and famous for collecting, like baseball cards. Several other photos with similar aesthetics to cartes-de-visite are also included.

The clusters seem to fall into three general categories:

1. Clusters of photos of natural and architectural landscapes (clusters 1 and 2, "landscape clusters")
2. Clusters of portraits (clusters 3, 4, 6, 9, "portrait clusters")
3. Clusters which that been grouped based on the framing or digitization method (clusters 4, 7, 8, "framing clusters")

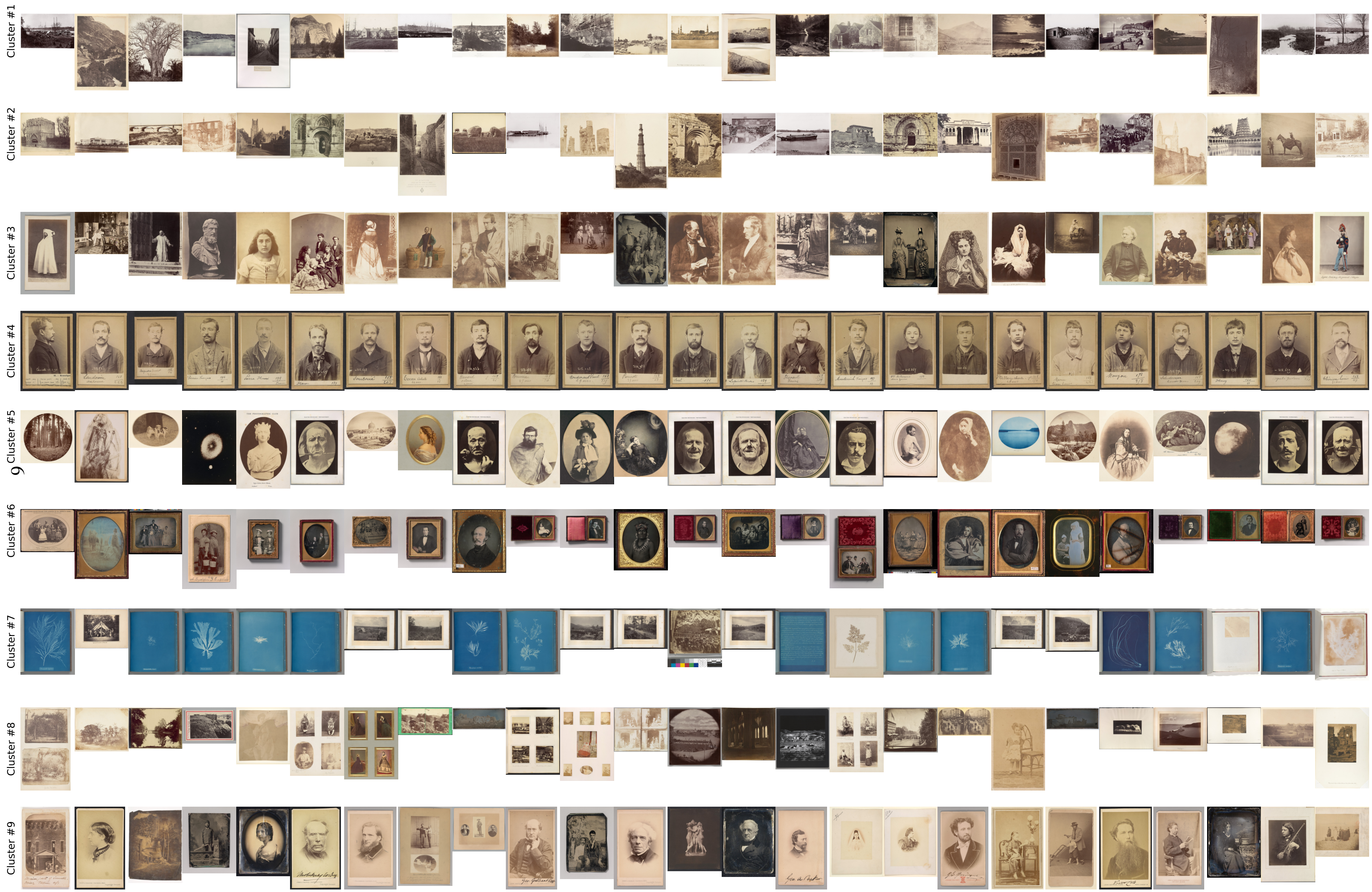

Figure 6: A random sample of photos from each cluster

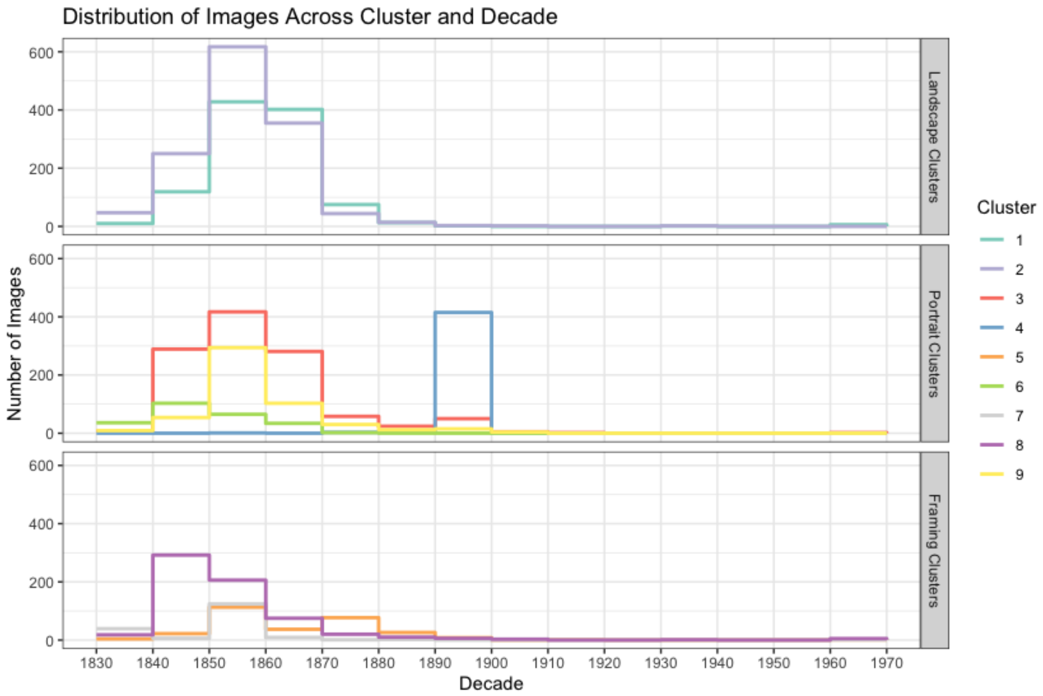

Figure 7: The distribution of photographs by cluster, grouped by cluster type, across decades

4.2 Distribution Across Time

How have these clusters varied with time? Figure 7 shows the distribution of the contents of each cluster by decade. Nearly all of the Met's Open Access program's available photographs are between the invention of practical photography in 1839 to the end of the 19th century.

We can see some trends in this plot; for example, natural and building landscapes were some of the first applications of photographic technology. The number of images in those clusters rises precipitously as the technology becomes more widespread. We also see that natural landscapes overtake architectural images in the 1860s, perhaps due to the number of photographs taken of the U.S. Civil War and the Western United States during that time.

We also see that fragile daguerreotypes and tintypes (cluster 6) fall out of favor; images created using this method were unique direct positives, meaning that each image was unique and could not be copied. The alternative were negatives, where the camera captures a negative image that could then create many positive prints from the same capture. As negative photography technologies (such as albumen prints) continued to develop, we see these easily-reproducible photos replace direct positive methods and dominate the market, as represented by carte-de-visites (cluster 9).

The exception, of course, is Alphonse Bertillon's mugshots (cluster 4). He invented the standardized mugshot in 1888 as part of the "Bertillon system" of criminal identification. Therefore, the Met collection includes hundreds of mugshots of Bertillon's French criminals over the early 1890s.

The third group's trends — clusters that have been grouped based on the framing or digitization method — are much less meaningful. It appears that the zoomed-out style of digitization (cluster 8) applies most readily to early photographic works from the 1840s and 1850s.

4.3 Preliminary Novelty-Transience Analysis

As noted before, each cluster can be interpreted as an aesthetic photographic concept. Then, the distribution of clusters for a given decade can be understood as a snapshot of the breadth of ideas popular during that decade. With these distributions, we can measure how "suprising" a decade is,

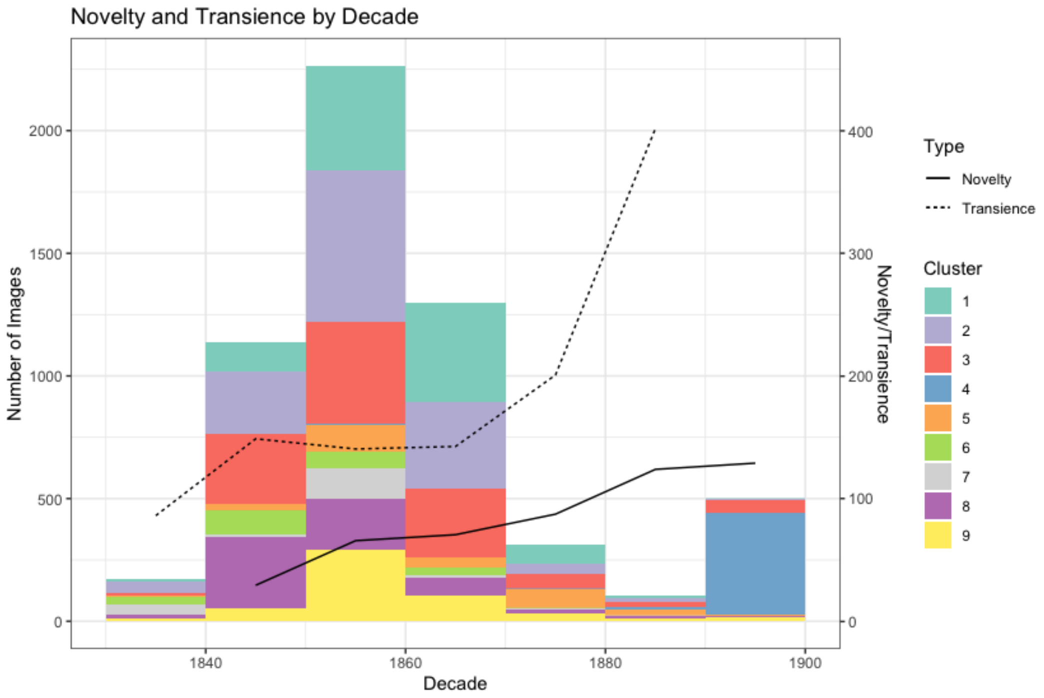

Figure 8: Novelty and transience across decades

compared to what came before or after, with the Kullback–Leibler divergence (KLD). It represents the degree to which a new probability distribution differs from a reference distribution.

Therefore, it is possible to measure a decade's conceptual "novelty" by calculating the average KLD of its cluster distribution, given each previous decade's cluster distribution. Even further, it is possible to find the decade's "transience" (that is, how quickly ideas dissipated) via the average KLD of the distribution given each subsequent decade [12].

An initial novelty and transience calculation was performed on the decades 1830 to 1900. The results are shown in Figure 8. There is a strong overall rising trend in both transience and novelty. However, it is not easy to know how much of this trend is due to the Met's curatorial choices or indeed seen across the medium.

If the latter is true, this could be interpreted as an increase in the rate of innovation (and failure) as photographic technology improves and becomes more accessible. Additionally, the spike in transience in the 1840s, soon after the invention of practical photography, adds an intriguing subtlety. Perhaps, so soon after the creation of such groundbreaking technology, not every application that was explored ended up working out.

5 Discussion

Overall, the embedding-clustering methodology performed surprisingly well grouping images into reasonable semantic and aesthetic categories. With zero prior knowledge of 19th-century photography, the embedding network and clustering algorithm were able to group the Met's photography collection into clusters, most of which represent a specific photographic technology, style, or subject type.

This analysis does suffer from the embedding network becoming "distracted," as it were, by the borders of some digitized photographs. Future work may require a manual sweep through the entire dataset, cropping digitized files to contain just the image itself without borders or backgrounds. (Alternatively, an automated frame-detection method may be developed.)

Additionally, the analyses of cluster contents and cluster sizes across time show that this methodology can highlight how collections represent changes throughout photographic history. However, ultimately, this analysis has suffers from having two occasionally-conflicting confounding factors: 1) the forces of history on photography, and 2) the curatorial work the Met has done to shape its collection.

To a certain degree, this methodology has profiled not the history of photography as a whole, but how the Met has valued and acquired their collection. By definition, curation does not aim representation of all artwork; the field instead strives to collect and display only the most salient and meaningful works.

We can see this at work most strikingly with Alphonse Bertillon's mugshots. There are only 16 mugshots in the Met's collection after 1895, since, as this criminal identification technique becomes more common, each mugshot becomes less notable. Therefore, the Met's collection of mugshots, a profoundly important application of photography, is significantly skewed toward the work of one man.

This problem points to two directions for future work. On the one hand, this shows that this methodology may be able to compare the contents of the collections of different institutions. Since all images map to the same latent embedding space, we can find what types of work are better represented in one collection than the other.

On the other hand, it may be possible to merge many institutions' collections to gain a more rounded, representative view of all photography. Examples of these merged datasets can be found with Google Arts & Culture [1] and Wikidata [4]. Analysis on these larger datasets might be able to quantitatively corroborate or disprove photographic historical narratives, or show how novelty and transience rise and fall in reaction to the creation of a new artistic medium and changes in technology.

References

[1] URL https://about.artsandculture.google.com/.

[2] URL https://images.nga.gov/en/page/openaccess.html.

[3] URL https://openglam.org/.

[4] URL https://www.wikidata.org/wiki/Wikidata:WikiProject_Visual_arts.

[5] Brendan Ciecko. Ai sees what? the good, the bad, and the ugly of machine vision for museum collections – mw20 | online. *MuseWeb*, MW20(MW 2020), Jan 2020. URL https://mw20.museweb.net/paper/ai-sees-what-the-good-the-bad-and-the-ugly-of-machine-vision-for-museum-collections/.

[6] Ahmed Elgammal, Yan Kang, and Milko Den Leeuw. Picasso, matisse, or a fake? automated analysis of drawings at the stroke level for attribution and authentication. *arXiv:1711.03536 [cs, eess]*, Nov 2017. URL http://arxiv.org/abs/1711.03536. arXiv: 1711.03536.

[7] Alise Fisher and Linda St. Thomas. Smithsonian releases 2.8 million free images for broader public use, Feb 2020. URL https://www.si.edu/newsdesk/releases/smithsonian-releases-28-million-free-images-broader-public-use.

[8] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26, pages 2121–2129. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/7cce53cf90577442771720a370c3c723-Paper.pdf.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv:1512.03385 [cs]*, Dec 2015. URL http://arxiv.org/abs/1512.03385. arXiv: 1512.03385.

[10] Matthew Lincoln, Golan Levin, Sarah Reiff Conell, and Lingdong Huang. National neighbors - cmu dh, Nov 2019. URL https://nga-neighbors.library.cmu.edu/about.

[11] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv:1309.4168 [cs]*, Sep 2013. URL http://arxiv.org/abs/1309.4168. arXiv: 1309.4168.

[12] Jaimie Murdock, Colin Allen, and Simon DeDeo. Exploration and exploitation of victorian science in darwin's reading notebooks. *Cognition*, 159:117–126, Feb 2017. ISSN 00100277. doi: 10.1016/j.cognition.2016.11.012. arXiv: 1509.07175.

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and et al. Imagenet large scale visual recognition challenge. *arXiv:1409.0575 [cs]*, Jan 2015. URL http://arxiv.org/abs/1409.0575. arXiv: 1409.0575.

[14] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv:1505.00855 [cs]*, May 2015. URL http://arxiv.org/abs/1505.00855. arXiv: 1505.00855.

[15] Babak Saleh, Kanako Abe, Ravneet Singh Arora, and Ahmed Elgammal. Toward automated discovery of artistic influence. *arXiv:1408.3218 [cs]*, Aug 2014. URL http://arxiv.org/abs/1408.3218. arXiv: 1408.3218.

[16] Loic Tallon. Introducing open access at the met, Feb 2017. URL https://www.metmuseum.org/blogs/digital-underground/2017/open-access-at-the-met.