Automatic Identification and Classification of Portraits in a Corpus of Historical Photographs

Taylor Arnold[1,*,†], Lauren Tilton[2,†] and Justin Wigard[2,†]

[1]Linguistics Program, Carole Weinstein International Center, 211 Richmond Way, Richmond, VA 23173, U.S.A
[2]Rhetoric & Communication Studies, 231 Richmond Way, Richmond, VA 23173, U.S.A

Abstract

There have been recent calls for an increased focus on the application of computer vision to the study and curation of digitised cultural heritage materials. In this short paper, we present an approach to bridge the gap between existing algorithms and humanistically driven annotations through a case study in which we create an algorithm to detect and and classify portrait photography. We apply this method to a collection of about 40,000 photographs and present a preliminary analysis of the constructed data. The work is part of the larger ongoing study that applies computer vision to the computational analysis of over a million U.S. documentary photographs from the early twentieth century.

Keywords

computer vision, cultural heritage, photography, public humanities

1. Introduction

1.1. Motivation

In this paper we present work that adapts and applies computer vision algorithms to aid in the discovery and use of historic digital photography [2, 9]. Rather than treating cultural heritage images as a monolith, whereby computational approaches are often developed and applied without attention to the form of cultural heritage in technical scholarship, we pursue technical research with computer vision that considers the specificity of photography as a medium, social practice, and source of evidence for humanistic inquiry [4]. We present work on a photography collection from the early 20th century, totaling nearly 40,000 photographs and held by the U.S. Library of Congress (LoC). The number of photographs in the LoC collections provides not only a scale that benefits from the use of computer vision, but speaks to the necessity of experimenting and developing approaches to computer vision for access and discovery of images. Our work is designed to support the LoC's work to "expand access" and "increase discoverability" to their collections.

CHR 2022: Computational Humanities Research Conference, December 12 – 14, 2022, Antwerp, Belgium
*Corresponding author.
†These authors contributed equally.
tarnold2@richmond.edu (T. Arnold); ltilton@richmond.edu (L. Tilton); jwigard@richmond.edu (J. Wigard)
https://statsmaths.github.io (T. Arnold); https://anon@anon.org/ (L. Tilton); http://justinwigard.com/ (J. Wigard)
0000-0003-0576-0669 (T. Arnold); 0000-0003-4629-8888 (L. Tilton); 0000-0003-0124-5934 (J. Wigard)

Our project builds on recent work at the intersection of machine learning and photography [7, 10]. Corrin, Davis, Lincoln, and Weingart's CAMPI paper offers a report on the use of computer vision for processing digital photograph collections [6]. The scope includes the assessment of back-end metadata generation for visual similarity search, duplicate or close-match detection, and utilizing image similarity for tagging. Our work looks at the feasibility of apply a larger set of algorithms for a broader approach to metadata classification. This new approach gives attention to the specificity of the features of photography such as composition and types of photography.

The case study described in this paper is part of the ADDI project, which makes interventions in several areas.[1] The project investigates the accuracy and appropriateness of different types of computer vision algorithms applied to historic photography. Because current algorithms are designed and trained on 21st century color images, potential challenges for these algorithms include identifying and labeling historic objects as well as analyzing black and white images, particularly legacy preservation digital formats [9]. The project models the feasibility of using algorithmically generated metadata features for automated search and discovery with attention to the ethical considerations of using computer vision. We investigate the use of algorithmically extracted features to directly create automated metadata to facilitate discovery. Specific attention will be to given metadata categories that reveal features of photographs such as composition. Questions include the usefulness of automated features for search and discovery and how to display automated features to general publics using best practices from the field of human-computer interaction (HCI). Finally, the project assesses the necessary technical architecture for running and storing the results of computer vision algorithms within a cloud-based architecture. In this paper, we present an analysis that models these four interventions applied to a specific set of photographs.

The *George Grantham Bain Collection* consists of nearly 40,000 black-and-white photographs taken by one of the earliest news picture agencies in the United States [11]. Most images come from the first two decades of the 20th century. They document a wide range of activities, including quotidian scenes of shops, gas stations, and lunch counters, major political rallies, University football games, weddings, and funerals. One type of image that we found to be particularly prominent when browsing through the collection are formal portrait photographs. There are clearly many of these in the collection, but they are not directly identified by metadata fields or a consistent description in the photograph titles. The goal in this paper is to determine how we can use computer vision annotations to identify and describe the the portrait photographs found within the Bain collection.

2. Detecting studio portraits

2.1. Locating people

As a starting point, we investigated the results of the application of a region segmentation algorithm to every image in the collection [1, 3].[2] This algorithm attempts to associate each pixel

[1] Information about the larger project can be found at https://github.com/distant-viewing/addi.

[2] We used a model trained on the COCO segmentation dataset using the R50-FPN architecture from the Detectron2 model zoo (137260431).

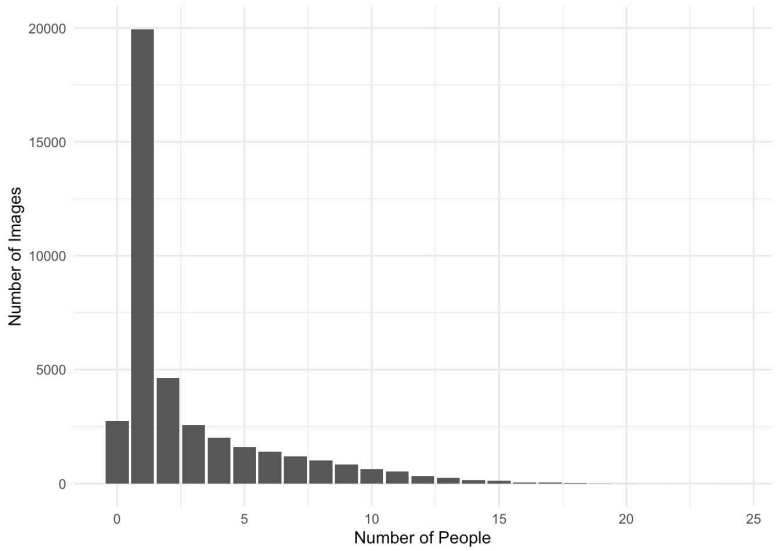

Figure 1: Distribution showing the number of detected people in each photograph in the *Bain* collection. Notice that the modal number of people is one; these correspond to (a superset of) the portraits in the collection.

in an image with a object type or background region (such as the ground or sky). Though technically not an "object," one of the object types detected by the algorithm are people. Figure 1 shows the number of images in the Bain collection based on the number of people detected by the region segmentation algorithm. Two interesting things stand out in the results. First, notice that there are very few images that contain no people. This leads us to assume that there are not many images that contain only the built or natural environment. It also indicates potential patterns about how Bain visually defined news through the centering of people. Secondly, we see that images with a single person can be clearly identified. There are far more images with one person compared to any other number of people. These, likely, are where most of the portrait photographs can be found.

2.2. Bimodal distribution

To investigate further, we need to understand more about the people detected in the images. We can do this by looking at the proportion of the image frame that is taken up by people. Specifically, we will look at a density plot of the proportions of the images taken up by people based on whether there is only one person or multiple people. Images without people are excluded. Our primary interest is the shape of the images with one person; the other images will help as a point of comparison. A density plot is shown in Figure 2. Interestingly, we see that for images with only one person, the proportion of frame image taken up by the person concentrates around two different values. One is around about 20% of the image and the other is around 50% of the image. As a comparison, notice that the density curve for images with two or more people has a sharp peak around 3%, with a steady decrease for larger proportions.

To understand more, we will look closer at the images with a single person from each of the

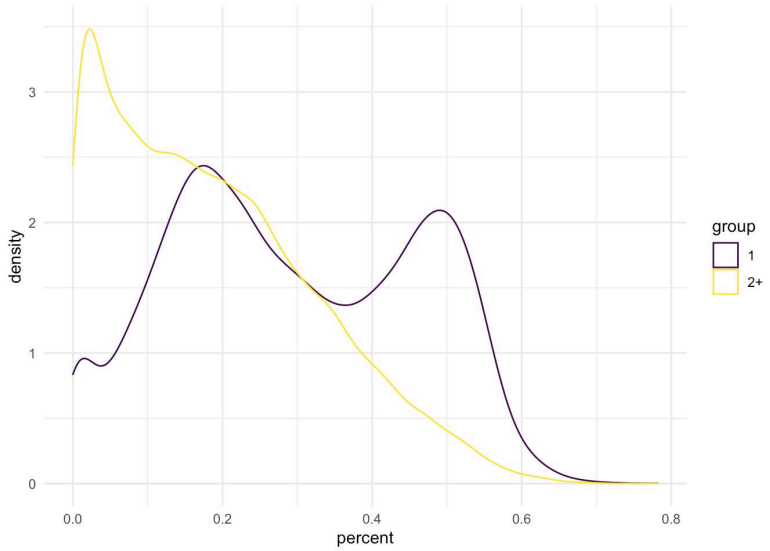

Figure 2: Density plot showing the percentage of the frame taken up by people based on whether there are one or more than one person detected. There is a bimodality of the single-person distribution, showing modes around 18% and around 47% of the frame.

modes of the density plot. Note that the process of moving back and forth between aggregating the data and looking at individual images is a common and fruitful mode of analysis throughout our work. Figure 3 shows 20 random photographs that have a single detected individual that takes up between 15% and 20% of the image frame. For comparison, Figure 4 shows 20 random photographs that have a single detected individual that takes up between 45% and 50% of the image frame.

Looking at the images, we can now understand the difference between the two types of photographs containing a single detected person. In the first mode, most of the images feature the entire body of a single person shown in an interesting place; for example, a baseball player on the field, a man in a radio station, or a woman in front of a sewing machine. In contrast, the second mode primarily contains portraits of people shown from the chest upwards. The majority of these images appear to be shot in a studio setting and a neutral background. The person in the frame is often dressed in a formal wear or an official uniform. They seem to be looking directly at, or just slightly off, the camera.

It seems, then, that we can use the number of people (1) and the percentage of the frame taken up by a person (around 50%) to determine if an image in this collection is a studio portrait. However, note that these general patterns of the two modes are generally accurate but not perfect. One image in the second mode is an interestingly framed image of a man playing the piano; two of the images in the first mode are full-length shots taken in a studio setting. So, in using these derived annotations, we should keep in mind that there will be some errors. This does not stop us from using the results in aggregate or as a general method for search and discovery. However, if one wanted to add this information directly into the archival metadata tag directly, we will want to be clear that the categorization is algorithmically generated.

Figure 3: Twenty randomly selected images from the first mode of the single person images, in which around 18% of the frame is taken up by the person.

The analysis also opens up interesting questions about news and visual culture. Are their social roles that are documented in certain ways compared to others? Our initial data suggests that activities such as dance and sports (which we could generalize to a category called performers) as well as women more generally are often photographed with visual information to clearly communicate to the audience the role of the person. In other words, the scene is their skill and helps the viewer understand why they are being featured. On the other hand, the studio portrait with a neutral or decorative background draws the eye to the face and clothes. There is little extra information to indicate exactly who the person is. Like the portraits that line government buildings with a name engraved in brass, the style of the portrait is designed to convey the person's prominence. It appears that certain roles in society such as military officials and politicians are being granted the visual power and cultural prominence of the close up. There is significantly more analysis to bolster this initial observation, but the initial differences are opening up questions and potential (historical) patterns regarding the relationship of framing, social position, and power [8].

Figure 4: Twenty randomly selected images from the first mode of the single person images, in which around 47% of the frame is taken up by the person.

2.3. Application

Now that we have identified the set of portrait photographs, what can we do with them? From an access perspective, we could identify these photographs and create an exhibit or digital public project focused specifically on these images. As a form of analysis, we might try to identify how other archival metadata compares to portrait photographs. As one example, we can look at the distribution of photographs from the Bain collection by time based on whether an image is a portrait or not. Figure 5 shows a density plot of these results.

Looking at the density plot shows that the dates of photographs seem to cluster around "round" dates, such as 1900, 1910, 1915, and 1920. This is likely an artifact of the data collection rather than an interesting feature of the data itself. Unfortunately, a significant portion of the collection was lost in a fire. Specifically on the topic of the portrait photography, what is most interesting is that what we have tagged as studio portrait photographs seem to be equally distributed across the same time periods of the rest of the collection. So, the set of portrait photographs are an important element of the Bain collection through the early 20th century rather than being a feature of only a few years.

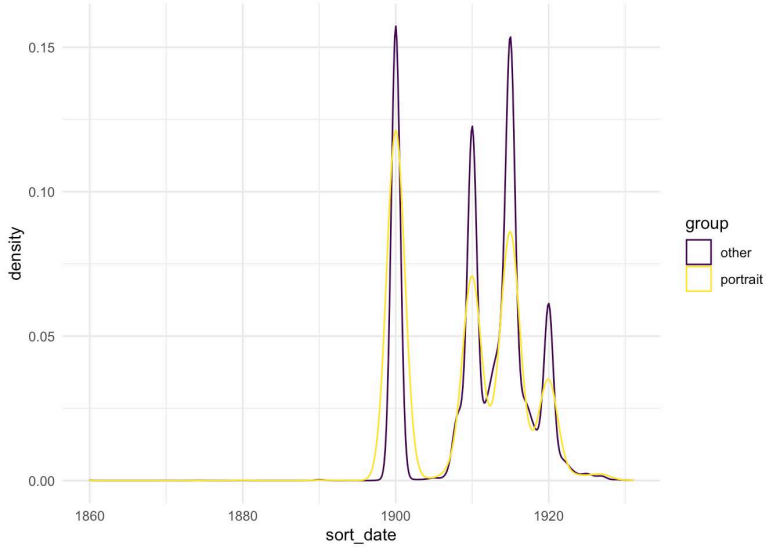

Figure 5: Distribution of the one-person photographs by the date listed in the archival data. Notice that there are bumps around particular clusters but that the mixture of the two modes seems to be relatively constant.

3. Portrait classication

3.1. Orientation using pose detection

Another way of looking at the portrait photography in the Bain Collection is by using a pose detection algorithm to estimate the orientation of people looking at the camera [5].[3] This will help us understand if people are rotated to the left, right, or squarely looking into the camera. The pose detection algorithm can help with this by allowing us to compare the position of different body parts relative to one another. As a starting point, we compared the distance between one's nose with their right and left ears. Calculating which ear is closer in two-dimensional space to the nose is a way of detecting how the face is framed relative to the camera. Applying this algorithm to the portraits in the Bain Collection shows that there seems to be no particular preference for poses to the left or right.

In order to understand these results, let's look at some of the images based on their pose. Figure 6 shows 20 randomly selected images that appear to be posed to the right and for comparison, figure 7 shows 20 randomly selected images that appear to be posed to the left. Looking at these images we can see that they do seem to correctly identify the orientation of people's faces. However, our algorithm only uses the location of the ears and therefore is unable to to detect which way people's actual eyes are being directed. One trope we see in the above images is that may poses have one's face directed to one side of the frame, but their eyes cutting across the frame in the other direction. Further exploration of this compositional pattern is necessary for it also has the potential to connect back to our earlier observations. We can see again that

[3]We used a model trained on the COCO Person Keypoint dataset using the R50-FPN architecture from the Detectron2 model zoo (137261548).

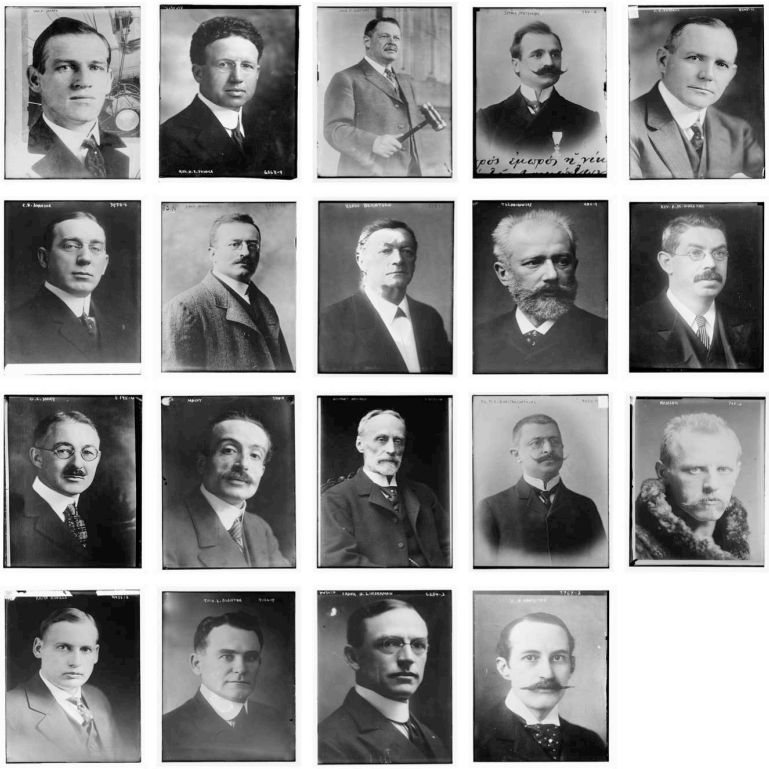

Figure 6: Twenty randomly selected portraits posed to the right, according to the post detection algorithm.

the portraits in this style are primarily men and many appear to be White, although we want to proceed with caution about assuming race by doing additional research. The analysis opens up more questions about the role of portraiture, gender, and race in early 20th century visual culture.

3.2. Orientation using face keypoints

We can repeat the same process using the locations of the eyes themselves relative to the location of the nose key points. Similarly, this algorithm does not display any strong preference for poses to the left or right. Looking at examples can once again be helpful. Figure 8 shows 20 randomly chosen examples of poses based on the eyes to the right. And in the final set, Figure 9 shows 20 randomly chosen examples of poses based on the eyes to the left.

Looking at these results, we see that the eye-based calculation does find poses which are more strongly oriented to one side of the image or another. In all of the example cases, we see that the entire person is oriented in the expected directly. Still, several examples show people who are looking off across the camera with their pupils. This highlights that orienting a person one way while gazing across the frame of the image is a common element of these studio portrait photographs. Further close and computational analysis, as well as a refinement of the classification of portrait photography, in order to better understand this phenomenon is

Figure 7: Twenty randomly selected portraits posed to the left, according to the post detection algorithm.

a planned topic for future work.

4. Future Directions

In this short paper we present prelimenary results from our work to identify and classify formal elements of photography using existing computer vision algorithms. Despite being used on historic, black-and-white photographs, the algorithms used in our application appear to perform well in terms of both precision and recall. However, significant work is required to construct rules for mapping low-level computer vision annotations into meaningful categories that are of interest to scholars of visual culture, archivists, and others looking to increase access and discover of digitised photographic corpora. Our application here shows the feasibility of this task on the specific case of portrature detection and classification in one relatively large collection. Future work will be able to identify and classify images based on additional formal features of photography and ensure that the techniques can be adapted uniformally across many collections and time periods.

Figure 8: Twenty randomly selected portraits posed to the right, according to the face keypoints.

Acknowledgments

The work in this paper was funded in part by a grants from the Library of Congress' *Computing in the Cloud Initiative* (BAA #LCCIO20D0112), the National Endowment for the Humanities (HAA-261239-18), and the Mellon Foundation.

References

[1] T. Arnold and L. Tilton. "Distant viewing Toolkit: A python package for the analysis of visual culture". In: *Journal of Open Source Software* 5.45 (2020), p. 1800.

[2] T. Arnold and L. Tilton. "Distant viewing: analyzing large visual corpora". In: *Digital Scholarship in the Humanities* 34.Supplement_1 (2019), pp. i3–i16.

[3] H. Caesar, J. Uijlings, and V. Ferrari. "Coco-stuff: Thing and stuff classes in context". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1209–1218.

[4] C. Dijkshoorn, L. Jongma, L. Aroyo, J. Van Ossenbruggen, G. Schreiber, W. Ter Weele, and J. Wielemaker. "The Rijksmuseum collection as linked data". In: *Semantic Web* 9.2 (2018), pp. 221–230.

Figure 9: Twenty randomly selected portraits posed to the left, according to the face keypoints.

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick. "Mask R-CNN". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[6] M. Lincoln, J. Corrin, E. Davis, and S. B. Weingart. "CAMPI: Computer-Aided Metadata Generation for Photo archives Initiative". In: (2020).

[7] T. Smits and S. Asser. "The Great Unseen. Photojournalism and the archive: from analogue to digital". In: *TMG Journal for Media History* 25.1 (2022), pp. 1–17.

[8] J. Tagg. *The disciplinary frame: Photographic truths and the capture of meaning*. U of Minnesota Press, 2009.

[9] M. Wevers and T. Smits. "The visual digital turn: Using neural networks to study historical images". In: *Digital Scholarship in the Humanities* (2019).

[10] M. Wevers, N. Vriend, and A. de Bruin. "What to do with 2.000.000 Historical Press Photos? The Challenges and Opportunities of Applying a Scene Detection Algorithm to a Digitised Press Photo Collection". In: *TMG Journal for Media History* 25.1 (2022), pp. 1–24.

[11] D. Yotova. "The Bain Collection: Created and maintained by the Library of Congress: http://www. loc. gov/pictures/collection/ggbain/. Reviewed September 2016". In: *American Journalism* 33.4 (2016), pp. 488–490.