Method Paper: Access & Discovery of Documentary Images (ADDI)

Taylor Arnold and Lauren Tilton

1. Introduction

Access & Discovery of Documentary Images (ADDI) is a project for LC Lab's Computing Cultural Heritage in the Cloud (CCHC) Initiative.[1] The initiative "pilot[s] ways to combine cutting edge technology and the collections of the largest library in the world, to support digital research at scale." [2] As a part of this work, LoC released a call for researchers to "use cloud computing services to explore the Library's digital collections at scale."[3] The project supports the library's strategic goals to "throw open the treasure chest," "maximize the use of content," "support emerging styles of research," and "cultivate an innovation culture" by supporting computational experiments to increase access to one of the nation's most important collections of cultural heritage.[4] We proposed ADDI to explore how computer vision can facilitate access and discovery of large image collections, focusing on photography. The project joins two exciting projects: "America's Public Bible: Machine Learning Detection of Biblical Quotations Across LOC Collections via Could Computing" led by Dr. Lincoln Mullen, and "Situating Ourselves in Cultural Heritage: Using Neural Networks to Expand the Reach of Metadata and See Cultural Data on Our Own Terms" led by Andromeda Yelton.[5] All of the projects benefited from the incredible support and openness of the Library of Congress, particularly LC Labs and the CCHC Team. We hope ADDI offers ideas, even a model, and cautions for opening access and interpretation of image collections using computer vision. Below we outline the project in further detail, including scope, methods, and next steps.

1.1 Project Goals

Our project Access & Discovery of Documentary Images (ADDI) adapts and applies computer vision algorithms to aid in discovering and using digital collections, specifically photography. Rather than treating cultural heritage images as a monolith, whereby computational approaches

[1] "Computing Cultural Heritage in the Cloud," Library of Congress LC Labs, https://labs.loc.gov/work/experiments/cchc/; "Library Receives $1M Mellon Grant to Experiment with Digital Collections as Big Data", Press Release, October 9, 2019, https://www.loc.gov/item/prn-19-098/.
[2] "Computing Cultural Heritage in the Cloud," Library of Congress LC Labs, https://labs.loc.gov/work/experiments/cchc/; "Library Receives $1M Mellon Grant to Experiment with Digital Collections as Big Data", Press Release, October 9, 2019, https://www.loc.gov/item/prn-19-098/.
[3] Manchester, E. (2021). "LC Labs Welcomes Computing Cultural Heritage in the Cloud (CCHC) Researchers!" LC Labs Letter. June 20, 2021, https://blogs.loc.gov/thesignal/2021/06/lc-labs-welcomes-computing-cultural-heritage-in-the-cloud-cchc-researchers/.
[4] "Digital Strategy for the Library of Congress," Library of Congress, https://www.loc.gov/digital-strategy/. For the full report, see FY2019-2023 Digital Strategic Plan of the Library of Congress Version 1.1.2; April 26, 2019, https://www.loc.gov/static/portals/digital-strategy/documents/Library-of-Congress-Digital-Strategy-v1.1.2.pdf.
[5] Castellanos, S. (2021). "Library of Congress Looks to AI to Help Users Sift Through Its Collection," *Wall Street Journal*, June 24, 2021. https://www.wsj.com/articles/library-of-congress-looks-to-ai-to-help-users-sift-through-its-collection-11624552197

are often developed and applied without attention to the form of cultural heritage in technical scholarship, ADDI pursues technical research with computer vision that considers the specificity of photography as a medium, social practice, and source of evidence for humanistic inquiry.

The project focuses on five photography collections from the early 20th century, totaling over a quarter of a million photographs held by the Library of Congress (LoC). Specifically, we choose the following collections:

Detroit Publishing Company (25,172 photos)
Farm Security Administration-Office of War Information (170,907 photos)
George Grantham Bain Collection (41,447 photos)
Harris & Ewing Collection (41,542 photos)
National Photo Company (35,619 photos)

The number of photographs in the LoC collections provides a scale that benefits from the use of computer vision and speaks to the necessity of experimenting and developing approaches to computer vision for access and discovery of images. While collections such as the FSA have extensive metadata such as photographer, date, and captions, tens of thousands of images still have minimal metadata. Collections like Detroit Publishing Company and the Bain Collection often have short captions comprising a short phrase like “In Central Park, New York” or “Harriet Quimby."[6] A query on the text captions would not convey the action in the first image — children sledding and walking with their dog on a snowy day — or the second image — a woman at the helm of an early airplane. Adding context to the photos at this scale is a significant and timely undertaking, particularly the important and rich metadata generated by subject experts and projects such as crowdsourcing initiatives.[7] Computer vision is *not* a replacement for this expertise nor the kind of knowledge it produces. At the same time, components of an image like people, dogs, and planes are amenable to computer vision. They offer an *additional* approach to metadata to facilitate search, discovery, and analysis of image collections.

In this document, we assess existing algorithms for generating metadata of historic image collections as a public methods paper, identify how generated visual annotations can be mapped onto formal features and approaches to studying photography, and develop an experimental public interface for search and discovery with the metadata. This work is designed to support the LoC’s work to “expand access” and “increase discoverability” to “the largest collection of human knowledge ever assembled.”[8] It is conducted as a part of the LC Lab’s Computing Cultural Heritage in the Cloud Initiative.[9]

---

[6] Byron, "In Central Park, New York," https://www.loc.gov/pictures/collection/ggbain/item/2001704137/. Publishing Company, https://www.loc.gov/pictures/item/2016808726/; "Harriet Quimby," Bain Collection,

[7] For example, LoC undertook a Flickr crowdsourcing project in 2014 that led to important new metadata in the Bain Collection. For example, see “Kath. Stinson & Dario Resta” and a field called "Summary" at https://www.loc.gov/pictures/collection/ggbain/item/2014701554/.

[8] Library of Congress Strategic Plan: https://www.loc.gov/strategic-plan/

[9] "Computing Cultural Heritage in the Cloud," Library of Congress LC Labs, https://labs.loc.gov/work/experiments/cchc/; “Library Receives $1M Mellon Grant to Experiment with Digital Collections as Big Data”, Press Release, October 9, 2019, https://www.loc.gov/item/prn-19-098/.

1.2 Our Approach

The project's methods are shaped by an approach that brings together data science and digital humanities. We draw on two definitions as guides. For data science, we understand the area of inquiry as the interdisciplinary study of collecting, analyzing, and communicating data.[10] For digital humanities, we draw on Kathleen Fitzpatrick's definition: "a nexus of fields within which scholars use computing technologies to investigate the kinds of questions that are traditional to the humanities, or…ask traditional kinds of humanities-oriented questions about computing technologies".[11] The two fields are increasingly a Venn diagram and mutually constitutive. Their interconnectivity is such that recent work is further identifying and articulating these connections and shaping hiring across cultural heritage and institutions of higher education.[12] We bring them together through the set-up of this collaboration as well: Arnold identifies as a Data Scientist, and Tilton identifies as a Digital Humanist, and vice versa. ADDI sits in the middle of the Venn diagram by engaging in the critical use and development of computational methods to analyze data to further humanistic inquiry.

The blend of our expertise and interests led to our method of distant viewing, which guides this project.[13] The framework offers an approach to computer vision analysis with careful attention to how these algorithms are trained to see and view, which we further detail in Section 1.4. Along with providing an approach to computer vision for image analysis, distant viewing is an approach to access and discovery. Developing annotations about features of a photo offers a way for people to explore a collection, particularly when the number of photos is in the tens of thousands. The annotations for ADDI provide another layer of information to augment discoverability and expand access to "the largest collection of human knowledge ever assembled." [14]

Our work joins an ongoing conversation across the digital humanities, informational retrieval, and cultural heritage about facilitating access to a collection, from working within existing organizations to remixing collections to introducing serendipity.[15] Emerging calls to think about

---

[10] Donohue, D. (2017). "Fifty Years of Data Science," *Journal of Computational and Graphical Statistics* 26 (4).
[11] Fitzpatrick, K. (2012). "The Humanities, Done Digitally." *Debates in the Digital Humanities,* ed. Gold, M. and Klein, L. University of Minnesota Press. https://dhdebates.gc.cuny.edu/read/untitled-88c11800-9446-469b-a3be-3fdb36bfbd1e/section/65e208fc-a5e6-479f-9a47-d51cd9c35e84.
[12] For example, see "The challenges and prospects of the intersection of humanities and data science," White Paper from The Alan Turing Institute, https://www.turing.ac.uk/research/publications/challenges-and-prospects-intersection-humanities-and-data-science. Information schools such as those at the University of Illinois Urbana-Champaign and the University of Michigan are hiring digital humanities scholars into data science programs.
[13] Arnold, T. & Tilton, L. (2019). "Distant Viewing: Analyzing Large Visual Corpora." *Digital Scholarship in the Humanities* 34.: i3-i16; Arnold and Tilton, *Distant Viewing*, The MIT Press (Forthcoming).
[14] Library of Congress Strategic Plan: https://www.loc.gov/strategic-plan/
[15] Deuschel, T., Heuss, T., B Humm, B., Fröhlich, T. (2014). "Finding without Searching-A Serendipity-based Approach for Digital Cultural Heritage." *Digital Intelligence*, Nantes; *Seguin, B*. (Spring 2018). "The Replica Project: Building a visual search engine for art historians." *XRDS* 24, 3: 24–29. *DOI:https://doi.org/10.1145/3186653*; Cooper Hewitt Labs, (2013). "All your color are belong to Giv," Cooper Hewitt Labs blog, https://labs.cooperhewitt.org/2013/giv-do/; Darms, Lisa. (Fall 2015). "'Radical Archives' Introduction by Lisa

"collections as data" have led to a reframing of library collections.[16] By thinking of materials such as photographs as data, and images as data more broadly, the range of analytical possibilities expands particularly computational methods.[17] Ryan Cordell's *Machine Learning + Libraries*: *A Report on the State of the* Field for the Library of Congress identifies promising machine learning applications, including the possibilities of visual data annotation as an emerging direction.[18] Our work takes up the call in the report to demonstrate how machine learning paired with knowledge about the medium, documentary photography as cultural and social practice in our case, can produce annotations that guide publics through the collections with attention to the specificities of the medium. In more and more cases, the challenge is not if the algorithm finds the features that it is trained to view within the data set; it is whether that feature is appropriate, tells us something about the interpretable images, and if that information is meant for the intended publics. In other words, it is less and less an accuracy issue (i.e., identifying cars) but an analytical issue (i.e., is this a feature that furthers our ability to find and interpret the image). To what purposes do we apply computer vision? How does this shape which computer vision algorithms we use? Which visual annotations map onto concepts and ideas that guide how publics are looking to access and what they are looking to discover? These questions guide the methods below.

Our approach focuses on assessing how well current computer vision algorithms are positioned to facilitate access and discovery. We focused on two ways. One is thinking about public engagement with the collections, particularly facilitating information retrieval processes such as browse, search, and recommendations. For example, we chose to explore face and pose detection algorithms as a potential way to identify genres such as portraiture. We also approached discovery from another angle: harnessing computational methods to analyze and understand the collections to further research on visual culture. For example, we decided to see if we could identify ways of seeing in portraiture (see Data Analysis Paper). The two ways are

---

Darms," *Archive Journal* 5, http://www.archivejournal.net/issue/5/archives-remixed/radical-archives/; Masson, Eef et al. (2020). "Exploring Digitised Moving Image Collections: The SEMIA Project, Visual Analysis and the Turn to Abstraction.," *DHQ: Digital Humanities Quarterly* 14, no. 4. This kind of work is also modelled by projects such as *The Civil War Sleuth* project developed by a team at Virginia Tech University (https://www.civilwarphotosleuth.com), Newspaper Navigator developed for LC Labs by Ben Lee (https://labs.loc.gov/work/experiments/newspaper-navigator/), Photogrammar developed by teams at the University of Richmond and Yale University (https://photogrammar.org), PixPlot developed by the Yale DH Lab (https://dhlab.yale.edu/projects/pixplot/), Serendip-o-Matic developed by a team during the One Week | One Tool Institute sponsored by the NEH (http://serendip-o-matic.com/index.html), and Ukiyo-e Search developed by John Resig (https://ukiyo-e.org).

[16] Padilla, T. (2017). "On a Collections as Data Imperative." *UC Santa Barbara*. https://escholarship.org/uc/item/9881c8sv.

[17] For another example, see our project "Images as Data: Processing, Exploration, and Discovery at Scale" with Carol Chiodo and Lidia Uziel (Harvard University) https://collectionsasdata.github.io/part2whole/cohortone/. The white paper is available at distantviewing.org.

[18] Cordell, R. (2020). *Machine Learning + Libraries: A Report on the State of the* Field. Commissioned by LC Labs, Library of Congress. July 14, 2020. For another important report, see Padilla, T. (2019). *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*. Dublin, OH: OCLC Research. https://doi.org/10.25333/xk7z-9g97. For another example of visual annotations in a library workflow, see Lincoln, M, Corrin, J, Davis, E, Weingart, S. B. (2020). "CAMPI: Computer-Aided Metadata Generation for Photo archives Initiative." Carnegie Mellon University. Preprint. https://doi.org/10.1184/R1/12791807.v2

mutually constitutive. How we design information retrieval systems can be guided by and guided by our understanding of the collections. Finally, we selected algorithms that are broadly relevant to historic still photography and are well studied with popular open-source implementations because the CCHC initiative is about the potential for the library and computational researchers to use these methods across collections. We did not train custom algorithms, although we suggest this as a future direction in Section 6.

1.3 Overview of Computer Vision Methods Used

Computer vision offers a computational method for analyzing images. The field focuses on using computer algorithms to approximate human visual systems. Training computers to see and view has been shaped by tasks from detecting colors and shapes to objects and motion. While we more often recognize our encounters with these technologies on our phones, at airports, and in cars, these same technologies also shape how we access and learn about the sources that impact how we understand the past, present, and future. An important domain is access to essential institutions such as the nation's library, the Library of Congress, a timely issue amid the challenges of a global pandemic where digital access became the primary form of access.

Access to image collections primarily relies on text to describe the image, which has most often been manually imputed by a person. Computers can now analyze the pixels that comprise an image and, depending on the algorithm that we select, can provide information about the image. Computer vision, therefore, offers an opportunity to look directly at the images. This is an exciting development for image collections, particularly large collections, since it allows us to analyze the photographs themselves and at a large scale. Computer vision is not a replacement for people such as metadata librarians and curators whose expertise is an important guide to computer vision applications. The metadata data provided about images such as administrative, descriptive, structural, and provenance metadata is necessary and, for the most part, kinds of data computer vision cannot generate. Rather, computer vision offers another angle to view the images that adds an *additional* layer of data to help increase access to images.

A challenge is how to turn how the computer sees an image – as pixels – into a form that is amenable to the information retrieval process, such as search, browse, and recommender systems. To accomplish this task, we will use what we call "annotations." They are structured data produced by computer vision algorithms. Examples include summary numbers, predicted categories, bounding boxes, and embeddings. An added benefit is that we can aggregate annotations and pair them with existing metadata to conduct data analysis on a collection (See Data Analysis Paper). They can also be the basis of a discovery interface, which we demonstrate in the ADDI prototype that we address inSection5.

A key question that we are engaged in by using computer vision is one that is foundational to the study of visual culture: what is this an image of?[19] From trying to understand the messages of a photo to identifying genres and their conventions, this question guides the interpretation. Analyzing a single image across a set (as in our case by photo organizations such as the Detroit

[19] For a great piece about this topic, see Cara Finnegan's piece about searching the FSA archive. Finnegan, C. (2006) "What is this a picture of? Some Thoughts on Images and Archives". *Rhetoric and Public Affairs* 9 (1).

Publishing Company) offers a range of interpretations that can be augmented as we shift our focus from a few images and zoom out to a larger set and zoom back in again. What we view and interpret, using just computer vision or combined with additional information (such as the powerful metadata that comes with the FSA collections), then shapes how people understand the images, which then can serve as a guide to access and discovery.

To work through this question, we focused on two kinds of algorithms. The first kind focused on identifying elements of an image: faces (face detection), bodies (pose detection), objects (object detection), and background elements (region segmentation). In this case, the algorithm looked at each image individually. A way to begin to answer the question is through the people and objects that comprise the image. Understanding the message of a photo not only often involves looking at a single photo but understanding the large ecosystem of images that the photo is in conversation with.[20] To compare images, the second kind focused on identifying similar images (image embeddings). By creating a distance measurement between images, this approach allows us to see similar images without having to be explicit about what makes them like one another. In this project, we compare within individual collections and across the collections (see Data Analysis Paper).

We applied and adapted three classes of computer vision algorithms to the selected collections. The first class of algorithms we used was the face- and pose-detection methods, which are used to identify the position of bodies and faces within an image. Competitive open-source algorithms exist for pose detection, but these have been trained and applied almost exclusively to high-definition images that have clearly defined people fully in the frame.[21] The second class of algorithms that we used is the classification of images into thing and stuff categories.[22] A popular task in computer vision research is the identification of specific objects within an image. Recently, a newer area of research has evolved to identify regions in an image that contain unenumerable "stuff," such as roads, trees, mountains, and the sky. From one perspective, these algorithms may be potentially difficult to apply to black and white images as they seem to rely heavily on color information. At the same time, they identify many elements of an image that should be relatively constant through time and, if feasible, offer an excellent surrogate for the study of historic images at scale; we explore this opportunity in Section 4. The third and final category of computer vision algorithms is image similarity.[23] Given their out-of-the-box power, these algorithms have been applied successfully to historic photographs. The novel question

[20] Photography and media theory points out how images make meaning through the visual cultures that they draw on as well as how they circulate. For example, John Berger calls these "ways of seeing," Stuart Hall theorizes how images make meaning as “encoding and decoding," and Marita Sturken and Lisa Cartwright discuss "practices of looking." See Berger, J. (1990). *Ways of Seeing*, Penguin Books; Sturken, M. and Cartwright, L. (2009). *Practices of Looking: An Introduction to Visual Culture*, Oxford University Press; Stuart Hall. (September 1973). “Encoding and Decoding in the Television Discourse," Paper for the Council of Europe Colloquy on "Training in the Critical Reading of Television Language."
For an example of a work that thinks about the development and social and cultural impact of a particular genre of photography, see Hirsch, M. (1997). *Family Frames: Photography, Narrative, and Post Memory*. Harvard University Press.
[21] Cao, Zhe, et al. (2019). “OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields.” In *IEEE transactions on pattern analysis and machine intelligence* 43.1: 172-186.
[22] Caesar, H., Uijlings, J., & Ferrari, V. (2018). “Coco-stuff: Thing and stuff classes in context." In *Proceedings of the IEEE conference on computer vision and pattern recognition*: 1209-1218.
[23] Barz, B., & Denzler, J. (2019, January). “Hierarchy-based image embeddings for semantic image retrieval." In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*: 638-647.

investigated by this project is the feasibility of pairing image similarity algorithms with the region of interest algorithms to find connections between sub-regions of different images, which we turn to in Section 4.2.

In the sections below, these three classes of algorithms are further introduced. After a long discussion of their motivations and history, we give references to current state-of-the-art models and implementations. Then, we discuss the algorithm's performance on the selected collections and highlight potential areas for application in the discovery and use of digitized collections of visual culture. We finish with a discussion of any potential social and ethical issues to the application of each algorithm.

1.4 Our Process

There are important theoretical and practical challenges to working computationally with digital images. Understanding visual materials must consider the ways in which knowledge is produced through visual media. Our process focuses on the automated creation of *annotations*, which are structured data produced by computer vision algorithms that describe one or more elements of an image. The method that we use to create, explore, and communicate these annotations is what we refer to as *Distant Viewing*.[24] This approach is an adaptation of a standard data science pipeline that considers the unique ways that images convey meaning. The pipeline breaks the process into several discrete steps: collecting the data, creating annotations, exploring the annotations, and communicating the results. By breaking the process into concrete steps, we can clarify the humanistic and algorithmic assumptions that underlie the study of large collections of digitized media. We can integrate, for example, how algorithmic biases in the assumptions of standard machine learning algorithms affect the exploratory visualizations. We can identify points of tension between loosely defined formal elements, such as composition and genre, and the need for computational algorithms to precisely define and delineate labels and categories. And, we can understand how the priorities of computer vision research have shaped the kinds of data aggregations that can be reliably calculated without direct human intervention.

In this paper, we focus on a discussion of the collection of the data, the creation of computer vision annotations, and the communication of results when applying the Distant Viewing approach to the five collections of documentary photography outlined above. The exploration step is covered in the accompanying Data Analysis paper.

1.5 Document Outline

The goal of this document is to be a resource to researchers looking to apply computer vision algorithms to collections of photography, with a particular focus on those looking to use collections housed at the Library of Congress.[25] As a resource, our main goal has been to include sufficient references and to document the details about methodology, such as how we choose

[24] Arnold, T. & Tilton, L. (2019). "Distant Viewing: Analyzing Large Visual Corpora." *Digital Scholarship in the Humanities* 34: i3-i16; Arnold, T. & Tilton, L. *Distant Viewing,* The MIT Press (forthcoming).

[25] The research comes in many forms and occurs at many levels. We define researchers broadly, including library staff, domain experts in visual culture, students, and members of the general public interested in exploring and learning about computer vision and photography.

data formats and programming languages. We have also tried to highlight places for possible extensions, outline the methods used in the project, highlight successes, flag areas of difficulty, and provide advice for future approaches to computational analyses of large image data archives.

The document is organized around the steps of the data science pipeline. First, we describe the technology and data formats used in ADDI. Then, we describe the process of collecting and organizing the data. We then proceed to describe each of the classes of algorithms used in the analysis. For each method, we describe the algorithms, their uses, and a short history of their development. These are followed by a discussion of how the computer vision algorithms provide annotations that can be used in applications and then highlight any difficulties we found in the results. Finally, we finish by explaining the methods used to communicate the results through the interactive visualization application.

2. Technology

The following section outlines the technologies used in ADDI, including programming languages, software libraries, and data formats, along with the tech stack required for the prototype. All decisions were guided by a commitment to open source and open access. All code is available at https://github.com/distant-viewing/addi.

2.1 Programming Languages for Data Analysis

One of the first decisions to consider when starting a computational project is the choice of technologies to use and the formats for storing the input, intermediate, and final versions of the project's data. Making clear decisions about how to organize and manipulate data in the early stages can greatly simplify the data analysis process in later stages. In this section, we will describe the technologies that we used in this project, along with a brief description of our rationale.

We have decided to use open-source software for all the data collection and analysis in this project. There are several reasons for our selection of open-source software. First, we believe it is important to understand how our tools are working; ideally, nothing should be hidden within a proprietary box that is beyond our ability to understand and critique. When dealing with proprietary software, there is often no way to ensure that the code is working as documented. Biases and misspecifications are easy to miss, even with open, well-documented code. However, at least with open-source software, there is an opportunity to look inside the box. Secondly, using open-source software ensures that the methods and techniques that we use will be accessible to anyone with some access to moderate computing power without the need to attain an expensive software license. While challenges certainly remain in terms of access and expertise, using open-source libraries and tools eliminates at least one difficulty in replicating and extending our results. Finally, the community of researchers working on state-of-the-art computer vision algorithms primarily release results as open-source software.[26]

[26] There are many companies that produce and use novel computer vision software for internal use, but these are not typically released in any form that can be accessed or purchased for personal use.

Therefore, we are not limiting ourselves by our choice of using open-source software; rather, we are engaging with many of the most accurate and advanced methods currently available.

In addition to choosing open-source software, we have specifically chosen to use programming languages rather than GUI-based point and click tools for our analysis. There were several reasons that this was advantageous for our work. First, it coincides with our desire to make sure that results would be reproducible and understandable. By writing the entire analysis pipeline in code, we have a reproducible script that explains each step and decision along the way (See our GitHub repository at https://github.com/distant-viewing/addi for the code). In addition to including extensive comments in the code, this makes it possible for others to have a well-documented way of understanding what we did in the project. Secondly, even if we did want to use a GUI-based tool, there are currently very few options for this approach. Those that do exist are proprietary, and therefore not open source. Existing GUI-based tools for computer vision typically only work to apply one type of computer vision algorithm and often lag software libraries accessible through programming languages in terms of their accuracy and speed.

Our desire to use an open-source programming language suited for data analysis led us to select a mixture of R and Python. Both languages are heavily used within the field of data science. R has more libraries for statistical modeling and data visualization, whereas Python is the tool of choice for text and image analysis. We primarily used R to collect our dataset by calling the Library of Congress API and website (see the next section) and as a helpful scripting language. R was also used for the computational analysis presented in our corresponding Data Analysis Paper. The Distant Viewing Toolkit (DVT)—the software we built for the analysis of visual culture—is written in Python; therefore, we used the language for generating the automatically produced computer vision annotations (See our GitHub repository at https://github.com/distant-viewing/dvt or the latest version of DVT).[27]

2.2 Software Libraries for Data Analysis

Within the R programming language, we used a collection of standard and popular third-party packages. These include dplyr for data manipulation, ggplot2 for data visualization, xml2 for parsing XML data, jsonlite for parsing JSON data, and stringi for parsing raw strings.[28] The choice of a specific Python library for computer vision required some careful consideration.

There are several libraries for doing computer vision in Python, but these are often difficult to install and particularly difficult to install on one machine at the same time.[29] Two of the most popular low-level libraries at the time of this writing are TensorFlow and PyTorch.[30] Slightly

[27] Arnold, T. and Tilton, L. (2020). "Distant Viewing Toolkit: A Python Package for the Analysis of Visual Culture." *Journal of Open-Source Software*, 5(45): 1800.

[28] For an overview of these tools, see Wickham, H. and Grolemund, G. (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* O'Reilly Media.

[29] The reason for the difficulty is that each library often requires a very specific version of many other dependencies. These will not be the same for different tools, making it hard to work with multiple tools within the same runtime environment.

[30] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., and Kudlur, M. "Tensorflow: A system for large-scale machine learning." In *12th USENIX symposium on operating systems design and implementation*: 265-283; Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G.,

higher-level libraries include Keras and Detectron2.[31] Our initial version of the Distant Viewing Toolkit was built around TensorFlow. However, as part of this project, we transitioned to working with Detectron2. While slightly limiting in the ability to build new models, Detectron2 was found to be much easier to install and had a much better set of models ready to use out-of-the-box. For our application goals, it was better to be able to apply a wide range of models rather than being constrained to those available directly for TensorFlow or the need to build every model from scratch. Since we hope that this project might serve as a model and guide to applying CV to other image collections, balancing ease of installation and the scope of available models shaped our decision making, to realize these goals, we also rebuilt the Distant Viewing Toolkit around Detectron2 as a part of this project.

2.3 Data Formats

As an additional technological decision, we needed to decide on formats for the data constructed from our analysis. Two popular options are the Extensible Markup language (XML) and JavaScript Object Notation (JSON).[32] These forms have several advantages from the perspective of software engineers building reliable internal software and exchanging data between a variety of programs running in a diverse technological stack. However, these formats are not ideal for data science applications because they break the relational database model.[33] As an alternative, to use the language of Hadley Wickham, we preferred for our data to be stored as "tidy data " in the form of a set of interconnected tables.[34] We have chosen to store our data in the form of comma-separated value (CSV) files, one of the most common formats that can be read by almost all software designed for data science applications.

For the prototype, we converted the relevant fields of the CSV fields into a JSON format, which is more easily manipulated by the JavaScript language (see next section for details). To reduce the amount of data that needs to be loaded when a visitor first opens the prototype, the annotations are stored with one file for each image. In this way, only one small file needs to be loaded to get started; a new file is opened and parsed each time a visitor clicks to a new image. The use of flat files, rather than a formal database, makes it possible to serve the entire application of GitHub pages. This greatly reduces the cost and potential for bugs in the code and possible downtime. To streamline the JavaScript code, we created a single format for describing bounding boxes that could be used for the annotations from the face detection and two object detection algorithms.

Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. and Desmaison, A. (2019). "Pytorch: An imperative style, high-performance deep learning library." *Advances in neural information processing systems*, *32*: 8026-8037.

[31] Chollet, F. (2018). "Keras: The python deep learning library." *Astrophysics Source Code Library*.

[32] For the standards of these formats, see https://www.w3.org/standards/xml/core and https://www.ecma-international.org/publications-and-standards/standards/ecma-404/.

[33] Technically, it is possible to store tabular, relational data in either XML or JSON. The core difficulty is that it is also possible to break this model and that typically becomes overly tempting for anyone using either format. Also, there is no uniform way to store tabular data in XML or JSON, making it difficult or impossible to load these formats directly with tools such as Excel and Tableau.

[34] Wickham, H. (2014). "Tidy Data." *Journal of Statistical Software*. 59(10): 1–23.

2.4 Technology for the prototype

In addition to the data analysis, our project also included building a web-based, interactive, proof-of-concept application to visualize the automatically generated annotations extracted by our computer vision algorithms. To simplify long-term maintenance, we designed the application to run locally on a user's machine within their web browsers. Web browsers are built to run code in JavaScript, and so this is the language we choose to use for the visualization application.

Our specific technology stack for the visualization engine uses several popular libraries and frameworks in JavaScript. Specifically, we wrote a mixture of code in JavaScript and marked-up files in HTML and CSS. These elements were put together within the React.js framework, which provides an easy way of building one-page websites that coordinate a mixture of static data and user interactions. The ADDI prototype further demonstrates our commitment to open source and open access (See our GitHub repository at https://github.com/distant-viewing/addi or the codebase.)

3. Data Collection

The following section outlines the data collection process. There are two primary kinds of data: image and text. We discuss the process of accessing the data, deciding on which file types, and organizing the data, including identifying opportunities to streamline the data collection process.

3.1 Structured Metadata

The Library of Congress has excellent metadata attached to each of the collections that are part of our study. Descriptive metadata such as captions, photographers, and dates offer historical context. Structural metadata such as film type provides insight into the materiality of the image (i.e., negative or print) as well as the technology that produced the image. Administrative metadata such as file formats (i.e., jpeg and resolution) gives technical information that informs what kind of computational methods are possible. The metadata offers important context that informs the application of computer vision as well as the prototype.

LoC provides a public API that is available for researchers looking to access this metadata. The API currently includes several endpoints for accessing data, including a full-text search of record fields and the ability to grab all records from a single collection. When called, the API returns a set of up to 150 records, with one record per object (a photograph, in our case). The records include links to images at various resolutions and provenance data. Many records contain additional information, such as location, date, and title. However, some of these fields are missing for some of the images. According to the existing documentation, it was not clear which elements of returned JSON objects are required (i.e., always returned by the API) and which were optional. We started by trying to use this API but did run into some issues, which we will briefly describe here. Note that we are giving a description of our experience using the API in the summer of 2021; the API team has generously worked with us through these challenges and is in the process of making changes, so some of these issues may no longer apply.

One challenge is that there were several different sources of documentation for using the Library of Congress' API, and it was not clear which one was the authoritative source of information. At

the time, none of the sources referenced one another. First, there was a very brief page contained on the loc.gov site titled "JSON/YAML for LoC.gov." This page provided information about making a request and the various endpoints. It did not have any tutorials or information about the format of the returned data.[35] Secondly, there was a website hosted on GitHub pages.[36] The information was more detailed and very helpful. In addition to the formal specifications, there were friendly tutorials that explained how the API was designed and what kinds of data were returned. However, this version did not include information about all the available endpoints nor provide the rate limits that were included on the main loc.gov site. Finally, there is also the site at labs.loc.gov; this is a mixture of the information from the other two sources and was not our primary source of information during the project.

Our initial pipeline for working with the LoC data was adapted from the GitHub pages tutorials. Because of the limited information about all the available endpoints, we assumed that it was necessary to make an API request that first searched the collection and gave us a list of all the available items. Then, we would need to call the API again for every record in the collection. This was easy to code, but we quickly ran into issues with the rate limits that were not documented on the GitHub site. Given the scale of our requests, this was a huge issue. With some trial and error, we determined that it was going to take us several weeks just to query one collection. With the need to make some quick progress, we decided to bypass the API and scrape the data directly from the Library of Congress website. It took minimal effort to figure out the URL scheme, and then we were able to download all the data. Surprisingly, the rate-limiting seemed to not apply to the main Library of Congress website. It took several days still because we had to fetch each record one by one, but this was still substantially shorter and less complicated than working with the API.

Later in our work, we decided to go back and grab data from the official API. One reason for this is that the data on the website is in the MARC record format, which had inconsistent field tags for different collections. At this point, we learned about the documentation on the loc.gov site and were at first able to get the API working relatively well.[37] However, when working with the larger FSA-OWI collection, we ran into another issue. We were paginating through the collection but learned that the API would only allow us to go a certain depth through the results. We were only able to get to about page 700 and around 90,000 images, which is only half of the collection. Unfortunately, there was no easy solution to this issue, and we were never able to get the full dataset through the official API.[38]

The other challenge that we had with the data was determining which of the various version of the data on the Library of Congress was the most official version of the record. The easiest way

[35] Several helpful changes were made between our collection of the data and the writing of this paper. The website now contains much more information about the returned data formats.
[36] https://libraryofcongress.github.io/data-exploration/index.html
[37] There is an endpoint that allows the querying of up to 150 records at once, which significantly reduces the number of requests that are needed.
[38] This was not just a buffering or waiting issue. The URL will always fail when requesting too many pages. For those interested, here is the query that we were not able to process (at least as of January 2022): https://www.loc.gov/collections/fsa-owi-black-and-white-negatives?c=150&fo=json&sp=900.

to explain this issue is through an example. For one record, here are four different ways of accessing the record's metadata:

https://www.loc.gov/item/2017798400/
https://www.loc.gov/pictures/collection/fsa/item/2017798400/
https://www.loc.gov/pictures/item/2017798400/marc/
https://lccn.loc.gov/2017798400/marcxml
https://www.loc.gov/item/2017798400/?fo=json

These links are not redirects that point to the same location; each is a different page with slightly different fields and formats. Most of these differences are minor, but these small differences make it hard to algorithmically compare data from different sources.[39] The API has lowercase names, and the locations are different formats from the others, for example. Each of the website pages is different too. One has each location separated, and the other one does not. One has an actual "city" field, and the other does not. In our case, the differences were small but did mean that we needed to pay very careful attention to make sure that the differences were a matter of differing field names for the same data or the data entry such as capitalizations rather than one data set including information that another set did not include. The Prints & Photographs Division kindly clarified the difference between the data sources, which we shared would be helpful to document to further support future researchers using images as data.[40]

Ultimately, we primarily used the data from our initial scraping of the Library of Congress website. During the final preparation of the visualizer, we included the data from the API for the two collections that were part of the prototype (FSA-OWI color images and the Bain Collection). From our workflow, it would have been significantly easier to go to a site or make a request to LOC, where we could have received all the data at once. As an example, consider the Rijksmuseum's website, which has a "download" section with zip files with multiple formats for the entire collection.[41] Similarly, The Metropolitan Museum of Art releases all its open-access data as a single CSV file.[42] A clear, official, authoritative data set available as a CSV would streamline access for data analysis and clarify which version of the data is primary. As data is updated, documenting the changes, and providing past snapshots of the data, would also be a major asset.[43] We are excited to see how the work of LoC's AWS Dataset Pilot Working Group might facilitate streamlined access to the collections as data.

[39] As another change during the duration of this project, the JSON returned by the API now has a version number which makes keeping track of changes much more manageable.
[40] Thank you to the Prints & Photographs Division team for their time helping us understand the data. Our understanding following these conversations is that data behind Prints and Photographs Online Collection (PPOC) and LOC.gov is organized slightly differently. This is the result of shifts over time in data management and systems serving the data to the public. PPOC was an early pioneer of online digital access and, therefore, slightly differently structured than the more recent aggregation of data within LOC that serves LOC.gov.
[41] https://data.rijksmuseum.nl/object-metadata/download/
[42] https://github.com/metmuseum/openaccess
[43] For example, the FSA collection metadata has been updated based on information such as crowdsourced data from Flickr as well as our own work through Photogrammar (Photogrammar.org). As a part of that project, we provided metadata for tens of thousands of more photos based on the negatives. For more about, see Arnold, T., Maples, S., Tilton, L., and Wexler, L. (2017). "Uncovering Latent Metadata in the FSA-OWI Photographic Archive." *Digital Humanities Quarterly* 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000299/000299.html. LoC

3.2 Images

Once we gathered the structured metadata, there was also the need to download all the images associated with our selected collections. Each of the data sources included direct URLs for the images, and there did not appear to be any associated rate-limiting that prevented directly downloading each of the images with a small script of code. Given the size of the images and the number of requests, however, it took several days to download all the files. We used an external server to make sure that the process could run slowly and steadily, with extra checks to make sure that the files were correctly downloaded.

The one challenge we had in terms of downloading the images was deciding which of the image versions to work with. Most records had two different files in an uncompressed TIFF format as well as three or four JPEG images at different resolutions. The TIFF files are very large, but we wanted to make sure that we were not losing any important information by using the JPEGs. The collection website contains almost no information about the files and how they were created. We ran a few experiments by grabbing the largest TIFF file, converting them locally to a JPEG, and then comparing the available JPEG images. Once converted to JPEG, it was a similar size to the other JPEG versions. Therefore, it appeared that the largest JPEG was either a converted TIFF or at least close to it. From this experimentation, we decided to use the largest JPEG image for our application.

For the five collections, the Library of Congress provides one or more archival uncompressed TIFF files and several JPEG files of various resolutions. For computer vision algorithms, it is usually best to use the highest resolution available. However, we determined the larger archival TIFF images were of the same resolution as the compressed JPEG images. Therefore, to optimize for storage space and speed, we downloaded and used the largest JPEG. There is no need to use the largest file just because it is available if it isn't necessary. In addition to substantially increasing the amount of time that it takes to apply the computer vision models, the environmental costs only increase.[44]

Thinking of images as data is enabled by understanding digitization decisions. We are greatly appreciative of the time that Prints & Photographs staff spent helping us understand the digitization histories of the collections. Understanding that part of the FSA-OWI collection, for example, was re-scanned with better technologies whereas other parts were not, and when they might be, is important for the application of computer vision algorithms. Their knowledge and expertise were crucial to ADDI. A new section, "Digitizing the Collection," is being added and

---

has now added metadata based on the method. They use brackets [ ] to indicate metadata that is based on assumptions from the negatives. Another use is related to labor. By providing previous snapshots of the data, one can go back and look at the revision history of the LoC. As Ben Schmidt has demonstrated, studying the history of MARC records opens analytical possibilities to study how the government shapes public knowledge (see https://creatingdata.us/datasets/marc-history.htm). D'Ignazio and Klein discuss his work in Chapter 7: Show Your Work of *Data Feminism* (The MIT Press, 2020) to demonstrate how metadata can be used to acknowledge labor, which reveals the amount of work that it takes to create digital collections.

[44] Crawford, K. and Joler, V., "Anatomy of an AI System: The Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources." *AI Now Institute and Share Lab*, (September 7, 2018) https://anatomyof.ai; Dhar, P. "The carbon impact of artificial intelligence." *Nature Machine Intelligence* 2**,** 423–425 (2020). https://doi.org/10.1038/s42256-020-0219-9.

provides incredibly valuable information that contextualizes and situates the collection. A way of sharing with the public what has been digitized, what is next to be digitized, and the expected timeline, could also help researchers. It would also be useful if the API contained metadata about the digital images themselves, such as when they were scanned and what process was used for their creation.

Thinking of images as data amenable to computer vision also brings to the fore questions about exactly what to digitize. There are a plethora of considerations when digitizing, including the capabilities of the digitization technology, state of the collection, and intended audience. Debates ensue, and best practices have changed over time. A major factor when working with computer vision is that they are often designed to examine all the pixels in the image. For example, the difference between a frame, matte, and the print of a photograph is not accounted for (e.g., studio portraits in the Bain Collection). To add one more example, the difference between the sprockets and the photographic image produced by the exposure to light is also not accounted for (e.g., digitized film negatives from the FSA Black & White collection). To the computer, the set of pixels that constitute the image file is the image. If one is interested in comparing the features only of the image seen through the lens, for example, one must edit or write custom code to remove features such as a frame. In multivalent collections like Bain, where there is a mix of studio portraits and news photos, writing a custom algorithm to account for various features that we may want to analyze (or not) becomes even more challenging. This is also especially pertinent to decisions about over-scanning, a process that helps capture information about the materiality of the film. To facilitate images as data for computer vision, considering the ways that objects are digitized and how that shapes the ability to do computer vision will need to be a consideration in digitization. We delve further into this challenge in Section 4 about the algorithms (particularly Section 4.3 on image embeddings) and Section 5 about the prototype below, as well as demonstrate how these features affect results in the Data Analysis paper.

4. Applying and Assessing Computer Vision Algorithms

Returning to a key question that we introduced in Section #.#., identifying the components that are included such as objects like cars, in focus such as the face of a person, and how much of the frame the feature occupies, such as a monument is a way to answer this question. How and who we document is not only a major area of study in visual culture studies but the key to identifying the kinds of evidence and messages that an image provides. These visual annotations may be in a position to support access to the collection through elements such as keywords and descriptions that can become a part of features such as search, browse, and recommendation, which we discuss in more depth below.

The following section looks at the three classes of algorithms – Face and Pose Detection, Object Detection, Region Segmentation, and Image Embeddings - that we identified as potentially amenable to the study of photography collections. We provide a description of the method, address relevant social and ethical concerns, discuss the process of applying the algorithms, and evaluate the results, including identifying opportunities and remaining challenges.

4.1 Face and Pose Detection

Many formal elements from photography and film theory can be described in terms of the locations of bodies within a frame. The framing of an image is often defined explicitly in terms of how much of the visual space is framed by a person's face. For example, in film studies, a medium close-up shot contains only a subject's head and chest, whereas a medium-long shot contains almost all of a subject's body.[45] In photography, the close-up of a person is often categorized as a portrait and an important and powerful genre in visual culture. Who and how people are documented is imbued with cultural, social, and political values. Other formal elements, such as an indication of low-angle shots and birds' eye view, are typically described in relation to the subjects on the screen. In photography, this is often referred to as camera angle and steeped in power relations such as who and what the camera looks up to as well as looks down on. They can likewise be inferred by the angle of a person's body in relationship to the frame. While many of these formal definitions are grounded in film theory, they can be applied to still photography, television, and other forms of visual materials. Face and pose detection offer one way to use computer vision to explore composition such as portraiture and framing such as angle.

4.1.1 Method Description

There has been a significant amount of computer vision research on the location and description of people within an image. Beginning with the studies of Woody Bledsoe and Helen Chan Wolf in the 1960s, facial recognition has had a particularly long history in computer vision.[46] Early successes led to further approaches and competitions, such as the eigenfaces of Sirovich and Kirby, the Department of Defense's Face Recognition Technology program (FERET: 1997-2003), and Iris Evaluation Challenge for the National Institute of Standards and Technologies (NIST).[47] Through this government-backed research, increasingly powerful methods were developed to consistently identify well-framed faces, at least as well as most human experts by the end of the 1990s. Within the last decade, the creating of very large training sets and the development of deep-learning techniques has produced models, such as Mask R-CNN for face detection and VGGFace2 for face recognition, that are able to find and identify faces in a variety of poses and lighting conditions.[48]

While earlier research primarily focused on the detection and recognition of faces, the success of deep learning has allowed for computer vision research to accurately accomplish additional

[45] Slightly different definitions exist, but almost all reference the relative size of a subject's body in the frame. Our definitions come from Butler (2018).
[46] Bledsoe, W. W., and Chan, H. (1965). "A Man-Machine Facial Recognition System-Some Preliminary Results." *Technical Report PRI 19A*, Panoramic Research, Inc., Palo Alto, California.
[47] Sirovich, L., Kirby, M. (1987). "Low-dimensional procedure for the characterization of human faces." *Journal of the Optical Society of America A*. 4 (3): 519–524. Doi:10.1364/JOSAA.4.000519; Rauss, P.; Philips, J.; Hamilton, M.; DePersia, T. (February 26, 1997). "FERET (Face Recognition Technology) program." In *25th AIPR Workshop: Emerging Applications of Computer Vision*. 2962: 253–263. Doi:10.1117/12.267831; Phillips, P. J., Bowyer, K. W., Flynn, P. J., Liu, X., & Scruggs, W. T. (2008, September). "The iris challenge evaluation 2005". In *2008 IEEE Second International Conference on Biometrics: Theory, Applications, and Systems*: 1-8
[48] Cao, Q., Shen, L., Xie,, W., Parkhi, O.M., and. Zisserman, A. (2018). "VGGFace2: A dataset for recognizing face across pose and age". In *International Conference on Automatic Face and Gesture Recognition*, https://arxiv.org/abs/1710.08092.

tasks regarding the detection of bodies within an image. Object detection challenges often include an object category for "person," which can include any part of a human being and their clothing within a frame. A person category was included, for example, in three of the most well-known image classification tasks: Pascal VOC (2005-2012), ImageNet Large Scale Visual Recognition Challenge (ILSVRC, 2010-2018), and Microsoft Common Images in Context (MS COCO, 2015-2020). The tasks originally focused on simply detecting whether a person is present in an image, while more complex subtasks from MS COCO have addressed locating a bounding box for each person and identifying the exact pixels that directly correspond to a person or their clothing. Increasingly complex tasks now focus on the identification of the *key points* within an image. Key points are specifically identified pixels that correspond to a particular body part, such as the left corner of a subject's eye or the pivot point of their right ankle.[49] Accurate models such as CMU's OpenPose[50] and Facebook AI's Detectron2 exist for identifying over one hundred of these points for human bodies, hands, faces, and feet.[51]

The confluence of interest in identifying bodies within an image from both humanities scholarship and research in computer vision forms a good starting point for building a reusable, general-purpose set of annotations for distant viewing. Finding links between our application domain and computational models minimizes the gap between the properties of interest in disciplines of visual culture and the computational study of digital media. Specifically, we propose building a system that locates each person within a frame, estimates the location of key points, and makes a prediction about the identity of any detected faces. These can be done relatively accurately with currently available systems and, given continued interest within computer vision, will be further improved on in the years to come.

4.1.2 Applications

It is possible to use the low-level derived data extracted from images to retrieve measurements that mirror the formal elements from photography. We can directly map the size of the largest detected face relative to the entire image to classify types such as portraits. We often find that it can be helpful to modify the exact values based on the aspect ratio of the source material. By combining the face sizes with other information about detected people in the image, we can build simple models that detect more complex tropes. If we detect that most of an image is taken up by a face, it might be labeled as a portrait. When defining these classifications, it is useful to ensure that cut-off values are clearly defined. Specifics may change between applications and source materials. An application of using the sizes of faces and people is further explored in the Data Analysis paper. We show how these metrics can distinguish individual collections as well as identify sub-clusters within individual collections.

[49] The location of facial key points has a longer history cotemporaneous with facial recognition systems, as early models were built on the idea of finding faces vis-à-vis key points. However, the usage of facial key points as an important subtask itself is a relatively new concept.
[50] Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2018). OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. arXiv preprint arXiv:1812.08008.
[51] Wu, Y., Kirillov, A., Massa, F., Wan-Yen., L., and Girshick, R. (2019). Detectron2. https://github.com/facebookresearch/detectron2

4.1.3 Evaluation and Challenges

After applying the pose and face recognition algorithms to the data in our five collections, we manually investigated several hundred of the images to determine any patterns of how well (or not) each algorithm seemed to perform. As our focus was on a broad qualitative understanding rather than a narrowly defined numeric one, we did not perform a formal testing procedure.

Figure 1. Example of results from a face detection algorithm, showing the ability to detect faces at different angles to the camera. [2014687262]

Figure 2. Example of results from a face detection algorithm, showing how only a subset of faces are picked up in a larger crowd. [2014689609].

The face detection algorithm performed relatively well on all the images that we investigated. The algorithm we used almost always detected faces that were relatively clear, not too small, and that were oriented no more than 90-degrees away from the camera **[Figure 1]**. It sometimes found very small or rotated faces, but these were less consistent. The algorithm seemed to avoid detecting too many faces in one part of the image. Some images of crowds detected many faces but only a subset of those that were clearly visible **[Figure 2]**. Setting a reasonable cut-off score, the algorithm made no false detections that we were able to detect in our exploration. When faces were detected, the bounding boxes seemed to consistently capture the size of a person's head well. It seems that using these detected faces to classify images is a reliable method, keeping in mind that it may miss people that are turned significantly away from the camera and may miscount large crowds.

We believe the results are because of the way that face detection is used in many applications. It is intended to be the step before face recognition, the process of identifying a person. The goal isn't to find every person; the goal is to find any faces in the image to then identify the person. Using this algorithm to provide a tag or keyword for search and browse could provide a productive way into a collection, but the specificity of language is important here. This is "faces." For a collection like FSA, where an explicit goal of the project became to create a portrait of America, offering search by "faces" as a kind of visual trope in the collection (and there is perhaps no more iconic photo than Dorothea Lange's portrait of Florence Owens Thompson known as "Migrant Mother") offers a way to bring together computer vision with the context of a photo collection. For future research, we plan to use face detection and then recognition to see if this might be a way to identify prominent figures within and across collections. For finding people, we discuss in Section 4.2 how region segmentation returns better results and opens other possibilities for discovery.

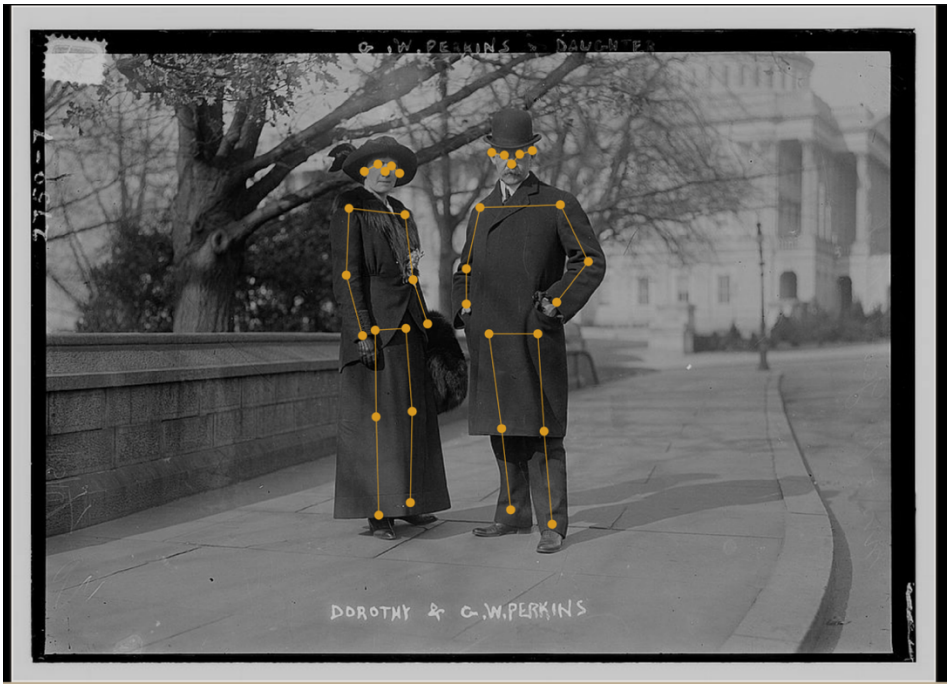

Figure 3. Example of pose detection, showing good performance on a black-and-white image with people wearing period specific clothing. [2014693509]

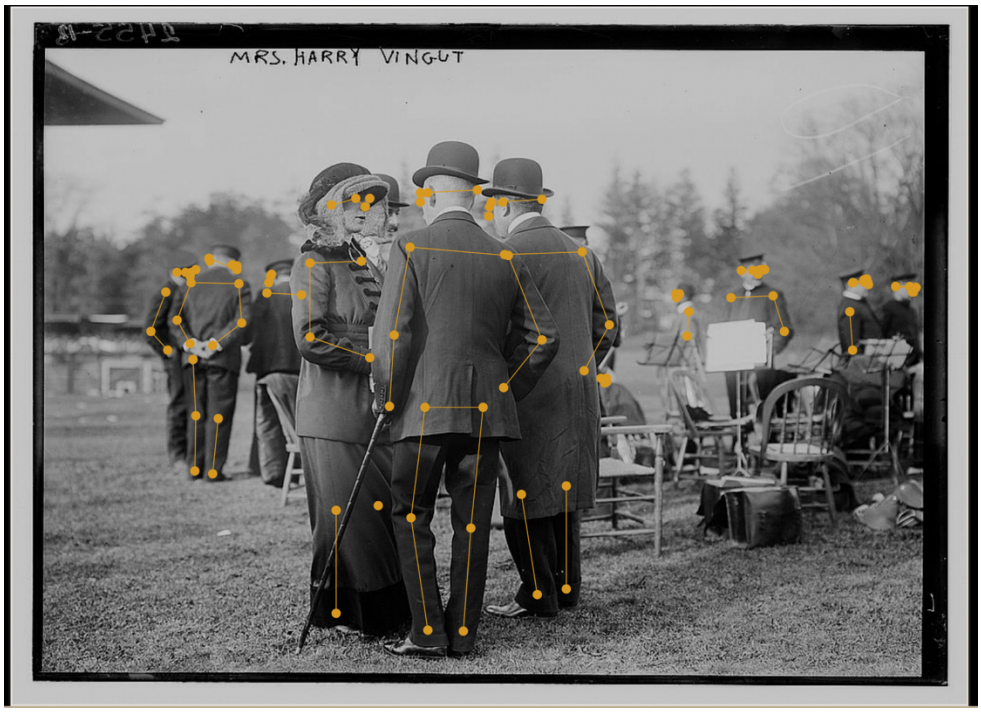

Figure 4. Example of pose detection showing the ability to pick up people in the background. Notice, though, that the arm of the man in the right middle of the foreground's arm in confused with the arm of the man in the middle of the frame. [2014690917]

Figure 5. Another example of pose detection. Notice that the algorithm trying to find the lower arms, torso and legs of the person shown only in portrait. The confidence score for the facial features is much higher than the lower features. [2014716249]

The pose detection algorithm also performed relatively well on our collection **[Figure 3]**. As a particularly surprising result, it was able to pick up many complete poses from people far off in the distance **[Figure 4]**. In a few cases, we did not initially notice the individuals within the frame and thought these were false detections. However, the pose algorithm did pick up some false positives, even when selecting a very high confidence level. We noticed three different types of relatively common errors. First, it was confused when given an image of a person whose body was significantly cut-off by the frame. The algorithm would sometimes try to fit the entire body into the clothing even though clearly an individual was only photographed from the waist and up **[Figure 5]**. Secondly, the algorithm would sometimes confuse the body parts of individuals who were standing close to one another. For example, it would swap the two arms of people standing very close together **[Figure 4]**. Finally, the algorithm occasionally would detect a pose with very high confidence where there was no person to be found. Usually, this was a dark shadow or strangely shaped object somewhere in the background.

Given the number of high-quality poses, it seems that there is a potential to use the detected poses in computational work. However, the large number of errors does suggest some caution. We suggest that the pose information could be used if combined with the detection of people through region segmentation in the next section or when it is known that only one person is in the shot. It can be helpful for analyzing angle and framing, which we also explore more in the next section. It could also be used in a human-in-the-loop model, where poses are manually annotated to indicate whether they are correct or not.

4.2. Object Detection and Region Segmentation

Identifying the contents of an image is an important part of analyzing and recognizing what information is in a photo. For a person interested in the history of technology or excited about fashion, they may want to find all the photos with a car or necklace. Figure 8 offers an example. The image title is "Galli-Curci," who was one of the most popular opera singers of the 20th century. She is noteworthy for sure, and the relatively intimate photo of her at the piano staring into the camera is arresting. For those interested in the history of music, a quick query of her name will return the photo. For a person interested in fashion, looking closely at how a musical celebrity of the time composed herself for the camera offers a lens into contemporary fashion. Object detection offers a way to sort the collection in unexpected ways.

4.2.1 Method Description

The identification of objects within an image constitutes one of the most popular and prominent tasks in computer vision. Early tasks focused on relatively simple objectives, such as the classification of hand-written digits in the MNIST dataset, which used small 28-by-28 black and white pixels.[52] Tasks such as CIFAR, STL, and ILSVRC, have offered increasingly complex objectives that require hundreds or thousands of object categories from higher-resolution images. Categories in recent competitions feature highly abstract concepts such as a "grocery store" or difficult to distinguish categories such as 200 bird species and 102 categories of

[52] Platt, J. C. (1999). "Using analytic QP and sparseness to speed training of support vector machines." In *Advances in Neural Information Processing Systems*: 557-563.

flowers.[53] Within the confines of each of these specific tasks, novel models have been shown to outperform most human annotators. While the resulting models can produce false assertions when applied to new corpora, they do tend to produce reasonable and potentially useful estimates of certain object types.

Despite the popularity and seemingly powerful results of object detection algorithms, we have found that current state-of-the-art models require careful consideration when used as a general-purpose set of annotations for the analysis of visual culture. Currently, available models feature categories are very specific and only cover a very small number of the object types that could be seen within the frame of modern, western-centric film and photography. When considering historical or more diverse datasets, the coverage is even worse. For example, the COCO dataset contains only three types of fruits (apple, orange, banana), two vegetables (broccoli, carrot), and five other food items (cake, donut, pizza, hotdog, sandwich). There are no generic catch-all food categories for other items falling outside of these lists. Applying these object detection models indiscriminately to a large corpus without understanding its limitations will result in biased results. They will find certain kinds of food items, animals, and clothing but will completely ignore examples outside of a narrowly curated list of categories. This is particularly important to consider when identifying and quantifying objects in a photo and therefore making claims about whether a collection does or does not address a certain topic. In general, object detection is best understood as about presence. It tells us what a certain algorithm sees but is less effective at telling us what is not in the photo, which is a common analytical move in the study of visual culture and one that we should approach with caution using computer vision.

Object detection models do have an important place within the distant viewing framework. Using them to detect a small set of specific objects can be very useful. For example, we have found that the specific category of "people" is particularly robust when applied out-of-sample and can lead to useful ways of recording formal elements. Other categories, such as "vehicles" or "televisions," could also be useful for an analysis specifically designed to study these classes and when users are aware of the potential limitations. Training a new object detection model for specific categories of interest is also possible; while potentially time-consuming, these custom models can lead to very useful conclusions. Likewise, applying a model trained on one of the relatively exhaustive specialized datasets can be very productive, provided there are relevant research questions related to the chosen subcategories.[54]

We applied two different algorithms for object detection in the project based on two different sets of objects. The first uses the 80 object classes from the COCO dataset. These categories cover common objects in modern photography with a particular emphasis on industry applications. It includes tagged objects of cars, bikes, and stop signs. We used these categories for two reasons. First, this is one of the best-studied collections in computer vision research. If any generic algorithms work well on historic documentary photography, it is likely a model based

[53] Lin, T., RoyChowdhury, A., & Maji, S. (2015). “Bilinear CNNs for fine-grained visual recognition." arXiv preprint arXiv:1504.07889. Nilsback, M. E., & Zisserman, A. (2008, December). “Automated flower classification over a large number of classes." In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*: 722-729.

[54] For a well-known example of a niche tagged image classification dataset, see Welinder P., Branson S., Mita T., Wah C., Schroff F., Belongie S., Perona, P. (2010). “Caltech-UCSD Birds 200”. California Institute of Technology. CNS-TR-2010-001.

on this dataset would be one of the best performing. Secondly, the categories include common images in busy street urban scenes. These are featured prominently in several of the news journalism collections, and therefore the ways of viewing seemed like they might align with our data. The second model we used is called LVIS (Large Vocabulary Instance Segmentation). This dataset has 1000 different categories, including very specific and relatively rare objects such as a stapler, pineapple, and donut. We applied this algorithm because of its potential, with many categories, to provide a serendipitous set of tags that could aid in the search and discovery of the collections under consideration.

For this project, we also worked with an exciting new approach to object detection that was recently introduced, called region segmentation. In 2018, a research team from the University of Edinburgh and Google AI released a newly hand-tagged version of the COCO dataset that contained 91 new categories. The dominant focus of computer vision research at that time had been the identification and localization of objects. This, they argued, had overshadowed the automatic identification of other categories that constitute the "amorphous background regions" within an image.[55] They described the existing categories within the COCO dataset as consisting only of "thing" categories. That is, these categories consist of objects that have a specific size, well-defined shape, concrete parts, and can be enumerated. In other words, what has been called objects in object detection algorithms is relabeled as “things” in this approach.

Region segmentation adds another way of viewing the content of an image. Features of an image that are not enumerable are known as regions. Regions that do not correspond to things, such as the sky, water, and the ceiling, often compose most of an image, but at the same time, they lack the data and models needed for automatically identifying them. The team described these regions, in contrast to things, as "stuff." Their work resulted in a comprehensive ontology for "stuff" regions. Their approach proposes to split all regions under two super-categories: "indoor stuff" and "outdoor stuff." These groups are further divided into mid-level categories, which include "water", "ground", "sky", "furniture", and "floor". Finally, these are split into 91 fine-grained categories such as "sea," "mud," "clouds," and "carpet." The team labeled the entirety of the COCO dataset with these classifications. The joint task of identifying these labels alongside the "thing" labels, collectively known as the panoptic task, has been one of the primary competitions sponsored by the COCO dataset in 2018 and 2019.[56] As a result, there are now many accurate models for automatically labeling these regions within an image.

The development of this approach offers a few things. It allows for computational analysis to semantically zoom out and move beyond individual objects. This is accomplished by no longer needing to be specific about discrete objects but rather working with classes of stuff and entire regions of the image at once. This opens the study of compositional elements such as horizon and vantage point for our analysis. Categories like indoors and outdoors can matter for documentary photography; clouds, water, and sky are useful for environmental photography. Information from the entire scene collapsed to general categories can be used to organize and

[55] Caesar, H., Uijlings, J., & Ferrari, V. (2018). “Coco-stuff: Thing and stuff classes in context." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*: 1209-1218.
[56] Kirillov, A., He, K., Girshick, R., Rother, C., & Dollár, P. (2019). Panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition:* 9404-9413.

suggest photos. This helps address the object detection issues above and provides exciting areas of application, which we explore in the next section.

4.2.2 Applications

The stuff segmentation task offers a sufficiently general-purpose set of annotations for use within the distant viewing framework. While no classification scheme can be free of cultural assumptions nor account for all possible scenarios, the stuff categories are significantly more generic than the object categories. The "stuff" is more likely to be features that are less likely to change over time. For example, the design of a phone has changed significantly over time. Detecting a phone in our collections using a contemporary algorithm trained on the design of a phone in the 21st century will not render great results. However, detecting the sky is possible. This is particularly true of the high- and mid-level categories. The higher-level categories avoid some of the material-specific designations from the lowest-level categories, such as wood-flooring, that may not be applicable with images that significantly depart from the available training data. By aggregating information about detected stuff categories, we can make intelligent guesses about whether an image was taken inside or outside, how the people in the image are placed relative to the background, and the location and role of the horizon in framing the image.

There are several ways that we can make use of the stuff-category segmentations to build a set of annotations. Our general philosophy is to provide the results of deep learning models in their most unaltered format. This approach is unfortunately not very practical when detecting image segmentation for larger datasets because the results are too large. Because we need to identify the category of each pixel in the image, the segmentation results are of a similar size to the original dataset.[57] Therefore, we need a way of summarizing the results of the image segmentation for downstream processing. A simple approach is to count how many pixels exist of each category within the image. This is sufficient for predicting whether the image is inside or outside and the general objects that may be present. More information about the spatial relationships within the image can be achieved by dividing the image into a course grid system and recording the total number of each category that exists in each grid point.[58] We will make use of this aggregation technique in the following applications.

4.2.3 Evaluation and Challenges

As with pose and face detection (see Section4.4), we manually investigated several hundred of the images to determine any patterns of how well (or not) the object detection, and region segmentation algorithms seemed to perform. We noticed several different patterns across the different algorithms but very similar results across the five collections.

[57] The current format for image segmentation supported by COCO is, in fact, to store the segmentation as an image file where a fixed color is assigned to a specific category. This is a clever approach because the image file can be easily opened by most software programs, and it is able to make use of well-researched and supported algorithms for image compression.

[58] A third option is to assign each large region to the dominant category, which works well when the grid is made to contain a sufficiently large number of columns are rows (such as a 20-by-20 grid).

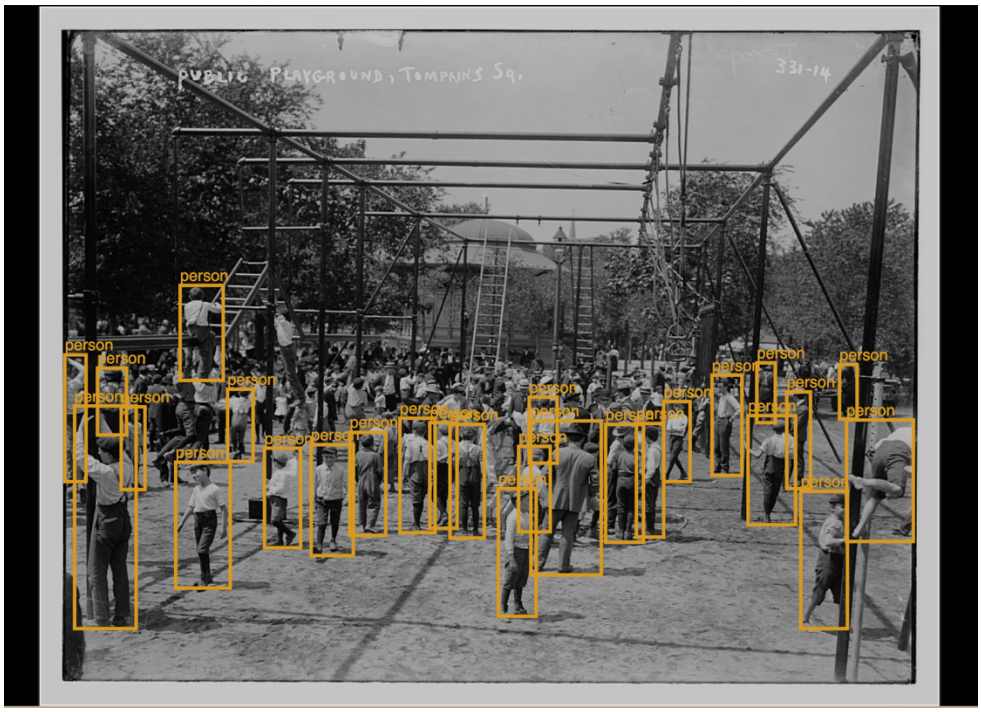

Figure 6. Object detection showing the people detected in a crowd. Notice that not all people are detected, but there are many people that are identified, even when obfuscated or very small. [2014681564]

Figure 7. Object detection results on a color FSA-OWI image. It detects the person and some of the oranges, but most of the oranges are missed even when setting a very low confidence score. [2017878149]

Figure 8. Example showing the LVIS image detection algorithm. It correctly identifies the flower arrangement, necklace, and ring in the historic black-and-white photograph. [2014712262]

For the object detection, we found that the detection of people had a very high precision and recall **[Figure 6]**. We noticed very few errors, and most of those were false negatives of small individuals in the background. Other object classes produced interesting results while also indicating that caution should be used. In many cases, objects were not detected even when they were clearly present in the image. For example, one image has dozens of oranges of all the same size and color, but the algorithm only detects five of them even when the confidence score is set very low **[Figure 7]**. Also, some types of errors seemed to be the result of objects that were only partially present or technological objects (such as cars) that have changed significantly from the time period of the photographs to the training data used for the algorithms. In other cases, object classes were confused. Usually, these were in understandable ways, such as mistaking cows for horses or the front half of a car as a motorcycle. Many object types were not included in the vocabulary of the object detection algorithms, meaning that using the resulting categories produces a biased version of what is contained in the collection. At the same time, the object types do exist; they do often produce very accurate results **[Figure 8]**.

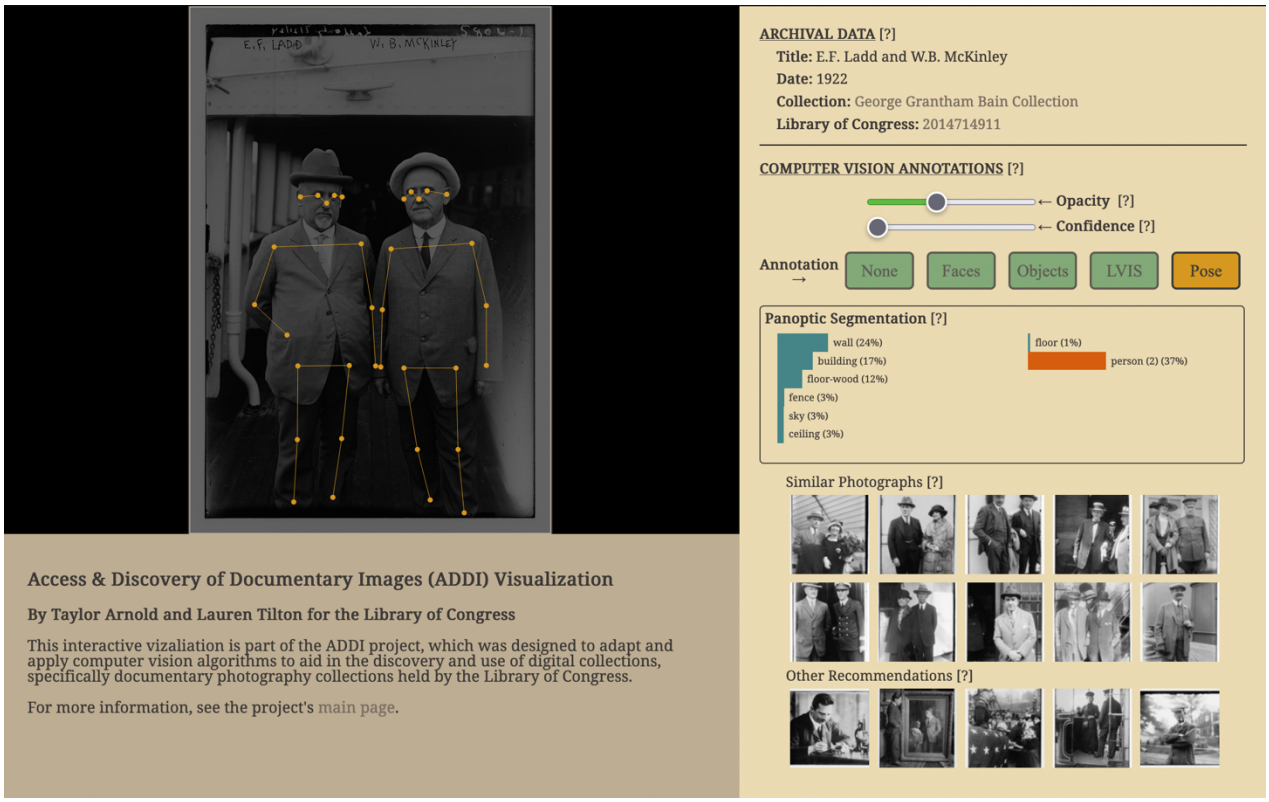

Figure 9. Example of the interactive visualization prototype showing detected poses. The opacity is set to dim the image and highlight the annotations. [2014714911]

The image region segmentation produced stronger results. Despite our initial hesitation that the lack of color would be a difficulty for detecting different regions, the algorithm made predictions that almost always covered the entire image and were generally reasonable for the depicted scene **[Figure 9]**. As always, when working with automatically generated annotations, care should be taken to avoid misinterpreting the results of stuff-segmentation algorithms. There are some categories that have some ambiguity between them, such as "dirt" and "sand" or "mat" and "rug." Also, the stuff categories were designed pragmatically for the task of assigning all the pixels in an image to a fixed set of classifications. The distinction between stuff and things is not a sharp epistemological distinction. Several categories overlap between the two, such as "furniture" and "door"; the difference in labels is a result of how big an object is within the frame of the larger image rather than a fundamental property of the objects themselves. Given time constraints for this project, we plan to conduct future work about specific stuff categories such as "sky" and "ground."

There are several applications of region segmentation that we found useful in our data analysis. The detection of people in the region segmentation was easier to use for detecting how much of an image was taken up by a person. It was good for identifying portraits and for classifying portraits based on their scale. Further, we can use segmentation elements for regions such as “sky” and “ground” to suggest photos that are outside. Eventually, this could lead to classifications for genres, such as environmental photography. Such a method also has potential power for formal analysis. A close analysis of Esther Bubley’s photos in DC through region segmentation brings attention to her framing. She literally looks up to the built

environment of DC, conveying the monumentality of the city and thereby communicating the largess and power of the nation's capital. It is literally a place to look up to.

From our assessment of the object and region segmentation algorithms, we feel that both algorithms can be used to make automatically constructed tags or recommendations. However, care needs to be used in doing aggregative computational analyses. The use of the people tags seems relatively safe, but the counts of other object categories should be considered questionable at best. The region image segmentation results can be used in an aggregative analysis, though the ambiguities between categories should be accounted for in the analysis. We recommend avoiding making claims that may come down to relatively arbitrary distinctions between categories—for example, claiming that Photographer A took more photos with dirt backgrounds whereas Photographer B preferred sand backgrounds—without carefully evaluating the appropriateness of the distinction and the accuracy of the automatic identification in a particular application. For an example of how to analyze this data and bypass this potential pitfall, see the Data Analysis Paper.

4.3 Image Embedding

While the previous algorithms were focused on being explicit about what we're looking for, such as people and objects, we also do not always know what we are looking for. Letting the data reveal patterns that we expect and did not expect can help guide us to new questions and avenues of inquiry. For example, we might see a pattern such as the use of a specific framing kind of framing in portraiture and a theme that we didn't expect to be prominent in the collection, such as the large focus on sports. Embeddings can also be a way to see within and across collections, including a way to introduce serendipity.

4.3.1 Method Description

Our previous sets of annotations have been developed by considering features at the intersection of things that are of research interest in the study of visual culture and the possibilities of existing computer vision algorithms. Systems at this intersection that are able to automatically capture elements of humanistic interest are ideal because they ease connections between a distant viewing analysis to prior scholarship. By recording explicit features, it is also easier to manually validate that the model is behaving as expected across a new corpus and to explain the meaning behind observed patterns and outliers based on concrete elements of the source material. The meaning of these patterns should ideally be possible without, at least initially, having to reference the algorithms that automatically annotated our corpus. The downside, however, is that it does not allow us to study features that are difficult to describe with an existing computational model, are very specific to a specific domain, or find patterns that fall outside of any prior academic study.

Image embeddings generate annotations that provide a way of finding patterns within a corpus without having to make explicit what underlying features are being captured. Embeddings are algorithms that place images into a high-dimensional space by assigning any image to an ordered collection of numbers. The individual dimensions, or numbers, do not correspond to any characteristic. Rather, the embedding produces an oppositional system, where certain unlabeled patterns can be seen when we look at relationships between images. For example, similar

images will often be placed in a similar position within the embedding. Subsets of images that all share a common trait—such as being outdoors, consisting of portraits, or occurring in specific locations—generally can be split out from the rest of the set in an easy way.

A typical approach for building an image embedding consists of a three-step process. We start with a large corpus with tagged labels, such as the COCO or ILSVRC collections. Next, we build a deep neural network that automatically learns a sequence of transformations that convert the original image into a representation of the objects contained within it. We then take images from our specific corpus of images and pass them through the neural network. However, instead of being concerned about the output of the entire model (i.e., an explicit set of annotations for the categories the model was training on), we save the intermediate results from one of the internal layers. It turns out that, when using a sufficiently powerful neural network, the internal representations provide an excellent image embedding that allows us to view a variety of patterns and features, even if the original dataset and supervised learning task differ substantially from our corpus of interest.

It is not necessary to manually retrain a general neural network each time we want to do an image embedding. Typically, one takes a published model that has already been built and uses it as-is. The only choice is to determine what model to use and which intermediate layer to project into. At present, we are aware of no exhaustive comparison of the recommendations between different models. Commonly, applications use the second-to-last layer of models trained on the ILSVRC dataset. The second-to-last layer is used because it, in theory, contains the most semantically rich information as it has passed through all but the final explicit set of transformations. However, it is also the most likely to be biased by the dataset used for training. Therefore, it may be useful to consider using shallower layers when applying to data such as digital scans of artwork.[59] Using models trained on the 1000-category ILSVRC dataset has been popular because it was for a long time the most well-known image-classification task and because it presents perhaps the most diverse set of objects to analyze.[60] In our analysis, we make use of the penultimate layer of the ResNet-50 model. [61] It is reasonably fast, very popular, and generally produces results that can be easily interpreted. Because the goal is to create an implicit set of annotations through the image embedding, our general framework should be largely unchanged when using a different model or layer. By taking this approach, we have a set of annotations that can be used to illuminate connections across the collection of images.

[59] There have been several studies looking at the transference of visual style between works of art. We are not aware of any existing attempt to use this approach for image embedding, but in theory, this offers an avenue for a more specific application of image similarity in the domain of art history. See, for example, Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*: 2414-2423.

[60] There are two other technical reasons that models from ILSVRC continue to be used for image embedding. First, popular machine learning libraries such as Keras and PyTorch currently contain implementations of these models, making the barrier to their application low. Secondly, the architecture required for the image segmentation tasks from the more recent COCO dataset (in contrast to object detection) requires a more complicated and harder to directly apply internal architecture. These concerns are likely to change over time; we expect to see new waves of popular image embedding algorithms over the coming years.

[61] He, K., Zhang, X., Ren, S. and Sun, J. (2016). "Deep Residual Learning for Image Recognition," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition:* 770-778.

4.3.2 Applications

As described above, there are two primary sets of applications for image embeddings. They can be used as inputs to a subsequent model in a process known as transfer learning. Alternatively, they can be used as a way of finding similar images based on images that are close together in the embedding space. In our applications, we did not have a specific secondary training algorithm to use and focused on the use of embeddings as distance metrics. Image embedding distances can be used to build recommendation systems by recommending images that share similar features. We eventually implemented these recommendations in the prototype visualization system described below and can be explored at https://distant-viewing.github.io/addi/06_interactive_viz/build/. A list of the fifty nearest neighbors for each image in our dataset, both within a collection and across collections, is given in the produced metadata.

Another usage of the image embedding as a distance metric would be to produce a visualization of the embedding space. This method has been implemented in several digital humanities tools, including PixPlot.[62] Such plots were not part of our workflow but would be a good follow-up application for future work with these collections.

4.3.3 Evaluation and Challenges

Image embeddings offer a powerful way to identify patterns and connections in a corpus that we may or may not have expected. This can lead to identifying formal patterns (such as portraits) and content (such as baseball) that can then be used to offer entry points into a collection. This could be paired with a recommender system that provides audiences a way to explore the collection and see patterns for themselves, which we implemented in the ADDI Prototype. Another way that we used these results was to see how similar photos were across collections (See Data Analysis paper). This could serve as a guide for which collections to put in conversation with each other based on visual features.

[62] See https://github.com/YaleDHLab/pix-plot.

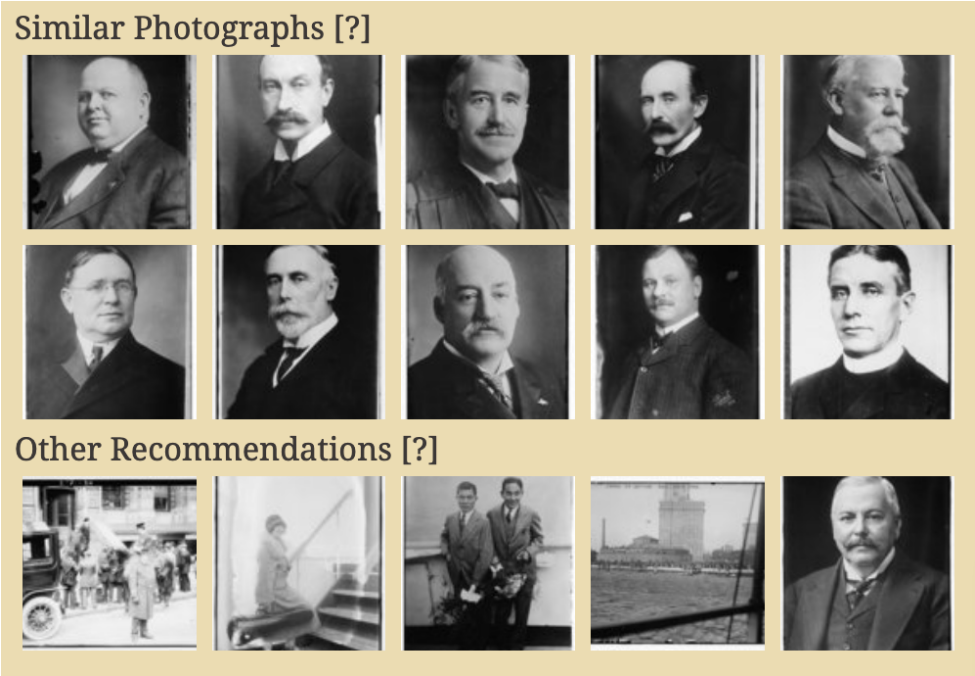

Figure 10. Top two rows show the nearest neighbors to a starting image. It does a good job of finding similarly framed portraits in the collection. Bottom row shows five random recommendations from the collection for comparison. [Source image is 2014686845]

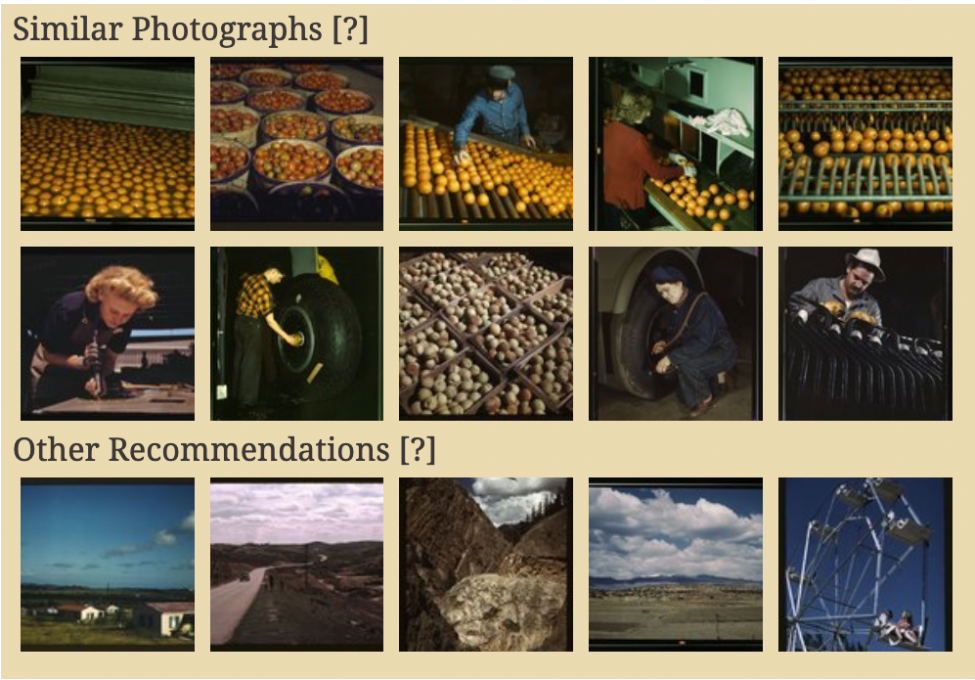

Figure 11. Top two rows show the nearest neighbors to a starting image. It finds examples of other images with orange or other fruit. Bottom row shows five random recommendations from the collection for comparison. It is not entire clear how some of the images on the second row relate to the starting image. [Source image is 2017878147]

The opportunities can also be a challenge. Each of the collections in our set contains many images that are very similar. Often these are photos of similar scenes taken from different angles or similarly posed individuals taken by the same photographer in the same location. Because of these similarities, it is easy to assess whether the image embeddings can associate images to others that are very similar **[Figure 10]**. And, in our visualization prototype, we did find that many of the recommendations returned images that were very similar to each other. They were so tightly connected that we needed to add a set of random suggestions to the recommendation system (see description below) to make sure that users did not get too stuck in a part of the collection. The random suggestions are included in the example figures. Finally, it was difficult to assess from this collection whether the distance metric would be able to find more subtle connections across a more diverse set. We noticed that occasionally some recommendations would surface that we could not necessarily explain, though these were the minority of cases **[Figure 11]**. However, it does not mean that another person might not see the connections, and therefore sharing the results through a recommender system still offers analytical possibilities.

Figure 12. Top two rows show the nearest neighbors to a starting image. Notice that the algorithm is heavily influenced by the round shape on the images. Bottom row shows five random recommendations from the collection for comparison. [Source image is 2014684875]

Another challenge for consideration is how to define an image. The image embedding seemed to be heavily affected by borders and the shape of images. For example, the algorithm would associate images with a round border to other images with a round border **[Figure 12]**. Identifying framing patterns with and across collections is a potential way to organize and remix the collection for access and discovery, as well as analyze composition and framing. Are there particular ways that certain groups (such as photo studios) framed their portraits? Did this differ by place and overtime? Are there patterns to how certain groups were framed that might indicate how framing played into visual social cues such as class, race, and status? (See the Data Analysis paper for more about the analytical possibilities.) Being able to ask and address such questions is opened by the way that the algorithm is viewing (i.e., similarity) and the scale of this analysis. This returns to the issue of digitization and data preparation highlighted previously. Depending on what one is studying, how an image is defined, such as the inclusion of the frame, becomes a key decision and shapes the results of the algorithms.

4.4 Social and Ethical Concerns

The usage of face and body recognition systems is not without its challenges. One reason for the historic and continued usage of these systems is the numerous applications that exist for use with tracking systems from within both the government and industrial applications. The importance of these applications can be seen by the heavy governmental resources put into face recognition through agencies such as DARPA, IARPA, the U.S. Department of Defense, the Department of Homeland Securities' Science and Technology Directorate (S&T), and the United Kingdom's Defense and Security Accelerator. The importance of industry applications, such as within digital security, marketing, and analytics, is similarly shown by the dominance of teams such as Facebook's AI Research, Google AI, Baidu Research, and Microsoft Research, across many of the face recognition and pose detection tasks.[63] The potential for troubling applications of these technologies should not be forgotten and need to be considered when applying algorithms and analyzing and releasing the results, particularly when engaged with the analysis of sensitive materials.

While many of the internal applications of face recognition algorithms are likely unknown to the public, there have been several recently published applications of this technology that are troubling. One research team at Stanford, for example, built an algorithm that attempted to distinguish between straight and gay men's profiles.[64] Despite the dataset coming from a purely observational study, and without citing any scholarship within the fields of gender and queer studies, the scholars concluded from their model's predictive power that "faces contain much more information about sexual orientation than can be perceived or interpreted by the human brain." They further suggested that their results offer strong support that homosexuality is the result of genetic and environmental factors that are present before birth. In another study from Shanghai Jiao Tong University, researchers tried to build a classifier for predicting whether an individual would become a criminal. [65] The authors claimed that their approach was "free of any

[63] This dominance is with only the small fraction of results that are publicly available. It is likely that the large technology companies have many more models that are released only for internal applications.
[64] Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. Journal of personality and social psychology, 114(2), 246.
[65] Wu, X., & Zhang, X. (2016). "Automated inference on criminality using face images." arXiv preprint arXiv:1611.04135, 4038-4052.

biases of subjective judgments of human observers" and was able to "discover very delicate and elusive nuances in facial characteristics and structures that correlate to innate personal trails."[66] In addition, there exist many research papers attempting to predict subjective and social categories from images such as emotion and gender.[67] While most of these studies temper their results, there is still the potential for problematic applications of these models in everyday lives that may disproportionally affect certain groups of people. Again, these are features that must be considered whenever we apply algorithms for face and body detection.

Even in cases where the goal of a facial recognition model seems relatively straightforward, issues often arise from skewed training data and when used in applications that differ from the type of data used to build the models. Most computer vision algorithms are trained on modern, high-definition inputs. We have found that some face detection models perform particularly poorly on older films and television shows that were recorded only in standard definition. Even when applying algorithms to modern datasets, biases in training datasets can cause unexpected challenges with the application of face recognition algorithms. Joy Buolamwini, the founder of the Algorithmic Justice League, analyzed several popular face detection algorithms and found that they routinely misclassified darker-skinned subjects.[68] Furthermore, they were particularly bad at performing the classification of dark-skinned women. Additional computer vision biases seen in practice include the inability for some people to unlock their mobile phones and darker-skinned people having trouble with automatic soap dispensers.[69] Careful attention to the results, including the training sets, is needed. This is one reason that we propose in Section 6 the creation of training data and training algorithms according to features animated by the study of photography and the goals of cultural heritage institutions.

It is challenging (perhaps even impossible) to find an algorithm that does not come with cautions. Our general approach has been to explore how we can reimagine how and for whom they are used. Yet, we always center the second half of Fitzpatrick's definition: "ask[ing] traditional kinds of humanities-oriented questions about computing technologies." [70] By actually using computational techniques to study the images, an added benefit is that the data that we are interested in often also illuminates challenges with these algorithms with attention to ethics and power. We do not apply these models without a critical perspective, particularly guided by attention to ethics and, even more importantly, power. [71] We do not apply these models without a critical perspective. This includes asking what ways of viewing the algorithm claims to engage

[66] Outside of the numerous socio-economic covariates that could explain any detected signals in the dataset, there were also fundamental methodological issues with the paper. For example, the photographs of criminals came from a government service, whereas the "law-abiding" individuals were taken from a random crawl of personal websites.

[67] Tarnowski, P., Kołodziej, M., Majkowski, A., & Rak, R. J. (2017). "Emotion recognition using facial expressions". *Procedia Computer Science*, 108: 1175-1184."

[68] Buolamwini, J., & Gebru, T. (2018, January). "Gender Shades: Intersectional accuracy disparities in commercial gender classification." In *Conference on fairness, accountability, and transparency*: 77-91.

[69] Fussell. S. (2019). "How an Attempt at Correcting Bias in Tech Goes Wrong." *The Atlantic*, https://www.theatlantic.com/technology/archive/2019/10/google-allegedly-used-homeless-train-pixel-phone/599668/.

[70] Fitzpatrick, K. (2012). "The Humanities, Done Digitally." *Debates in the Digital Humanities,* ed. Gold, M. and Klein, L., University of Minnesota Press. https://dhdebates.gc.cuny.edu/read/untitled-88c11800-9446-469b-a3be-3fdb36bfbd1e/section/65e208fc-a5e6-479f-9a47-d51cd9c35e84.

[71] For a more general discussion about the relationship between power and data, see D'Ignazio, C. and Klein, L. (2020), *Data Feminism,* The MIT Press.

in and then actually how did it view. We always do our own evaluation of the models within our corpus to try to identify any unintended biases that arise because of the differences between our corpora and the data used to train the computer vision models. To add, we try to minimize the amount of work that is being performed by black-box deep learning models. We use the models to detect people, identify faces, and tag their key points. Otherwise, we try to build all subsequent models using these extracted features as inputs to an interpretable model. Overall, we believe that bringing the perspective of libraries and researchers to assess the use of existing models is an important voice in the debates over computer vision and its applications.

5. Communication

It can be difficult for non-technical users to assess the performance of computer vision algorithms because most annotations are returned in a numerical format that cannot be easily related to the source image. For example, while the list of objects found by an object detection algorithm can be understood, the exact location of the object in the image will be reported in pixel coordinates. Figuring out where these are in relation to the image can be challenging without the use of a computer programing language. As one way to address this difficulty, we have built a prototype image annotation visualizer as part of the ADDI project. The prototype is designed to be a public search and discovery interface. It builds on work on generous interfaces and recommender systems for information retrieval, particularly work in the digital public humanities.[72] The visualizer is available through a web browser.

Our prototype visualizer was written in JavaScript using the popular React.js framework. This framework is specifically designed for single-page applications that update with respect to data and user interaction. These features, along with its extensive documentation and active community, made it an excellent choice for this project. We used CSS and HTML to control the style of the page and tested it on several different computers and operating systems. As described in Sections 2.3 and 2.4, we created JSON versions of the tabular CSV files for the visualization engine. This format is better suited for the needs of web applications.

The prototype visualizer is hosted on GitHub pages along with all the annotations and metadata. Due to their larger size, the images themselves are linked to the permalinks at the Library of Congress.[73] We included the Bain Collection and the color FSA-OWI images in the prototype. These can be accessed at the following links:

Bain Collection Visualization
FSA-OWI Color Images Visualization

[72] Arnold, T., Ayers, N., Madron, J., Nelson, R., and Tilton, L. (2020). "Visualizing a Large Spatiotemporal Collection of Historic Photography with a Generous Interface." In *5th Workshop on Visualization for the Digital Humanities*; Sherratt, T. and Bagnal, K. (2019). "The People Inside." *Seeing the Past: Experiments with Computer Vision and Augmented Reality in History*, eds Kevin Kee and Timothy Compeau. University of Michigan Press: 11-31; Whitelaw, M. (2015). "Generous interfaces for digital cultural collections." *Digital Humanities Quarterly*, 9 no 1. 2015. http://digitalhumanities.org/dhq/vol/9/1/000205/000205.html; Mitchell Whitelaw, M. (2012). "Towards generous interfaces for archival collections." In *Proceedings of International Council on Archives Congress*.

[73] This design was discussed with our contacts at the Library of Congress. At present, there are no known restrictions on embedding links from the LoC's collections on a third-party application.

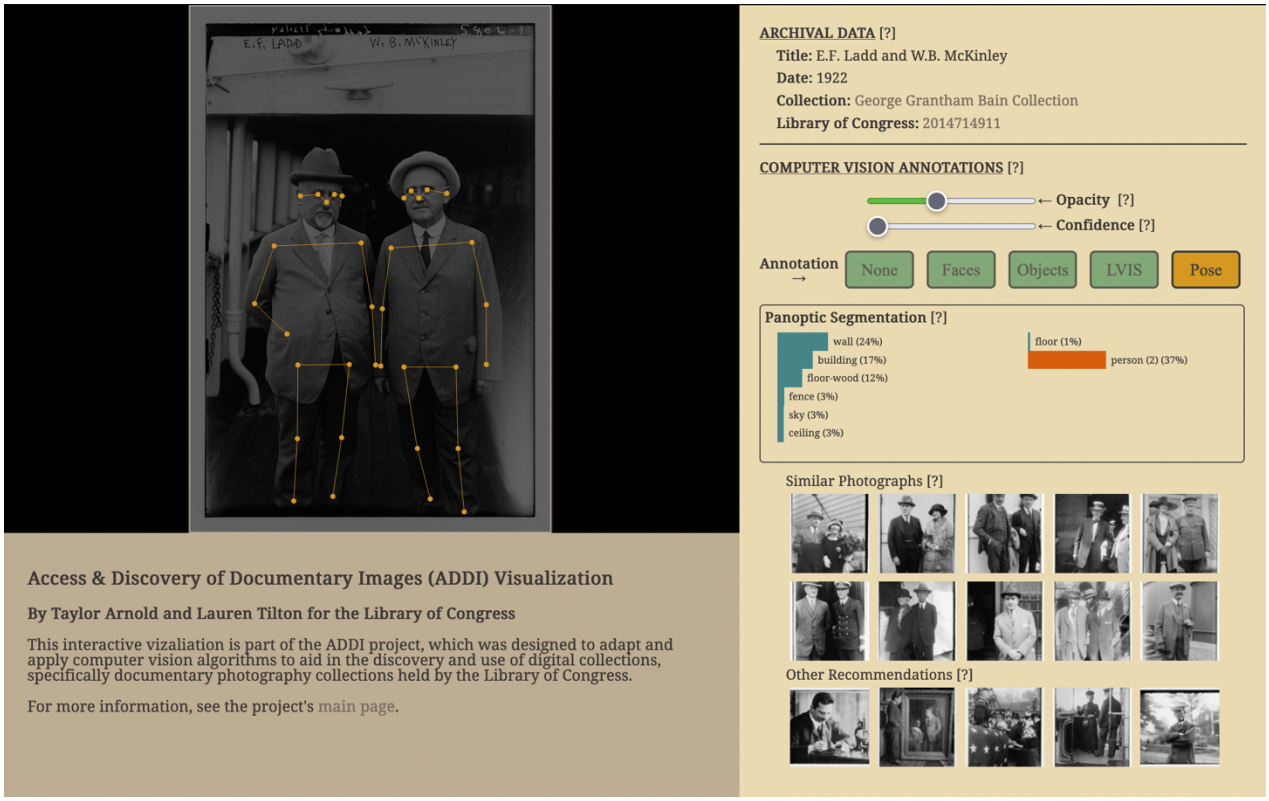

Figure 12. Example of the interactive visualization prototype showing detected poses. The opacity is set to dim the image and highlight the annotations. [2014714911]

A static screenshot of the visualization is included as a figure in this document **[Figure 12]**. All the other annotation figures in this document were taken from screenshots of the visualization prototype.

The ADDI visualization prototype offers several different kinds of interactivity to visitors. One of four different annotations can be turned on to illustrate what the computer vision algorithms detected in relation to the image itself. These include a face detection algorithm, an object detection algorithm based on the small set of COCO classes, an object detection algorithm based on the larger LVIS categories and the detected poses of people in the frame. Clicking on a button shows bounding boxes with labels, or in the case of the pose detection, the identified body key points and links between them.

Each of the annotations is accompanied by a confidence score. Visitors can drag a slider to the left to include more, possibly less accurate, annotations or right to include only annotations with a very high confidence score. This allows visitors to see less confident annotations, which are frequently wrong but often interesting when they are correct, while also being able to see the annotations with high confidence. The visualization also helps other computational researchers choose a cut-off score to use in their own applications.

A second slider changes the opacity of a black filter that sits between the image and the annotations. Off by default, visitors can make this darker to highlight the annotations. This is useful from a practical standpoint when trying to read the annotations on top of a busy image.[74] Additionally, we like to encourage visitors to slowly turn the opacity all the way to the right; when the image is entirely obscured, we can see the image the way that the computer vision algorithm sees it in terms of just the annotations. This helps highlight was the annotations could identify and what they are missing.

As our goal was to create a tool for a general U.S. public, the visualization prototype includes several additional design and information features to explain the components of the prototype as well as communicate degrees of authority and precision. There are hover tooltips that explain in a few sentences what each of the annotations and sliders is meant to be used for.[75] A link to the full project site with the project's full documentation, including this paper, is given at the bottom of the screen. We also wanted to be careful to explain which elements in the visualization are directly from the carefully curated archival data and which are automatically generated from the computer vision algorithms. We do not want the noisy and often inaccurate nature of the computer vision algorithm outputs to bring into question the archival information such as the title and year. These two groups are distinguished with a horizontal line as well as separate titles (Archival Data and Computer Vision Annotations) and tooltips to explain the differences.

The current version of the visualization algorithm is fully usable and functional but still in the prototype phases with limited features. A fuller version of the tool would allow users to search the collection by metadata fields and by the computer vision annotations. Also, we include only a limited number of archival metadata fields; a more complex interface would be needed to display more fields within the limited space of the browser window. Transitions between images result in a reload of all the page elements, something that would be ideally updated in a full version. Also, while the application implements a full browser history of all previously visited images, using the back button requires refreshing the page. We hope to be able to further refine, polish, and extend the current prototype in future work.

6. Conclusion

The Methods Paper offers an in-depth look into the processes and decisions behind ADDI. Computer vision offers a range of approaches to viewing images. Attention to the ways of viewing alongside the specificity of the kind of images being viewed guided this project. The algorithms that we applied offer opportunities and challenges as outlined above and in additional documentation on our GitHub site.[76] While all of them have potential uses, image

[74] The issue of readability is less of a problem with the black-and-white images because the color of the annotations offers a helpful contrast.

[75] Brennan, S. (2016). "Public, First." *Debates in the Digital Humanities,* ed. Gold, M. and Klein, L., University of Minnesota Press. https://dhdebates.gc.cuny.edu/read/untitled/section/11b9805a-a8e0-42e3-9a1c-fad46e4b78e5

[76] Many of the points central to our discussion here were dependent on certain technical specifications of the underlying models. Where necessary, attempts were made to describe and discuss these features. As our focus has been on a high-level description of the challenges and potentials of our approaches, many of the more technology choices and details that would be needed to put this system into place were avoided. Readers interested in these specific details can consult the code contained in this text's supplemental materials for a precise description detailing exactly what models and parameters are used through each of our analyses.

segmentation and image embeddings particularly stood out to us as exciting directions for search and discovery. The more generalized categories of image segmentation offer a level of abstraction that is amenable to search, such as indoor and outdoor. The results with identifying people through image segmentation also offer promising data for categorization such as individuals and crowds. Finally, image embeddings offer a way to find expected and unexpected patterns in the collection. All are amenable to remixing within and across collections to facilitate access and discovery.

As the ADDI Prototype models, making it clear to audiences that they are engaging with computer vision results is important. The results are probabilistic and come with uncertainty, as we make explicit in the prototype. We want audiences to explore with a productive skepticism and ask questions about the results rather than take them for a given. The purpose of the results isn't to outright answer the question "what is this an image of?" but to offer what we think *might* be an image. The final answer has animated debates over the meaning of images for centuries, and computer vision won't answer that question. It can only offer potential ways to answer this perennial question.

As we look toward the future, making it clear that the visual annotations are probabilistic and generated by algorithms is not just a matter of the design of the visualization platform but also the metadata itself. If we are to add such data to images in their metadata records, we argue that we need to make it explicit where these annotations come from. This includes extensive documentation of data decisions and their sources. We also look towards the possibility of developing custom annotations specifically designed for humanistic research. There is significant future work ahead to figure out standards and best practices for this practice.

We end with more immediate future directions for research. We hope other researchers will draw on our experience to study images as data. The Library of Congress is doing exciting work to streamline access to the data, in part shaped by the experiences and challenges that we discussed above as well as those of colleagues involved in the CCHC initiative. We will continue to do our part by developing the Distant Viewing Toolkit to lower the barrier to computer vision for humanistic inquiry (GITHUB URL). We hope you will reach out to collaborate and share how the toolkit can support your research. We also plan to keep delving into these collections to see how data analysis can further the study of early 20th century visual culture (See Data Analysis Paper). Together, we can harness and remake computer vision for a new purpose: to throw open the treasure chest to better understand the past, present, and imagine the future.

Acknowledgments

We are grateful to the Library of Congress and LC Labs for the opportunity to be a part of this innovative initiative. Thank you to the teams across LoC that met with us and graciously shared their knowledge and expertise. In particular, the Prints & Photographs Division team took significant time to guide us through their collections and histories of digitization, which was invaluable knowledge that shaped ADDI. This project is only possible because of the library's pioneering work to collect, preserve, and digitize photographs. Thank you as well to Meghan Ferriter, Alice Goldfarb, and Olivia Dorsey for meeting with Lauren bi-weekly to share ideas, ask questions, and collaboratively learn. There is extensive hidden labor in keeping an initiative like this on track, and Jaime Mears kindly helped us navigate the process. Finally, thank you to our fellow researchers in-residence. Along with their exciting projects, Lincoln Mullen’s always astute assessments of the state of the field and Andromeda Yelton's careful and critical use of machine learning helped us think bigger and broader about the impact of this initiative. Thank you to everyone involved for letting us be a part of a culture committed to collaboration, openness, experimentation, innovation, and most importantly, support and kindness.