Computer Vision and Cultural Heritage

Case Study 2

AEOLIAN

Artificial Intelligence for Cultural Institutions

Contents

Computer Vision and Cultural Heritage: A Case Study

Author: Catherine Nicole Coleman

We would like to acknowledge the information provided by interviewees, Jeff Steward, Elaine Treharne, Benjamin Albritton *and those who reviewed this case study prior to publication. We are very grateful for their contribution.*

This case study on computer vision applied to cultural heritage looks at critical points of intersection between research questions, the affordances of the technology and curatorial desires. The primary focus of this case study is Stanford Global Currents, a project completed in 2017 that applied computer vision techniques to medieval manuscripts. The discoveries and outcomes of that project are used as a point of departure to touch on related work at other institutions and independent work with computer vision applied to cultural heritage that has influenced how we think about search and discovery in libraries, archives, and museums.

This second case study for the AEOLIAN project is written based on interviews, project reports, conference papers, and published research. Some key terminology is defined and core concepts of computer vision that are essential to understanding the project are explained, but this is not a study of how computer vision works, nor does it address in any detail the methods or techniques applied in the Stanford Global Currents project. Global Currents took place from 2014 to 2017 and in the intervening years, visual feature extraction, which for Global Currents required custom-built algorithms, can now be done much more quickly and easily using commercially available services. Nor is the case study about computer vision as a field of study. It is about what can be learned from computational approaches to archival research that rely in some way on computer vision for information retrieval. The reason Stanford Global Currents remains an important case study today is not the technology they use, but what emerged from the researcher's and curator's engagement with the technology.

The case study is organized into four sections and a conclusion:

- The first section, "Stanford Global Currents" gives an overview of the project, its intended outcomes and the approaches used. It also addresses the research questions that drove the project and the unexpected discoveries that were enabled by the technologies the team used.
- The second section, "Segmentation of Images" delves into the way the technique of segmentation itself transforms the presentation of archival materials. This section looks back to an important project from 2011 that made use of computer vision turned on the archives to create windows onto materials that were previously invisible.
- Section three, "Reconfiguring the Collection" considers the tension between what researchers want to see and what our systems of discovery make possible. The systems and services that govern access to collections are grounded in textual description and classification. What happens when we just see the visual instead?

- The fourth section, “Classification Lessons” looks at computer vision as a prosthetic that allows us to see differently. Computer vision algorithms are human-made, but they enable super-human vision similarly to the way binoculars allow us to see clearly things that are far away. How has this technology also encouraged a critical examination of the assumptions built into classification systems?
- Finally, the conclusion, “Democratizing Access” considers how computer vision is influencing new modes of discovery and delivery of cultural heritage and why it plays an important role in re-imagining the possible uses of digital collections.

Introduction

Images of all kinds live particularly complicated lives within the organizational and information retrieval systems of libraries, archives and museums. Discovery in these systems depends upon metadata, part of which includes descriptive text. This forces a visual medium into a textual system. Work done in 2014 by John Resig on applying image similarity to anonymous Italian works at the Frick revealed the many ways that existing metadata is incomplete and inaccurate. The implications of this mismatch between the image and its metadata are being explored anew as we adopt computer vision methods to augment description and introduce new modes of discovery.

In Visual Studies, this shift is understood as an algorithmic reconfiguration of subject-object relations, specifically an algorithmic intervention between the viewing subject and the object viewed. William Uricchio writes, "… it is this algorithmic layer that stands between the calculating subject and the object calculated, and that refracts the subject-centered world charted by Descartes, that merits closer inspection".[1] Though Uricchio is particularly interested in very intentionally algorithmically constructed environments using techniques like Photosynth and Augmented Reality, his underlying point that algorithmic processing fundamentally changes or realigns subject-object relations is also relevant to the re-ordering of images within archives. The work documented in this case study demonstrates that visual organization of the past not only enables new ways of seeing, but changes, in turn, how we see and understand our archives. This algorithmic turn for the sake of image exploration and discovery has entered slowly into libraries, archives, and museums. Even in the digital humanities— the vanguard of re-visiting cultural heritage through computational methods—there has been an emphasis on the analysis of text rather than images. Optical Character Recognition (OCR) brought about a revolution in research and reading that has not been equaled for images.[2] There has been no practical or obvious technological intervention in image exploration that has led libraries to wide adoption as OCR has done. Computer vision, and specifically the development of Convolutional Neural Networks (CNNs) has changed that equation, opening up the possibility of non-textual search and discovery while also opening handwritten text to transcription to rival the OCR of print materials.

Digital humanists have been exploring applications of computer vision to better understand and engage with digitized materials from the past. In *Seeing the Past with Computers*, the editors Kevin Kee and Timothy Compeau argue that 'seeing technologies' are becoming essential tools for historians.[3] This move is inevitable, in part due to the quantity of material being generated but also because it presents new avenues for investigation of the past that scholars want to explore. In the early 2010s this work was still considered experimental, now it is becoming essential and, as this case study will demonstrate, cultural heritage institutions are helping to shape the technology in collaboration with researchers. Libraries, archives, and museums are often doing the digitization work. Even in the cases where libraries are acquiring

[1] Uricchio, p 27.

[2] Print OCR is by no means an entirely solved problem. The heterogeneity of early print and elements of print documents including tables and footnotes remain challenging to transcribe effectively using machine learning. See, for example, Zhalehpour et al (2019), *Visual information retrieval from historical document images.*

[3] Kee and Compeau, p. 3.

materials that have already been digitized, since they are responsible for storing and preserving those materials, they are also responsible for making them accessible in ways that are aligned with these new modes of inquiry.

Stanford Global Currents

The Stanford Global Currents project began in February 2014 as part of an international, inter-institutional research project with team members in the US, Canada, and the Netherlands. The overarching multi-institutional project was known as "Global Currents: Literary Networks, c. 1090-1900" and was spear-headed by Professor Andrew Piper at McGill University. That larger project was looking at different aspects of book production over time, across geographies and languages. The Stanford team, led by Professors Elaine Treharne and Mark Algee-Hewitt, and Dr Benjamin Albritton, received funding from the National Endowment for the Humanities' 'Digging into Data' Program.[4] The Stanford project, which is the focus of this case study, looked at British manuscripts. Expertise in computer science, humanities research, and library technology were brought into conversation to explore image processing and machine learning applied to textual and codicological analysis. The team sought to understand what computer vision could reveal about handwritten literary communication.

The Stanford project benefited from digitization work that had already been done. The corpus of manuscripts the team used came from the Parker Library of Medieval Manuscripts (http://parkerweb.stanford.edu/), a collaborative project between Stanford University and Corpus Christi College, Cambridge, consisting of 210 manuscripts dated between 1060 and 1220 with 63,000 total page images. In the print world, the availability of large, digitized text corpora in libraries soon led to the availability of large, searchable, and analyzable text corpora. Optical Character Recognition (OCR), a machine learning approach to text transcription from images of print was commonly used in academic libraries when Stanford Global Currents began. The technology, which has continued to improve in accuracy, can predict the words represented in an image of scanned print material successfully enough that OCR has become integral to the digitization workflow in libraries. OCR brought about a revolutionary change in how all of us read and search text; one that we now take for granted. Researchers working with digitized print texts are not limited to searching for keywords in catalogs, they can search the full text of multiple books at once. In 2014, when Stanford Global Currents began, no such benefit was afforded to researchers working with handwritten literature.

The Stanford Global Currents project set out, in 2014, to find a means to see within and across handwritten textual objects. Transcription would seem to be an obvious starting point with a text-based project, but handwritten objects are not amenable to OCR.[5] What makes OCR so successful at prediction is the standardization of printed text, it did not, however, work well with the variation inherent in handwritten text. The Stanford corpus was made up of medieval manuscripts from a range of genres, spanning two centuries, from 1080-1220, and included text

[4] Benjamin Albritton and Elaine Treharne, 'Medieval Manuscripts through new eyes: Automated Feature Recognition and *Mise-en-page*' (forthcoming).

[5] Since 2014, text recognition for handwritten materials has improved significantly. See, for example, *Transkribus* which was developed in 2016 as part of the Horizon 2020 "READ" EU project.

in three languages: Latin, English and French. Even printed material across a 200-year span varies significantly in letterforms. Medieval manuscripts present a uniquely different set of challenges because each object is unique. As Treharne explains in *Perceptions of Medieval Manuscripts: The Phenomenal Book*, the materials upon which the scribes have written will show variations from one square inch to another. In addition, there are idiosyncrasies at the level of the format, which could be sheets or scrolls, and differences between the stylistic patterns and choices of the manuscript compilers and scribes. Rather than OCR, Global Currents set out to experiment with two different approaches of Visual Language Processing (VLP). One approach was to identify similarities of lexical formation in handwritten materials. This investigation had, initially, similar goals to OCR for print. The idea was to automatically discover and isolate particular words, which might allow manuscripts to be machine-read.

To achieve automation of this kind, it is necessary to collect enough varied examples of the same word, or token, so that the lexical recognition software could learn to recognize that token when it encounters an example it has not seen before. This is a fundamental concept of machine learning: given enough examples, it is possible to train a model to identify with some measure of accuracy another example of the same kind of thing. Producing the necessary training data for this effort proved difficult and time consuming. One issue, as mentioned above, was the distinctive characteristics of the folios themselves. A number had to be weeded out of the process because of damage, intentional annotations, marginalia, and other marks that interfered with reading the selected words. In addition, the texts contained characters that are not in contemporary use in English or Latin such as, ð, Ð, þ, æ, Þ, Æ, ꝑ, ꝓ, 7. And, of course, there was the inherent variability resulting from the fact that the documents are handwritten rather than typeset. The clear frustration with the process is written into the team's report. They wrote, "the software initially showed few signs of learning: even at the end of processing one of the manuscripts, when nearly sixty examples of the word 'thing' had been entered, the software was still not able to reliably recognize the token."[6]

Though the transcription effort was not successful, the process yielded new insights into the kinds of variation in the manuscripts. The lexical recognition software used, MONK,[7] developed by Lambert Schomaker's team at the University of Gronigen (Netherlands), presented the group at Stanford with an analysis of the selected images of words for review and verification. By taking individual images of words selected from the folio and looking at those visual fragments side by side, the Stanford team discovered that certain scribal characteristics were identifiable across manuscripts. Abbreviations, for example, have distinctive characteristics that make it possible to distinguish between scribes not only within the same codex, but across different codices. This was just the type of cross-textual analysis the project was hoping to discover. This offered clues to many possible investigative paths. To identify a distinct scribal hand, for example, across a corpus that spans two centuries and multiple geographic locations, one can learn a tremendous amount about scribal practices. It was at this point that the team's attention shifted

[6] Treharne, E. (2016), p.3.This is the National Endowment for the Humanities Project HJ-50187-14 Stanford Global Currents White Paper. Applications of computer vision have improved significantly since Stanford Global Currents project ran. See, for example, *In Codice Ratio*, a transcription project focused on the Vatical Secret Archives that was presented at the Fantastic Futures Conference at Stanford University, 2019.

[7] See https://www.ai.rug.nl/~lambert/Monk-collections-nl.html

from trying to extract words held within the texts to what could be learned from the material qualities of the manuscript.

The study of the relationships between elements in the page layout or *mise-en-page* was undertaken in collaboration with Professor Mohammed Cheriet at the Synchromedia Laboratory at École de Technologie Supérieure (ETS) in Montreal. Rather than attempting to recognize words as they did with MONK, with Cheriet they turned to identifying the information retrieval tools used by medieval scribes and designers.[8] They considered a number of these features: running-headers, catchwords, writing grid format, *litterae notabiliores*, enlarged initials, minor flourishes and decorative devices, rubrics, intertextual space, ink-filled graphemes, and interlexical space. Over the course of the project, these four became their focus: *litterae notabiliores*, notable letters that mark the start of a section; enlarged initials, which, as the title suggests, are large initials usually drawn two or three lines high in red, blue, green, yellow, or purple; rubrics, which are titles of new texts or important sections of text, almost always in red in the body of the text; and intertextual space, which is white space within text, often found around rubrics and enlarged letters. Intertextual space may seem trivial, but it is an important component of page design. For example, the amount of space in a manuscript often reflects the resource available to the scribe-compiler.

This shift in attention to *mise-en-page* proved to be a particularly fruitful study of manuscript production in the long twelfth century. Cheriet's team was successful at extracting the four visual features the Stanford team was interested in. As a result, they were able to identify, as they anticipated, trends in the evolution of those important page elements. As recorded in the white paper, "palaeographical and codicological developments in the second half of the twelfth century are critical and include notable shifts in the complexity of folio design (double- or triple-column from single; introduction of running heads; systematization of rubrication; introduction of more navigational aids, including capitals, *capitula*; and recognition of the significance of clearly demarcated textual boundaries)."[9] The study of *mise-en-page* also allowed them to make important discoveries about localization. "Localization remains one of the most vexed, but important, aspects of manuscript studies in modern scholarship: fewer than one-third of manuscripts can be assigned to a place of origin."[10]

Identifying and labeling the visual features was essential, but not sufficient to the project's success. The other key element was the Stanford team's ability to analyze the results of ETS's work in galleries. The galleries provided a link from the visual feature out to the full codex so they could see that feature in its native context. This was new. Research questions that had emerged from traditional scholarship, which, as their report explains, involves looking through the material folio-by-folio, quire-by-quire, codex-by-codex, could now be tested by seeing all types of a feature together, side by side. This view on the material, which is simply impossible whether working with the physical objects or even digital surrogates that are arranged only for page-by-page online viewing, became the catalyst of a cascade of questions answered and the formulation of new questions.

[8] See Arabnejad, E. et al. (2016) for a technical account of the process. The paper was presented at the "Second International Conference on Natural Sciences and Technology in Manuscript Analysis, Centre for the Study of Manuscript Cultures (CSMC), Hamburg 29 February - 2 March 2016".

[9] Treharne, E. (2016), p. 3.

[10] Id., p. 3.

Segmentation of images

"Focusing on singular components aligned, often fortuitously, really does show this old material in a new light."[11]

Within the field of Computer Vision, image segmentation describes partitioning an image into meaningful regions or objects for processing. Images can be segmented, for example at the pixel level or at the level of the bounding box. At the pixel level, you can more precisely capture the shape and contours of a region, whereas the bounding box is, as it sounds, a simple box. The bounding box approach is often used during the process of creating labeled training data. As in the preparation of training data for MONK mentioned above, a person draws a line around the word 'thing' in a manuscript and labels it as 'thing'. Algorithms trained to identify objects, regions, and faces are now familiar not only because of their use in policing and surveillance but also in the commercial products used by owners of smartphones and people browsing the internet. Facial recognition technology—the ability to identify and classify faces based on biometrics—is extremely controversial in part because the classification systems are discriminatory and also because the use of biometrics to identify individuals has serious implications for personal privacy.[12] In this section, attention is given not to identification and classification but to the computer vision task of detecting a face in an image. To the extent that detecting a face and distinguishing it from other elements in an image is successful on a given set of images, it makes it possible to draw a bounding box around the face and then segment the image based on those bounding box coordinates to produce a gallery of faces just as the Stanford Global Currents team did with visual elements of medieval manuscripts.

The web project, "The Real Face of White Australia" originally developed by Tim Sherrat and Kate Bagnall in 2010 is an example of a creative application of face detection to transform engagement with an archive. The online project presents the visitor with a page of 100 faces, no text. Scrolling down the page produces more and more faces. The images of faces were selected from thousands of immigration documents held in the National Archives of Australia; the result of the "White Australia" policies of the nineteenth and twentieth centuries intended to limit and discourage immigration by non-Europeans.[13] A practice that was first instituted in the port city of Sydney in New South Wales to keep track of convicted criminals, which involved taking mug shots accompanied by descriptions of distinctive physical traits, was later applied to people crossing all borders into and out of Australia. The intent of "The Real Face of White Australia" was to reveal the people inside systems of historical record-keeping; because the photographs in these archival documents identify race as well as face, this gallery of faces very intentionally confronts Australia's claim of being a white country.[14]

---

[11] Id., p. 15.

[12] Joy Boulamwini and Timnit Gebru have demonstrated that the failure of algorithmic systems to both detect and classify faces is a result of racial and gender discrimination in classification systems. See "Gender shades: Intersectional accuracy disparities in commercial gender classification" by Joy Boulamwini and Timnit Gebru (2018) and the forthcoming book by Kashmir Hill, The Face Race, about facial recognition and personal privacy.

[13] See www.realfaceofwhiteaustralia.net

[14] Sherratt, T., & Bagnall, K. (2019), p. 13.

What made “The Real Face of White Australia” so compelling was that it offered an entirely different way into archival records. It was not only a visually striking collage of faces; it was also a document browser. “We know that the records, the photographs, the handprints, all carried emotive weight,” wrote Sherratt and Bagnall, “—it was the very reason we sought to expose them. What we did not quite realize was the effect of scale. Bringing all those photos together, without interpretation or intermediation, created a different type of experience.”[15] Sherrat and Bagnall scraped the document images from the National Archives of Australia and then used an open-source python computer vision library to detect the faces. The facial detection algorithm returned coordinates to define a bounding box where a face is detected in the image. Based on those coordinates, they could crop the original image, save the selection as a new file, and present a wall of faces — large thumbnail images that link through to the full document image.

The Stanford Global Currents project galleries of *litterae notabiliores*, enlarged initials, rubrics, and intertextual space do not carry the immediate social and political power that the faces in Sherrat and Bagnalls work do, but they had a powerfully transformative effect on the research process, leading to unanticipated questions. The outcomes described in the project white paper explains how, by taking elements out of the page and placing them side-by-side in gallery view provided an all-at-one-view experience that was entirely novel:

> The gallery has had useful consequences in permitting the team to formulate and begin to answer globally significant research questions. For instance, from experience of working with medieval manuscripts, it might be assumed that green is a prevalent color in the embellishment of large capitals. Our results indicate that this is not the case, and that where green does occur, it may have important information to provide about date and place of origin of the manuscript. Our rapid overview of manuscript *mise-en-page*, facilitated by the gallery of images, also intimates that it is possible to offer a chronological typology of features of decoration; of the introduction of running headers; of the uses of rubrics; of the tendencies towards effects, like diminuendo display scripts, by particular scriptoria at particular times.

Stanford Global Currents used image segmentation as part of the process of defining regions of interest on manuscript folios that would be used by their partners as training data for the machine learning models. The Stanford team drew bounding boxes around the *litterae notabiliores*, enlarged initials, rubrics, and intertextual space. As mentioned above, the bounding box defines coordinate space on an image which makes the selection of a part of an image possible. The Stanford team used IIIF, the International Image Interoperability Framework, to enable the entire process of viewing the digitized folios, annotating them with area selections, delivering the annotations to the team in Montreal, receiving the results, and viewing them in galleries. They explain this process in the white paper:

> A secondary, but significant, research goal was to test the mechanism for large-scale image processing to be done on a corpus of digital resources held by an institutional repository in such a way that all new knowledge produced through analysis of those resources could be re-incorporated into the repository to enhance the digital

[15] Id., p. 21.

resources themselves. This "virtuous circle" of scholarly communication, where a project consumes and then enriches re-usable repository data, has proven to be an ongoing challenge in the information sciences and library communities. Using the protocols specified by the International Image Interoperability Framework (IIIF), the project provided images via API (rather than the more usual exchange of hard-drives through the post) and requested returned data be provided to conform to the IIIF specifications as well, insuring full re-usability of the results outside of the context of this particular project.

By using IIIF protocols, Cheriet's team at ETS had freedom to determine the size of image they wanted to use. This is an important control for the computational team to have because the resolution of the image can have a significant effect on the success of the model. Too much information can take too long to process but it can also sometimes add unnecessary noise when considering visual saliency. And removing the long interruption of sending hard drives back and forth through the mail added to the thrill and satisfaction in the collaboration. Even in the very early stages of the collaboration results could be viewed by the Stanford team almost immediately through simple html galleries pulling again from the archived image files. As Albritton described it, "…pulling the images on the fly as the processing is happening, as the presentation is happening, and as the re-presentation is happening."[16]

Reconfiguring the Collection

"It's like looking at the world through a kaleidoscope. You know what it looks like and then you put the kaleidoscope up to your eye and it's a whole new world."[17]

Both "The Real Face of White Australia" and "Stanford Global Currents" used the capabilities of computer vision to fragment the whole of an archive in order to see it in a new way. With the physical archive, cutting images out of documents would be a destructive act and an illegal one. Medieval manuscripts have been particularly susceptible to this kind of damage. But the plasticity of the digital image makes what was a destructive act into a generative one. These transformative engagements with the digital surrogates are, in some ways opportunistic applications of a technology that was built for another purpose. Segmentation is intended for computational analysis, not human viewing. But these projects break down the barriers that the technological systems of information delivery in our libraries and archives impose. "We are deeply in love with the records and the stories they reveal" wrote Sherratt and Bagnall, "[w]e cannot say the same about the National Archives' collection database, RecordSearch."[18] Discovery systems reflect the underlying data models and long-standing data management practices within the institution rather than research practices. These systems, built to aid in discovery can often hinder discovery when their organizing principles dictate the questions one is required to ask to find objects.

Since machines trained to "see" images do not see the way humans see, the results provide opportunities to see differently things we thought we understood well. The algorithmic layer that stands between the subject and the image object relies entirely on an abstraction of

[16] Albritton, B. (2022), Interview.
[17] Treharne, E. (2022), Interview.
[18] Sherratt, T., & Bagnall, K. (2019), p. 17.

the visual object into numbers.[19] Digital or digitized images are processed as a matrix of pixel values, effectively converting a semantically complex whole, as the human being sees it, into a grid of numbers that can be filtered, or broken up into sections, and analyzed to identify subtle patterns and collections of patterns. As discussed above, much of the research in computer vision has focused on distinguishing objects represented in an image, known as object detection. Training an algorithm on labelled examples makes it possible to learn the features that make those examples similar to each other and, on that basis, find other visually similar objects. But some of the most important discoveries come from what could be understood as errors. Reflecting on the project years later Treharne noted, "[w]orking with images the way that we did created contiguities that I would never have otherwise seen, but also strange juxtapositions."[20] Some of the ways that discoveries led to new research questions are captured in the Stanford Global Currents white paper.

> Inductive research questions leapt off the galleries put together by Dr Albritton from the raw data sent from Professor Cheriet's team. We were surprised to see how dissimilar particular *litterae notabiliores* are from others in the gallery. Dissimilarity might be attributable to national trends in color use; to the 'rusticity' of specific initials in manuscripts not produced at major writing centers; or to the idiosyncrasy of scribe-artists, who we might now be able to trace with greater precision. We were delighted to discover that manuscripts never before associated with one another might, in fact, be related in terms of their production methods. We saw this emerge through the serendipitous juxtaposition of initials in the gallery.

The serendipitous juxtapositions were made possible in part because of the segmentation described in the previous section, but also because the collections were presented based on visual features, not based on search terms. The galleries of visual elements created by the Stanford Global Currents team were intentionally separated into the four classes that they were seeking so that they could compare similar items side by side. Since that time, a number of projects have applied computer vision to heterogeneous image collections in order to find similarities without pre-defining classes. There are echoes of the discoveries of the Stanford Global Currents project in these other projects that reveal sometimes unexpected visual patterns in image collections.

In 2017 the National Endowment for the Humanities funded a collaboration between the Frank-Ratchye STUDIO for Creative Inquiry at Carnegie Mellon University and the Carnegie Art Museum to experiment with computer vision applied to the Charles 'Teenie' Harris Photography Archive.[21] The collection includes 80,000 exposures by Charles Teenie Harris (1908–1998) who photographed Pittsburgh's African American community for about forty years in the mid 20th Century. It is described on the website as one of the most detailed and intimate records of the black urban experience. Working with 60,000 digitized images from the archive, they used a classifier, InceptionV3, which was pre-trained on labeled images from the ImageNet benchmark

[19] See Zhalehpour, S. et al. (2019), 'Visual information retrieval from historical document images'.
[20] Treharne, E. (2022), Interview.
[21] See https://cmoa.org/teenie

dataset to generate labels for each image. Then they took the top five labels from each photo and compared them with the top five labels for every other photo as a method of identifying similarity.[22] The experiments with this type of automated classification revealed groupings like women in fur coats and car crashes— image sets that, according to collection archivist Dominique Luster, could never have been discovered via the existing metadata.[23] The results were surprising and intriguing. They also revealed the limits of applying an algorithm trained on twenty-first century photos to images from the mid-twentieth century. InceptionV3 is one of a set of convolutional neural network architectures developed by Google that was intended to automate captions for images. It is a pre-trained supervised model, meaning that it has already learned to how to classify images into many pre-defined categories. As Peter Leonard explained in his description of the Yale DH lab's experiments with the Inception algorithm, there "are likely be to labels such as 'cup of coffee', 'cat', and 'automobile' – but you're unlikely to find 'parasol' or 'steam engine.'"[24]

Convolutional neural networks like Inception are multi-layer architectures. The input to the process is the image and the output is the score indicating how well the image matches the different pre-defined classifications. Layer by layer, it builds a more and more complex understanding of visual elements like edges, textures, and shading, refining, and aggregating that information such that, at the penultimate layer, just before the image is rated by its similarity to specific classes, the algorithm has captured a sophisticated understanding of the image based on high level features that place it into multi-dimensional vector space. The image's position in vector space, makes it possible to relate it to other images that are similar. Since the classification strategies of libraries, museums, art historians, and other researchers do not match well with those of InceptionV3, they instead use these measures of similarity to present the images in a display, eschewing the descriptive terms. The result is a visual gallery like those that Stanford Global Currents produced, but in which, even across very large heterogeneous collections, the contiguities and strange juxtapositions appear.

One of those experiments, again involving the STUDIO for Creative Inquiry, explored visual similarity in digitized works in the National Gallery of Art (NGA) in order to compare the visual distribution of different collections within the National Gallery.[25] It is important to note that the type of similarity that the algorithm produces, as mentioned above, is complex and multi-dimensional, based on 2,048 features. To create a grid view for people to study easily like the Stanford Global Currents galleries, the team working with the NGA collection needed to reduce that dimensionality (a process known as dimensionality reduction), necessarily losing

[22] This process was confirmed in conversation with Zaria Howard, a student at the time, who worked with Kyle McDonald in 2017 at the Studio for Creative Inquiry on the Teenie Harris Photography Archive at the Carnegie Museum of Art.

[23] Conversation with Dominique Luster, Golan Levin, and Caroline Record as part of a speaker series on AI in libraries, archives, and museums hosted by Stanford Libraries. See an example of an image cluster here: https://www.flickr.com/photos/creativeinquiry/34456612652/in/album-72157681593431871/

[24] See https://www.pleonard.net/semantic-image-clustering-with-neural-networks/

[25] The project team included Sarah Reiff Conell (University of Pittsburgh, Department of History of Art and Architecture), Lingdong Huang (Carnegie Mellon University, STUDIO for Creative Inquiry), Golan Levin (Carnegie Mellon University, STUDIO for Creative Inquiry), and Matthew Lincoln (Carnegie Mellon University Libraries). See https://nga-neighbors.library.cmu.edu

some information along the way.[26] In an essay describing their findings, they explain how they interpreted the sometimes surprising juxtapositions:

> "If you look closely in the portraiture section, you'll glimpse a black and white Robert Motherwell painting amidst the varied portrait heads. Although the Motherwell is an abstract painting, its forms do bear some resemblance to a silhouette, explaining why it ended up in the same visual neighborhood. Inception's fixation on broad geometric qualities can eclipse more important features, though. For example, it is quick to cluster together circular paintings, prioritizing the general overall outline shape over the fact that the Holy Family inside one *tondo* might be more appropriately placed next to other images of robed groups of figures."[27]

The different applications of computer vision described in this section all provide radically new ways to examine visual culture, from trends and changes over time, to similarities that would not appear in a metadata search, to the overview of patterns of collection development. An important distinction between the approach used by Cheriet's team on Stanford Global Currents as compared to the approach used by the STUDIO of Creative Inquiry is that the former was developing a very targeted model, optimized for four carefully curated and domain-specific classes of object while the latter applies a general model trained on a decidedly heterogeneous and even haphazard set of images drawn from the bounds of contemporary internet culture. Both approaches introduce important lessons for a critical examination of classification in libraries, archives, museums, and academic fields of study.

Classification Lessons

There are two primary genres of classification problem that the project teams discussed in this case study have encountered. One appears when trying to train an algorithm to understand existing classification schemes and it fails because those existing classification schemes are incorrect or inadequate. The other appears when a pre-trained model is used that has —built into it— assumptions about how things ought to be classified and organized that are problematic in many ways. A close look at the lessons learned about classification by the Stanford Global Currents team and those of other projects in libraries, archives, and museums that make use of the techniques of segmentation and reconfiguration of collections based on visual features, point to ways that computer vision can contribute to a critical assessment of discovery systems.

Treharne noted in an interview that the Stanford Global Currents project revealed to her how unhelpful contemporary classifications of data can be. "You have to have [the algorithm] distinguish between things that we categorize as the same thing when, in fact, they are not at all the same thing."[28] An example of this problem arose when determining which examples should define the class *litterae notabiliores*. *Litterae notabiliores* are often decorated with flourishes and are visual cues for the beginning of a new textual item. The team soon discovered that the class

[26] You can see the results here: https://nga-neighbors.library.cmu.edu/essay
[27] Lincoln, M., Levin, G., Conell, S. R., & Huang, L. (2019). National Neighbors: Distant Viewing the National Gallery of Art's Collection of Collections. https://nga-neighbors.library.cmu.edu
[28] Treharne, E. (2022), Interview.

litterae notabiliores contained at least three different types of visual cues for the start of a new paragraph. And *litterae notabiliores* were difficult to distinguish from enlarged initials because both are enlarged initials. They can be very large initials when at the beginning of a text or they can be smaller, pen-drawn initials indicating a new 'paragraph' or section.
Another class that challenged their assumptions about which information retrieval devices are most important in medieval manuscripts was rubrics. Rubrics are titles of new texts or important sections of text, almost always in red in the body of the text. The algorithm trained on examples of rubrics ended up finding other similar elements, like numeration systems, that were red, but not rubrics. As with the application of the *litterae notabiliores* class, a computer vision algorithm in the hands of a subject expert becomes a versatile instrument that not only speeds up the process of identification of visual elements, making study on a much larger scale possible, but also allows the subject expert to see difference and distinctions with more precision.

Computer vision can, similarly, serve as an instrument in the hands of information professionals who are actively seeking ways to improve cultural awareness in curatorial practices in order to address the legacy problems with textual classification systems and the way they define what can be discovered.[29] It can begin, as in the Stanford Global Currents project, with examining the classification practices of the field. There are striking examples of racial bias in classification made obvious by comparing a search term to the results. Writing about collecting infrastructures, Yanni Loukissas recounts a presentation in 2015 by Marya McQuirter in which she reveals the way the academic descriptions of artwork that drive the search engine reveal the racism in the curatorial practice. Searching a Smithsonian online image catalog for the term *black* brings up examples of work by African American artists because the curators document racial identity in those descriptions whereas a search for the term *white* brings up little about race. Dominique Luster describes this in terms of the dual problems of white normativity, in which whiteness appears neutral/natural/right and the white gaze in which the descriptive practices assumes that the viewer is white.[30] Applying pre-trained classifier algorithms produce similar biases, even if, as in the examples above, the classifier is used for its ability to produce a visual representation of similarity. That layer of high-level visual abstraction that considers a Motherwell painting similar to a silhouette is also capable of propagating the bias that McQuirter revealed in a keyword search in the Smithsonian because the similarity of the visual elements are ultimately defined by the label that a human being has given them.

Biases in pre-trained commercial models are difficult to trace because the practice of tracking provenance and documenting data collection are not part of the process. Pre-trained computer vision models, as described above, are intended to assign labels or categorize images. Not only are the criteria for selection of those categories entirely different than those applied in libraries, archives, and museums, the data collection practices are, too.[31] Thomas Smits and

[29] See, for example, Engseth, Ellen. "Cultural competency: A framework for equity, diversity, and inclusion in the archival profession in the United States." *The American Archivist* 81, no. 2 (2018): 460-482.
[30] See AI, Metadata Creation and Historical Bias, at the 2021 National Information Standards Organization Annual Conference. https://niso.cadmoremedia.com/Category/73fcc4e2-7f1b-493a-9cec-1d4f85fe3afe
[31] As noted in the first Aeolian case study, *The National Archives (UK),* Eun Seo Jo and Timit Gebru have highlighted the potential advantages of bringing archival data collection practices into conversation with the machine learning community. See Jo, E.S. and Gebru, T. (2020) 'Lessons from Archives: Strategies for Collecting Sociocultural Data in

Melvin Wevers looked closely at six of the widely used benchmark datasets to understand how they were collected and how they were used to train computer vision models. What they discovered is that the image collection was not rooted in any theory of visuality. Rather, it was based on matters of economic convenience including the availability of images, perceived practical applications, and a favoring of categories that can be unambiguously described by text.[32] In other words, they did what was expedient. This approach reflects the big data conceit that not only is more data better but that lots of data renders theory dead. It is particularly problematic when it drives image search in commercial services like Google that are guided by business interests rather than operating, as libraries, archives, and museums more often do, in the interest of the public good.[33] But libraries, archives, and museums are also often driven by expediency and expense, too. And legacy problems with classification systems are expensive to solve. With visual materials, as Benjamin Lee has argued, biases are also propagated through the long history of digitization practices even before machine learning enters the process.[34]

Conclusion: Democratizing Access

Computer vision and other applications of artificial intelligence are understood in the context of computer science in terms of automation and optimization. But these tools in the hands of librarians, curators, artists, designers, and scholars more often drive critical encounters with the systems that organize, classify, restrict, and confine access to cultural heritage. In our attempts to train machines to see as we do, our own biases and normative assumptions are revealed. This tension between the 'dumb' machine and the very segregated and specialized academic training that we impart to it is, as in the Stanford Global Currents project, encouraging a more pluralistic approach to interpretation with an underlying motivation of liberating the objects of study to better provide access to cultural heritage.

The Stanford Global Currents project challenged assumptions about the interpretation of manuscript materials. The collaborative nature of the project meant that people who had never encountered medieval manuscripts before were seeing them and sharing their experiences with the research team. The Stanford team reflected on this as a design opportunity in the white paper: "The team at Stanford will determine if these initial audience responses can be employed in the design of better interpretative frameworks for digital repositories that present complex early textual materials, often to interested viewers who have little or no expertise in paleography, codicology, and modern methods of curation and display."[35]

At the Harvard Art Museums, Jeff Steward, the Director of Digital Infrastructure and Emerging Technology, has similar motivations. His institution holds about 250,000 art objects.

Machine Learning' in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 2020, pp. 306-316. Available at https://dl.acm.org/doi/10.1145/3351095.3372829.

[32] The Smits and Wevers (2021) study also addresses the temporal bias in the image datasets. A critical and damning point they make is that the way these benchmark datasets and the technology they are enabling are presented in the literature obscures the power and subjective choices of its creators.

[33] See Safiya Noble's *Algorithms of oppression* (2018) for an examination of the racialization of classifying people both in commercial and public information systems.

[34] See Benjamin Charles Lee, Compounded Mediation: A Data Archaeology of the Newspaper Navigator Dataset, 2020.

[35] Treharne, E. (2016), p. 8.

"When you run the numbers," says Jeff, "it's less than 1% that is ever physically on view."[36] Much of the material has been digitized, but the cataloging remains very thin. This problem of the cataloging backlog is common at institutions that hold unique objects. At https://ai.harvardartmuseums.org, Steward has created a space where the digital images of items in the collection can be searched based on terms applied by pre-trained computer algorithms. But rather than attempting to reduce the machine-generated labels to the top three terms, Steward allows the patron to see competing results from four different commercial services, including the confidence levels for each tag. When machine-generated labels are used in existing information retrieval systems, an opportunity to engage with 'seeing' the image is lost. As Steward describes it, the academic descriptions that accompany objects in the museum are very subjective. They reflect the interpretation of the curator, based on their expert training, but nonetheless subjective. Exposing the variety of tags assigned by the commercial algorithms, including the confidence level, reveals uncertainty. In Steward's words, this "exposes people to the idea that it's alright to have an opinion".[37]

Making the wealth of cultural heritage objects accessible, whether to research or public engagement more broadly, requires much more than digitizing them and presenting them as one-to-one virtual replicas of each piece online. For Jeff Steward, it also means extracting visual elements, maximizing the capabilities of digital display, cutting through the narrowly academic descriptions of the objects in the catalog, and exposing the subjectivity of human and machine descriptions. While we might assume that the concerns of an experimental group within a university art museum with a mandate to provide access to the full extent of an art collection would be worlds away from the interests of a Stanford professor and Welsh medievalist specializing in manuscript studies, they intersect around the possibilities of computer vision to influence interpretation. Since the Stanford Global Currents project concluded, Treharne has continued her work exploring digital interpretative frameworks, the phenomenology of the digital environment, and the phenomenology of the book.

[36] Steward, J. (2022), Interview.
[37] Id.

Bibliography

Arabnejad, E., Nafchi, H., Treharne, E., Allen, C., Albritton, B., and Cheriet, M. (2016). Visual Saliency for Visual Feature Analysis of Historical Manuscripts. purl.stanford.edu/gp006cz9645.

Albritton, B., Henley, G., & Treharne, E. (Eds.). (2020). *Medieval Manuscripts in the Digital Age*. Routledge.

Albritton, Benjamin. (2022) Interview for the "Stanford Global Currents" project, Stanford University.

Arnold, T., & Tilton, L. (2019). Distant viewing: analyzing large visual corpora. *Digital Scholarship in the Humanities*, *34*(Supplement_1), i3-i16.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Caceres, A., Weber, A., & Schomaker, L. (2020). MONK in practice: Indexing heterogeneous handwritten collections. In *7th Digital Humatities Benelux 2020*.

di Lenardo, I., Seguin, B. L. A., & Kaplan, F. (2016). *Visual patterns discovery in large databases of paintings* (No. CONF).

Firmani, D., Merialdo, P., Nieddu, E., & Scardapane, S. (2017, November). In Codice Ratio: OCR of Handwritten Latin Documents using Deep Convolutional Networks. In *AI* CH@ AI* IA* (pp. 9-16).

Hitchcock, T. (2017). Digital searching and the re-formulation of historical knowledge. In *The virtual representation of the past* (pp. 81-90). Routledge.

Lee, B. C., & Weld, D. S. (2020, October). Newspaper navigator: Open faceted search for 1.5 million images. In *Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology* (pp. 120-122).

Lee, B. (2020). Compounded Mediation: A Data Archaeology of the Newspaper Navigator Dataset.

Lee, B. C. G., Baco, J. O., Salter, S. H., & Casey, J. (2021). Navigating the Mise-en-Page: Interpretive Machine Learning Approaches to the Visual Layouts of Multi-Ethnic Periodicals. *arXiv preprint arXiv:2109.01732*.

Luster, D. (2021 February 23). *AI, Metadata Creation and Historic Bias* [https://nisoplus2021.cadmore.media/Category/0338ac98-5024-4e3f-aa75-13933b97f5bd]. NISOPlus, 2021. https://nisoplus2021.cadmore.media

Manovich, L. (2009). Cultural analytics: visualising cultural patterns in the era of "more media". *Domus March*.

Lincoln, M., Levin, G., Conell, S. R., & Huang, L. (2019). National Neighbors: Distant Viewing the National Gallery of Art's Collection of Collections. https://nga-neighbors.library.cmu.edu

Kee, K., & Compeau, T. (2019). *Seeing the Past with Computers: Experiments with Augmented Reality and Computer Vision for History* (p. 254). University of Michigan Press.

Leonard, Peter. "Semantic Image Clustering with Neural Networks" https://www.pleonard.net/semantic-image-clustering-with-neural-networks/

Loukissas, Y. A. (2019). *All data are local: Thinking critically in a data-driven society*. MIT Press.

Noble, S. U. (2018). *Algorithms of oppression*. New York University Press.

Resig, J. (2014). Using computer vision to increase the research potential of photo archives. *Journal of Digital Humanities*, *3*(2), 3-2.

Sherratt, T., & Bagnall, K. (2019). The people inside. *Seeing the past: Experiments with computer vision and augmented reality in history*, 11-31.

Smits, T., & Wevers, M. (2021). The agency of computer vision models as optical instruments. *Visual Communication*, 1470357221992097.

Snydman, S., Sanderson, R., & Cramer, T. (2015, May). The International Image Interoperability Framework (IIIF): A community & technology approach for web-based images. In *Archiving Conference* (Vol. 2015, No. 1, pp. 16-21). Society for Imaging Science and Technology.

Stern, R., Emanuel, J. P., Harward, V. J., Singhal, R., & Steward, J. (2021, May). IIIF as an Enabler to Interoperability within a Single Institution. In *Access to the World's Images: The 2016 International Image Interoperability Conference*.

Steward, Jeff. (2022) Remote interview for Aeolian Case Study 2.

Treharne, E. (2016). Global Currents: Cultures of Literary Networks, 1050-1900.

Treharne, Elaine. (2022) Interview for the "Stanford Global Currents" project, Stanford University.

Treharne, E. (2021). *Perceptions of Medieval Manuscripts: The Phenomenal Book*. Oxford University Press.

Treharne, E. (2012). "The Good, the Bad, the Ugly: Old English Manuscripts and Their Physical Description". *The Genesis of Books: Studies in the Scribal Culture of Medieval England in Honour of A. N. Doane*, 261-83.

Uricchio, W. (2011). The algorithmic turn: Photosynth, augmented reality and the changing implications of the image. *Visual Studies*, *26*(1), 25-35.

Wevers, M., & Smits, T. (2020). The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities*, *35*(1), 194-207.

Zhalehpour, S., Arabnejad, E., Wellmon, C., Piper, A., & Cheriet, M. (2019). Visual information retrieval from historical document images. *Journal of Cultural Heritage*, *40*, 99-112.