White Cube / Black Box: Investigating Bias in Museums and Algorithms

Sophia Brueckner, Shannon Yeung, Jing Liu, David Choberka, Kerby Shedden, John Turner, Isabelle Marie Anne Gillet, Mingchen Lu, Xingwen Wei

University of Michigan
Ann Arbor, MI, USA
sbrueckn@umich.edu / shannysw@umich.edu / ljing@umich.edu /
dchoberk@umich.edu / kshedden@umich.edu / jmturner@umich.edu /
igillet@umich.edu / mingchlu@umich.edu / weixw@umich.edu

Abstract

White Cube / Black Box seeks to identify bias and the many ways bias gets introduced into and amplified within systems. A highly interdisciplinary team of data scientists, curators, designers, and artists used face detection and race classification algorithms to explore bias in algorithms and University of Michigan Museum of Art's collection of artworks.

Keywords

Machine learning, face detection, race classification, bias, museums.

DOI

10.69564/ISEA2023-45-full-Brueckner-et-al-White-Cube-Black-Box

Introduction

White Cube / Black Box is a collaboration between artists, designers, curators, and data scientists at University of Michigan Museum of Art (UMMA), the Michigan Institute for Data Science, and the University School of Art and Design that attempts to shed light on the opaque decision-making processes within museum collecting practices and machine learning algorithms.

White Cube / Black Box seeks to identify bias and the many ways bias gets introduced into and amplified within systems. In art, the phrase "White Cube" references the history of exclusionary practices within museums and galleries. Using sterile white walls and decontextualized spaces, works of art are divorced from the outside world, making them less approachable and accessible. In technology, the "Black Box" is a controversial metaphor used to describe automated systems where the decision-making process is very difficult or even impossible to understand.

The resulting art installation featured some of the interesting, curious, and troubling findings that our research has uncovered about both facial-recognition technology and about the history of representation in the University of Michigan Museum of Art's collection of approximately 24,000 works.

We applied one of the most widely used facial detection algorithms to UMMA's art collection. After detecting faces in UMMA ' s artworks, we used a race classification algorithm to look at the diversity of subjects in the collection. We used the FairFace Dataset for examples of faces belonging to different races. We used these results to characterize and visualize the racial diversity of the acquisitions made under all of UMMA's directors.

We used a technique called "eigenfaces" to explore variation within faces found in UMMA's collection and to understand which features are most important in detecting a face.

By applying facial detection algorithms to UMMA's art collection, we visualize bias in the museum's collecting practices throughout its 150-year history. We can also see the ways algorithms amplify human bias. Our research makes more transparent the opaque decision-making processes within museum collection practices and machine learning algorithms as these rapidly evolving technologies are being deployed across the world.

Background

Museum Bias

Art museums have a long history of racial and gender bias. A recent study looking at 18 major US art museums found that 85% of its collected artists are white and 87% are men.[1] Who is depicted in these artworks is not only an issue of numbers, but bias is also evident in how people are depicted. Racialized caricature is one obvious example. Furthermore, museums have historically excluded certain groups of people from visiting museums in both overt and subtler ways.[2] Museums are now reckoning with how they may have reinforced prejudices in the past and what responsibility they have in confronting prejudice going forward.[3,4]

Algorithmic Bias

Face recognition algorithms are increasingly adopted for commercial use, for public safety, and in other applications. However, flaws in the current algorithms not only limit their effectiveness, but also have adverse consequences for certain demographic groups that are the "usual suspects" of being marginalized or victimized by new technology. The algorithms' significant flaws in race and gender recognition can be attributed partially to the lack of diversity in the training set—white male being the overrepresented face.[6] While researchers have repeatedly pointed out such flaws and are improving the training sets, there may be other limitations of the algorithms that have not been adequately addressed. For example, the algorithms rely on faces in the training set that are mostly photographs of full frontal view faces. How well do the algorithms work when the faces are sideways, partially visible, and so on? In our study, we did not aim to develop a new algorithm or improve a current one; instead, we focus on the use of a highly unconventional dataset (UMMA's art collection) to test the limit of existing algorithms trained on human photographs and understand what features are essential for the correct or incorrect face and race recognition.

Related Works

In 2018, Google Arts & Culture released the Art Selfie phone app as a playful way to discover art. The user takes a selfie and the app searches thousands of artworks to find one with a similar face.[5]

The 2020 film *Coded Bias* summarizes MIT Media Lab researcher Joy Buolamwini's research on how facial recognition algorithms do not see dark-skinned faces accurately and demonstrates the need for legislation to reduce bias in algorithms.[6]

UK Research and Innovation recently funded a project titled *Transforming Collections: Reimagining Art, Nation and Heritage* led by a team of researchers at the University of the Arts London. The project aims to "build on decolonial feminist approaches and creative machine learning (ML) development: to enable digital cross-search of collections to surface patterns of bias, and to uncover hidden and unexpected connections, and to thus open up new interpretative frames and potential narratives of art, nation and heritage."[7]

Process

1) We selected YOLOv4 as the main algorithm to test on the art collection. We also used a second algorithm, Dlib, to a more limited extent. The two algorithms both returned some successful face and race detections and some unsuccessful ones. We focus our paper on results with YOLOv4. [11,12]

Instead of using artworks (whether within UMMA's collection or elsewhere) to train the algorithm, we simply used pretrained weights. Our rationale for not using any art collection as the training set is that we do not have the resources to manually inspect and label the artworks that can be used for training, and that the size of such an arts training set could be prohibitively large given the much larger variation in faces in the artworks than in photographs. We did not evaluate the efficacy of a perfectly customized algorithm but rather mimicked the realistic practice of brittle deployments despite limited training data.

2) After identifying faces in UMMA's collection, we applied the VGG-Face CNN network with pretrained weights from FairFace Dataset to assign race to the faces in UMMA's art collection. [8,9,10]

3) We used a technique called "eigenfaces" to explore variation within faces found in UMMA's collection and to understand which features are most important in detecting a face.[13]

4) We created an exhibition to share results with museum visitors.

Algorithms and Data Sets

Face Detection

For face detection we used the algorithm YOLOv4 and we trained the artificial intelligence (AI) with the Google Open Images Database, which is comprised entirely of photographs.

Race Classification

After detecting faces in UMMA's artworks, we used a race classification algorithm, the VGG-Face CNN network, to look at the collection's diversity. We used the FairFace Dataset for examples of faces belonging to different races. FairFace was created to measure and mitigate racial bias. It contains 108,501 Flickr images of faces categorized as Asian, White, Middle Eastern, Indian, Latino-Hispanic, or Black. [8] , [9]

Overview of Results

Of the 21386 UMMA collection objects that we used with the algorithm, 6026 objects (28%) were classified as having at least one face. For race classification, we "forced" the algorithm to choose among the seven racial groups defined in the FairFaces dataset, but combined their definition of East Asian and Southeast Asian into one group, and obtained the following: White (69.1%), Black (10.5%), Indian (2.9%), Asian (10.7%), Middle Eastern (3.9%), and Latino (2.9%). The racial classification algorithm we used lacks a category for Native Americans.

Limitations and Failures

In addition to being unable to classify Native Americans at all, the algorithm had difficulty identifying faces in several cases: faces in profile; tilted heads; highly abstract faces; caricature. On the other hand, non-face objects with round/oval shapes and symmetric features were often classified as faces, such as many vases.

Acquisition Patterns by Different Museum Directors

We sought to understand the acquisition patterns of different UMMA directors, focusing on the predicted races of people depicted in the acquired works. Although we did conduct some benchmarking to establish the performance of the algorithms that we employed, an important caveat of this analysis is that it is based on race classification of detected faces in the

artworks by algorithms. It was not humanly possible to validate that all these predictions were correct. With this proviso, we constructed a contingency table showing the number of works depicting individuals of each race acquired by each director.

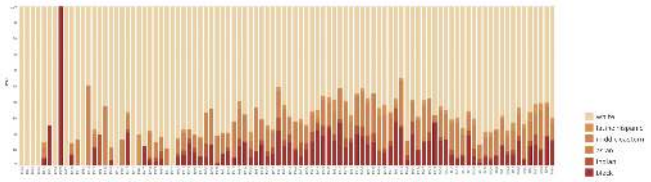

Figure 1. Racial diversity in UMMA's Acquisitions by year. The results from the race classification algorithm showed that UMMAs collection became more diverse over time. Each bar breaks down the racial makeup of that year's acquisitions.

To aid interpretation of this contingency table, we performed a standardization. Let N_ij denote the number of works acquired by director i that depict individuals of race j. Then, if N is the total number of works depicting any race acquired by any director, and p_i denotes the proportion of all works acquired by director i, and q_j denotes the proportion of all works depicting individuals of race j, then N*p_i*q_j is the reference point for N_ij. We can interpret N*p_i*q_j as the number of works acquired by director i depicting individuals of race j in the event that all directors purchased works depicting the races with the same frequencies. The residual R_ij = N_ij - N*p_i*q_j is the excess (if positive) or deficiency (if negative) of works depicting race j acquired by director i. The standardized residual S_ij = R_ij / sqrt(N*p_i*q_j) aims to place these residuals on a common scale that is fairly comparable between directors with small and large numbers of acquisitions, and between races with small and large overall representation in UMMA collection.

Conventionally, values of S_ij smaller in magnitude than 2 are viewed as unimportant. It is not easy to conclude definitively that a given large value of $|S_{ij}|$ is large enough to be important, but in many cases values exceeding 2.5 or 3 are likely to reflect a specific cause and not occur randomly due to variation of small numbers.

We noticed a significant uptick in the diversity of the collection in 2019. After taking a closer look at the race classification results from that year, we found that the trend seemed to be specifically tied to the *Take Your Pick* exhibition where museum visitors voted to select 250 everyday photographs to add to the collection, suggesting that a single exhibition can have a notable impact on the overall diversity in the collection. Though imperfect, we found that the algorithms generated results (such as the uptick in diversity in 2019) that offered new perspectives and points of entry for further manual investigation into smaller, manageable subsets of the collection.

Eigenfaces

We used "eigenfaces" to explore variation within faces found in UMMA's collection and to understand which features are most important in detecting a face. Eigenfaces represent axes of variability in a collection of images of faces. This technique was first developed in the 1990's. In prior work, the eigenfaces have been found to capture factors such as lighting, pose, the presence of eyeglasses and beards, and anthropometric features such as dimensions of the jaw, nose, forehead, and spacing between the eyes. Eigenfaces can be used to understand the principal ways that faces in a collection vary, and can also be used as a data compression technique, in that they represent a "high dimensional" face using a relatively low-dimensional vector of "scores.".

Eigenfaces show the most important ways that individual faces differ from the mean face. Each eigenface corresponds to a spectrum along which variation occurs. Each eigenface below represents one feature important in detecting faces in our collection.

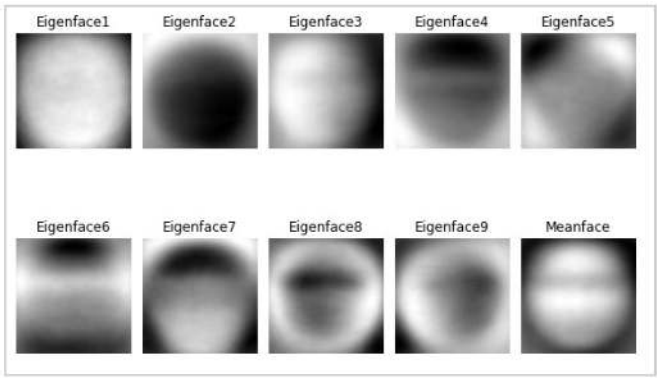

Figure 2. Each of these eigenfaces represents one feature important in detecting faces in our collection.

The blurriness shows how that feature varies in the faces found in UMMA's collection. For example, eigenface 3 corresponds to a spectrum along which the face is lit from directions varying from the left to the right. Eigenfaces 1 and 2 correspond to variation in the overall size and shading of the face. Eigenfaces 6, 7, and 8 correspond to different patterns of shading at the top of the head, forehead, and around the eyes.

To begin, we first try to limit the extraneous variation by scaling and cropping each face in our collection to approximately the same position in a fixed-size image (# 224*224*3 # pixels). We then use a mathematical technique called the "singular value decomposition" to identify the eigenfaces. Specifically, an eigenface is a pattern represented by signed (positive and negative) weights. Each eigenface assigns one weight to each pixel location in the images. These weights represent a common pattern of deviation from the mean face. We note that the mean face itself generally appears "ghost-like" and does not resemble a human face, but the deviations from this mean face are informative about the unique characteristics of an individual face.

Since there is one weight for each pixel, the eigenfaces can be visualized in the same way that the face images are themselves visualized. For example, an eigenface corresponding to illumination on the left side may be bright (positive) on the left side of the image and dark (negative) on the right side of the image; an eigenface corresponding to spacing between the eyes may have alternating bright and dark regions of weights in a band located at the level of the subjects' eyes.

An eigenface represents a spectrum of variation. For example, illumination from the left is part of the same spectrum as illumination from the right, and hence this variation can be represented by one eigenface; similarly, wider-than-average eye spacing may be part of the same spectrum as narrower-than-average eye spacing. Since a spectrum has no defined beginning or end, each eigenface is equivalent to its additive inverse, i.e., F and -F represent the same eigenface, with the spectrum of variation represented by -F being the same as the spectrum of variation represented by F, traversed in the opposite direction.

As noted above, the eigenface technique has often been used with collections of highly standardized images, like passport photos. Even in such a standardized collection, the eigenface technique is generally found to be influenced by lighting and pose as much or more than it is influenced by anthropometry, which is a drawback to the approach. Moreover, while some level of dimension reduction is achieved, it often is necessary to use 100-200 eigenfaces to represent most of the variation in a collection of faces. Using the eigenface technique on UMMA collection is even more susceptible to this issue, since artists represent humans in every possible pose, and it is not possible to standardize these faces beyond simple translation and scaling.

Using the eigenfaces, we identified a painting of a clown, with makeup caricaturing a face, as having the most representative face in UMMA's collection.

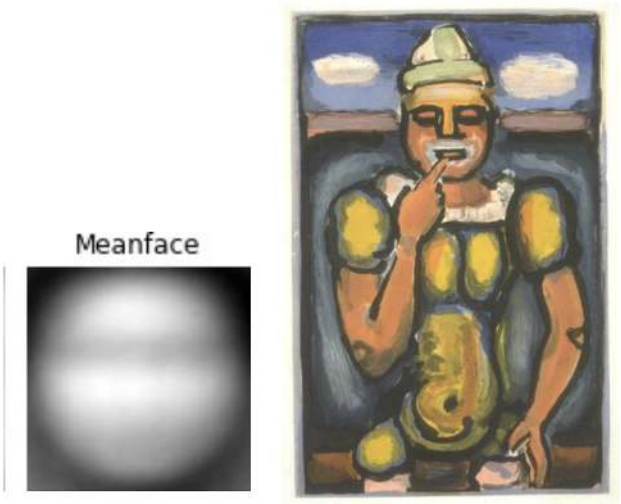

Figure 3. The mean eigenface and the clown painting that the computer identified as having the most representative face in the museum's collection.

Exhibition

Using data from our research, we created two video explainers that explain parallels between biases in art museum systems and in algorithm systems. They consisted of data visualizations that highlighted race representation in UMMA collection over time, influence of certain exhibits, and notable research findings and challenges. The videos and select paintings were exhibited in UMMA to self-reflect and critique the museum's past in full transparency.

We displayed these two videos along with actual paintings from the collection as part of the *You Are Here* exhibition, which invited museum visitors to consider where they *are* and where they *aren't.* Above our videos, we displayed the question, "Are you here?" inviting the viewers to consider how they are represented within the museum's collection and exhibitions.

Video Documentation

https://vimeo.com/641632433/942e533cf2

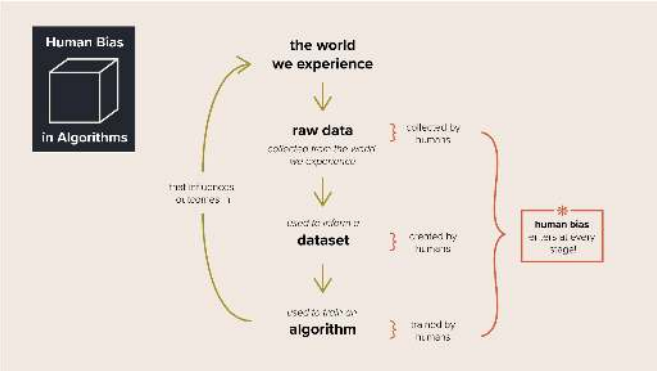

Figure 4. Video still explaining how human bias plays a role in algorithms.

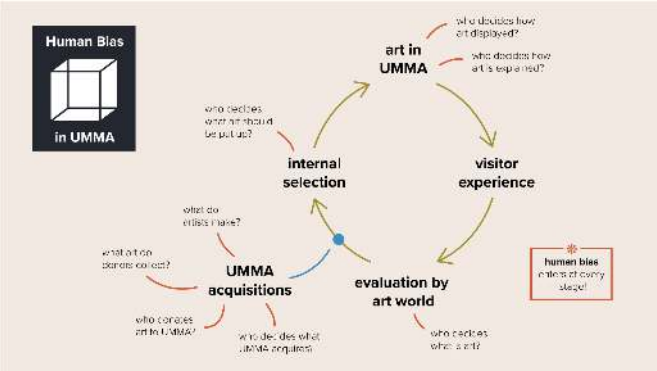

Figure 5. Video still explaining how human bias plays a role in UMMA.

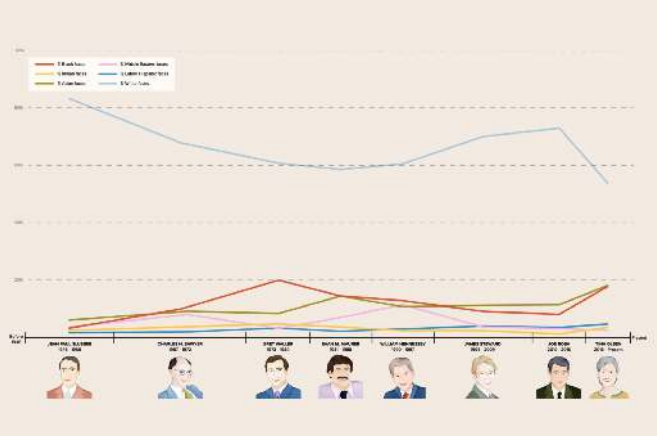

Figure 6. Racial diversity in acquisitions by museum directors over time

Future Work

We are currently planning a second exhibition scheduled for Fall 2024 that will invite UMMA visitors to observe and evaluate algorithmic facial detection. We plan to visualize the data by exhibiting actual artworks sorted by the algorithm's confidence in recognizing faces within. Additionally, we plan an interactive wall projection that will contrast the algorithm's confidence in recognizing faces with confidence judgments submitted live by museum visitors.

Conclusion

An interdisciplinary team of artists, designers, data scientists, and curators applied face detection and race classification algorithms to UMMA's collection of approximately 24,000 artworks that were collected over 150 years.

When we began this project, we asked, "How can the application of machine learning expose or amplify human bias?" We wanted to know if our AI could reveal the bias in artists, collectors, donors, curators, and society in general over time. What biases did our AI learn when it was trained on datasets of example faces? What biases are embedded in the algorithm itself?

In addition to learning about and visualizing how the diversity of UMMA's collection changed over time (for the better), we experienced these artworks through the lens of machine learning for the first time. We were not navigating the collection through the usual categories like who created the artwork, artistic movement, artistic medium, date created, or the artwork's origin. We were not experiencing these works as part of a curated exhibition. We were encountering the artworks in buckets such as "91-100% confident it's a face" and "non-face" or through simplistic labels like "has face, White" and "has face, Indian". Going through this process changed our own perception and sensitivity to certain aesthetics as we wondered why the AI made certain decisions. In some cases, the AI's decisions caused us to question our own understanding of certain artworks.

In addition to exploring biases within both algorithms and museums, this research invites museums and museum goers to reflect on ideas of transparency, self-reflection, and critical thinking about collecting and curatorial practices. How does our understanding of art, curation, and history change when artworks are algorithmically curated?

References

1 Topaz CM, Klingenberg B, Turek D, Heggeseth B, Harris PE, Blackwood JC, Chavoya CO, Nelson S, Murphy KM. Diversity of artists in major U.S. museums. PLoS One. 2019 Mar 20;14(3):e0212852. doi: 10.1371/journal.pone.0212852. PMID: 30893328; PMCID: PMC6426178.

2 Olivares, A., Piatak, J. Exhibiting Inclusion: An Examination of Race, Ethnicity, and Museum Participation. *Voluntas* 33, 121–133 (2022). https://doi.org/10.1007/s11266-021-00322-0

3 Sandell, R. (2006). Museums, Prejudice and the Reframing of Difference (1st ed.). Routledge. https://doi.org/10.4324/9780203020036

4 Shirley Li, "American Museums Are Going Through an Identity Crisis," *The Atlantic,* November 28, 2020, accessed Dec 1, 2022, https://www.theatlantic.com/culture/archive/2020/11/american-museums-are-going-through-identity-crisis/617221/

5 Google Arts & Culture, "Art Selfie," accessed Dec 1, 2022, https://artsandculture.google.com/camera/selfie

6 Joy Buolamwini, *Coded Bias*, accessed Dec 1, 2022, https://www.codedbias.com/about

7 "Transforming Collections: Reimagining Art, Nation and Heritage", accessed Dec 1, 2022, https://www.arts.ac.uk/ual-decolonising-arts-institute/projects/transforming-collections

8 Karkkainen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1548-1558).

9 "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age," accessed Dec 1, 2022, https://github.com/dchen236/FairFace

10 "VGG Face Descripto," accessed Dec 1, 2022, https://www.robots.ox.ac.uk/~vgg/software/vgg_face/

11 "Dlib," accessed Dec 1, 2022, http://dlib.net/

12 Bochkovskiy, Alexey, Chien-Yao Wang and Hong-Yuan Mark Liao. "YOLOv4: Optimal Speed and Accuracy of Object Detection." *ArXiv* abs/2004.10934 (2020): n. Pag.

13 Matthew Turk, Alex Pentland; Eigenfaces for Recognition. *Journal of Cognitive Neuroscience,* 1991; 3 (1): 71–86. doi: https://doi.org/10.1162/jocn.1991.3.1.71