Measuring Representation of Race, Gender, and Age in Children's Books: Face Detection and Feature Classification in Illustrated Images

Teodora Szasz
University of Chicago
tszasz@uchicago.edu

Emileigh Harrison
University of Chicago
harrisone@uchicago.edu

Ping-Jung Liu
University of Chicago
pliu5@uchicago.edu

Ping-Chang Lin
University of Chicago
pcclin@uchicago.edu

Hakizumwami Birali Runesha
University of Chicago
runesha@uchicago.edu

Anjali Adukia
University of Chicago and NBER
adukia@uchicago.edu

Abstract

Images in children's books convey messages about society and the roles that people play in it. Understanding these messages requires systematic measurement of who is represented. Computer vision face detection tools can provide such measurements; however, state-of-the-art face detection models were trained with photographs, and 80% of images in children's books are illustrated; thus existing methods both misclassify and miss classifying many faces. In this paper, we introduce a new approach to analyze images using AI tools, resulting in data that can assess representation of race, gender, and age in both illustrations and photographs in children's books. We make four primary contributions to the fields of deep learning and social sciences: (1) We curate an original face detection data set (IllusFace 1.0) by manually labeling 5,403 illustrated faces with bounding boxes. (2) We train two AutoML-based face detection models for illustrations: (i) using IllusFace 1.0 (FDAI); (ii) using iCartoon, a publicly available data set (FDAI_iC), each optimized for illustrated images, detecting 2.5 times more faces in our testing data than the established face detector using Google Vision (FDGV). (3) We curate a data set of the race, gender, and age of 980 faces manually labeled by three different raters (CBFeatures 1.0). (4) We train an AutoML feature classification model (FCA) using CBFeatures 1.0. We compare FCA with the performance of another AutoML model that we trained on UTKFace, a public data set (FCA_UTK) and of an established model using FairFace (FCF). Finally, we examine distributions of character identities over the last century across the models. We find that FCA is 34% more accurate than FCF in its race predictions. These contributions provide tools to educators, caregivers, and curriculum developers to assess the representation contained in children's content.

1. Introduction

The images in the books presented to children not only teach skills but also more broadly impart understanding of how people fit into the world around them. The inclusion and exclusion of characters of different identities in the images of books transmit implicit and explicit messages about children's roles in society and their potential, in addition to that of others different than them. The messages sent from images are particularly salient to children from the youngest of ages, especially before they are able to read words. However, we have little systematic understanding about the actual representation of different identities in images to which children are exposed. Content analysis of children's books has traditionally been conducted manually, involving human parsing of content [7, 25, 17]. Such methods are time-consuming and limited to small sample sizes, thus neglecting to provide a broader understanding of exposure to representation of identities such as race and gender.

Per the adage "a picture is worth a thousand words," images convey numerous messages. Automating the measurement of their content can provide previously unquantifiable information about the messages implicitly and explicitly being sent through these visual depictions. In particular, with the emergence of convolutional neural networks (CNNs) in recent years, there have been meaningful advances in research and application of object detection and feature extraction. Most established state-of-the-art face detection prediction models are trained using photographs to ensure their robustness in the real world [27]. Because over 80 percent of images in children's books are illustrated, it is important that the models incorporate such images in their training. However, the few existing face detection data sets based on artwork and cartoons do not capture the diversity of different drawing styles used in illustrations for children's books, which often contain anthropomorphic fea-

January 2022, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

tures that are similar to real-world human faces (e.g., caricatures, cartoons, comic books). Indeed, because these established face detection methods do not capture the large diversity of drawing styles, they resulted in high numbers of both false positives and false negatives when we applied them to children's books [39].

In this paper, we design and implement an original method to detect illustrated faces and classify their features in order to understand the representation of gender, age, and race in children's books. Our key contributions include:

1. We curate a new data set of 5,403 faces from illustrations which captured a broad range of drawing styles. We name this data set IllusFace 1.0.

2. We develop a face detection model using AutoML-based approach to detect faces in illustrations from children's books, using the IllusFace 1.0 data set. We name this method Face Detection using AutoML for Illustrations (FDAI). We also train a face detection model using an AutoML-based approach using the iCartoon data set [39]. We name this method Face Detection using AutoML for Illustrations (FDAI_iC) and compare it to FDAI.

3. We curate a novel data set of the race, gender, and age including an "unsure" label of 980 manually labeled faces from children's books (CBFeatures 1.0), separately labeled by three different raters.

4. We develop a feature classification model using AutoML-based approach for classifying gender, age, and race from faces in illustrations, using CBFeatures 1.0, our curated data set of labeled faces. We name this method Feature Classification using AutoML (FCA). We also train a model based on AutoML, using the UTKFace data set [38]. We name this method Feature Classification using AutoML using UTKFace data set (FCA_UTK) and we compare it with FCA. We then compare predictions of gender, age, and race of characters in our trained FCA and FCA_UTK models to a more established model (named Feature Classification using FairFace (FCF)), and evaluate their accuracy using a sample of manually labeled features.

The paper is organized as follows. In Section 2, we describe prior work related to face detection and feature prediction in images. In Sections 3 and 4, we present our developed methods for detecting faces and classifying features to measure representation of gender, age, and race in images. In Section 5, we summarize the results of our analysis, and in Section 6, we conclude. For the remainder of this paper, we define (1) a photograph (or photo) as an image of an existing object or person, captured using a camera, and (2) an illustration as a drawing generated by an artist or machine.

2. Related Work

The problem of inequality and unjust representation of people has endured through history and has become more salient in recent years. [9] amplified and raised issues related to unfairness in the field of computer vision to the public's attention. The authors investigated unbalanced representations of skin color and gender in state-of-the-art face detection methods and further introduced a balanced data set with photos. In our work, we aim to expand the effort of advocating for equality in technology to analyzing the different representations in children's books.

2.1. Face Detection

The task of detecting human faces in content has been a well-studied field in computer vision. Scale Invariant Feature Transform (SIFT) [21] and histogram of oriented gradients (HoG) [10] were two influential object detection methods using traditional machine learning. While they have the advantage of explainability and easy-to-understand model structures, CNN-based methods surpassed them in performance metrics such as precision and recall, particularly due to the emergence of large data sets [32]. [34] conducted a comprehensive survey of existing methods for face detection and find that most state-of-the-art methods are based on different CNN structures typically trained on data sets that are heavily comprised of photographs [37]. For example, FaceNet, a high-performing end-to-end face-detection model, was built using CNNs and trained with traditional back propagation on 260 million training faces [28]. FaceNet was evaluated using several large data sets of photographs and considered to be a premier product at the time of its inception in 2015. It was an early open-source version of FDGV. Although they can achieve performance of over 99% in terms of area-under-the-curve (AUC, or AuPRC) receiver operating characteristics (ROC), these methods fail to detect faces in illustrations.

Detection of illustrated faces has been a widely discussed problem, and algorithms dealing with cartoon face detection can be applied to address this issue. For example, [6] created a data set with face annotations of characters in 109 Japanese comic books. [35] used artwork sampled from a data set from Painter By Numbers (PBN) using facial landmark detection and style transfer techniques to transform photographs of faces into artistic images using a particular style (e.g., that of Picasso). They used a MultiTask Cascaded Convolutional Network (MTCCN) to detect illustrated faces, but it only detected 75% of faces in artwork.

[39] introduced iCartoonFace, a data set of illustrations that can be used for face detection and recognition. Their training set consists of 50,000 images with 91,163 faces. Trained using a RetinaNet architecture, this model achieved 89% mAP (mean Average Precision). [36] proposed an asymmetric cartoon face detector (ACDF), based

on a VoVNetV3 network trained using iCartoon, obtaining a mAP of 88.9% at IoU of 0.35. In [26], a Faster R-CNN architecture was built for face detection of comic characters using a data set with 3375 comic faces, obtaining precision of 75.2% and recall of 49.8%.

In this paper, we use the iCartoon data set to train an Auto-ML based model for face detection (FDAI_iC) and compare the results with the AutoML-based model trained on our face annotated data set IllusFace 1.0 (FCAI).

2.2. Feature Classification of Faces

Deep neural network structures capture subtle features that cannot be easily observed by human eyes (e.g. emotions) [23, 24]. Feature classification tasks often use transfer learning through CNNs [31], because it decreases the cost of training new models by drawing on existing networks that have already been trained on millions of labeled images (e.g. ImageNet data set [11]). [18], [33], [30], and [12] demonstrated the potency of training transfer learning models for predicting the gender, age, and race of humans detected in photographs. Because pre-trained VGG architectures [29] can have over 90 percent accuracy, we trained our own transfer learning models for prediction purposes.

3. Face Detection in Illustrations Methods

3.1. Data Source: Children's Books

For developing our tools, we draw from children's books commonly found in school libraries and classrooms, and thus, likely to have been presented to children. Specifically, we chose books that received awards administered by the Association for Library Service to Children, a division of the American Library Association. These books include those selected for the oldest children's book awards in the US – Newbery and Caldecott awards – in addition to other awards selected because they highlight the experiences of people from historically underrepresented identity groups, for example, race or gender, as used in [5].[1] These other awards include: American Indian Youth Literature, Américas, Arab American, Asian/Pacific American for Literature, Carter G. Woodson, Coretta Scott King, Dolly Gray, Ezra Jack Keats, Middle East, Notable Books for a Global Society, Pura Belpré, Rise Feminist, Schneider Family, Skipping Stones, South Asia, Stonewall, and Tomas Rivera Mexican American Awards. These are books often placed on "diversity lists" such as during Black History Month or Women's History Month. Our corpora includes 1,130 books and 162,872 pages of content, published over the last century between 1923 and 2019.

Figure 1: Examples of illustrations depicted in children's books and their complexity

The number of pictures (e.g., illustrations, photographs) used in children's books has increased over time. While many image analysis tools exist for photographs, their performance is limited when applied to illustrations. Many images from our database contain distinct artists' influence, non-human characters, unusual color schemes, and different body poses. Examples of such illustrations are shown in Figure 1. As Edgar Degas said, "Art is not just what you see but what you make others see." The drawings of characters naturally reflect the biases of the illustrators or authors of books, potentially moreso than photographs because the artist necessarily dictates every aspect of a drawing. This motivates our curation of an original data set composed of illustrated faces, for the purpose of training a model that is more likely to detect faces in illustrations.

3.2. Labeled Data Sets

We created an original face detection data set consisting of 1,963 scanned pages comprising 5,403 labeled faces selected from children's books with a wide variety of color schemes, character poses, and contexts. We name this data set IllusFace 1.0. While labeling, we tried to took the perspective of an average reader (in this case, a child) and maintained the mindset that if humans cannot easily observe a specific face, then the network should not be able to do so as well, and then in such cases we would discard certain indistinct faces or head poses. We also intentionally include some personified animal characters so the models can learn general characteristics of faces without relying on standard human features and skin colors.[2] The illustrations were selected from Newbery and Caldecott books because of their variety in characters.

In addition, we created a smaller data set which we name CBFeatures 1.0 containing manually annotated gender, age, and race of 980 faces from children's books using the following categories:[3] (1) gender: female, male, unsure; (2) age: infant, child, teenager, adult, senior, unsure;[4] and (3) race: Asian, Black, Latinx, White, other, and unsure.[5] The addition of the "unsure" label is an improvement beyond

[1] These books are primarily written in English and reflect US culture, but they are frequently translated into many different languages due to their ubiquitous nature in children's literature in US schools and libraries.

[2] Children's books feature characters – both human and non-human – with a wide variety of colors: monochromatic (e.g., black and white), non-typical (e.g., blue or green), and human skin colors.

[3] 78% of these faces are illustrated.

[4] We also analyze age by aggregating these smaller age categories to two larger age categories: child and adult.

[5] Our Asian label is a combination of two labels: Asian and Indian.

existing data sets, which do not include this uncertainty. Due to the subjectivity of facial features found in illustrated faces, we had three different research assistants manually label each face. We assigned each face its modal label.

3.3. Face Detection: Google AutoML Vision in Illustrations (FDAI, FDAI_iC)

We develop two face detection methods using Google AutoML Vision: one we trained on our IllusFace 1.0 data set (FDAI) and the other on the iCartoon data set (FDAI_iC). AutoML is one of the most prominent developments in deep learning in recent years [4, 14]. It automates the complicated process of designing network structures and fine-tuning hyperparameters for specific tasks. Google's AutoML Vision [2] platform applies transfer learning to first extract significant features from input images and utilizes its undisclosed Neural Architecture Search (NAS) [40] to search for the most ideal neural network structure. While the resulting models are undisclosed and forbid further optimization, the platform allows researchers to efficiently train models with minimal time spent on searching for network structures and hyperparameters. To use the trained model for prediction, we deployed the model in Google Cloud and used Python-based representational state transfer (REST) APIs to access the face detection model.

We use the precision, recall, and area under precision/recall curve (AuPRC) metrics to measure the performance of the two models. The higher the precision, the fewer false positives (FP) the model produces. On the other hand, the higher the recall, the fewer false negatives (FN) the model produces. The formula for precision is TP/TP + FP and for recall is TP/TP + FN. TP stands for true positives. As the number of false positive goes up, the denominator of the precision equation increases, resulting in lower precision. The same goes for the recall equation; when the number of false negative goes up, recall drops. AuPRC measures the tradeoff between precision and recall across different decision thresholds [8]. We set the Intersection over Union (IoU) threshold (a measure of the overlap between the predicted and actual bounding box around the detected face) to 0.5.

3.4. Face Detection: Google Vision (FDGV)

We compare the results of our FDAI model to an industry standard: Google Vision API for detecting faces. We named this method Face Detection using Google Vision (FDGV). Multiple platforms offer face detection application programming interface (API) services [3, 1]. [16] compared the performance of detecting faces between Microsoft Azure API (MZAPI) and Google FDGV and concluded that FDGV performs better in detecting faces compared to MZAPI. Thus, we chose to use FDGV as a benchmark comparison model to our FDAI model. FDGV provides access to pre-trained machine learning models through REST and remote procedure call (RPC) API requests.[6] The method for detecting faces inside FDGV offers the capability of detecting multiple faces within a given image together with features for each detected face (e.g., eye location, emotions).

4. Feature Prediction: Gender, Age, Race

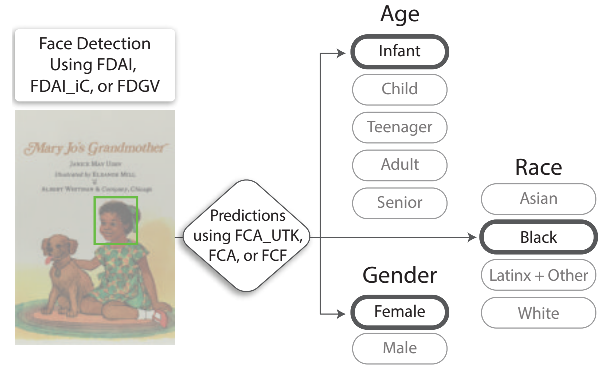


Figure 2: Pipeline for classifying gender, age, and race using FDAI, FDAI_iC, FCA, FCA_UTK, and FCF methods

In this section, we discuss how we classify gender, age, and race of detected faces in images of illustrations. There are three feature classification models that we use. (1) We use our manually labeled data set (CBFeatures 1.0) containing gender, age, and race labels to train a Google AutoML Vision model, which we call "Feature Classification using AutoML," or FCA; (2) We use the UTKFace [38] data set to train a model using Google AutoML Vision, which we call "Feature Classification using AutoML with UTKFace data set," or FCA_UTK; and (3) we use the FairFace data set [19] trained on a ResNet-34 architecture [15], to which we refer as "Feature Classification using FairFace data set," or FCF. While both the UTKFace and FairFace data sets have been widely applied in different research studies [13, 22], they are solely based on photographs. To the best of our knowledge, there is no related work that uses these data sets and methods for predicting features using faces in illustrations. We apply FCA, FCA_UTK, and FCF on the faces detected from the FDAI method. Figure 2 illustrates the entire pipeline for applying FDAI, FDAI_iC, FDGV, FCA, FCA_UTK and FCF on our data.

4.1. Feature Classification: Google AutoML Vision in Illustrations

We perform feature classification by training a multi-label classification model using Google AutoML Vision on

[6] Other features available through Google Vision API include optical character recognition (OCR), label detection, multiple object detection, face detection, emotion recognition, and others.

our manually labeled feature data, which we call FCA. We compare this method to the results of a Google AutoML Vision model on UTKFace data set, which we call FCA_UTK.

To train the FCA model, we use our CBFeatures 1.0 data set which contains manually assigned age, gender, and race labels for a sample of faces detected in our children's books using FDAI. The faces contained in CBFeatures 1.0 were not balanced on gender, age, and race. For gender, 392 faces were labeled as Female, 454 Male, and 134 Unsure. For race, 197 images were labeled Black, 86 Latinx, 84 Asian, 16 Indian, 377 White, 59 Other, and 161 Unsure. For age, 16 Infant, 261 Child, 87 Teenager, 448 Adult, 52 Senior, and 116 Unsure. 700 images were used for training and 280 for validation and testing.[7]

To train the FCA_UTK model, we used the UTKFace data set, which contains over 20,000 photographs of faces with manually verified gender, age, and race labels.[8] We selected a total of 10,000 images from the data set to ensure a balanced data set among labels. We extracted 1,000 images for validation (500 images) and testing (500 images).

For both our labeled data and UTKFace, we then used Google AutoML Vision's multi-label classification framework to train our data. The resulting model predicts a confidence score for each individual label. We then assign a label based on the highest predicted score within a given feature.

4.2. Feature Classification: FairFace Data (FCF)

We use another feature classification model to help benchmark the performance of our new FCA model: a Feature Classification Model using the FairFace data set (FCF). FairFace is a data set containing 108,501 labeled images that was trained on a ResNet architectures. The creators of this data set made deliberate attempts to balance the data set on gender, age, and race to minimize potential for bias that exists in more popular face data sets. We adopted a pre-trained ResNet-34 model and reshaped its original race categories (White, Black, Indian, East Asian, Southeast Asian, Middle East, and Latino) into the four race categories that we use in FCA to enable comparison: Asian, Black, Latinx + Others, and White. Similarly, we also regrouped the age categories into the five categories of infant, child, teenager, adult, and senior to allow for comparison with FCA, while keeping the gender classification as female or male.

Method	Number of Detected Faces
FDAI	22,198
FDAI_iC	17,657
FDGV	2,957

Table 1: Total number of detected faces in Newbery and Caldecott books, by method.

5. Results

In this section, we show our main results. We first describe the performance of each face detection model. We then show the results of each feature classification method.

5.1. Face Detection in Illustrations

We first compare the three face detection models (FDAI, FDAI_iC, FDGV). Evaluated on testing data of illustrations, the resulting FDAI model has 93.4% precision, 76.8% recall, and AuPRC of 84% at IoU of 0.5. The slightly low recall of this model indicates that there are many illustrated faces that are still difficult to recognize, even for human eyes, after training on our data set. FDAI_iC resulted in an AuPRC of 85%, with 87.8% precision and 78.35% recall at IoU of 0.5. When tested on the children's book data, we discovered a higher number of false positives using FDAI_iC compared to FDAI.

We present the comparison of each face detection model in Table 1. Using the FDAI method, we detect 5.5 times more faces compared to FDGV and 1.25 times more faces compared to FDAI_iC. We observe that FDGV is more likely to fail to detect faces in children's books, potentially because it was trained on photographs but is being applied to illustrations. Even if FDAI_iC shows a great improvement over FDGV (since it was trained on cartoon images), it greatly exceeds the number of false positives compared to FDAI. In Figure 3, we observe the three methods' performance on the Caldecott corpora, which demonstrates the significant advantage of FDAI on illustrated face detection. This compels us to use FDAI to extract faces for our subsequent feature prediction tasks.

Figure 4 shows the distribution of faces detected per decade in the Newbery and Caldecott award winning books using FDAI, FDAI_iC, and FDGV. We see that the number of faces is increasing over time, which may be due to increased inclusion of images in children's books over time. FDAI outperforms both FDAI_iC and FDGV in terms of faces detected in these books.

5.2. Feature Prediction: Gender, Age, and Race

In this section, we show results from classifying gender, age, and race in all corpora. We compare results obtained using FCA, FCA_UTK, and FCF. Note that these results are based on the assumption that gender, age, and race can be

[7] Google Vision AutoML requires a training data set to have a minimum of 10 images for each label. If a particular label was not represented in at least 10 images, we supplemented our CBFeatures 1.0 data set by adding a few hand-selected faces (detected in our children's books using FDAI) with the necessary labels.

[8] The labels in the data set include: gender (female or male), age (infant (0-3), child (4-11), teenager (12-19), adult (20-64), senior (65+)), race (Asian (a combination of the Asian and Indian labels), Black, White, and Latinx + Others (such as Middle Eastern)).

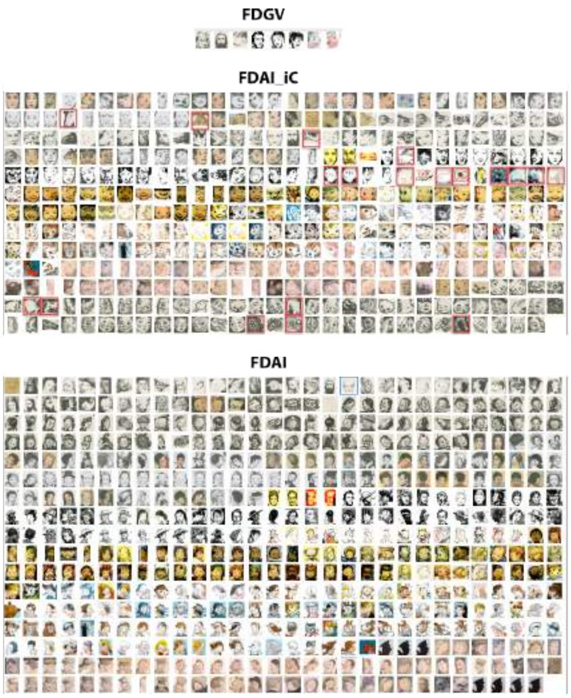


Figure 3: Face detection model comparison using a sample of books from the Caldecott corpus. In this sample, FDGV detects 8 faces, FDAI_iC detects 389 faces, while FDAI detects 510 faces. We demarcated in red the false positives.

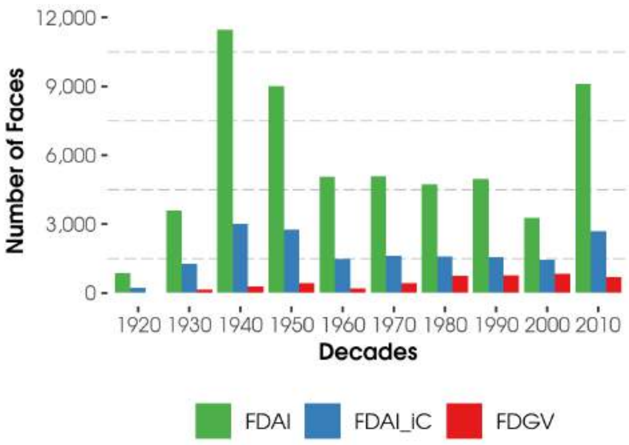


Figure 4: Total number of detected faces in all award books by decade.

predicted solely based off of one image of a face.

When tested on CBFeatures 1.0 data set, FCA model has precision of 69.42%, recall of 44.42%, and the AuPRC of 63%. The low accuracy and recall are due to the small number of faces for each of the gender, age, and race categories and the imbalance of the data set. When tested on UTKFace data set, the FCA_UTK model has 90.64% precision, 88.98% recall, and AuPRC of 95%. Similarly, when tested on FairFace data, FCF has an accuracy of about 94% [15].

We compare each models' predictions to the manual labels of the 280 faces from our CBFeatures 1.0 data set preserved for testing. Table 2 presents the accuracy for each of our models calculated using this manually labeled sample. Note that manual labeling reflects human biases and can result in subjective inconsistency among the manual raters. Additionally, many illustrated characters do not have clearly categorized identities, making it more difficult to label their identities and interpret the accuracy of our models when tested on our corpora that predominantly includes faces in illustrations. [9] To estimate the extent of agreement about the "ground truth" among the manual coders, we use the full CBFeatures 1.0 to calculate the inter-labeler reliability of each feature which measures the degree to which the three manual labelers were in agreement over the latent features belonging to each detected face.[10] We show the inter-labeler reliability calculations in Table 2. The range of these calculations goes from 0-1 where 0 means complete disagreement and 1 means perfect agreement between labelers.

Table 2: Model Predictions vs. Manual Labels

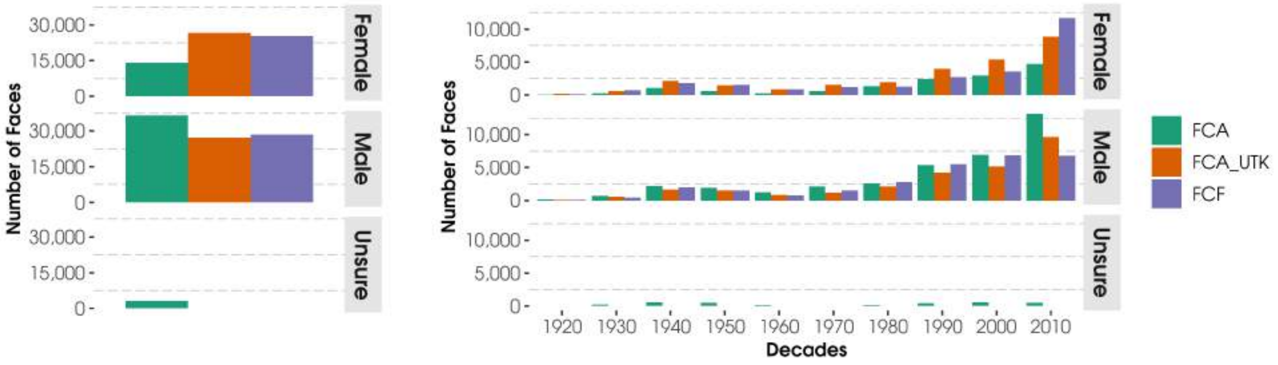
	Model Accuracy			Inter-Labeler
	FCA	FCA_UTK	FCF	Reliability
Age (2)	71.8%	66.1%	73.4%	0.66
Age (5)	59.3%	29.4%	56.5%	0.52
Gender	53.6%	65.2%	66.8%	0.75
Race	54.3%	44.4%	40.5%	0.60

Figures 5a and 5b show the results of gender distribution overall and per decade using FCA, FCA_UTK, and FCF. We see the distribution of female faces and male faces differs between feature classification methods. The number of female and male predictions per decade is comparable when using FCA and FCA_UTK. However, male prediction seems to be mostly predominant in FCF across decades, except for 1930, 1940, 1960, and 2010. This observation is consistent with Figure 5a, where we show the total distribution of female and male predictions. We can see that FCF and FCA_UTK predict more female labels than FCA.

Figure 6a shows the total distribution of the age prediction results. We observe that age prediction results using FCA_UTK appear biased towards seniors. When looking at the results obtained using FCF_UTK across decades (Figure 6b), senior representations consistently surpass that of the younger ages, while teenagers are the least repre-

[9] Moreover, gender labels are limited to binary classifications (female and male), disregarding potential non-binary or gender-fluid identities.

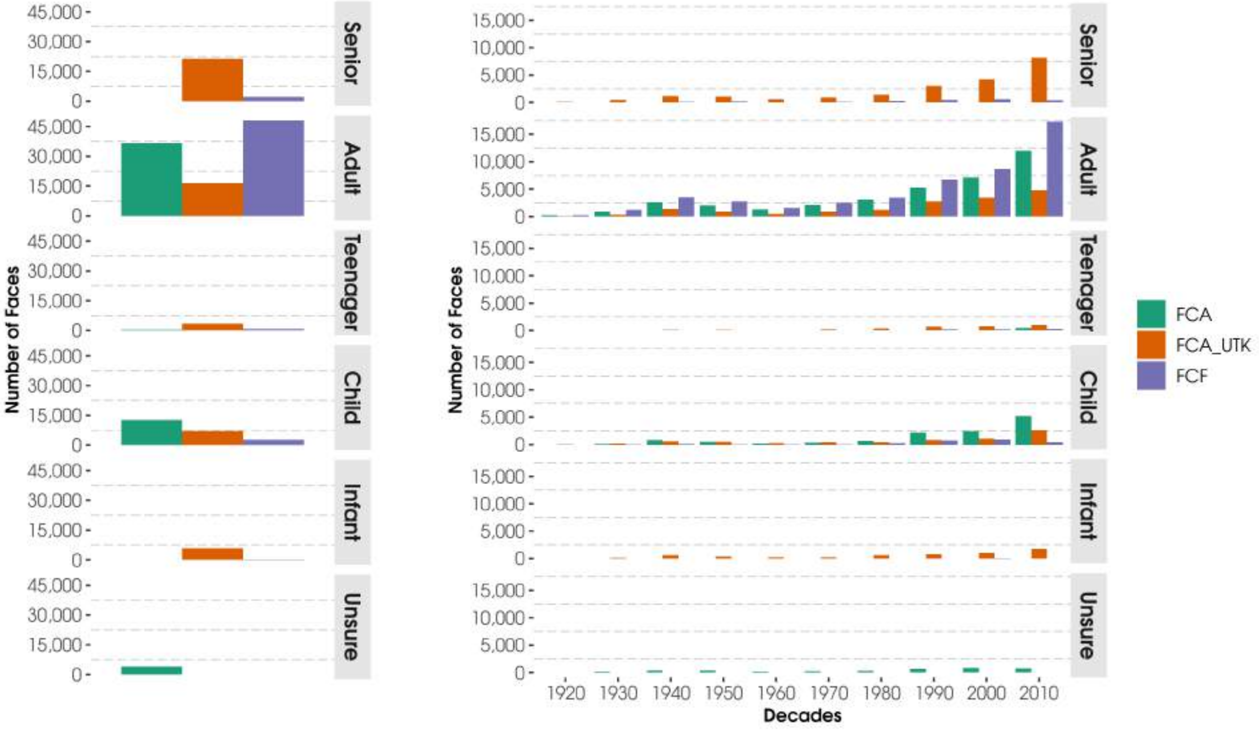
[10] Inter-labeler reliability was calculated using the Fleiss' kappa measure commonly used to find the agreement between categorical labels given by a fixed number of labelers.



(a) Overall

(b) Over Time

Figure 5: Panel (a) shows the total number of female (top) and male (middle) faces along with faces whose gender is unsure (bottom) detected via FDAI and classified using FCA, FCA_UTK, and FCF. Panel (b) shows these estimates by decade.



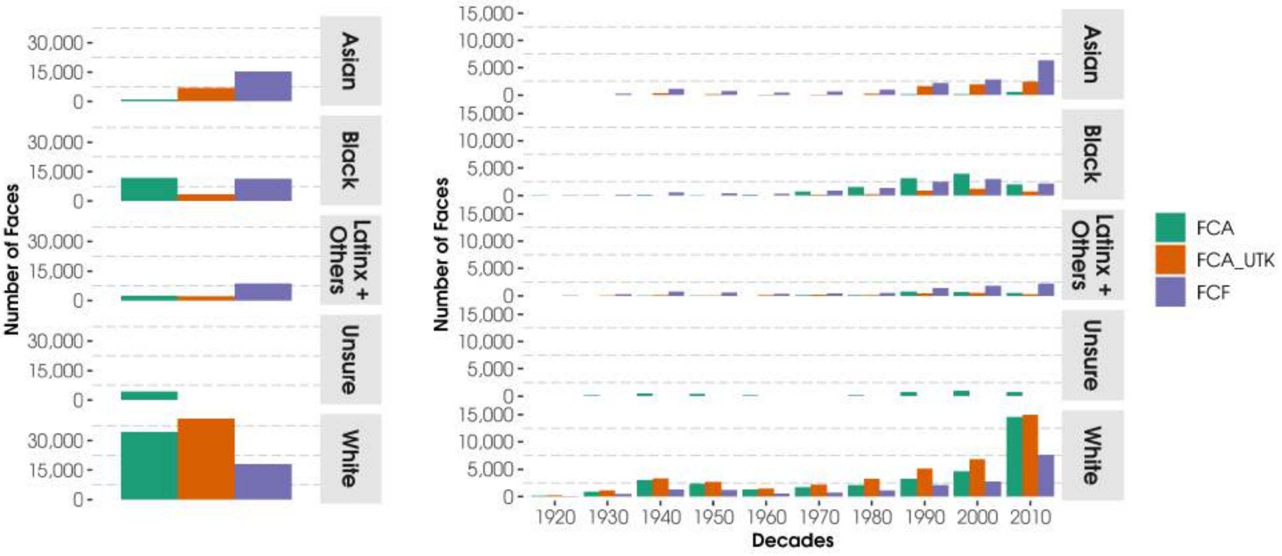
(a) Overall

(b) Over Time

Figure 6: Panel (a) shows the total number of faces detected via FDAI and classified from each age group using FCA, FCA_UTK, and FCF. Panel (b) shows these estimates by decade.

sented category over time. When we compare these results to our manually labeled data (which we consider our "ground truth"), FCA_UTK provided an accuracy of only 29%, while FCA and FCF resulted in accuracies of 59% and 57% respectively. While none of the models performed well with five age labels, when we regroup the age labels into two larger groups (children and adults), accuracy substantially improved. The accuracy of FCA_UTK more than doubled at 66%. FCA and FCF performed even better with 72% and 73% accuracies, respectively.

We show in Table 2 that the FCA model is 34% more accurate in race classifications than the established FCF model. Figures 7a and 7b show the race prediction results. Results from each model show that the number of faces



(a) Overall

(b) Over Time

Figure 7: Panel (a) shows the total number of faces detected via FDAI and classified from each racial group using FCA, FCA_UTK, and FCF. Panel (b) shows these estimates by decade.

labeled as White are always greater than or equal to the number of faces of other race labels over the decades. The prediction results from FCA and FCA_UTK are especially skewed toward labeling faces as being White, followed by faces being labeled as Asian and Black, which is also shown by the FCF predictions. For all models, faces are least likely to be given a label of Latinx + Others.

6. Conclusion

In this paper, we make four primary contributions. First, we curate a face detection data set (IllusFace 1.0) comprising 5,403 manually labeled illustrated faces with bounding boxes which can be used for training data sets. Second, we train two AutoML-based face detection models to detect faces in illustrations using the IllusFace 1.0 data set and the publicly available iCartoon data set; we call these models FDAI and FDAI_iC, respectively. We find that our FDAI model detected 2.5 times more faces in prominent children's books than the existing state-of-the-art face detection model (FDGV). Third, we curate a data set with manually labeled features of 980 faces, coded by three different labelers, including "unsure" labels for race, gender, and age (CBFeatures 1.0). Finally, we train two models to predict the race, gender, and age of illustrated characters - one using an existing data set (FCA_UTK) and one using an original data set (FCA). We compare the performance of these new models to the performance of FCF, an established model, in which we examined the distributions of character gender, age, and race representation in children's books across the different models over time.

In the future, we will extend this work designing measurements for representation of different identities. Models can be expanded to detect the full character rather than just the face. In order to evaluate the biases introduced by our trained model, one can implement interpretability methods (such as gradient class activation maps), which will help interpret the predictions from deep learning methods, creating visual explanations of where the CNN looks in the image, at different depths, to predict a label [20].

Across the world, caregivers and educators use books to teach children about the world, society, and social norms. Because schooling is the greatest non-parental influence in children's lives, the content to which children are exposed in school conveys messages about identity, society, and values. We introduce tools to help measure these messages so that decision-makers can better understand the representation to which they are exposing children.[11]

[11] Acknowledgements: For financial support, we thank UChicago BFI, UChicago CDAC, and UChicago Career Advancement. The research reported here was also supported by Google Cloud Research through research credits and the Institute of Education Sciences, U.S. Department of Education, through Grant R305A200478 to the University of Chicago. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. For helpful feedback, we thank Alex Eble. For excellent research assistance, we thank Ruth Beaubrun, Oishee Chakrabarti, Jarvis Lam, Erica Lin, Simon Mahns, Brandon Mathu, Jan Morgan, Kristof Turan, and Jichao Yang. We also thank Ashiyana and Kairav Adukia-Hornbeck for manual coding assistance. For access to important resources, we thank the Research Computing Center at UChicago.

References

[1] Amazon rekognition – video and image - AWS. URL: https://aws.amazon.com/rekognition/.

[2] Cloud AutoML custom machine learning models | google cloud. URL: https://cloud.google.com/automl.

[3] Facial recognition | microsoft azure. URL: https://azure.microsoft.com/en-us/services/cognitive-services/face/.

[4] What is AutoML? URL: https://www.ibm.com/cloud/learn/automl.

[5] Anjali Adukia, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz. What we teach about race and gender: Representation in images and text of children's books. *National Bureau of Economic Research Working Paper 29123*, 2021.

[6] Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. Building a manga dataset "manga109" with annotations for multimedia applications. 27(2):8–18. URL: http://arxiv.org/abs/2005.04425, arXiv:2005.04425, doi:10.1109/MMUL.2020.2987895.

[7] Philip Bell. Content analysis of visual images. In *The Handbook of Visual Analysis*, chapter 2, pages 11–34. Sage, Thousand Oaks, CA, 2001.

[8] Kendrick Boyd, Kevin H. Eng, and C. David Page. Area under the precision-recall curve: Point estimates and confidence intervals. In Camille Salinesi, Moira C. Norrie, and Óscar Pastor, editors, *Advanced Information Systems Engineering*, volume 7908, pages 451–466. Springer Berlin Heidelberg. Series Title: Lecture Notes in Computer Science. URL: http://link.springer.com/10.1007/978-3-642-40994-3_29, doi:10.1007/978-3-642-40994-3_29.

[9] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 2018.

[10] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. CVPR, 2005.

[11] Jia Deng, Wei Dong, Richard Socher, Li Li-Jia, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. ISSN: 1063-6919. doi:10.1109/CVPR.2009.5206848.

[12] Amit Dhomne, RanjitKumar, and Vijay Bhan. Gender recognition through face using deep learning. 2018.

[13] Antonio Greco, Gennaro Percannella, Mario Vento, and Vincenzo Vigilante. Benchmarking deep network architectures for ethnicity recognition using a new large face dataset. 31(7):67. doi:10.1007/s00138-020-01123-z.

[14] Xin He, Kaiyong Zhao, and Xiaowen Chu. AutoML: A survey of the state-of-the-art. 212:106622. URL: http://arxiv.org/abs/1908.00709, arXiv:1908.00709, doi:10.1016/j.knosys.2020.106622.

[15] Kimmo Karkkainen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. page 11.

[16] Salik Ram Khanal, João Barroso, Nuno Lopes, Jaime Sampaio, and Vítor Filipe. Performance analysis of microsoft's and google's emotion recognition API using pose-invariant faces. In *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion*, DSAI 2018, pages 172–178. Association for Computing Machinery. doi:10.1145/3218585.3224223.

[17] Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[18] Kimmo Kärkkäinen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age. URL: http://arxiv.org/abs/1908.04913, arXiv:1908.04913.

[19] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. 2019.

[20] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A review of machine learning interpretability methods. 23(1). doi:10.3390/e23010018.

[21] David G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004.

[22] Avuthu Sai Meghana. Age and gender prediction using convolution, ResNet50 and inception ResNetV2. 9(2):1328–1334. URL: http://www.warse.org/IJATCSE/static/pdf/file/ijatcse65922020.pdf, doi:10.30534/ijatcse/2020/65922020.

[23] Wafa Mellouk and Wahida Handouzi. Facial emotion recognition using deep learning: review and insights. 175:689–694. URL: https://www.sciencedirect.com/science/article/pii/S1877050920318019, doi:10.1016/j.procs.2020.07.101.

[24] Dimitris Metaxas, Sundara Venkataraman, and Christian Vogler. *Image-Based Stress Recognition Using a Model-Based Dynamic Face Tracking System*.

[25] Kimberly A. Neuendorf. *The content analysis guidebook*. Sage, 2016.

[26] Xiaoran Qin, Yafeng Zhou, Zheqi He, Yongtao Wang, and Zhi Tang. A faster r-CNN based method for comic characters face detection. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1074–1080. ISSN: 2379-2140. doi:10.1109/ICDAR.2017.178.

[27] Inioluwa Deborah Raji and Genevieve Fried. About face: A survey of facial recognition evaluation, 2021. arXiv:2102.00813.

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. CVPR, 2015.

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2015.

[30] Philip Smith and Cuixian Chen. Transfer learning with deep cnns for gender recognition and age estimation. 2018.

[31] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. URL: `http://arxiv.org/abs/1808.01974`, `arXiv:1808.01974`.

[32] Qin-Qin Tao, Shu Zhan, Xiao-Hong Li, and Toru Kurihara. Robust face detection using local CNN and SVM based on kernel combination. 211:98–105. URL: `https://www.sciencedirect.com/science/article/pii/S0925231216305665`, `doi:10.1016/j.neucom.2015.10.139`.

[33] Thanh Vo, Trang Nguyen, and C. T. Le. Race recognition using deep convolutional neural networks. 2018.

[34] Mei Wang and Weihong Deng. Deep face recognition: A survey. 2020.

[35] Jordan Yaniv, Yael Newman, and Ariel Shamir. The face of art: landmark detection and geometric style in portraits. 38(4):1–15. URL: `https://dl.acm.org/doi/10.1145/3306346.3322984`, `doi:10.1145/3306346.3322984`.

[36] Bin Zhang, Jian Li, Yabiao Wang, Zhipeng Cui, Yili Xia, Chengjie Wang, Jilin Li, and Feiyue Huang. ACFD: Asymmetric cartoon face detector. URL: `http://arxiv.org/abs/2007.00899`, `arXiv:2007.00899`.

[37] Kaipeng Zhang, Zhanpeng Zhang, Hao Wang, Zhifeng Li, Yu Qiao, and Wei Liu. Detecting faces using inside cascaded contextual CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3190–3198. IEEE. URL: `http://ieeexplore.ieee.org/document/8237606/`, `doi:10.1109/ICCV.2017.344`.

[38] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. URL: `http://arxiv.org/abs/1702.08423`, `arXiv:1702.08423`.

[39] Yi Zheng, Yifan Zhao, Mengyuan Ren, He Yan, Xiangju Lu, Junhui Liu, and Jia Li. Cartoon face recognition: A benchmark dataset. URL: `http://arxiv.org/abs/1907.13394`, `arXiv:1907.13394`.

[40] Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. URL: `https://arxiv.org/abs/1611.01578`.