Toward Interactively Balancing the Screen Time of Actors Based on Observable Phenotypic Traits in Live Telecast

MD NAIMUL HOQUE, Stony Brook University, USA
NAZMUS SAQUIB, MIT Media Lab, USA
SYED MASUM BILLAH, Pennsylvania State University, USA
KLAUS MUELLER, Stony Brook University, USA

Several prominent studies have shown that the imbalanced on-screen exposure of observable phenotypic traits like gender and skin-tone in movies, TV shows, live telecasts, and other visual media can reinforce gender and racial stereotypes in society. Researchers and human rights organizations alike have long been calling to make media producers more aware of such stereotypes. While awareness among media producers is growing, balancing the presence of different phenotypes in a video requires substantial manual effort and can typically only be done in the post-production phase. The task becomes even more challenging in the case of a live telecast where video producers must make instantaneous decisions with no post-production phase to refine or revert a decision. In this paper, we propose *Screen-Balancer*, an interactive tool that assists media producers in balancing the presence of different phenotypes in a live telecast.

The design of Screen-Balancer is informed by a field study conducted in a professional live studio. Screen-Balancer analyzes the facial features of the actors to determine phenotypic traits using facial detection packages; it then facilitates real-time visual feedback for interactive moderation of gender and skin-tone distributions. To demonstrate the effectiveness of our approach, we conducted a user study with 20 participants and asked them to compose live telecasts from a set of video streams simulating different camera angles, and featuring several male and female actors with different skin-tones. The study revealed that the participants were able to reduce the difference of screen times of male and female actors by 43%, and that of light-skinned and dark-skinned actors by 44%, thus showing the promise and potential of using such a tool in commercial production systems.

CCS Concepts: • **Human-centered computing → Interactive systems and tools**; • **Social and professional topics → Gender**; **Race and ethnicity**.

Additional Key Words and Phrases: Bias; Media; Live Telecast

ACM Reference Format:
Md Naimul Hoque, Nazmus Saquib, Syed Masum Billah, and Klaus Mueller. 2020. Toward Interactively Balancing the Screen Time of Actors Based on Observable Phenotypic Traits in Live Telecast. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 154 (October 2020), 18 pages. https://doi.org/10.1145/3415225

1 INTRODUCTION

Visual media play an important role in our society. Visual content in media can be seen as a reflection of our societal beliefs, but at the same time societal beliefs can be influenced by media [55]. Prior

Authors' addresses: Md Naimul Hoque, Stony Brook University, USA, mdhoque@cs.stonybrook.edu; Nazmus Saquib, MIT Media Lab, USA, saquib@mit.edu; Syed Masum Billah, Pennsylvania State University, USA, sbillah@psu.edu; Klaus Mueller, Stony Brook University, USA, mueller@cs.stonybrook.edu.

2573-0142/2020/10-ART154 $15.00
https://doi.org/10.1145/3415225

research has shown that the abundant representation of *passive*, *sexual* and *supporting* female characters can increase the likelihood of sexual misconduct in society [29]. Researchers have also shown that if visual content mostly revolves around characters of a specific gender or race, children may learn to think less of marginalized groups in society [12]. Therefore, it is imperative that visual media should not promote gender or racial stereotypes through conscious or unconscious representations of actors and characters.

In an attempt to make content-creators more conscious to these issues, organizations, such as UNESCO, the Geena Davis Institute, and UN Women, as well as researchers across many domains have made enormous efforts to bring forth the implicit biases and stereotypes in visual media [1, 39, 50, 51, 63]. Recent reports suggest that onscreen gender and skin-tone representation has been improving [37, 52, 53], but the problem is far from being solved as biases are wired inherently in different parts of the media production [37]. One of the primary ways media promote certain gender and racial stereotypes is by allocating imbalanced screen time to actors of different gender and race [67]. Thus, screen time, defined as the amount of time an actor appears on screen in a movie or television show, has become an important measure of (mis)representation of different minority groups in visual media [50].

In this paper, we concentrate on how technology can assist media producers in the task of allocating different phenotypic groups equal screen time in visual media. We identify that current commercial video production software provides little support for quantifying the screen time of actors based on their observable phenotypic traits, but argue that this can be addressed by taking advantage of recent image analysis and facial recognition technologies. These technologies have become sufficiently robust such that they are now being deployed in many mission-critical applications, such as smartphone authentication [5], airport, and home security systems [31, 62]. Of particular note is the GDIQ tool, an analytical tool from the Geena Davis Institute that utilizes these technologies to determine the screen time of different genders in a video [50].

Tools like GDIQ can be effectively adopted as production tools to measure the gender representation (in terms of the screen time) in videos that undergo editing phases. For example, a movie usually passes through several editing phases, each of which can use GDIQ to measure and refine the gender representation in the edited movie. Even without any technical support, a movie director can refine the phenotypic representation over the movie's production period which usually spans several months.

A live telecast, on the other hand, is unique in terms of the production process as it is created in real-time with no post-production or editing phase. A video producer takes instantaneous decisions to produce a live telecast in a dynamic and uncertain production environment, making offline tools such as GDIQ unsuitable for the task of balancing the screen time of the actors based on their phenotypic traits.

We introduce ***Screen-Balancer*** to address this void. Screen-Balancer allows media producers to interactively modulate the composition of observable phenotypic traits of the actors in a live telecast. Screen-Balancer currently supports gender and skin-tone as observable phenotypic traits, but the interface is not restricted to these two. It augments the current live telecast systems by (1) automatically extracting the screen time of the different actors based on these phenotypic traits in the live camera-feeds, and then (2) visualizing these statistics in real time, offering different visual cues to help media producers balance the screen time of the actors.

Our design of Screen-Balancer has been informed by a field study we conducted in a commercial production studio. In this study, we observed that during the production of live telecasts, the production studio typically uses multiple cameras capturing several different perspectives of the scene. One camera might zoom into a character, another might keep a master angle that shows all of the actors, and yet another might only capture a subset of individuals in a frame. Consequently,

these camera-feeds are expected to have different distributions of screen time of the featured actors and their phenotypic traits. We also observed that choosing a particular camera-feed for telecasting at any given time typically occurs fully at the producer's discretion.

These observations informed the design of Screen-Balancer. Screen-Balancer gives access to an enhanced form of situational awareness by proactively displaying the gender and skin-tone distributions for each camera-feed using a set of easy-to-compare bar-graphs. In addition, we also present the impact of choosing an individual camera-feed in the near future, enabling producers to make an informed decision as they choose the next camera-feed.

In summary, we make the following contributions:

- The design and development of *Screen-Balancer*, a prototype system to balance screen time of actors based on phenotypic traits (e.g., gender and skin-tone).
- A user study with 20 participants to evaluate the effectiveness of Screen-Balancer in balancing the screen time in simulated telecasts.
- Feedback from professional video producers and social science researchers regarding the implications, challenges, and potential deployment issues of Screen-Balancer.

The remainder of this paper is organized as follows. In Section 2 we describe prior research related to gender and skin-tone bias in visual media, the limitations of the current production software, and the potential of adapting facial recognition technologies in media production. In Section 3, we present a field study that leads to the design of Screen-Balancer in Section 4 and implementation in Section 5. We present the evaluation of Screen-Balancer in Section 6 followed by a discussion in Section 7. Finally, conclusions are drawn in Section 8.

2 RELATED WORK

The notion that there exists gender and skin-tone bias in visual media has been shown both in quantitative and qualitative manners. In this section, we discuss the studies conducted to evaluate and determine the effect of gender and skin-tone bias in media, as well as movements against such biases organized by different NGOs and organizations.

2.1 Gender Bias in Media

A fair share of existing literature on the relationship between gender and media revolves around identifying how gender bias affects societal beliefs [54, 67]. More specifically, researchers strived to identify the effects of gender bias posed by media on children or teenagers since they are more susceptible to inherit gender stereotypes at an early age [10, 60, 61, 65]. Gender portrayal in media has been shown to affect career choices in later stages of life [28]. Other studies reported effects of media on acceptance of violence against women [49], low self-esteem of women [6, 27], sexual socialization among American teens [66], acceptance of different genders and racially aware TV shows [11, 12]. Several articles analyzed the imbalanced representation of male and female screen time in live telecasts and discussed implications of such misrepresentation [16, 18, 45].

Considering the crucial role of visual media in society, *UNESCO* has declared *Gender Sensitive Indicators for Media (GSIM)*, a list of indicators that ensures equality and women's empowerment in media. GSIM states in a report that the balanced representation of men and women in media is one of the primary indicators of gender equality.

The *Geena Davis Institute on Gender in Media* introduced *GDIQ* tool [50], the first-ever automated tool to analyze male and female screen time and speaking time in videos. Using the *GDIQ* tool, the institute has analyzed 200 top-grossing Hollywood movies released in 2014-2015 and discovered that men appear more than females in almost all types of movies, even in movies with female lead characters. Similar kinds of tools and machine learning-based methods have also been developed

to analyze visual content from other countries such as Bollywood movies [48], and Bangladeshi TV-shows [36]. Recently, Jang et al. [39] discussed in detail the various adverse effects of gender bias in media, their implications, and attempts to quantify such biases. The authors argued why the Bechdel Test, a popular test used as a measure of gender bias in movies, is not sufficient to encapsulate the complex notion of gender bias in visual contents and proposed eight quantitative indices to quantify the gender bias posed by visual media. They analyzed 40 movies using off-the-shelf computer vision tools to reaffirm the existence of gender bias based on those 8 indices. Such tools could help find historical gender stereotypes, and only detect biases in already released visual content.

Our tool is different from these, as we assist content-creators to balance the screen time of male and female actors during the production of a live telecast, with the hope of making content-creators more aware of the existence of potential gender bias in their product.

2.2 Skin-tone Bias in Media

Intersectionality [58] is a framework often used to define a population through different identities, such as gender, race, and class. In this paper, we concentrated on balancing screen time of genders and race in live telecast videos. But race is not an observable phenotypic trait and the task of detecting a person's race solely from a face image can be challenging. Neural networks have shown good performance on detecting race from images for people from some specific geographic regions [64] but not in general. We thus opted for skin-tone as a feature instead of race because we wanted our tool to be generally applicable. Skin-tone is easy to recognize visually for humans and computers alike and there is very little chance of disagreement among different users on this matter. It has been used to identify race in many cultures, albeit it is a method that has seen some debate in the literature [17, 43].

While gender bias adversely affects women, discrimination based on skin-tone disadvantages dark-skinned people [30]. This phenomenon is often known as Colorism or Shadeism. The idea behind colorism and racism is quite similar since in both phenomena darker-skinned people are discriminated against. Ben-Zeev et al. (2014) found that educated African-American men appear lighter in the mind of their peers [8]. Other researchers linked darker skin-tone to smaller incomes, lower marriage rates, longer prison terms, and fewer job prospects [30, 33, 35].

A number of studies in Colorism in media found the existence of discrimination against dark-colored people in movies, tv-shows, and news presentation. Travis Dixon [19–21] authored a series of publications to prove that media outlets consistently portray black people as poor, violent, and dysfunctional, whereas white people are portrayed as stable and welfare-oriented. Several advertisements from companies such as L'Oreal[1], Elle Magazine[2] have been accused and criticized for whitening skin-color of their models.

2.3 Facial Recognition and Commercial Video Production Software

Adobe After Effects [2] supports face tracking in videos. Through face tracking, users can apply effects on facial landmarks, such as nose, mouth, and pupils. Adobe Photoshop [3] uses facial recognition to organize and search images in a catalog. Similarly, Adobe Premiere allows masking and tracking of moving objects in videos through facial and object recognition. Final Cut Pro has a facial detection system to detect faces in a video. To the best of our knowledge, none of the

[1]https://www.theguardian.com/media/2008/aug/08/advertising.usa1
[2]https://www.telegraph.co.uk/news/celebritynews/8005734/Elle-magazine-in-Gabourey-Sidibe-skin-lightening-controversy.html

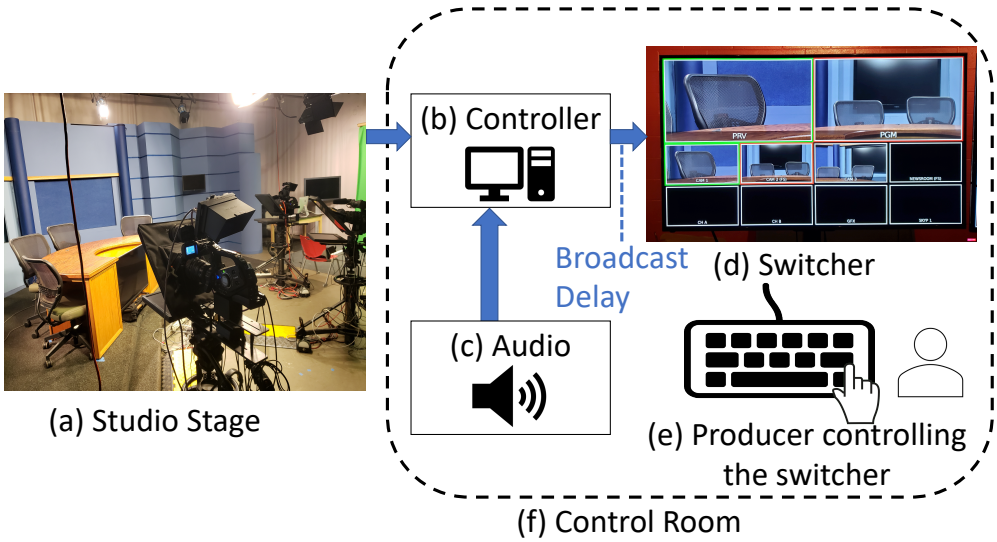


Fig. 1. A live telecast setup: (a) the studio stage; (b) the controller, a computer equipped with production software to process and filter video feeds and audio; (c) the audio mixer; (d) the switcher, a visual interface that shows different camera-feeds; (e) the switcher controller, a console which a video producer can control to interact with the Switcher and select a camera-feed.

current video production software provides any kind of functionality related to gender or skin-tone distribution in a video. This is also true for live telecast or broadcasting software.

Although automated technologies are becoming important features of production software, the integration of such technologies needs verification since the underlying models are often trained using historically biased datasets which can lead to the mistreatment of minority groups. For example, recent research has shown that three automated facial recognition systems are more likely to miss-classify darker women than white men or white women [13].

3 FIELD STUDY

To understand how live videos and telecast are produced, two of the authors visited a professional production studio, where a video producer and several technical crews accompanied them. The video producer gave them a guided tour to the studio stage and control room, and explained the setups for a live telecast.

3.1 Studio Stage

Figure 1a presents the studio stage at the time of our visit to the studio. The actors perform in this stage and this is what the audiences see on their TVs. This stage is customizable and the decoration varies from show to show. At the time of our visit, the stage had a table discussion setup with three cameras pointing to the table from three directions.

3.2 Control Room

The control room (Figure 1f) is the production hub for creating live telecasts. This room has the hardware and software panels necessary to process the audio and video feeds coming from the studio stage. Short descriptions of those panels follow next.

3.2.1 Controller. As shown in Figure 1b, the Controller is a computer equipped with modern video and audio editing software. The studio we visited had a powerful desktop computer running Windows 10 and Ross Live Production, a software to manage other studio hardware.

3.2.2 Audio Mixer. It acts as a hardware hub for sounds coming from the stage (see Figure 1c). During live telecasts, a crew is stationed at the Audio Mixer for audio production.

3.2.3 Switcher. The Switcher is a software that provides a visual interface (Figure 1d) for previewing all available camera-feeds. The producer uses this interface to author live telecasts. The studio we visited had three camera-feeds on the stage, namely, CAM1, CAM2, and CAM3 (shown in the second row in Figure 1d); and an optional *Preview (PRV)* frame (shown in the first rectangle in Figure 1d), in which the producer could preview a camera feed before choosing it for telecasting. Right next to the preview frame is the broadcast frame (PGM), although the positions may vary from system to system. We incorporated this side by side, stacked representation of camera-feeds and frames of a Switcher in our design.

3.2.4 Switcher Controller. It is a hardware (Figure 1e) with multiple buttons, each of which is assigned to a camera-feed in the Switcher. At any time, the producer can press a button to preview the corresponding camera-feed in the PRV frame or to select that feed for broadcasting in the PGM frame.

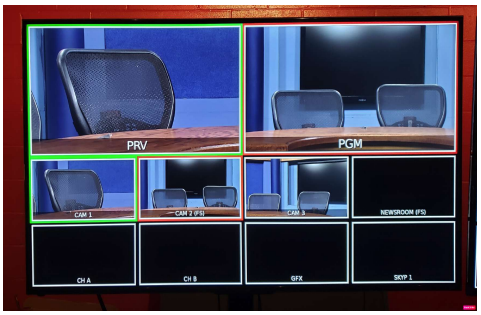
3.2.5 Broadcast Delay. The broadcast delay (BD) is often a 7 to 10 seconds delay deliberately integrated into the signal path to censoring live telecasts. Interestingly, the producers are oblivious to this delay, as the censoring happens in the Controller computer (Figure 1b). As a result, the camera-feeds appearing in the Switcher interface seem live to the producer, even though the feeds are delayed by some time defined by the broadcast delay. This delay plays a critical role in our design since we envision that *broadcast delay could let us run computer vision packages in the Controller machine to analyze video frames before they are fed to the Switcher.*

4 DESIGN

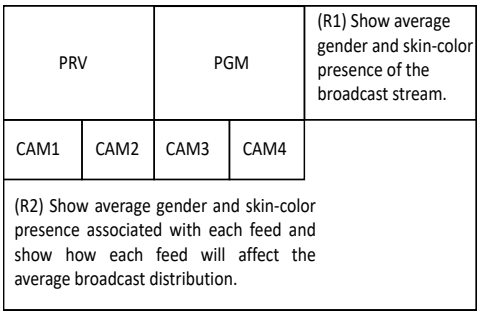
4.1 Guidelines

"How to enable the screen time balancing functionality in the existing live telecast setup?"—this was the key challenge to our design. During the field visit, we observed that the video producer constantly changes camera angles to produce the final telecast. The video producers use the Switcher interface (described in paragraph 3.2.3 and shown in Figure 1d) to select camera-feeds. Interestingly, these camera angles capture different perspectives of a live show and thus have different phenotypic distributions. The presentation of phenotypic distributions related to each camera angles could allow video producers to make an informed decision in terms of screen time while changing camera-feeds. Therefore, we focused on extending the current visual interface of *Switcher* by presenting gender and skin-tone distribution of each camera-feed in the interface.

"*How to convey the information related to gender and skin-tone distribution to the video producers?*"—this was our second design challenge. Given the dynamic nature of live telecasts and the fact that video producers constantly take quick decisions while producing a telecast, it is imperative that our system presents these distributions in a way that is easily perceivable. It is also important to design an interface that could easily be extended to moderate other observable phenotypic traits. To satisfy these two requirements, we opted for glanceable information visualization over other means (e.g., textual explanation). We set the following design guidelines in order to incorporate the screen time balancing feature in the *Control Room.*



(a) Professional Video Switcher Interface



(b) Initial design of Screen-Balancer

Fig. 2. Initial design of Screen-Balancer

- G1. The interface should emulate Switcher functions. Producers would be able to observe the camera-feeds in the interface and interact with the feeds easily.
- G2. The Controller would exploit broadcast delay and use computer vision packages to analyze video frames before they reach the switcher. Thereafter, the interface should visualize the distribution of gender and skin-tone of the live telecast, as well as all available camera-feeds in the Switcher.
- G3. The visualization should update periodically to reflect the gender and skin-tone changes in the live telecast and camera-feeds. The update cycle should match the broadcast delay, which would allow Controller to analyze the video frames on the fly.
- G4. Due to short broadcast delay, the visualization should allow the producers to quickly and conveniently compare different feeds quantitatively.

4.2 Interface Design

The interface was developed iteratively. During each design iteration, we held formal and informal meetings with a video producer (referred to as P1) to discuss different aspects of the tool. Figure 2b shows our initial design and how it was inspired by the current switcher interface. Following G1, the interface has a Preview (PRV) frame, a Broadcast (PGM) frame, and several Camera-feeds (CAM). Further, based on G2, we integrated two requirements to the Switcher interface at our first iteration: (R1) visualize statistics related to overall gender and skin-tone presence in the broadcast stream; and (R2) visualize statistics related to gender and skin-tone presence in each camera-feed, as well as the after-effects of selecting a camera-feed on the overall broadcast distribution.

Figure 3 shows Screen-Balancer in a simulated live telecast setup. The interface is divided into five regions. The functionalities of each regions are illustrated in the next few sections. Along with the illustration, we discuss some of our design choices, how our design evolved in each iterations, and what alternatives we considered to fulfill G3 and G4.

4.2.1 Countdown Timer. Region (a) shows a 10-second countdown timer to indicate when the charts will be updated next. The timer resets to 10 upon reaching 0. Interval of this countdown timer must match the duration of broadcast delay (and vice versa) (G3). To determine a comfortable duration for broadcast delay, P1 experimented with different number of camera-feeds and with different amount of broadcast delay, and settled on 10 seconds.

4.2.2 Screen time Distribution (R1). Region (d) in Figure 3 shows a bar chart and a stream graph to present the cumulative distributions of gender and skin-tone in a live telecast.

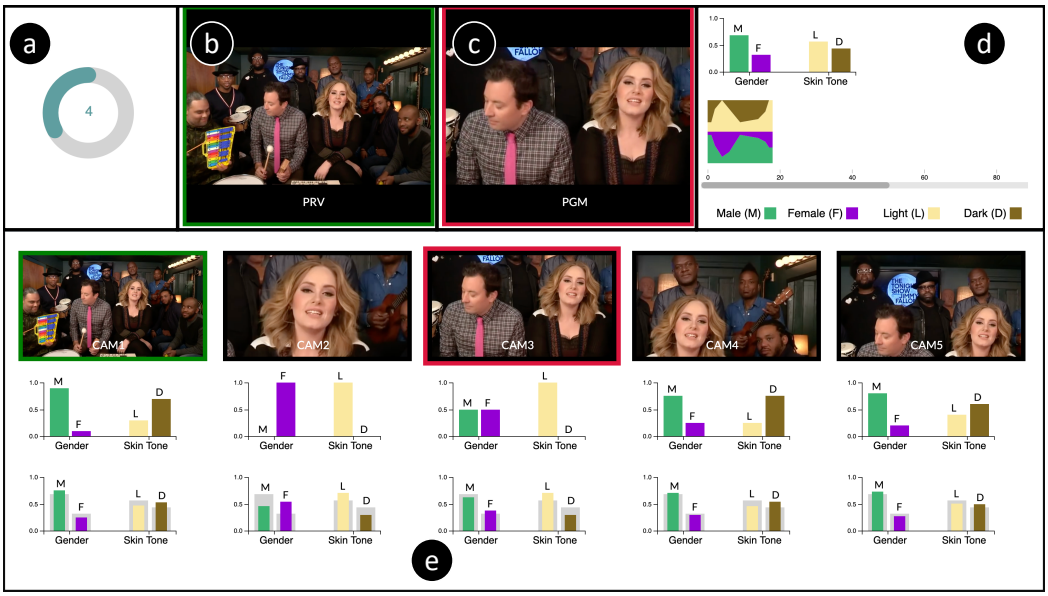


Fig. 3. Components of *Screen-Balancer*: **(a)** a countdown timer showing when the charts and graphs will be updated next; **(b)** the preview stream, which video producers use to isolate a particular camera feed from (e) before selecting it for the output stream; **(c)** the output stream; **(d)** bar-charts and a stream graph showing the cumulative distribution and the timeline distribution of sensitive attributes (e.g., gender, skin-tone) in output stream until now; **(e)** input streams (e.g., cameras with different angles and views) along with one bar-chart and one bullet chart each, one showing screen-time of different genders and colors in that stream 10 seconds in advance, and another showing how choosing that particular stream will affect the overall screen-presence distribution of genders and colors in (d) in the next 10 seconds.

Bar charts with different heights are easy to compare visually (requirement G4). Pie charts, on the other hand, require users to make comparisons based on angles which can be difficult when the sectors have similar values [25]. Using stacked bar charts in place of horizontal distribution of bars has similar problems to pie charts, i.e., difficult to compare stack segments with similar values. Cleveland & McGill [15] reported that people performed substantially worse on stacked bar charts than on aligned bar charts, and comparisons between adjacent bars were more accurate than stacked bars. We hence conclude that presenting the phenotypic variables in the form of simple bar charts, placing the categories (such as Male, Female) side by side would allow users to quickly evaluate the overall distribution.

The color legends for each variables are also presented in region (d). We avoided the stereotyped use of blue and pink for gender data, and opted for *green* and *purple* for male and female respectively [14]. The color green conflicted with the border of the preview frame (PRV), but P1 was comfortable with that.

4.2.3 Timeline (R1). The bottom part of Region (d) in Figure 3 shows the timeline view. We use a *(streaming) 100% Stacked Area Chart* because it can visualize the change of skin-tone (top) and gender (bottom) proportions in a live telecast over time (stretching along the y-axis). A 100% Stacked Area Chart is essentially a set of stacked and filled line charts, normalized to fit in a box. To eliminate small-scale jitter from the display, we applied *Moving Average* filtering to smooth the area chart. That is, for any time t, each embedded line graph has a value that is the average of the

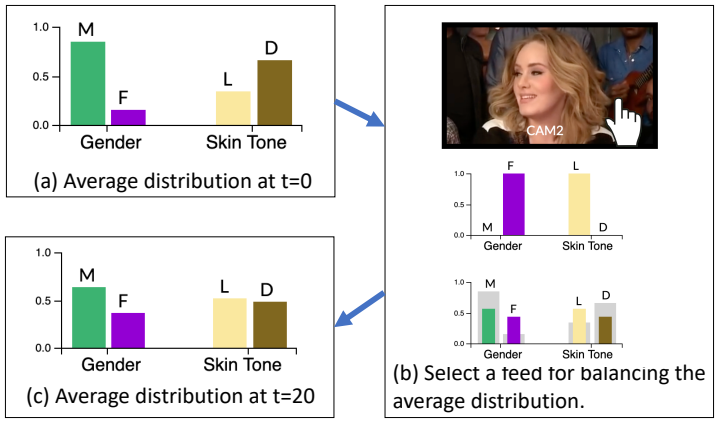


Fig. 4. Interaction with *Screen-Balancer*. (a) During a live telecast the producer observes that both the gender and skin-tone distributions are unbalanced. The producer decides to select the camera-feed in (b) for broadcast, as it will balance both the gender and the skin-tone distribution. (c) The distribution of gender and skin-tone 20 seconds later, which are more balanced than (a).

values from time $t-3$ to time $t+3$. The horizontal axis of the area chart dynamically increases as the time increases in the live telecast. This chart was included in the second design iteration as P1 suggested that a historical representation of gender and skin-tone distribution may help understand the overall trend.

4.2.4 Input Camera Feeds (R2). The bottom row of Screen-Balancer (region (e) of Figure 3) shows the set of input camera feeds in a horizontal arrangement. During a live telecast, the producer can select any of these camera-feeds which will connect it to either the preview (PRV) or the broadcast frame (PGM) in regions (b) or (c). As stated in Section 3.2.3, producers often use the PRV to isolate a camera-feed from the pool of camera-feeds available before selecting it for broadcasting. In Screen-Balancer, the broadcast feed is highlighted with a red border while the preview feed is highlighted with a green border. To select a camera-feed for previewing, a user can either use designated keys assigned for each camera-feed or click on the camera-feeds. After previewing, a user can select that feed for broadcasting by hitting the "ENTER" button. This mechanism of changing camera-feeds is identical to that of a professional switcher controller, except we use a Keyboard to facilitate the functionality of changing camera-feeds. The interface also allows switching a camera-feed without previewing it, by simply pressing a designated key, or clicking on a feed will change the broadcast feed to that particular feed.

Each camera feed in region (e) is accompanied by two bar charts. The *first* chart associated with a feed shows the distributions of gender and skin-tone in the next 10 seconds, thanks to broadcast delay. The *second* chart associated with each camera-feed is a variant of bullet graph [24], which shows how the selection of a feed affects the average distribution in region (d). The bullet graph is known for its usefulness in comparative analysis and can show both the reduction and increment of values in the bar [24]. The current value of any category is shown in grey bars in the bullet graph. The effect of choosing a camera feed is shown as colored bars. At any moment, the producer only needs to see the colored bar to determine the after effect, and the direction of impact, when choosing a camera-feed.

Figure 4 demonstrates how Screen-Balancer balances the phenotypic presence in a live telecast.

5 DEVELOPMENT OF SCREEN-BALANCER

5.1 Face and Gender Detection

Face detection is crucial to our analytical pipeline, as most of our analyses depend on facial attributes. We used dlib, an open-source Python/C++ library, to detect faces from a video frame, at a rate of 1 fps, because faces on the screen hardly change within a second.

For detecting gender from facial images, we used a trained convolutional neural network proposed by Levi and Hassner [46]. The model is claimed to have an 86% accuracy rate on the Adience benchmark [22]. We created our own test set of 1000 faces extracted from our video sources and found the accuracy rate to be 89%. This test set is also utilized to measure the fairness of the gender and skin-tone detection models.

To measure the fairness of our gender detection model, we evaluated the accuracy rate for detecting males and females from the aforementioned curated test set. We found that the model did perform better in detecting male faces (93%) than female faces (87%), exposing potentially biased behaviors. The gap remained at 3-4% as we increased the test set size gradually which suggests a moderate disparity. We opted for this model as this is a widely used open-sourced gender detection model and other commercial gender detection models have been shown to have disparate impacts [13]. The model is limited to detecting males and females, as the current state-of-the-art detection models are unable to identify non-binary genders [57]. An alternative approach to devise a diverse gender-inclusive design is discussed in Section 7.2.

5.2 Skin Tone Detection

For detecting skin-tone from facial images, we leveraged an unsupervised learning algorithm, k-means. Our method is similar to the one Hoque et al. [36] used to estimate skin-tone in TV serials. We considered two skin-tone labels: *Light* and *Dark*, in accordance with the skin-tone labels used in [13, 36]. These labels are based on the Dermatologist approved Fitzpatrick's six-point skin type scale [26], where Type I, II, and III are labeled as light skin-tone and Type IV, V, and VI are labeled as dark skin-tone [13].

To train our skin-tone detection model, we used the *CelebA* dataset [47], which contains 202, 599 face images of 10, 177 different celebrities. For each image, we used dlib's landmark detection framework [41] to draw a bounding box around a face excluding the hair area. The pixels around eyes and mouths were also excluded as those pixels might disrupt the skin-tone estimation if a person wears a sunglass or shade, or has lipstick in the mouth or has a mustache.

We applied k-means on the selected pixels with $k = 2$ as used in [36]. Since the images only contain face pixels and we excluded possible disrupting pixels from the faces – the biggest cluster among the 2 clusters should contain the pixels that would indicate the skin-tone of the faces. We took the centroid of the biggest cluster as an estimation of the skin-tone of the faces. Afterward, we estimated the skin-tones (hexadecimal values), and applied k-means with $k = 2$ again on these skin-tone estimations to find the skin-tone class of each image. Based on the predefined classes and the clustering result, we assigned the appropriate skin-tone label (*Light and Dark*) to each image.

To evaluate the performance of the labelling task, one of the authors measured the accuracy on a set of one thousand images. The accuracy rate for the labelling task was 98%. Finally, we trained this labeled *CelebA* dataset on a two hidden layered Convolutional Neural Network. The network achieved an accuracy rate of 93% on the testing set of *CelebA*.

Similar to the gender detection model, we evaluated the *CelebA* dataset for potential disparate representation. CelebA has shown to have excellent demographic parity, and good equality of opportunity in terms of skin-tone [56]. For evaluating the fairness of the detection model, we calculated the accuracy rate for both lighter and darker skin-tone faces from the curated test set

from Section 5.1. The accuracy rates were comparable for both categories (89% and 91% for lighter and darker toned faces respectively).

6 EVALUATION OF SCREEN-BALANCER

We evaluated Screen-Balancer in two phases: first, we conducted a user study with 20 users to assess the effectiveness and usability of Screen-Balancer; second, we interviewed 5 industry professionals and researchers to gather expert feedback and real-world potential of Screen-Balancer.

6.1 User Study

We aimed at validating the following two hypotheses in the study:

- **H1**: A live video production software with Screen-Balancer will be more effective in balancing screen time, than without it.
- **H2**: The Screen-Balancer tool will be easy to use.

6.1.1 Participant Demographics. We recruited 20 participants (11 males, 9 females) through local mailing lists, university mailing lists, and public posts on Facebook groups. The participation was voluntary with no compensation. Our inclusion criteria included familiarity with video editing, content creation, post-production, news broadcasting, and live streaming. The participants varied in age from 19 to 35 ($M = 25.9$, $SD = 3.81$), gender ($male = 11$, $female = 9$), skin-tone ($light = 8$, $dark = 12$), and professions ($filmmaker = 2$, $camera\ operator = 1$, $video\ advertisement\ maker = 1$, $journalist = 3$, and $YouTuber = 13$). All of them had undergraduate degrees in different majors.

6.1.2 User Task and Study Setup. The participants were asked to control the switch of a 3-min (simulated) live broadcasting from 5 different camera-feeds, while balancing the distribution of male and female actors in the telecast, as well as their skin-tones. In live telecasts, since it is important to focus on the actor who is currently performing (e.g., focusing on the person currently speaking in a news debate), the participants were instructed to capture this element to maintain coherency.

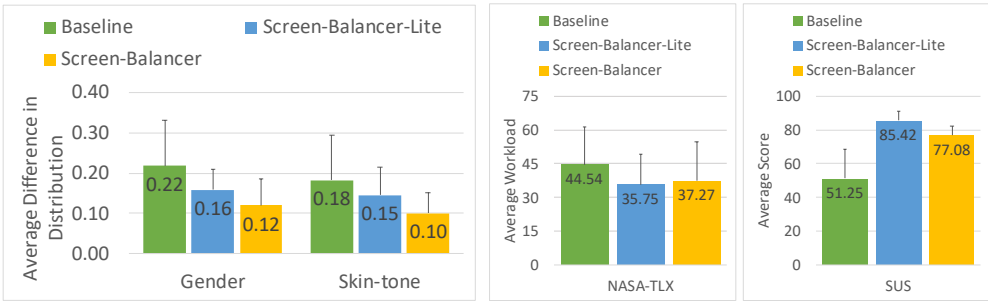
To acquire real-world, pre-production videos for our study, we sought to collect videos that were available online and were captured from different camera angles to simulate a live-stream setup. Finding such videos was difficult, because content creators already edited and compiled these videos from different camera angles. We specifically searched for videos that (i) were shot from a single camera angle, or (ii) had multiple angles stacked together in a single frame.

We found several musical shows meeting criterion (i); we cropped out different parts from those shows to simulate different camera angles. We also found several discussion shows meeting criterion (ii), for which we cropped each portion showing one or more character to re-crate the live studio setting. While cropping, we ensured that the gender and skin-tone distributions in each cropped feed were different, and no single feed provided absolute parity to the average distribution of gender and skin-tone. In addition, feeds were chosen to be dynamic, so that the distribution of gender and skin-tone changed over time in different camera feeds. In total, we prepared 6 videos from 3 different genres, such as talk-shows, news, and live music performance.

To facilitate remote participation, we deployed our prototype on the web, hosted on a MacBook Pro laptop running an Apache web server. Five participants conducted the experiment remotely, and the rest conducted it on the aforementioned MacBook Pro laptop in our lab. When administrating remote studies, the experimenter communicated over Skype.

6.1.3 Study Design. We designed a repeated measures within-subject study with the 3 conditions:

C1. Simple Video Editor, presented a basic video switching software, where users could choose a camera feed from a list of available feeds at any time, to produce the final live telecast. This was our *baseline*.



(a) Quantitative Measures (b) NASA-TLX Scores (c) SUS Scores

Fig. 5. Study Results. (a) The average differences in distributions of gender and skin-tones in the final telecasts produced using 3 study editors. (b) The average NASA-TLX, (c) and SUS scores reported for 3 study editors. Error-bars show +1 SD.

C2. Simple Video Editor + Screen-Balancer-Lite, presented a basic video switching software, along with the average and timeline distributions of screen time in the broadcast feed, as shown in figure 3d. We included this condition to better understand the effectiveness of our proposed visualization.

C3. Simple Video Editor + Screen-Balancer, presented a basic video switching software, along with the full prototype, as shown in Figure 3.

Each session lasted for 60 minutes, including ~20 minutes for practice at the beginning. To minimize the learning effect, we counterbalanced the ordering of study conditions and task videos. At the end of each condition, we administrated the NASA-TLX [34] questionnaire to measure the participants' perceived workload, and the SUS questionnaire [9] to measure user experience and usability of that condition. The experimenter took notes during the session. All sessions were video recorded and transcribed. Each session culminated with participants making suggestions and recommendations.

6.1.4 Data Collection. We analyzed the experimenter's notes, logs and recorded videos to measure the following metrics: (i) the distribution of males and females in the produced video clips; (ii) the distribution of skin-tones (e.g., light, and dark); and (iii) the number of times participants changed camera feeds. We also calculated the following subjective measures: (i) coherency of the produced video clips when using Screen-Balancer, and (ii) perceived workload and SUS scores.

6.1.5 Study Results. We performed a repeated-measures ANOVA with the conditions being independent variables. We used the Greenhouse-Geisser correction for correcting violations of sphericity, and post-hoc tests using a paired t-test with the Bonnferroni correction.

Distribution of Gender (H1). We found a significant effect of study condition on the differences in gender distribution, i.e., gender presence, in the produced video clips, $F(1.392, 26.453) = 6.602, p < .010$. As shown in Figure 5a, the averages were 0.22 ($SD = 0.112$) for baseline (C1), 0.16 ($SD = 0.049$) for condition C2, and 0.12 ($SD = .066$) for condition C3. While using C3, the participants reduced the gap between male and female screen-time by 43% compared with baseline (C1), and by 27% compared with C2. However, only the former was found to be statistically significant ($p < .032$), and Cohen's effect size value ($d = 1.027$) suggested a high practical significance. Even

though using C2 yielded a 27% reduction of screen-time between males and females compared to baseline (C1), this reduction was not statistically significant.

Distribution of Skin-tone (H1). We also found a significant effect of study condition on the differences in skin-tone distribution in the produced video clips, $F(2, 38) = 5.683, p < .007$. As shown in Figure 5a, the average difference between light and dark toned actors was 0.18 ($SD = 0.111$) for baseline (C1), 0.15 ($SD = 0.069$) for C2, and 0.10 ($SD = .0518$) for C3. Between baseline (C1) and C3, the difference of screen time was reduced by 44%, which was found to be statistically significant ($p < .027$), as expected. Further, Cohen's effect size value (d = 0.928) suggested a high practical significance. No other comparisons were significant.

Camera Switching Frequency (H2). We anticipated that the participants would frequently switch cameras when using Screen-Balancer. Surprisingly, we found that switching camera occurred most frequently when they used baseline ($M = 15.250$, $SD = 8.503$), followed by Screen-Balancer-Lite ($M = 11.550$, $SD = 5.443$), and Screen-Balancer ($M = 13.650$, $SD = 7.169$). However, these differences were not significant, as reported by a one-way repeated measures ANOVA. With baseline (C1), participants stated that they tried to remember what camera-feed they had chosen in the past. But at some point, they became clueless and started to choose a camera-feed arbitrarily and frequently.

Perceived Workload (H2). The NASA-TLX scores, as shown in Figure 5b, revealed a significant difference among the workload under three conditions, $F(2, 38) = 3.262, p < .049$. Balancing screen time without any visual aid in baseline yielded a mean NASA-TLX score of 44.540 ($SD = 16.583$). Surprisingly, they reported that their workload decreased the most (19%) with condition C2 ($M = 35.750$, $SD = 13.312$) compared to C1, rather than using condition C3 ($M = 37.270$, $SD = 17.269$). The decrement between C1 and C2 was only statistically significant ($p < .040$, Cohen's $d = 0.585$). Individual scores for each questions in NASA-TLX are shown in the supplemental materials, where all but the "Performance" one are negative impact metrics, thus the results mean that the lower, the better.

Usability Assessment (H2). The SUS scores (see Figure 5c) followed a similar trend: it increased the most (66%) for condition C2 ($M = 85.42$, $SD = 5.103$) compared with baseline ($M = 51.25$, $SD = 17.084$), followed by condition C3 ($M = 77.08$, $SD = 5.685$, $increment = 50\%$). Upon further inquiry, the participants mentioned that glancing over multiple bar-chars at once was the primary reason why their workload marginally increased and the SUS score marginally decreased with condition C3, compared with condition C2.

Video Coherency. To assess the quality and coherency of the videos produced by using Screen-Balancer, we asked three human evaluators to rate such videos on a scale of 3 (1=incoherent, 2=partially-coherent, and 3=coherent). The inter-annotator agreement was substantial (Fleiss' $\kappa = 0.63$). The final verdict was made via majority voting. Out of 20 videos, 16 were rated as coherent, 3 as partially-coherent, and only 1 as incoherent. One of our evaluators who was a professional film-maker made the following comment: *"It is very difficult to measure how good a video is. In terms of coherency, it is people's natural instinct to go with a video that has activity in it. Having said that, you often see live shows that frames one person and someone else is actually talking in the show. No one complains about them. Yes, I saw some minor inconstancy in the videos, but that is probably because the gender and skin-tone distributions among the videos were itself disproportional which is also true for many real-life television shows."*

These results validated our hypotheses H1 and H2. The participants were able to decrease the difference of screen time of male and female actors, as well as their skin-tones. Even the limited visualization provided in Screen-Balancer-Lite was found to be helpful in decreasing such differences. Although these decrements were not significant, they indicated the usefulness of our proposed visualization during the production of live telecasts.

6.2 Interview with Industry Professionals and Domain Experts

In this subsection, we present interviews with 3 commercial live video producers (anonymized as VP1-VP3) and 2 academicians (AC1 and AC2) who work in the field of gender studies and diversity inclusion. Our goal was to better understand the issues raised in the user study, such as cognitive load and video coherency, as well as the real-world challenges and potential of Screen-Balancer.

None of them were involved in the development of Screen-Balancer. VP1 and AC1 were female, and VP2 and AC1 had a dark skin-tone. Others were male having a lighted skin-tone. Each interview session started with a brief explanation of Screen-Balancer, followed by a demonstration, and ended with a semi-structured interview. We sorted the feedback into the following five thematic categories:

6.2.1 Resemblance to Video Switcher. All three video producers quickly related Screen-Balancer with a real-world Switcher (VP2's initial reaction was "*Ohh! that is a switcher!*"). They also recognized the value of bar charts but were confused at first realizing that these bar charts depicted the after-effects of selecting a feed. In addition, they were comfortable with the complexity of our system. VP2 made the following comment: "*My mind is trained to operate in a manner that allows me to glance multiple feeds simultaneously and act dynamically to produce a video. Your system is quite similar to what I am used to, so, no, I do not think it is a complex system.*". VP3, on the other hand, suggested a training session might be necessary to adopt Screen-Balancer.

6.2.2 "Exciting" and "Timely Solution". All three video producers expressed their admiration for the automated nature of Screen-Balancer. They tagged Screen-Balancer as "exciting" and "a timely solution". Two of the video producers mentioned that they were very cautious about how they portray different groups in their work and this tool could help them achieve that goal more effectively. According to VP1, Screen-Balancer might be more useful in dynamic telecasts and less in live newscasts, as the latter are often scripted, thereby lacking uncertainty.

6.2.3 Potential of Real-life Deployments. VP2 inquired about the technical difficulties of detecting gender and skin-tone automatically. He made this remark: "*Commercial production software are always competing with each other and looking for new features to integrate into their system. I believe, they would be very excited at the prospect of this tool.* "

6.2.4 Recommendations. The video producers recounted several aspects of video production which they thought if automated could decrease their cognitive load. For instance, they often have a high-level idea of how much screen time to allocate for each actor, but find it tedious to track each actor during a live telecast. According to them, Screen-Balancer could help them in this regard. In addition, Screen-Balancer could extend to other visually observable or detectable features such as lighting, actors' postures and gestures. VP1 suggested that the integration of such features into the system would make it more "lucrative", and would "motivate" video producers to use it.

6.2.5 Gender and Skin-tone Bias in Media. Both AC1 and AC2 who studied stereotypes in visual media, tagged Screen-Balancer as an "Action Plan" against media stereotypes and biases. However, AC1 had some reservations about the potential misuse of this tool. He asked *"what if someone uses Screen-Balancer to increase bias in the produced video?"* There is no straightforward answer to this question, but probably the best answer is to increase the consciousness among content creators.

7 DISCUSSION AND LIMITATIONS

7.1 Impact of Screen-Balancer on Gender and Skin-tone Biases in Media

Biases are deeply rooted in society and are reflected in both the structure and the creation of visual media, where minority groups are often neglected—from writing scripts to casting actors, to directing, to designing costumes, to portraying characters (a complete list for Hollywood entertainment industry is available in [37]). Given these complexities, it is unlikely that a single solution will fully solve the overall bias problem in visual media. Therefore, we consider Screen-Balancer as a probe for reducing biases from visual media and a catalyst for future efforts in this direction. As an example, our participants successfully used Screen-Balancer to balance the gender presence in one of the more challenging telecasts in our study, a music show that had a highly imbalanced gender ratio ($M : F = 6 : 1$). Further, the design of Screen-Balancer can be extended to several different production scenarios; we list three of these in the following:

First, Screen-Balancer could be integrated into the recording phase of a movie since movie directors use *Monitor*, an interface similar to Switcher, to observe a particular *take*, defined as a single continuous acting performance. A take is usually recorded multiple times. Screen-Balancer could visualize the presence of different phenotypic groups in each take in real-time in the director's monitor. A similar process can be carried out in other editing and post-production software.

Second, Screen-Balancer could be extended to include contextual factors such as postures, gestures, speech time, and emotion to gauge the actor portrayals quantitatively. This will require additional computer vision and natural language recognition techniques to estimate these factors.

Finally, while quantitative metrics are useful in understanding some dimensions of biases quickly, visual media often portray minority groups in insignificant contexts (discussed in Section 2), which are too subtle to be measured quantitatively. Video producers and directors could share the contents and the quantitative measures obtained from Screen-Balancer to a focus group to gather qualitative reviews.

7.2 Limitations of Automatic Phenotypic Recognition and Mitigation Strategies

The automatic phenotypic recognition systems have several limitations, such as dependency on lighting conditions that affect actors' skin-tone in video frames, not recognizing transgender and other minority groups [42, 57], and not being fair in some settings [4, 7, 32, 38, 40]. Although we evaluated the fairness of the models used in our system, the analysis was not exhaustive as the definition and measurement of fairness vary from system to system [44].

One mitigation strategy would be to bypass the phenotypic recognition systems and let the actors self-identify themselves in advance based on their gender and racial preferences. Screen-Balancer would then track the actors during the telecast, and utilize the self-identification labels to detect their phenotypes. In addition, the accuracy of facial recognition could be boosted by applying Active Learning [59] or One-shot Learning [23] on the images provided by the actors during the self-identification phase.

Such a tool, in practice, may pose challenges (such as labeling the performers beforehand) in live telecasts that feature many actors, but we believe it has the potential to increase the diversity and fairness of the system as the tool would be able to recognize all gender and racial identities alike with high accuracy. This approach also has the potential to include the actors in the overall bias balancing process and enable the actors to provide consent to their images being analyzed.

As a final note, we acknowledge that using facial recognition could be a concern from a policy standpoint since Screen-Balancer empowers the producers, rather than the actors appearing on the screen. The self-identification approach discussed above and an appropriate collaboration between the producers and the actors could minimize such concerns.

7.3 Cognitive Load and Video Coherency

None of the professional video producers we interviewed found Screen-Balancer overwhelming. To put that into perspective, producing live telecasts demands substantial cognitive workload, as the producers need to switch between multiple cameras in realtime to create a coherent show. We note that some of our study participants (particularly, the YouTubers) were not familiar with the *Switcher* interface, which might had contributed the issues of video coherency and cognitive load.

It is possible that the cognitive load using Screen-Balancer could increase in a highly dynamic set, such as in a live telecast with 15-20 cameras. But given that the screen time is a dimension that video producers actively moderate on their own, it is also possible that Screen-Balancer could actually decrease the workload in such scenarios.

One way to decrease the cognitive work load would be to increase the broadcast delay, which is currently set as 10 seconds. Although a 7- to 10-second delay is usual, longer delays lasting for several minutes are also common. Increasing this delay would give the producers more time to make a switching decision.

8 CONCLUSION

We presented Screen-Balancer — an interactive visual analytics tool that assists content moderators to balance phenotypic human traits in a live telecast. Our system can be extended and integrated into any production system, and it can also be generalized to moderate other types of visual characteristics. Our methodology tackles a difficult problem as screen time of different gender and skin-tone has not yet been acknowledged to be moderated by the production systems in use. Screen-Balancer can bridge this gap.

ACKNOWLEDGMENTS

This research was partially supported by NSF grants IIS 1527200 and IIS 1941613. We thank Jan (Dini) Diskin-Zimmerman and the members of Stony Brook Video Production Team for their constant support throughout the development of this work. We also thank the anonymous reviewers for their thoughtful reviews.

REFERENCES

[1] 2019. UN Women. https://www.unwomen.org/en Accessed: 2019-04-03.
[2] 2020. *Adobe After Effects.* https://www.adobe.com/products/aftereffects.html Accessed: 2020-05-20.
[3] 2020. *Adobe Photoshop.* https://www.adobe.com/products/photoshop.html Accessed: 2020-05-20.
[4] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 199.
[5] Apple. 2018. *Use Face ID on your iPhone or iPad Pro.* https://support.apple.com/en-us/HT208109
[6] Jennifer Stevens Aubrey. 2006. Exposure to sexually objectifying media and body self-perceptions among college women: An examination of the selective exposure hypothesis and the role of moderating variables. *Sex Roles* 55, 3-4 (2006), 159–172.
[7] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. *NIPS Tutorial* (2017).
[8] Avi Ben-Zeev, Tara C. Dennehy, Robin I. Goodrich, Branden S. Kolarik, and Mark W. Geisler. 2014. When an "Educated" Black Man Becomes Lighter in the Mind's Eye: Evidence for a Skin Tone Memory Bias. *SAGE Open* 4, 1 (2014), 2158244013516770. https://doi.org/10.1177/2158244013516770 arXiv:https://doi.org/10.1177/2158244013516770
[9] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.
[10] Jane D Brown, Carolyn Tucker Halpern, and Kelly Ladin L'Engle. 2005. Mass media as a sexual super peer for early maturing girls. *Journal of adolescent Health* 36, 5 (2005), 420–427.
[11] Jane D Brown, Kelly Ladin L'Engle, Carol J Pardun, Guang Guo, Kristin Kenneavy, and Christine Jackson. 2006. Sexy media matter: exposure to sexual content in music, movies, television, and magazines predicts black and white adolescents' sexual behavior. *Pediatrics* 117, 4 (2006), 1018–1027.

[12] Jane D Brown and Carol J Pardun. 2004. Little in common: Racial and gender differences in adolescents' television diets. *Journal of Broadcasting & Electronic Media* 48, 2 (2004), 266–278.

[13] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[14] Lisa Charlotte. 2018. *An alternative to pink & blue: Colors for gender data.* https://blog.datawrapper.de/gendercolor/

[15] William S Cleveland and Robert McGill. 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association* 79, 387 (1984), 531–554.

[16] Cheryl Cooky, Michael A Messner, and Michela Musto. 2015. "It's dude time!" A quarter century of excluding women's sports in televised news and highlight shows. *Communication & Sport* 3, 3 (2015), 261–287.

[17] Nicholas G Crawford, Derek E Kelly, Matthew EB Hansen, Marcia H Beltrame, Shaohua Fan, Shanna L Bowman, Ethan Jewett, Alessia Ranciaro, Simon Thompson, Yancy Lo, et al. 2017. Loci associated with skin pigmentation identified in African populations. *Science* 358, 6365 (2017), eaan8433.

[18] Kelly K Davis and CA Tuggle. 2012. A gender analysis of NBC's coverage of the 2008 summer Olympics. *Electronic News* 6, 2 (2012), 51–66.

[19] Travis L Dixon. 2017. A dangerous distortion of our families: Representations of families, by race, in news and opinion media. *Color of Change and Family Story research report, December. https://colorofchange. org/dangerousdistortion* (2017).

[20] Travis L Dixon. 2017. Good guys are still always in white? Positive change and continued misrepresentation of race and crime on local television news. *Communication Research* 44, 6 (2017), 775–792.

[21] Travis L Dixon, Erica Bauer, and Christopher Josey. 2018. How News Frames Prime Non-Whites as Social Problems. In *Doing News Framing Analysis II*. Routledge, 343–361.

[22] Eran Eidinger, Roee Enbar, and Tal Hassner. 2014. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security* 9, 12 (2014), 2170–2179.

[23] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence* 28, 4 (2006), 594–611.

[24] S Few. 2010. Bullet graph design specification. *Perceptual Edge-White Paper* 5 (2010).

[25] Stephen Few and Perceptual Edge. 2007. Save the pies for dessert. *Visual Business Intelligence Newsletter* (2007), 1–14.

[26] Thomas B Fitzpatrick. 1988. The validity and practicality of sun-reactive skin types I through VI. *Archives of dermatology* 124, 6 (1988), 869–871.

[27] BL Fredrickson and T Roberts. 1997. Toward understanding women's lived experiences and mental health risks. *Psychology of Women Quarterly* 21, 2 (1997), 173–206.

[28] Reuma Gadassi and Itamar Gati. 2009. The effect of gender stereotypes on explicit and implicit career preferences. *The Counseling Psychologist* 37, 6 (2009), 902–922.

[29] Silvia Galdi, Anne Maass, and Mara Cadinu. 2014. Objectifying media: Their effect on gender role norms and sexual harassment of women. *Psychology of Women Quarterly* 38, 3 (2014), 398–413.

[30] Arthur H Goldsmith, Darrick Hamilton, and William Darity Jr. 2006. Shades of discrimination: Skin tone and wages. *American Economic Review* 96, 2 (2006), 242–245.

[31] Google. 2019. *Learn about familiar face detection and how to manage your library.* https://support.google.com/googlenest/answer/9268625

[32] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[33] Matthew S Harrison and Kecia M Thomas. 2009. The hidden prejudice in selection: A research investigation on skin color bias. *Journal of Applied Social Psychology* 39, 1 (2009), 134–168.

[34] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage Publications Sage CA: Los Angeles, CA, 904–908.

[35] Joni Hersch. 2011. The persistence of skin color discrimination for immigrants. *Social Science Research* 40, 5 (2011), 1337–1349.

[36] Md Hoque, Rawshan E Fatima, Manash Kumar Mandal, Nazmus Saquib, et al. 2017. Evaluating gender portrayal in Bangladeshi TV. *arXiv preprint arXiv:1711.09728* (2017).

[37] Darnell Hunt and Ana-Christina Ramón. 2020. *Hollywood Diversity Report 2020.* https://socialsciences.ucla.edu/wp-content/uploads/2020/02/UCLA-Hollywood-Diversity-Report-2020-Film-2-6-2020.pdf

[38] Jevan A Hutson, Jessie G Taft, Solon Barocas, and Karen Levy. 2018. Debiasing desire: addressing bias & discrimination on intimate platforms. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 73.

[39] Ji Yoon Jang, Sangyoon Lee, and Byungjoo Lee. 2019. Quantification of Gender Representation Bias in Commercial Films based on Image Analysis. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 198.

[40] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*. 325–333.

[41] Vahid Kazemi and Josephine Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1867–1874.
[42] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 88.
[43] Razib Khan. 2019. *Skin color is not race.* https://www.discovermagazine.com/the-sciences/skin-color-is-not-race
[44] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.
[45] Martha M Lauzen. 2017. Boxed in 2016-17: Women on screen and behind the scenes in television. *Center for the Study of Women in TV & Film. Disponível em: https://womenintvfilm. sdsu. edu/wp-content/uploads/2017/09/2016-17_Boxed_In_Report. pdf. Acesso em* 2 (2017).
[46] Gil Levi and Tal Hassner. 2015. Age and Gender Classification Using Convolutional Neural Networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) workshops*. http://www.openu.ac.il/home/hassner/projects/cnn_agegender
[47] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.
[48] Nishtha Madaan, Sameep Mehta, Taneea Agrawaal, Vrinda Malhotra, Aditi Aggarwal, Yatin Gupta, and Mayank Saxena. 2018. Analyze, detect and remove gender stereotyping from bollywood movies. In *Conference on Fairness, Accountability and Transparency*. 92–105.
[49] Sarah K Murnen, Carrie Wright, and Gretchen Kaluzny. 2002. If "boys will be boys," then girls will be victims? A meta-analytic review of the research that relates masculine ideology to sexual aggression. *Sex roles* 46, 11-12 (2002), 359–375.
[50] Geena Davis Institue on Gender in Media. 2015. *The Reel Truth: Women Aren't Seen or Heard.* https://seejane.org/research-informs-empowers/data/
[51] Geena Davis Institue on Gender in Media. 2017. *The Geena Benchmark Report: 2007-2017.* https://seejane.org/wp-content/uploads/geena-benchmark-report-2007-2017-2-12-19.pdf
[52] Geena Davis Institue on Gender in Media. 2019. *Historic Gender Parity in Children's Television.* https://seejane.org/research-informs-empowers/see-jane-2019/
[53] Geena Davis Institue on Gender in Media. 2020. *Historic Gender Parity in Family Films!* https://seejane.org/2020-film-historic-gender-parity-in-family-films/
[54] Karen Ross. 2010. *Gendered media: Women, men, and identity politics*. Rowman & Littlefield.
[55] Tara Ross. 2019. Media and Stereotypes. *The Palgrave Handbook of Ethnicity* (2019), 1–17.
[56] Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. 2018. Fairness gan. *arXiv preprint arXiv:1805.09910* (2018).
[57] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 144.
[58] Ari Schlesinger, W Keith Edwards, and Rebecca E Grinter. 2017. Intersectional HCI: Engaging identity through gender, race, and class. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 5412–5427.
[59] Burr Settles. 2009. *Active learning literature survey*. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.
[60] MJ Sutton, Jonathon D Brown, KM Wilson, and JD Klein. 2002. 'Shaking the Tree of Knowledge for Forbidden Fruit. *Sexual teens, sexual media: Investigating media's influence on adolescent sexuality* (2002), 25–55.
[61] Tom FM Ter Bogt, Rutger CME Engels, Sanne Bogers, and Monique Kloosterman. 2010. "Shake it baby, shake it": Media preferences, sexual attitudes and gender stereotypes among adolescents. *Sex roles* 63, 11-12 (2010), 844–859.
[62] TSA. 2018. *Biometrics Technology | Transportation Security Administration.* https://www.tsa.gov/biometrics-technology
[63] UNESCO. 2019. Gender-Sensitive Indicators for Media. http://www.unesco.org/new/en/communication-and-information/crosscutting-priorities/gender-and-media/gender-sensitive-indicators-for-media Accessed: 2019-04-03.
[64] Thanh Vo, Trang Nguyen, and C Le. 2018. Race recognition using deep convolutional neural networks. *Symmetry* 10, 11 (2018), 564.
[65] L Monique Ward. 2002. Does television exposure affect emerging adults' attitudes and assumptions about sexual relationships? Correlational and experimental confirmation. *Journal of youth and adolescence* 31, 1 (2002), 1–15.
[66] L Monique Ward. 2003. Understanding the role of entertainment media in the sexual socialization of American youth: A review of empirical research. *Developmental review* 23, 3 (2003), 347–388.
[67] Julia T Wood. 1994. Gendered media: The influence of media on views of gender. *Gendered lives: Communication, gender, and culture* 9 (1994), 231–244.

Received January 2020; revised June 2020; accepted July 2020