Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-023-02040-y OPEN

Ethnic representation analysis of commercial movie posters

Dima Kagan [1,3], Mor Levy[2,3], Michael Fire[1]✉ & Galit Fuhrmann Alpert[1]✉

In the last decades, global awareness toward the importance of diverse representation has been increasing. The lack of diversity and discrimination toward minorities did not skip the film industry. Here, we examine ethnic bias in the film industry through commercial posters, the industry's primary advertisement medium for decades. Movie posters are designed to establish the viewer's initial impression. We developed a novel approach for evaluating ethnic bias in the film industry by analyzing nearly 125,000 posters using state-of-the-art deep learning models. Our analysis shows that while ethnic biases still exist, there is a trend of reduction of bias, as seen by several parameters. Particularly in English-speaking movies, the ethnic distribution of characters on posters from the last couple of years is reaching numbers that are approaching the actual ethnic composition of the US population. An automatic approach to monitoring ethnic diversity in the film industry, potentially integrated with financial value, may be of significant use for producers and policymakers.

[1] Ben-Gurion University of the Negev, Beer-Sheva, Israel. [2] Afeka College of Engineering, Tel Aviv, Israel. [3] These authors contributed equally: Dima Kagan, Mor Levy. ✉email: mickyfi@bgu.ac.il; fuhrmann@bgu.ac.il

Introduction

The film industry often portrays a utopian image: cocktail parties, glorious scenes shot behind magnificent views, and movie stars. This ideal image generates appreciation and worship from the general audience toward the industry and its stars. These affections have been effectively harnessed for profitable advertising over the years (Kaikati, 1987; Kamins, 1990). Apart from the obvious financial factors of movie advertisements, they also play a critical role in shaping cultural and societal perceptions, as films present an image of society that viewers often perceive as realistic (Carroll, 1985). This image is influenced by various aspects, such as cast selection, key concepts, and visual stimulation, all of which may depict a reality that differs significantly from our day-to-day lives. Film content can influence people's actions in real-life, particularly children, who are more susceptible to such influence (Albert, 1957; Beaufort, 2019; Nieman, 2003). This shaping of mind may have a more significant impact than we realize, especially concerning diversity awareness.

With the recent rise of the "Black Lives Matter" movement, diversity awareness has gained worldwide attention (BBC News, 2020; The New York Times, 2020), highlighting the significant lack of ethnic diversity in one of the largest and most influential countries in the world, the USA. Recent reports indicate that similar issues are also evident specifically in the film industry (Hunt and Ramón, 2020; MarketWatch, 2020). MarketWatch, for example, reported that the lack of non-White lead characters in Hollywood films is driven by economic concerns since, in some movie studios, having African–American leads is considered non-profitable when distributing films overseas (MarketWatch, 2020).

Furthermore, movie advertisements, such as posters and trailers, can reach people who did not watch the actual film, exposing both viewers and non-viewers to the movie. Consequently, movie posters have a particularly high impact on shaping society's perception. While researchers have studied posters in the context of diversity (Aley and Hahn, 2020; Freire, 2019; Gabriel, 2012; Rahmasari, 2014), most of the studies conducted to date have been small-scale and involved manual annotation of posters. Studying specific case studies is important for gaining insights into the problem, yet it does not address generalization and identifying complex trends.

In this study, we aim to quantify diversity representation in the film industry using movie posters. We hypothesized that we would observe trends of improved representation of minorities in posters as have been found by other studies (McKinsey, 2021; Smith et al., 2020). In other words, posters would contain the underrepresentation of minorities, gradually shifting from a completely biased representation to a more balanced one. Moreover, we also speculated that the design aspects of the poster, such as size and location, would be affected by the actor's ethnicity, where minorities would be smaller and in less central locations.

To verify this hypothesis, we addressed the following questions:

Question 1: How has the minority representation on posters evolved over time?

Question 2: Does ethnicity play a role in the location and size of movie actors on posters?

Question 3: Does the ethnicity of the leading actor also impact who else appears on the poster?

Question 4: Are minorities equally represented on movie posters of different genres?

Question 5: Are there explicit indications that minorities were specifically added to posters in recent years in order to diminish biases?

To address these goals, we developed a novel approach to automatically analyze and quantify diversity in movie posters (see Fig. 1). Our posters dataset consists of nearly 125,000 unique posters that we collected from over 35,000 movies of various genres, with almost two-thirds of them featuring actors. The data was collected from IMDb and TMDB online movie databases. We used state-of-the-art machine learning-based algorithms in order to identify patterns of interest that can shed light on ethnic diversity through several steps, as follows. For each poster, we first apply a face detection algorithm to recognize actors, followed by face embedding to match actors to their actual identities provided in IMDb and TMDB photos. We then apply deep learning to extract actors' ethnicity from their photos. These analysis steps are used to generate a large-scale annotated poster dataset and an open-source code framework that we offer as publicly available resources for poster analysis.

The results of our large-scale analysis indicate that until recently White Caucasian actors were over-represented on posters compared to their relative numbers in the US population, especially in leading roles. While in the last 20 years, a tendency for improvement in the representation of ethnic minorities is observed, there are nevertheless over 79% White Caucasian actors depicted on movie posters, regardless of the genre (see Fig. 7), a fraction which is 1.14 higher than their relative demographic percentage in the total population. We also found that non-White actors are more likely to be placed on a poster in minor roles. Out of the top-3 actors from the cast list on posters, only 9.2% were minorities. In more minor roles (top 4–12), 16.2% were non-White minority actors.

That being said, we will show that today the film industry is actually taking into account the ethnic diversity of the US population when creating posters. Posters of movies from the last 2 years are in fact perfectly balanced according to the US population compound. We propose that our approach could be used to continuously monitor ethnic representation in the film industry in an automated fashion.

The key contributions presented in this paper are five-fold:

- We present a novel method that we developed in order to evaluate ethnic diversity automatically.
- This is the first computer vision-based study that uses poster images to analyze ethnic diversity.
- The described study is the most comprehensive study to date that utilizes posters to analyze ethnic representation in the film industry over decades.
- We curated and offered the largest publicly available poster dataset, with 125,439 posters of 35,000 movies. The dataset consists of public metadata, as well as analyzed metadata that we automatically generated by face recognition and ethnicity detection of the identified actors' faces in each poster.
- We provide an open-source framework developed for movie poster analysis. Our framework could be used to generate metadata about movie posters, and facilitate research by creating and analyzing larger amounts of data than has been available ever before.

Related work

The lack of diverse representation in the film industry is not unique to ethnic minorities. Similar under-representation may also be observed for other minority groups, including female actors, sexual minorities, religious groups, or various disability groups.

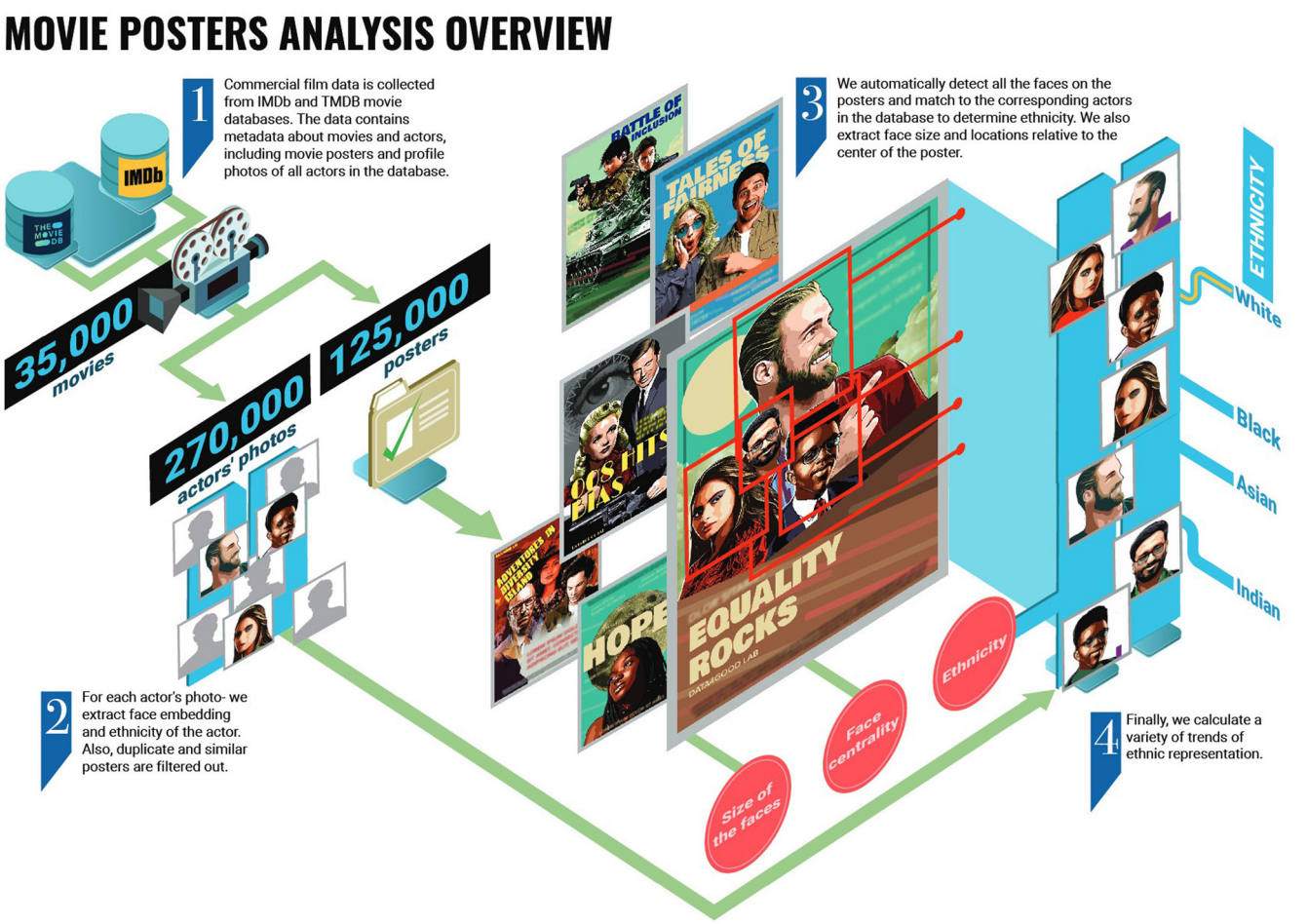

Fig. 1 Movie posters analysis overview.

Ethnic under-representation in the film industry. Ethnic under-representation is a prevalent issue in the film industry, which is part of a broader diversity problem (White, 2019). Lee (2014) studied this notion by investigating how a movie's racial composition affects its performance. The results indicate no significant differences in performance between films that star White versus non-White leading actors. Therefore, with the financial excuse out of the picture, it appears that the phenomenon of casting out minorities from lead roles in top movies results from the lack of diversity awareness.

According to Smith et al. (2014) in 600 popular films made between 2007 and 2013, only about a quarter (25.9%) of the speaking characters were from minority ethnic groups. Those characters are representatives of ethnic groups that comprise 37% of the US population and purchase 46% of movie tickets, thus are obviously underrepresented. Moreover, when looking at the next generation to come, nearly 50% of children under the age of 5 in the US are not White, meaning that both the current and future audience for films is much more diverse than what is shown on screen. Smith et al. (2014) also illustrated that existing cultural stereotypes may still influence how characters from different backgrounds are portrayed. For example, Hispanic females tend to be hypersexualized more often than females from other ethnic groups (Smith et al., 2014).

A more recent study by Smith et al. (2021) focused on inclusion in Netflix-produced films scripted in the US. They found that Netflix has made significant progress in representing minorities in its self-produced content. In terms of leads/co-leads in 2019, on average, they reached proportional representation. However, not all groups saw an increase in representation from 2018 to 2019, including Hispanic/Latino and Middle Eastern/North African groups.

Hennekam and Syed (2018) interviewed different workers in the film industry describing how they quickly realized that being "different" from the mainstream is a barrier to career advancement. Non-White women, for example, face significant obstacles to success due to the absence of role models in the form of successful ethnic minority women in the industry (Buunk et al., 2007). Smith et al. (2020) results strengthen this conclusion. They show that most decision-making on the filming set is made by white males. With such a lack of diversity, it should not be surprising if women and ethnic minorities are underrepresented in movies and advertisement media such as trailers and posters.

According to the common notion among directors in the industry, casting minorities in lead roles is considered a financial risk (Lee, 2014). Therefore, the ethnic background of the director may impact the actual casting decisions as—"Black directors conversely cast Black characters" (Smith et al., 2014). Director Ridley Scott, who casted White actors for the top leading roles in a story featuring Egyptian characters, justified his casting: "I can't mount a film of this budget, where I have to rely on tax rebates in Spain, and say that my lead actor is Mohammad so-and-so from such-and-such […] I'm just not going to get it financed. So the question doesn't even come up" (Lee, 2014).

Exploring race and gender biases using posters. Posters in general and movie posters specifically are a form of visual media

advertisements that contain various explicit and implicit data on movies, potentially influencing the audience. Over the past couple of years, posters have been analyzed for various studies on bias and discrimination.

Posters are used to advertise and promote various topics and ideas. The design of a visual advertisement is a complex process that can have various biases and side effects. For instance, posters were found to be an effective means to influence public opinion in fighting stigmas and discrimination related to HIV/AIDS (Johnny and Mitchell, 2006). However, Johnny and Mitchell (2006) also note that even though posters successfully improved the image of HIV/AIDS, they still reinforce some stigmas. De Run (2005) also shows how challenging it is to design an advertisement dealing with ethnicity. De Run (2005) found that targeting specific ethnic groups in advertisements using ethnic content positively affects the targeted group. However, the same campaign created negative reactions from other ethnic groups.

Baumgartner and Laghi (2012) explored how positive and negative content on movie posters affects adolescents and young adults. They found that highly positive images are more effective with young adults, while adolescents respond similarly in terms of affective responses to positive and negative imagery. Their results suggest that unlike what De Run (2005) found in regular advertisements, it is not necessarily behaving the same on movie posters. Portraying something negative on movie posters, such as ethnic groups, may not necessarily reduce the advertisement's effectiveness.

Aley and Hahn (2020) manually inspected the main characters in movie posters of 152 popular US animated children's films produced over the last 80 years. Their findings revealed that main characters were more likely to be male and males were portrayed as more powerful. Similarly, Gabriel (2012) conducted a human-based examination of nearly 150 posters, all of which were on the yearly top 30 movies, as measured by gross box office profit (2007–2011). One of their key findings was that male characters outnumbered female characters by a ratio of three to one. This was accompanied by a bias toward male characters in the fundamental composition of the posters.

Other studies focused on specific ethnic groups through related posters. For example, Freire (2019) investigated the current representation of the Latin American identity in mainstream media cinematic posters. Freire (2019) utilized Latin American visual design language to reinterpret the possibilities that film posters have in creating elaborate narratives that treat audiences with respect and complexity. Rahmasari (2014) analyzed the posters of "The Help," a movie that tells a story about racism. He found that the posters of this film represent a notion where "White people as the best race between others is pictured in the fifth of The Help movie posters".

Using machine and deep learning to study diversity. The current literature confirms that movie posters contain essential information about various topics (Aley and Hahn, 2020; Freire, 2019; Gabriel, 2012; Rahmasari, 2014). Moreover, how they are designed can affect different groups in various ways (Baumgartner and Laghi, 2012). However, only a few studies have relied on movie posters as a data source, and those studies have mostly been small-scale and relied on manual annotation processes. Movie posters are unstructured data in the form of imagery, which makes it challenging and expensive to study them at a large scale using traditional methods. Therefore, new methods are required to suit the current age and allow for studying fundamental societal problems using movie posters.

To conduct a deep and broad study of the topic of ethnic diversity in the film industry, we developed and applied data science tools, including machine learning and deep learning. For these methods to work efficiently and accurately large datasets are needed with individuals labeled by their ethnic origin. However, a recent study has shown that most existing large-scale face databases are biased toward "lighter skin" faces (around 80%), such as., White, compared to "darker" faces, e.g., Black (Merler et al., 2019). Biased training data is more likely to produce biased models that are trained on it (Mehrabi et al., 2021). This in turn raises ethical concerns about the fairness of those models, which may eventually affect decision-makers (Fletcher et al., 2021).

Therefore, when using machine and deep learning models to study problems such as diversity and bias, it is crucial to evaluate the models being used, even if they were created by someone else. Heavily biased models can produce results that do not represent reality but rather the model's bias.

Methods and experiments

Our main goal of the study is to examine how different ethnic populations are represented in the film industry and whether there are objective indications of inequality and under-representation. We analyzed movie poster images from multiple genres to explore the current state of ethnic diversity and to test whether representation diversity has changed over time. In the scope of this study, we focus on English-speaking movies produced in the US (referred to as English-speaking movies).

Datasets. We curated a large dataset of movie data and poster images, which were used for all of the analyses further described in this paper. We collected the images and metadata from publicly available online sources by the process described below. The data described in this section contains data both for English and non-English speaking movies. While for the analysis, we used the data related to English-speaking movies and released the whole data and source code to help advance additional studies in the field.

Movies dataset. To study bias as it is presented on movie posters, we assembled large-scale datasets of movies using the following publicly available data sources:

- *IMDb* (IMDB, 2022a) is an online database of information related to films, television programs, home videos, video games, and streaming content online—including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews. As of August 2021, IMDb has ~8 million titles (including episodes) and 11 million personalities in its database. It also contains over 580,000 movie records, along with 1,636,604 poster images.

  We downloaded IMDb's open movies dataset (IMDB, 2022b), selecting the following information:

  – General movie metadata such as primary title, genres, and release year.
  – Movies rating metadata such as average ratings and vote count.

  We then filtered the dataset to contain only non-animated movies rated by at least 1000 reviewers, thus obtaining a dataset that consists of about 35,000 movies. Of these 35,000 movies, 17,187 are English-speaking movies used in this study. Additionally, the dataset contains 13,286 non-English foreign movies; the rest are English-speaking movies produced outside the US.
- *TMDB* (TMDB, 2022a) is a that offers high-resolution posters in addition to the movie's metadata. As of August 2021, over 1000 images are added every single day on average. TMDb officially supports 39 languages along with extensive regional data and every single day TMDb is used

Fig. 2 An example depicting the complexity of directly inferring the ethnicity of actors from poster images. On the right: The actress Jennifer Aniston appears in a well-taken headshot profile photo. This figure is covered by the Creative Commons Attribution 4.0 International License. Reproduced with permission of gdcgraphics; gdcgraphics © NAME, all rights reserved. On the left: the same actress appears in an image where her face and skin color have been colorized. The image is based on the style of an actual poster of the movie "Blade Runner 2049" and is used not to violate copyrights. This figure is based on a work of an employee of the Executive Office of the President of the United States, taken or made as part of that person's official duties. As a work of the U.S. federal government, it is in the public domain.

in over 180 countries. As of August 2021, TMDb contains 683,671 movie records with 2,873,601 images overall. We used TMDb to obtain data about the movie's country of origin and posters and actor images for the movies obtained from the IMDb dataset.

Posters dataset. For each movie in the IMDb database, we fetched all its official posters from the TMDB database[1], and the main poster from IMDb using their APIs (Alberani, 2021; TMDB, 2022b). This process yielded a total of 286,654 posters. Manual inspection of random movies suggested that many posters of a particular movie were duplicates of the same image, except for the language of the text (title and actor's names). To overcome these duplicates, we applied the dhash algorithm(Dr. Neal Krawetz Kind of like that—the hacker factor blog, 2013) to identify visually similar images. For instance, many duplicate posters contained the same graphics with the title in a different location. We calculated each image's dhash value and computed the Hamming distance (Hamming, 1950) between all pairs of images of a specific movie. We removed all images with a distance smaller than 16 - an empirically set threshold.[2] This step resulted in the final dataset of 125,439 non-duplicate posters that we used for this study. Out of these 125,439 posters, 72,971 posters are of English-speaking movies.

Actor dataset. For each movie in the movie's dataset, and each actor in the movie's cast list, we used IMDb API to fetch the actor's name, the identifier "imdb id" and credits ranking (position in the cast list). Next, for each actor in the Actor dataset, we used IMDb and TMDB APIs to download up to three main profile pictures. We subsequently integrated all collected images with the collected related metadata into one unified actor dataset containing 118,136 actors and 217,575 related images of the actors. We filter all grayscale photos by calculating the mean squared error for each channel from the average pixel value. We filter grayscale images since it is harder to utilize them to identify the actor's ethnicity. After filtering grayscale images, the actor dataset contained 101,873 actors and 179,858 actor images.

US demographics dataset. The US demographic data is collected as part of the US census (United States Census Bureau, 2022) which takes place every 10 years. In terms of ethnicity, five categories are considered: White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian, or Other Pacific Islander. Since the census considers Indian people as Asian (as being part of the continent Asia), we accordingly merged our collected data of Indian and Asian when using the census data.

Feature extraction by image processing of the collected posters. In order to extract features representing actors in posters, we used image processing of posters and profile images of the actors in the following manner:

1. *Posters face detection*: For each poster image, we applied the RetinaFace face detection algorithm (Deng et al., 2020). We filtered out posters that did not contain at least one face, eventually holding a dataset of 77,192 poster images (about 60% of the collected posters contained at least one face image). Out of these posters, 45,613 belongs to English-speaking movies.
2. *Actors face detection*: For each actor in the established actor dataset, we applied the RetinaFace face detection algorithm (Deng et al., 2020) to identify the actor's face from the profile pictures. We then embedded each face using Arcface (Deng et al., 2019). Face embedding allows performing face verification by matching faces between multiple photos.
3. *Actors ethnic recognition*: For each poster, our goal was to determine the ethnicity of the actor faces depicted in the poster. However, using actor's faces from the poster is challenging and does not always yield a good classification. This is because advertisers that create the posters usually use visual methods to engage potential viewers, methods which include unnatural backgrounds, mixed colors, and faces positioned in unclear angles (see Fig. 2). We, therefore, flipped the question—for each poster, we first recognized the participating actors and classified their

ethnicity instead of ethnic-classifying the actual face we detected in the poster.
The ethnic classification was performed using FairFace-based models (Karkkainen and Joo, 2021). FairFace is a novel face image dataset with a balanced race composition for seven race groups: White, Black, Indian,[3] East Asian, Southeast Asian, Middle East, and Latino. This dataset enables better generalization and performance of classification for gender, race, and age, compared to model training (such as UTK (Zhang et al., 2017), LFWA (Liu et al., 2015), and CelebA (Liu et al., 2015)) on existing large-scale in-the-wild unbalanced datasets. FairFace provides two models of ethnic classification, one of which classifies faces into four[4] and the other into seven[5] racial groups.
For each profile picture, we computed the actor's ethnic scores. The models provide an ethnic score of an actor per ethnic group (the probability of being part of the ethnic group). Since we collected up to three profile pictures per actor, we averaged the ethnic scores computed per each profile picture and then selected the max averaged score as the voted ethnic group.

4. *Actors recognition in posters*: To match between posters and actors, we encoded each detected face (both from the actor's headshots and the posters) by utilizing the ArcFace (Deng et al. 2019). Using ArcFace, we encode all the faces detected in the posters and the actor's photos. Once we had all the faces encoded, we matched the encodings of the faces detected in the posters to the database of actor face encodings. We searched for the closest actor face by Euclidean distance. Then we evaluated two approaches for matching posters with actors (see the section "Poster actor matching").

Evaluation of the models in use. To evaluate the performance of the machine learning models used to generate the datasets for our study, we performed the following steps:

1. *Face detection algorithm*: To validate face detection, we manually inspected a random sample of posters containing a total of over 100 faces.
2. *Face recognition*: To validate the performance of the face recognition algorithm, we randomly selected 50 posters to extract 149 faces. For every face, we created a new picture as a collage of 2 images:

   (a) The original poster with a painted rectangle around a specific detected face.
   (b) The headshot of the actor who was matched to the detected face.

   We then manually evaluated the matches.
3. *Ethnic classification algorithm:* To validate FairFace (Karkkainen and Joo, 2021) performance on the data used in this study, we selected random images of actors and actresses from each race category considered in our models. Using IMDb website, we verified their ethnic origin and tagged them accordingly. A total of 40 actors were selected for evaluating the four-race model and 70 actors for the seven-race model (five actors and five actresses for each race category). We used the race predictions of the actors using the Ethnic Recognition algorithm (see Section Feature Extraction by Image Processing of the Collected Posters) and manually evaluated the predictions.

Poster actor matching. We evaluated two approaches for matching face to an actor:

Table 1 4-classes model evaluation.

Race	Precision (%)	Recall (%)
Asian	100.0	100.0
Black	100.0	90.0
Indian	100.0	70.0
White	71.42	100.0

Table 2 7-classes model evaluation.

Race	Precision (%)	Recall (%)
Black	100.0	70.0
East Asian	60.0	90.0
Southeast Asian	66.66	20.0
White	55.55	100.0
Indian	66.66	40.0
Latino Hispanic	30.76	40.0
Middle Eastern	50.0	40.0

1. Compare each face found in the poster to the whole cast list of the movie.
2. Compare each face found in the poster to only the top-10 actors from the movie's cast list.

Our line of thought for the second approach was that actors who appeared on the posters are most likely amongst the top-10 of the cast list; therefore, we believed the second approach would be more effective. We quantified the number of faces we detected across the posters and found that, on average, there are 3.81 actors on a poster with a standard deviation of 4.18. This indicates that most posters show actors from the top-10 most central roles. Hence, comparing faces to the complete cast list may add noise and cause matching errors. We validated this hypothesis (see the section "Evaluation of the model in use") and showed that it was incorrect - the approach of the whole cast list was better in terms of accuracy. We noticed that due to damaged data in IMDb's datasets, the order of the cast list in some movies was arbitrary and main actors who appeared in the posters were not among the first 10 actors. Moreover, Arcface detection was so accurate that trying to match actors from the bottom of the list did not add noise as we first suspected.

Analysis

To explore ethnic diversity from a new angle, we analyzed more than 45,000 posters of 24,062 English-speaking movies produced in the US from the 60s up to 2021. We searched for ethnic-related trends in actor casting across decades. We focused on data from the 60s since this area marked a change in the rights of minorities in the US with the legislation of the Civil Rights Act Legal Highlight: The Civil Rights Act of 1964 | U.S. Department of Labor (n.d.). The results of our analysis are as follows.

First, we evaluated the ML models used in this study. The RetinaFace face detection algorithm correctly detected 100% of the sampled faces. To match posters to actors, we have evaluated two matching approaches (see the section "Poster actor matching"), top-10 and all actors approaches. Both approaches for matching posters yielded 100% verification (every face was automatically matched to the correct actor, verified manually). The two methods differed slightly in their identification score (number of matched faces out of the total detected faces): 71% for the whole cast list approach and 69% for the top-10 approach. Evaluating ethnic classification methods (see Tables 1 and 2), we found that the four-race model provided an average precision of 92.85%, while the seven-race model only had an average precision of 61.37%.

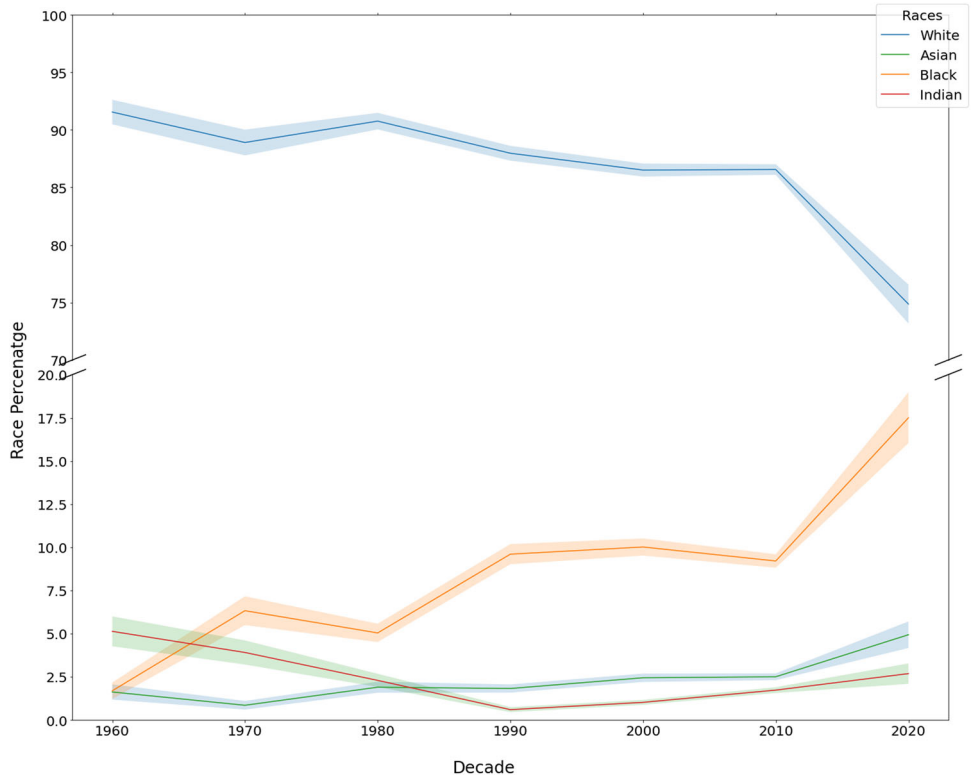

Fig. 3 Trends in the relative frequency of ethnic representation in posters of English-speaking movies. The relative representation of actors from each ethnic group is displayed as the percentage of ethnic face appearance compared to the overall number of faces detected.

Having tested the models, we explored race appearance trends on posters.[6] To study ethnically related trends in the relative frequency of actor representation on posters, we analyzed the number of faces from each category and displayed it as the percentage of the overall number of faces in a poster (see Fig. 3). We find that there is a clear White dominance. Even in the current decade, where the minimum value is seen, there are still 77% White actors on movie posters. Nevertheless, there has been a constant increase in the relative representation of other ethnic groups over the years. In the last decade, the percentage of black actors almost doubled, and there was also significant growth in Asian actors representation. Additionally, we normalized Fig. 3 data by demographic data per each ethnic group, according to USA ethnic population distribution,[7] as extracted from publicly available databases (see Section Datasets) (Gibson and Jung, 2002; Grieco et al., 2001; Humes et al., 2011). We normalized the data by dividing each ethnic group's percentage in the posters by its percentage in the US population. We found that nowadays (2020–2021), there is almost a perfectly balanced representation of minorities on posters similar levels to those in the US population.

We then studied the ethnic effects of actor face size and location on the posters. We specifically evaluated how the size of the actor faces relates to the size of the faces of the largest actors on the poster (see Fig. 4A). The incentive for exploring the representation of face sizes is that larger faces should represent more important actors or characters. We found that white actors' average relative face size is 25% larger than the other races. Also, we see similar trends in terms of distance from the poster center (see Fig. 4B). We also observed that the difference in the distance from the center had drastically reduced in the past couple of decades.

Next, we inspected the relationship between the number of different actors on movie posters and ethnicity (see Fig. 5). We have observed a growth in minority representation on the posters in the past 22 years. The growth is most evident for black actors. However, we see that minorities have higher representation, primarily in movies with more than six different actors on the movie posters. In other words, minorities have a higher chance of appearing on a movie poster, where many actors appear on the movie posters.

Additionally, we explored if the selection of the second largest character on a poster is race-biased, conditioned by the ethnic origin of the largest character on the poster. Thus, for each face, we analyzed its probability of belonging to each of the ethnic categories, given the race of the largest depicted face in the poster (see Fig. 6). We found that White actors have the highest probability of being second across all four graphs, meaning that no matter the ethnicity of the largest actor, the rest of the actors are most likely White. Moreover, when the largest actor (largest face size in the poster) is Non-White, the next most probable race category for actors (after White actors) is the same as the largest face on the poster.

We further analyzed the data for different movie genres. For each genre available in the IMDb dataset, we inspected the ratio of racial distribution in its posters (see Fig. 7A). White actors have the highest percent appearance across all genres and are particularly dominant in the Film-Noir category, holding a maximum value of 0.98 and in the Western and Mystery genres (with 94% and 93% respectively). The Sports, Music, Action, Crime, and Documentary genres include the highest percentage of Black actors, and the other two minority ethnic categories (Asian and Indian) have meager representation, with a maximum value of around 9%. We also inspected each race and its distribution across genres (see Fig. 7B). We observed that Asian actors are more apparent in the action genre and Black act than in other races.

Finally, we examined the first 12 rank positions in the cast list. We define rank as the sequential number within the cast list. We counted the number of actors and divided them into different race categories (Asian, Black, Indian, and White). For each rank, we calculated the ratio between the number of actors of each race with

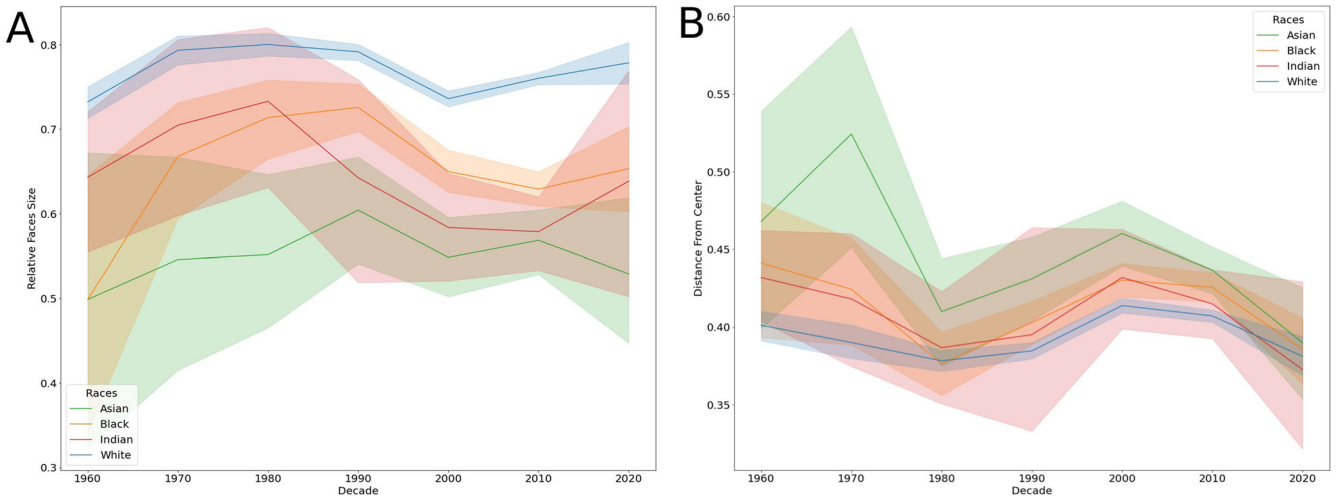

Fig. 4 Relation between race and face positioning on posters. A The change in the averaged face size relative to the largest face detected, for the different considered ethnic groups. **B** Normalized Euclidean distance from the actor to the middle of the poster.

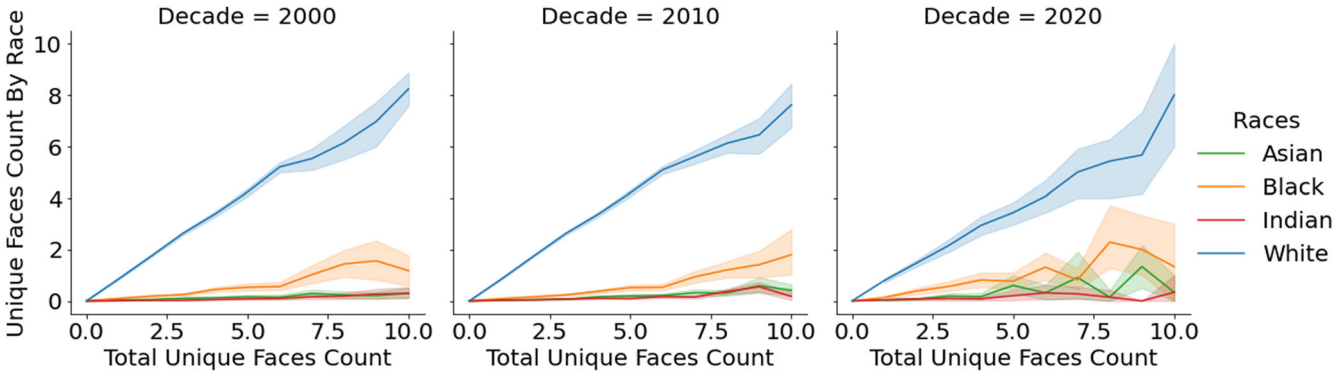

Fig. 5 The average number of unique actors on English-speaking movie posters 2000-2021. Each graph depicts the relative number of actors, as a function of the number of actors on a poster, for each of the considered races.

Fig. 6 The probability of actors of different races to appear, on a poster, conditioned on the the races of the most prominent actor. Changes in the conditional probability of a face in a poster to belong to each race category, given the race of the largest face in the poster.

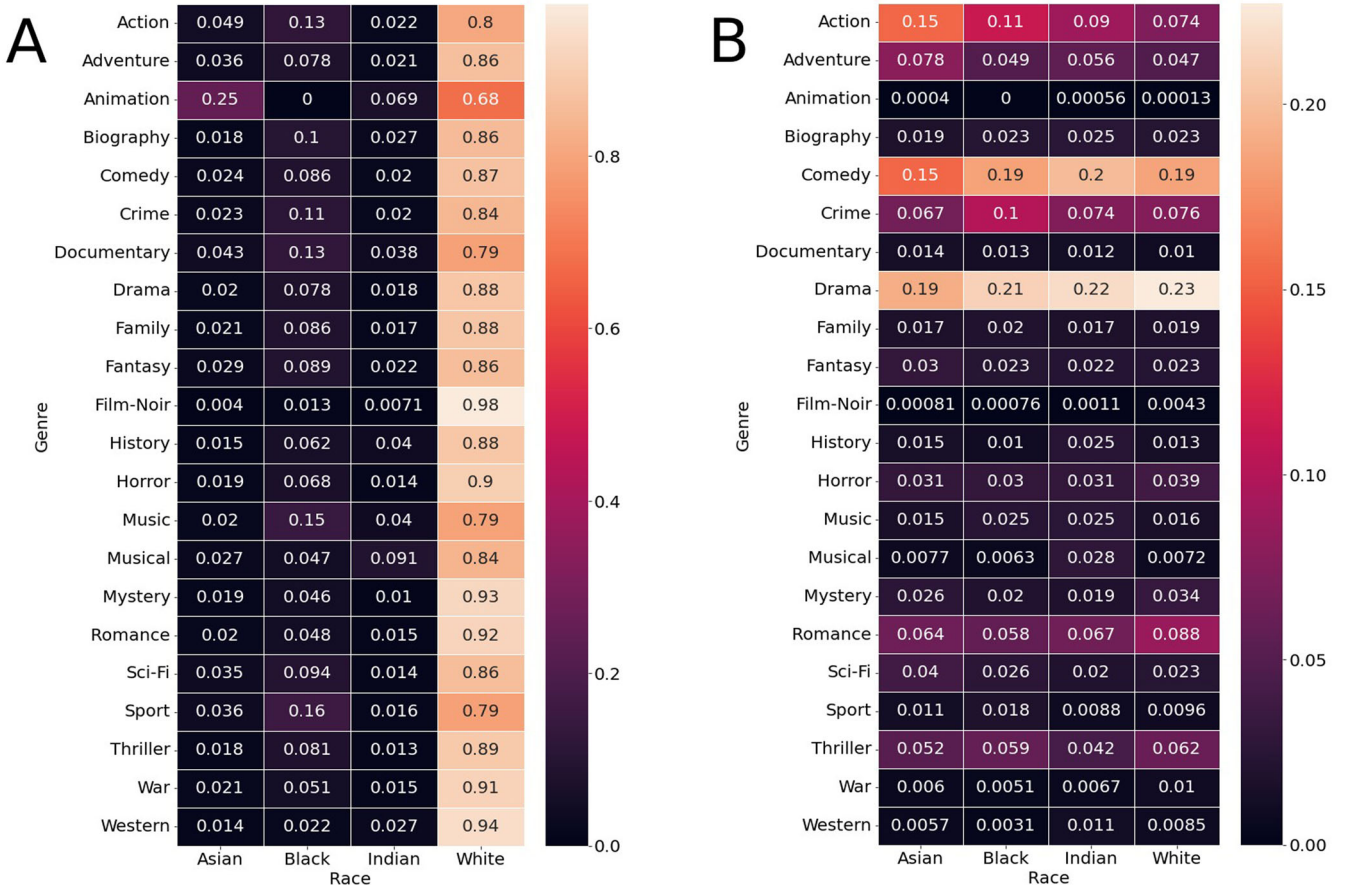

Fig. 7 Distributions of actor representations on posters over Race and Movie Genre. A Distribution of actor appearances for each Genre over the different races. **B** Distribution of actor appearances for each race over the different Genres.

respect to the total number of actors at that rank (see Fig. 8). The results demonstrate that White actors are the most common throughout ranks in posters. Moreover, in English-speaking movie posters, the White actor's relative appearance declines gradually with rank. In contrast, Black actor's relative appearance gradually increases with rank, indicating White actors are preferably positioned high up within the cast list, unlike Black actors who are positioned in more minor roles. Other ethnic groups demonstrate consistently low presence across all rank positions in the cast list.

Discussion

We developed an empirical analysis framework to study ethnic bias in the movie industry through the eyes of a poster viewer. To this end, we collected a large-scale dataset of movie posters. The temporal figure presents only results from the last 80 years since the number of non-white actors before the 60s was relatively small, resulting in highly biased plots. Using deep learning algorithms, we extracted features from the posters and explored racial bias in the film industry from multiple perspectives.

We studied trends in the ethnic-related representation of 26,069 actors in movie posters. We found that the representation of White actors in English-speaking movies tends to decrease with time, in parallel to an increase in the representation of minorities. These results are strengthened by a report on evidence of steady growth in minority representation in movies (Hunt and Ramón, 2020). We speculate that these changes result from rising cultural awareness. The BLM movement that started in 2013 can explain the drastic increase in black actors' representation in the past decade. Additionally, Hollywood has targeted the global market in the past decade, especially in China (Fan, 2015). As expected, we have seen a drastic growth in Asian actors' representation in the past decade. However, we see a different picture from the same data when normalized relative to the US population. We observe a similar increase in representation for all inspected ethnic minorities. Surprisingly, today, poster representation achieved equilibrium relative to the population in the US, and there is no over-representation to appeal to specific markets. These results suggest that today, the film industry makes an active attempt to put together posters that fairly represent the US population according to ethnic distribution.

We hypothesized that minorities' size and position on posters would depend on their ethnicity. We found that white actors are larger and closer to the center of the poster. This might be explained by the fact that white actors still get more lead roles since they are the biggest ethnic group of US actors. Leading roles usually translate to larger characters on posters and more central positioning.

Also, we were interested in testing whether, in recent years, actors of color were added to posters from more minor roles to appeal to minorities. As we observed on movie posters from the last couple of decades, minorities are prominently featured on movie posters with many different actors. We suspect that there are several possible reasons. The first option is that movie franchises with many stars, such as Marvel, DC, Fast & Furious, have become more common in the past two decades. Since these movies have many famous actors, many actors get placed on posters, including actors of color. Another option is that movie production companies feel that adding minorities to movie posters with many actors will result in less critique.

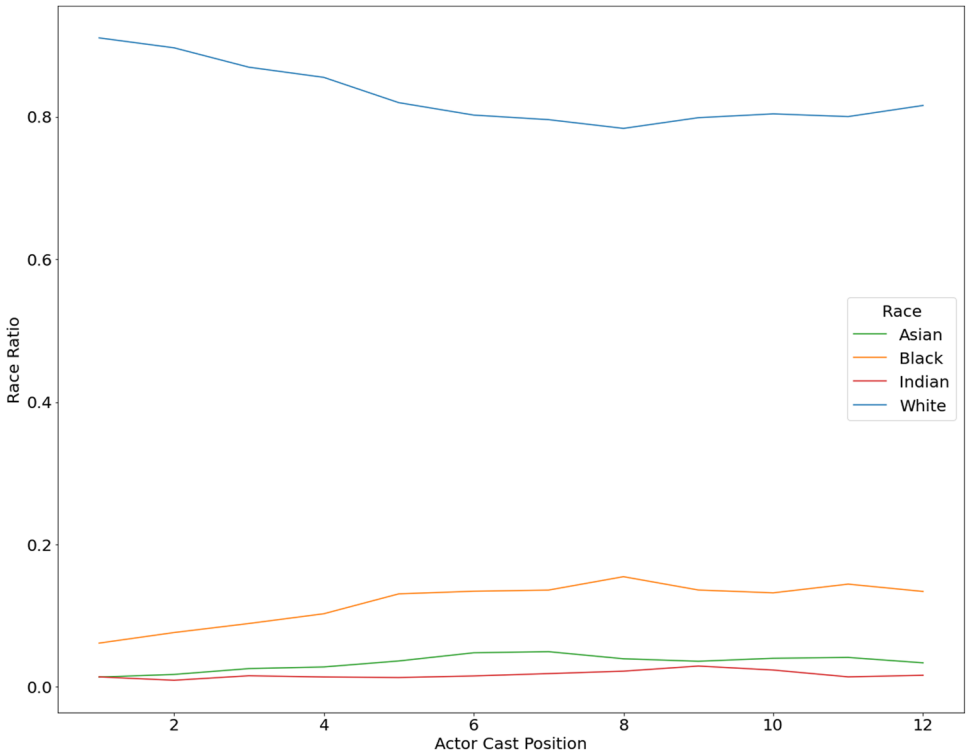

Fig. 8 Ethnic related appearance in each rank position. At each rank, the relative number of actors of each race category out of all actors is depicted.

Furthermore, we found evidence of homophyly on posters where the largest actor is from a specific ethnicity. There is a higher likelihood of finding actors from the same ethnicity than on other posters. We suspect that many actors of color are cast in movies with an ethnically related plot, requiring many actors from the same ethnicity.

Also, we found that the most diverse genre in English-speaking movies is Documentary, where the highest percentage of minorities on posters was observed. This is likely because documentaries are true stories, which leave little control over the cast to the producers. We did see indications for stereotypical representation of Asian and Black actors. Relative to other ethnicities, Black actors are more dominant in crime movies, while Asian actors are in action movies. We suspect it is a result of the stereotypes that all Asians know martial arts (Tran, 2021; tvtropes, 2022) and Afro-Americans are criminals (Welch, 2007).

Finally, we found that the number of actors appearing on posters generally distributes nearly exponentially with rank. The higher the actor's rank in the cast list, the higher their probability of appearing on the poster. This result indicates that actor appearance on posters has an excellent potential to act as a centrality measure of leading characters. In terms of ethnic-related effects on poster appearance at different ranks, we found that the percentage of non-White actors on posters is higher in lower-ranked positions on the cast list. With the data at hand, we can not conclude if it results from affirmative action or from the fact that non-White actors hold a ranking in the cast list that is lower than their actual importance in the first place. Diving into temporal trends in actor appearance on posters by cast list position and race, we find an overall increase in representation of non-White actors in most cases regardless of cast position in the past ten years.

As with any machine learning-based study, this work is mainly limited by the models used. In this study, we used the available state-of-the-art algorithms to reduce errors. Obviously, when better race prediction models will become available, also more accurate results could be obtained in the future. We found this to be an issue, mostly in race prediction on older black-and-white movies, where the sample size of non-White faces was small. The resolution of a poster is also a limitation, especially when there is a small face. A possible solution might be the use of super-resolution, but a further study should be performed to determine its effectiveness. Another limitation is that no data specify how many posters were produced for each movie and how many posters were printed from each variation which may portray a different picture.

Conclusions

Movie posters have an enormous potential for highlighting cultural biases, and in particular ethnic biases that appear in the film industry, and that eventually also shape our cultural perception as viewers. In this work, we created the first dataset of movie posters that contains metadata about actors' identities and ethnicity. We constructed this dataset using deep learning models and fused the poster data with multiple sources. We then analyzed the diversity in the film industry by examining multiple parameters.

Our results suggest that when it comes to poster design, on average white actors are larger and closer to the center of the poster. We also found an increase in the representation of Asian and Black actors in the past decade on posters. In fact, in English-speaking films, the actors on recent posters from the past two years represent exactly the ethnic distribution of the US. Additionally, we demonstrate that the main character's race affects the other characters' race on the poster, with a greater probability for their own kind. In a future study, we plan to explore differences between English-speaking and Non-English foreign movies and compare posters from different countries, looking for differences between the presented posters in different cultures. Also, it could be interesting to explore the relationship between the decision-makers identity and the posters' content. Additionally, we intend to present a centrality measure for the rank in the cast list based on movie poster appearance. In fact, it would be interesting to find cases of mismatch, where certain characters of minor roles are of greater

appearance in posters. This might indicate an unfair attempt for the correction of ethnic representation gaps.

Data availability

All code and data are available on the project GitHub (https://github.com/data4goodlab/PosterAnalyzer).

Received: 6 February 2023; Accepted: 28 July 2023;
Published online: 08 February 2024

Notes

1 Each movie may have multiple posters.
2 We randomly selected pairs of posters from the same movie and measured the distance between each pair. We then visually inspected the pairs and chose a threshold set by the average distance between duplicate posters, and validated it on the pairs of non-duplicates.
3 Indian refers to people who are originally from the Republic of India
4 Asian, Black, Indian, White
5 White, Black, Latino-Hispanic, East Asian, Southeast Asian, Indian, Middle Eastern
6 All the calculations are made at a movie level to reduce noise generated by movies with high posters.
7 There is no worldwide ethnic demographic data available.

References

Alberani D (2021) Imdbpy. https://pypi.org/project/IMDbPY/
Albert RS (1957) The role of mass media and the effect of aggressive film content upon children's aggressive responses and identification choices. Genetic Psychology Monographs 55:221–285
Aley M, Hahn L (2020) The powerful male hero: a content analysis of gender representation in posters for children's animated movies. Sex Roles 83:433–509
Baumgartner E, Laghi F (2012) Affective responses to movie posters: Differences between adolescents and young adults. Int J Psychol 47(2):154–160
BBC News (2020) George Floyd: What happened in the final moments of his life. https://www.bbc.com/news/world-us-canada-52861726
Beaufort M (2019) How candy placements in films influence children's selection behavior in real-life shopping scenarios—an Austrian experimental field study. J Children Media 13(1):53–72
Buunk AP, Peiró JM, Griffioen C (2007) A positive role model may stimulate career-oriented behavior 1. J Appl Soc Psychol 37(7):1489–1500
Carroll N (1985) The power of movies. Daedalus 114(4):79–103
De Run EC (2005) Does targeting really work? The perspective of a dominant ethnic group. Int J Bus Soc 6(1):27
Deng J, Guo J, Ververas E, Kotsia I, Zafeiriou S (2020) Retinaface: single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), Seattle, WA, USA, pp. 5202–5211. https://doi.org/10.1109/CVPR42600.2020.00525
Deng J, Guo J, Xue N, Zafeiriou S (2019) Arcface: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699
Fan H (2015) Hollywood big shots target China movie market. https://web.archive.org/web/20210315024046/https://www.telegraph.co.uk/sponsored/china-watch/culture/11763881/hollywood-film-industry-targets-chinese-movie-market.html. Accessed 31 May 2022
Fletcher RR, Nakeshimana A, Olubeko O (2021) Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. Front Artif Intell 3:116
Freire J (2019) Cultural identity in film posters. http://researcharchive.vuw.ac.nz/handle/10063/8549
Gabriel BP (2012) The rugged action hero and his sexy love interest: gender in popular movie posters. MA Theses, University of Texas at Arlington
Gibson C, Jung K (2002) Historical census statistics on population totals by race, 1790 to 1990, and by Hispanic origin, 1790 to 1990, for the United States, regions, divisions, and states. US Census Bureau, Washington, DC
Grieco EC, Grieco EM, Cassidy RC (2001) Overview of race and Hispanic origin, 2000, vol 8. US Department of Commerce, Economics and Statistics Administration, USA
Hamming RW (1950) Error detecting and error correcting codes. Bell Syst Tech J 29(2):147–160
Hennekam S, Syed J (2018) Institutional racism in the film industry: a multilevel perspective. Equality, Diversity and Inclusion: An International Journal 37(6):551–565
Humes KR, Jones NA, Ramirez RR et al (2011) Overview of race and Hispanic origin: 2010. US Department of Commerce, Economics and Statistics Administration, US
Hunt D, Ramón A-C (2020) Hollywood diversity report 2020: a tale of two Hollywoods. UCLA College of Sciences
IMDB (2022a) Internet movie database. https://www.imdb.com/
IMDB (2022b) Internet movie database datasets. https://www.imdb.com/interfaces/
Johnny L, Mitchell C (2006) "Live and let live": an analysis of hiv/aids-related stigma and discrimination in international campaign posters. J Health Commun 11(8):755–767
Kaikati JG (1987) Celebrity advertising: a review and synthesis. Int J Advert 6(2):93–105
Kamins MA (1990) An investigation into the "match-up" hypothesis in celebrity advertising: When beauty may be only skin deep. J Advert 19(1):4–13
Kärkkäinen K, Joo J (2021) Fairface: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, pp. 1547–1557
Krawetz N (2013) Kind of like that. the hacker factor blog. http://www.hackerfactor.com/blog/index.php?/archives/529-Kind-of-Like-That.html
Lee K (2014) Race in Hollywood: quantifying the effect of race on movie performance
Legal Highlight: The Civil Rights Act of 1964|U.S. Department of Labor (n.d.), https://www.dol.gov/agencies/oasam/civil-rights-center/statutes/civil-rights-act-of-1964. Accessed 01 May 2023
Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision ICCV), Santiago, Chile, pp. 3730–3738
MarketWatch (2020) https://www.marketwatch.com/story/this-simple-fact-may-explain-why-hollywood-films-arent-more-diverse-2020-02-11
McKinsey (2021) Representation of black talent in film and TV|McKinsey. https://www.mckinsey.com/Featured-Insights/Diversity-and-Inclusion/Black-representation-in-film-and-TV-The-challenges-and-impact-of-increasing-diversity. Accessed 5 Mar 2023
Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv 54(6):1–35
Merler M, Ratha N, Feris RS, Smith JR (2019) Diversity in faces. arXiv preprint arXiv:1901.10436
Nieman P (2003) Impact of media use on children and youth. Paediatr Child Health 8 5:301–17
Rahmasari I (2014) A semiotic analysis on the help movie posters. J Ilmiah Mahasiswa FIB 4(2). https://www.neliti.com/publications/202922/a-semiotic-analysis-on-the-helpmovie-posters#cite
Smith S, Choueiti M, Pieper K (2014) Race/ethnicity in 600 popular films: examining on screen portrayals and behind the camera diversity. Media, Diversity, & Social Change Initiative, pp. 2007–2013
Smith SL, Choueiti M, Pieper K (2020) Inequality in 1,300 popular films: examining portrayals of gender, race/ethnicity, LGBTQ & disability from 2007 to 2019. Annenberg Inclusion Initiative
Smith SL, Choueiti M, Yao K, Clark H, Pieper K (2020) Inclusion in the director's chair: analysis of director gender & race/ethnicity across 1,300 top films from 2007 to 2019. USC Annenberg Inclusion Initiative
Smith SL (2021) Inclusion in Netflix original US scripted series & films. INDICATOR 46:50–6
The New York Times (2020) Black Lives Matter May Be the Largest Movement in U.S. History. https://www.nytimes.com/interactive/2020/07/03/us/george-floyd-protests-crowd-size.html
TMDB (2022a) The movie database. https://www.themoviedb.org/
TMDB (2022b) The movie database API. https://developers.themoviedb.org/3/getting-started/introduction
Tran D (2021) Asians have long been stereotyped in martial arts roles. These shows are reclaiming combat. https://news.yahoo.com/modern-martial-arts-films-tv-164354094.html. Accessed 3 Mar 2022
tvtropes (2022) All Asians know martial arts—TV tropes. https://tvtropes.org/pmwiki/pmwiki.php/Main/AllAsiansKnowMartialArts#:~:text=All%20Chinese%20People%20Know%20Kung,to%20all%20kinds%20of%20monks. Accessed 3 Mar 2022
United States Census Bureau (2022) Measuring America's People, Places, and Economy. https://www.census.gov/
Welch K (2007) Black criminal stereotypes and racial profiling. J Contemp Crim Justice 23(3):276–288
White M (2019) Hollywood still has a diversity problem|thehill. https://thehill.com/opinion/civil-rights/455522-hollywood-still-has-a-diversity-problem. Accessed 29 Sept 2021
Zhang Z, Song Y, Qi H (2017) Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, USA, pp. 4352–4360

Acknowledgements
We thank Valfredo Macedo Veiga Junior (Valf) for designing the infographic illustration.

Author Contributions
Dima: conceptualization, methodology, software, formal analysis, writing—original draft; Mor: software, formal analysis, data curation, writing—original draft; Mickey: conceptualization, methodology, resources, writing—original draft, supervision, writing—review and editing; Galit: conceptualization, methodology, writing—original draft, supervision, project administration, writing—review and editing.

Competing interests
The authors declare no competing interests.

Ethical approval
This article does not contain any studies with human participants performed by any of the authors. Also, we did not train or develop a new ethnic-classification model. We used existing datasets and pre-trained models from papers that were already published.

Informed consent
This article does not contain any studies with human participants performed by any of the authors.

Additional information
Correspondence and requests for materials should be addressed to Michael Fire or Galit Fuhrmann Alpert.