Computational appraisal of gender representativeness in popular movies

Antoine Mazières, Telmo Menezes, Camille Roth

Computational appraisal of gender representativeness in popular movies

Antoine Mazières*[1], Telmo Menezes[1], and Camille Roth[1,2]

[1] *CNRS, Centre Marc Bloch, Computational Social Science team, Berlin, Germany*
[2] *CAMS, Centre d'Analyse et de Mathématique Sociales, CNRS/EHESS, Paris, France*

Abstract

Gender representation in mass media has long been studied by qualitatively analyzing content. This article illustrates how automated computational methods may be used in this context to scale up such empirical observations and increase their resolution and significance. We specifically apply a face and gender detection algorithm on a broad set of popular movies spanning more than three decades to carry out a large-scale appraisal of the on-screen presence of women and men. Beyond the confirmation of a strong under-representation of women, we exhibit a clear temporal trend towards a fairer representativeness. We further contrast our findings with respect to movie genre, budget, and various audience-related features such as movie gross and user ratings. We lastly propose a fine description of significant asymmetries in the *mise-en-scène* and *mise-en-cadre* of characters in relation to their gender and the spatial composition of a given frame.

Keywords: gender studies; image analysis; film theory; content analysis; face recognition

Introduction

There is assuredly a long tradition of scholarship in the description of sex roles on mass media of various types: already in her seminal review, Linda Busby (1975) described how instructional material, TV, films, advertising, newspaper, cartoons and literature have been used since the late 1950s to study gender-related representations such as sexual stereotypes, biases in occupational roles, body staging, marriage and rape. Back then, she further concluded that "media sex-role studies that have been completed in the 1960s and early 1970s can be used as historical documents to measure future social changes", emphasizing the need of replicating these analyses at several points in time to capture underlying mutations and trends. As empirical material, such sources provide the opportunity to grasp a certain state of affairs regarding gender representations, together with the intents and conflicts of interest at play in shaping them. Recent reviews of this research (Rudy et al., 2010; Collins, 2011) highlight the ubiquity of gender patterns, most notably the under-representation and sexualization of women, across multiple media and content types, even though some negative results may occasionally be found as well (Kian et al., 2009). Almost half a century after Busby's review, the roles of females and males in media and fiction have been a prominent domain of inquiry in content analysis and have been subjected to many quantitative analyses (Neuendorf, 2017), including for instance broadcast network programs (Lauzen, 2018), popular movies (Lauzen, 2019; Smith et al., 2019) and recurring TV show characters (Townsend et al., 2019).

Methodologically, this body of research principally relies upon qualitative assessments of text, images and scripts, which sometimes feature complex semantic concepts and possibly subjective interpretations. As a result, these approaches are difficult to scale to the large number of observations and, thus, human coders that are required to build sizable datasets in order to perform statistical and especially longitudinal analyses of the material itself. Some studies do rely on large-scale and automatically collated datasets, for instance through collaborative platforms such as IMDb, the Internet Movie Database, but they are by definition limited to the metadata that have been made available — such as film cast, crew, or budget (Yang et al., 2020). In other words, the automated construction and extraction from empirical media sources of relevant variables adequate for a given study and research question re-

*Corresponding author: antoine.mazieres@gmail.com

mains a challenge in media gender studies.

Recent advances in artificial intelligence and data science may significantly help in this regard, especially in terms of automated processing of text, image and video, where current technologies are sometimes capable of competing with humans in a wide array of specialized tasks, including automatic text summarization (Mani, 2001), topic detection (Chaney and Blei, 2012), or translation (Hassan et al., 2018); face recognition (Guo and Zhang, 2019; Dhomne et al., 2018), scene intensity estimation (Kataria and Kumar, 2016), narrative element extraction (Guha et al., 2015; Bost et al., 2016); or even at the interface of both, text description generation from images (Xu et al., 2015). At the moment, however, these methods have generally been applied on issues that remain quite close to the scientific fields from which they originate and are geared towards a technological rather than a scientific audience — including scientists from other fields. More precisely, they do not seem to have yet been used in sex role research.

Our contribution explores the possibility of using such advances to the construction of datasets relevant to sex role research. Firstly, we outline a field of inquiry by focusing on cinema, for which we identify a relevant subset of popular movies. We extract a representative set of frames from this dataset and applied machine learning models to detect human faces and infer their gender. We take the extra precaution of evaluating the performance and fairness of these inferences regarding the target categories (*female* and *male*), for these models are typically evaluated in all generality and their potential biases may vary with respect to data corpora. Secondly, we devise a metric to appraise women's presence in movies, the *female face ratio* (FFR). We compare it with another well-established measure, the Bechdel test. In aggregate, FFR markedly increases over time, to the point of approaching female-male parity. Also, there are significant differences in how its values are distributed for successive temporal periods. This indicates a noticeable mutation in the popular movie-making culture regarding women's representation. Thirdly, we explore several more sophisticated and experimental capabilities of automatic face detection to analyze how characters of distinct genders are framed on-screen. Interestingly, this yields mostly negative results in the sense that we observe very little variations. We nevertheless exhibit a few significant patterns related to gender-mixed environments.

More broadly, we contend that the systematic application of such techniques could lead to the formulation of ambitious research questions that would be hardly tractable with human workforce — introducing the notion of "distant viewing" (Arnold and Tilton, 2019) by analogy with the notorious concept of "distant reading" (Moretti, 2000). This could furthermore enable the creation of well-documented datasets featuring metadata adapted to sex role research for the community to thoroughly and conveniently reproduce and improve experiments. Tackling this challenge could indeed trigger new fields of interest, for instance on formulating theoretical understanding on the way representations are distributed on the whole spectrum of a specific media, or focusing on potential outliers showing very different patterns of representation, in order to unveil their possible contribution to forthcoming mutations.

Dataset and data processing

Corpus scope

Movie studies typically define the corpus scope by relying on box office data as a proxy for movie popularity (e.g., Follows, 2014; Lauzen, 2019; Smith et al., 2019). They essentially outline a selection based on the yearly top grossing movies over a period of time, i.e. a short-term commercial success in movie theaters admittedly related to popularity. We also aim at considering mainstream content, yet we assume that audience attendance tells only one part of the story. Popularity relies on complex behaviors: it relates as much to the value given by an individual to the content, as to the value an individual perceives, or anticipates, others will give. Intricate interactions of support, rejection, controversy, advocacy and imitation come into play to establish a cultural object's influence (Cillessen and Marks, 2011). Put shortly, attendance alone may not help fully capture movies that are both characteristic of cultural representations and influential in shaping them. We thus devised a different approach based on open collaborative platforms, assuming that it configures a broader filter than audience figures. Collaborative online environments such as peer-to-peer file sharing networks (Vassileva, 2002; Cohen, 2003) or wiki-based knowledge sharing systems (Rafaeli and Ariel, 2008; Yang and Lai, 2010) are fueled by interactions between a diverse and critical mass of users. Contributors are incentivized by the effort of others to in-

crease the system usefulness by creating and maintaining fashionable resources: they act from a variety of motives, including both the perceived value of the content they provide and the peer recognition that it entails. We argue that the intensity of such collaborative activity may be a more comprehensive proxy for the content relevance that we wish to capture.

Based on this, we focus on films for which data is available on two significantly distinct types of online platforms: (1) a peer-to-peer file sharing network, which is one of the major Torrent communities, YIFY (yts.mx); and (2) a movie-related knowledge-sharing platform, the above-mentioned Internet Movie Database (IMDb, imdb.com), which comprises records on about 500k movies, mostly stemming from user contributions. We first listed all 13,662 movies made available on YIFY, requiring that at least 3 people share them (seeders) as of December 2019. We then linked them to their respective record on IMDb, excluding documentaries and animation movies while requiring that key metadata be available: year of release, genres, users rating, parental rating, runtime, budget and world wide gross. We find that there are very few movies per year before 1985 (10 on average, no more than 48 for a given year): for the purpose of the longitudinal analysis, we decide to further focus on the period 1985-2019, wherefrom the yearly number of movies per year is always above 100. This yields a dataset of 3,776 movies. The average runtime is 109 minutes with a standard deviation of 18 minutes, indicating that we essentially gathered feature films. The budget distribution is broad, with a median of $23m while the first and third quartiles are at $10m and $45m, indicating that we focus on a quite diverse array of movie budgets. The same applies to world wide gross figures: median $43m, first quartile $11m, third quartile $122m. This further substantiates our approach for constructing a filter that is broader than when focusing on top audience figures only.

Face recognition and gender estimation

The computational extraction of artistic or semantic characteristics of a movie traditionally relies on the extraction of a number of significant images (Guha et al., 2015; Ko et al., 2019). This is commonly based on keyframes i.e., frames of a movie's timeline where new shots commence. This method results in better quality images, since keyframes are used as markers for video compression. Also, it likely captures narrative highlights, since a keyframe captures the first state of scenery —arguably an important one— from which the shot unfolds. However, the duration and pace vary very significantly from a shot to the other, and are also strongly influenced by shot type, movie genre and year of production (Cutting and Candan, 2015). In other words, keyframes are unlikely to provide a representative set of images from a film, even more so for the longitudinal analysis we aim at. To ensure the representativeness of our sample with respect to what spectators are shown, we simply extracted frames on a time frequency basis, one every 2 seconds, yielding a collection of more than 12.4 million images.

We processed each of these images with the help of face detection and gender estimation algorithms provided by a common scientific computing software, *Wolfram Mathematica 12*.

We eventually detect close to 10 millions faces over more than 6.6 million images, with an average of 2596 (σ = 1090) faces per movie. For every face, the algorithm provides the coordinates of a bounding box, enabling us to take into account both the position and the size of the surface occupied by the face with respect to the frame dimensions. It also provides an estimation of the likely binary gender of each face (male or female).

Both algorithms are built using conventional machine learning methods. Many questions have been raised over the recent years regarding the accuracy and potential bias of predictions based on these techniques, and our approach is no exception. Previous social scientific-oriented research specifically highlighted the issues associated with the construction of the datasets that are used to train machine learning algorithms (Crawford and Paglen, 2019). Put shortly, a dataset of human-labeled pictures is first gathered, such as ImageNet (Deng et al., 2009). Labels correspond to categories of interest that should be learned from this dataset, in order to predict them on any unknown dataset. In our case, these labels include the visible faces (presence and position) and their gender (male or female). Part of this human-labeled dataset is fed to a learning algorithm —such as a neural network— that will initially improvise predictions and then, iteratively, learn from its mistakes, readjusting and ultimately converging towards better guesses. The learned model is then tested on another part of the dataset to assess if the algorithm managed to *generalize* well — thereby measuring its *performance*.

Across the state of the art, both types of algorithms generally reach accuracies well above 90% (Guo and

Zhang, 2019; Dhomne et al., 2018). Yet, they also display a strong degree of performance variation depending on the type of dataset at hand and, plausibly, the context and type of images, for instance in medical imagery (Zech et al., 2018; McBee et al., 2018). Movie frames are likely a specific type of data. The work of Buolamwini and Gebru (2018) on designing *intersectional benchmarks* is also particularly relevant here, in that it highlights how face detection algorithms perform unevenly when tested on faces of specific genders or skin tones. In any event, we thus need to make sure that the algorithms perform sufficiently well with our dataset for our purposes.

To this end, we set up a simple experimental protocol: we randomly select 1000 frames each extracted from a distinct movie and on which the algorithm detected only one face, half of which female, the other male (so, 500 frames for each gender). We built the web interface shown in Fig. 1 displaying one random frame at a time with a bounding box around the detected face, followed by two questions, one aimed at checking whether the face detected in the bounding box and its gender are correct, the other aimed at checking whether the frame contains undetected faces. We sent the link to this website on our research center's internal mailing-list. Participants were invited to review as many frames as they could. Overall, 4,938 reviews were submitted with an average of 4.94 ($\sigma = 2.29$) reviews per frame. For every frame, we considered the most frequent answer. (Narrowing the evaluation only to pictures with identical answers over all reviews actually yielded very similar results). Raw results are gathered on Table 1. For face detection, there are 977+863=1840 correct inferences (true positives and true negatives) and 23+137=160 incorrect inferences (false positives and false negatives), thus a high accuracy of 92%, consistent with the literature. Note that there are much more false negatives than false positives i.e., the algorithm, when wrong, tends to rather fail to identify a face than erroneously detect one. Accuracy for gender inference is weaker, with 304+410=714 correct inferences and 162+75+7+8=252 incorrect ones (discarding the negligible "doubt" category which indicates that human participants were unable to be conclusive) i.e., a lower yet pretty high 73.9% accuracy. However, we also notice that gender inference performs quite differently between males and females. When it infers a female face, the face is actually of a women only 65% of the time, while of a man 35% of the time. Male faces are accurately identified 84.5% of the time, and are actually of a female for only 15.5% of the cases. Therefore, the model shows in aggregate a tendency to wrongly categorize faces as female more often than for male faces. It generally informs us that the *raw* inferences of woman faces and thus woman presence are overestimated by the machine learning algorithm that we used.

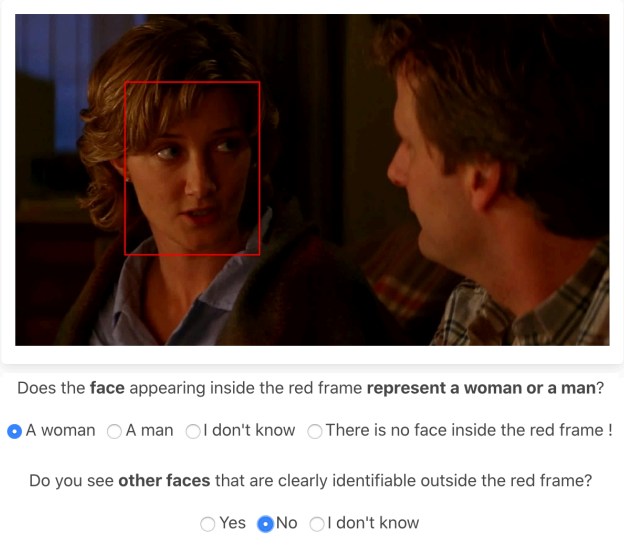


Figure 1: **Interface of the human evaluation experiment**

Thanks to this contextual validation step, we can now correct inference results appropriately. Knowing the shape and magnitude of model error makes it indeed easy to adjust face counts: for instance, if the algorithm detects a female face, we count .65 female faces and .35 male faces, using the confusion matrix of Table 1. The same applies for male faces. In a nutshell, we adjust the raw FFR using the following formula:

$$\text{FFR}_{\text{corrected}} = (1-\lambda) + (\lambda + \lambda' - 1)\text{FFR}$$

where λ and λ' are the proportions of true positives for male and female faces, respectively. Furthermore, we observe that algorithm error is not constant across time: female faces are over-estimated significantly more for the earlier than for the later years. In practice, we thus use time-dependent correction factors λ and λ' (based on time periods defined below for the longitudinal analysis).

Women's presence and its evolution

The content analysis literature has relied on diverse features to assess gender representation in media. It variously mixed field expertise, subjective perceptions and quantifiable variables. These endeavors often led to semantic characterizations such as

Table 1: **Evaluation of the detection models.**

(a) Face detection

		Humans	
		Positive	*Negative*
Model	*Positive*	977	23
	Negative	137	963

(b) Gender inference

		Humans			
		Female	*Male*	*Doubt*	*No face*
Model	*Female*	304	162	18	16
	Male	75	410	8	7

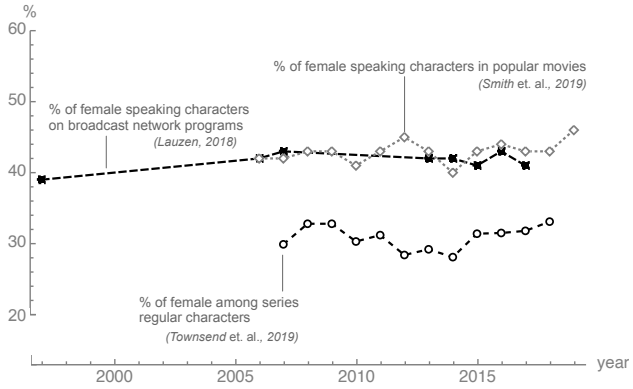


Figure 2: **Several metrics used in the literature, based on Smith et al. (2019); Lauzen (2018); Townsend et al. (2019)**

women appearing "as dependent on men", "unintelligent", "less competitive", "more sexualized" (Busby, 1975), which are identified, annotated and counted throughout the media for further comment. The more formal the feature, the easier it is to scale the analysis to more observations, either by increasing the number of observers or automating the process.

More recently, various academic and activist projects have undertaken large scale analysis of visual entertainment media. They often lessened the semantic complexity of the variables they rely on and mainly focused on presence ratios, while being able to increase sample sizes to a point that made longitudinal analysis possible. Figure 2 gathers some results from three of these projects (Townsend et al., 2019; Smith et al., 2019; Lauzen, 2018). They not only confirm the under-representation of women already widely observed across the literature (Busby, 1975; Collins, 2011), but they also invite the conclusion that this situation has not evolved markedly in any direction during the considered periods.

The face and gender detection algorithms we use provide us, for each movie frame, with three types of information of increasing complexity: number, gender and position of faces. In turn, we derive three types of variables. The first one is the most minimalist: the percentage of faces classified as female among all the detected faces on all frames of a given movie, or *female face ratio* (FFR).

Female face ratio (FFR)

The average FFR over all movies is 34.52% ($\sigma = 9.19$). This ratio is comparable to what is found in the literature, such as the ratio of female among characters in primetime television programming (39.6%) (Sink and Mastro, 2017) or among speaking characters in broadcast network programs and popular movies (see Fig. 2) (Smith et al., 2019; Lauzen, 2018). However, the FFR markedly differs from one genre to another: we find for example an average FFR of 31.3% for *Crime* movies while it reaches 37.1% for *Romance* movies.

To illustrate informally what the FFR means in practice, we provide a few examples of top grossing movies for some domains of this metric. First, among movies with a high percentage of male faces (*i.e.* FFR < 25%) we find movies such as *Pirates of the Caribbean* (2007), *Star Wars* (2005), *Matrix* (2003), *Independence Day* (1996) or *Forest Gump* (1994), all with a FFR of around 23%. Movies such as *The Hunger Games* (2014) and *Jurassic World* (2015), *Rogue One* (2016) and *Gravity* (2013) lie around a female-male parity, with a FFR of between 45% and 55%. Lastly, the movie with the highest FFR (68%) is *Bad Moms* (2016), closely followed by movies such as *Sisters* (2015), *Life of the Party* (2018) and *Cake* (2014).

Beyond these few examples, we further check how the FFR is correlated with narrative features by comparing it with the Bechdel (1983) test. This test is referenced and used in numerous studies (Selisker, 2015; Yang et al., 2020) and renowned for discarding around half of all reviewed movies with the simple criteria that two named women be present, speak to each other, about something besides a man. We rely on data produced by volunteers who manually evaluate if a movie passes or not the above cited conditions. This data is available at bechdeltest.com and only covers a subset of our dataset (n=2,454). As the FFR varies along movie genres, so does the test: we

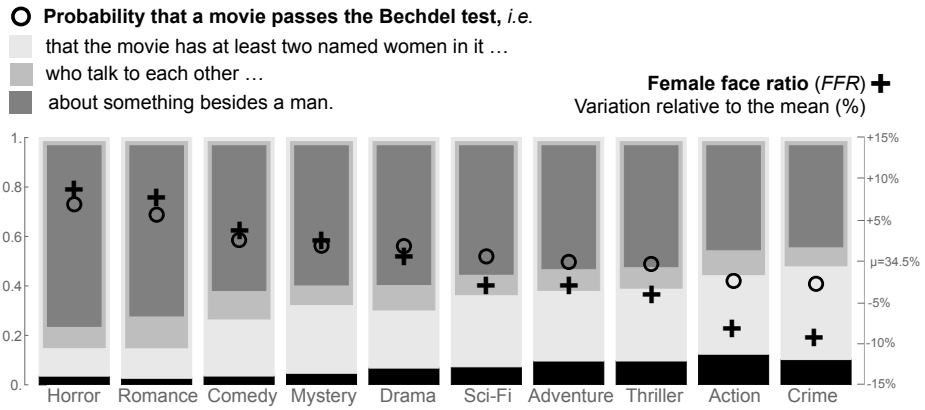


Figure 3: **Bechdel test and female face ratio (FFR) across a selection of popular movie genres**.

compared both metrics across the 10 most frequent movie genres, as shown on Figure 3. We find that they are ordered in almost the same manner (Spearman score > 0.93) even though the FFR varies somewhat less across genres in absolute values.

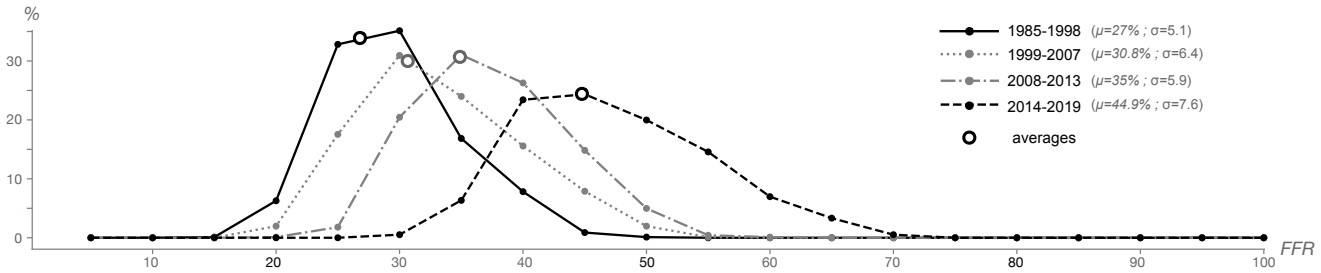
Longitudinal analysis. Our aggregate findings on the FFR since 1985 confirm women underrepresentation in terms of on-screen presence. Yet, they also show a significant trend toward less inequality. Our computational approach enables us to go into more detail by providing a relatively high resolution on the FFR distribution across the observation period which, in turns, reveals several features.

We temporally divided our dataset into quartiles, i.e. four consecutive periods featuring the same number of films. As shown in Figure 4a, the FFR markedly increases across time from an average 27% between 1985 and 1998 to a mean FFR of 44.9% for the last period (2014-2019), close to a female-male balance. The evolution of FFR ranges is equally significant: most movies shot over 1985-1998 exhibit an FFR of 20-45%, while movies of the most recent period 2014-19 generally cover the 35-65% range. Besides, the standard deviations of the underlying distributions increase overall (from 5.1 to 7.6). This probably denotes a higher diversity of situations with regard to on-screen gender presence. On the whole, it seems to be slowly evolving in favor of female representation as distributions appear to be increasingly right-skewed, i.e. towards a higher FFR. Furthermore, considering data from bechdeltest.com restrained to the films of our datasets, over the same periods, we also observe an increase in the percentage of movies passing the test: 51% between 1985 and 1998 up to 60% for the last period (2014-2019). This evolution is comparable to the increase of the FFR, albeit of a somewhat smaller magnitude – +9% vs. +18%.

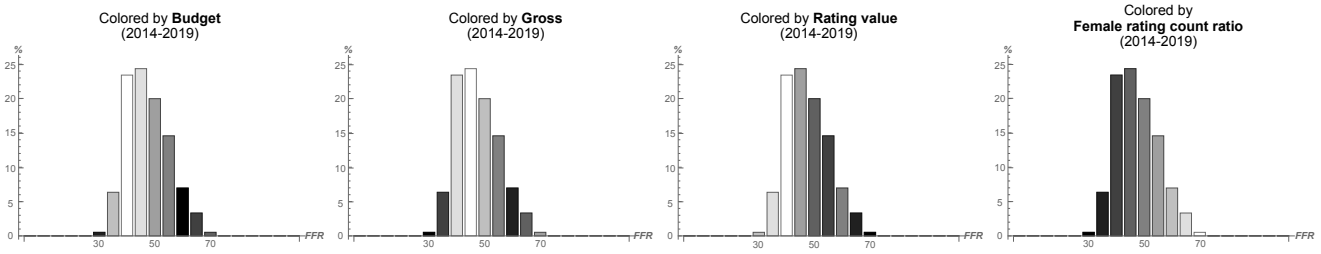
As previously mentioned, while the literature widely acknowledges that women are underrepresented in movies and, more broadly, in visual entertainment media, it usually states that this situation does not exhibit any significant evolution (see Fig 2). As it stands, we observe on our dataset a positive evolution over time of two distinct features, the FFR and the Bechdel test success probability, in apparent contradiction with the hitherto observed stable representation of women. Note however that we exhibit a correlation between the FFR and the Bechdel test, indicating that the FFR nonetheless captures at least in part some semantic features beyond the plain proportion of female faces.

We can think of two phenomena to explain the discrepancy between our study and the previous ones. The first one relates to the way we select content, whereby we focus on a selection of films that may be distinct from what is immediately available on prime-time TV and on-demand streaming platforms. In other words, both ours and the Bechdel test data are based on information contributed by users (on such and such website, relating to the interest of users for such and such content), while the traditional data is based on top-grossing films and/or programs (indicating what is offered to, or most successful for a given audience). The second one may be linked to the potential difference between on-screen presence (that we measure here) and more sophisticated features, such as effective speaking time or regularity of appearance (that is typically measured in the literature).

In essence, the discrepancy may demonstrate that there has been a significant evolution towards more on-screen female presence close to reaching female-male parity, but that this trend is only moderately



(a) Distributions of FFR for each period.



(b) Distributions of several features over the distribution of FFR.

Figure 4: **Distributions of female face ratio** (FFR): **(a)** *Percentage of movies with a given* FFR*, one data point every* 5*%;* **(b)** *Percentage of movies with a given* FFR*, colored by the given variable mean within the bin, the lighter the higher.*

related to the actual importance or influence of women in popular movies and their scenarios. In other words, put in perspective with the literature, the evolution that we uncover here may not be of sufficient influence on gender representation in popular movies. Figuring out to what degree the increase of female on-screen presence is potentially preludial to an upcoming fairer gender representation, or a subtle expression of "purplewashing", would require a deeper qualitative analysis which is beyond the scope of our study.

Relation between FFR and audience. We could see that distinct genres correspond to differing FFR values. Budget and audience-related metadata enable us to characterize more finely the type of films that correspond to certain areas of the FFR distribution. In Figure 4b we focus on FFR histogram for the most recent period (2014-19). On this histogram, we project the average rank of movie FFR with respect to budget, gross, rating given by users (rating value) or number of people having rated a movie (rating count). Note that we chose to color histograms from white to black using rankings rather than absolute values, for there are wide variations in the orders of magnitude of the underlying average values (for instance, budget spans several orders of magnitude – if a certain range of absolute values corresponded to a certain tone, we would almost have had either only white or only black bars, losing a significant resolution and missing the actual ordering and hierarchy between high-budget and low-budget movies). Lighter tones correspond to higher ranks: for instance, the white bar for the "budget" coloration (left-most histogram) denotes the highest movie budgets. It coincides with the main mode and specifically with the bar of the histogram featuring the highest proportion of movies, with an FFR of 35%. The darkest tones, on the other hand, are found for the most extreme values of FFR (very small or very high). Some exceptions are notable: there is a slightly light tone for an FFR value around 70%, indicating the existence of relatively high budget movies on that side as well. On the whole, the same phenomena are visible for world wide gross and rating. This suggests that the audience and their opinion resonate best with movies close to the main FFR mode, which corresponds to the average FFR under-representation of women. Interestingly, the higher FFR values that emerged over the recent years (around 60%) also correspond to relatively well-funded and successful movies. The last (right-most) histogram focuses on one of the best ordered tone scales (i.e., gray levels and FFR values are ordered similarly), with respect to the proportion of user ratings given on IMDb by females. In

other words, it reveals a virtually perfect agreement between movies featuring a high FFR and the engagement of women in rating these movies (regardless of the polarity of these ratings, positive or negative).

The framing of gender

Face-ism

From an experimental psychology perspective, little is known about the effect upon observers of visual composition and element framing in a picture (Sammartino and Palmer, 2012). A movie shot composition allegedly helps convey emotional attachment of viewers to characters and narrative elements, driving them through the plot. These elements have been widely discussed and commented since the early research on modern aesthetics (Fechner, 1871) and film (Eisenstein, 1949) theories, and taken as basis for a more socio-political critic of public *displays* of information such as gender (Goffman, 1979). While the features extracted in the present study are insufficient for recovering the highly qualitative nature of such editorial choices, they still enable us to discuss character framing, of interest in film theory and its history (Cutting, 2015). In particular, by focusing on simple elements such as face position and surface, we first explore the hypothesis of *face-ism* made by several gender studies. We further propose a more general appraisal of on-screen gender presence. This analysis is more sophisticated than the computation of FFR: in particular, propagating the above-mentioned inference correction of the algorithm to complex on-screen face positions (bounding box areas) and compositions (one or several faces) would prove to be quite arduous. For this reason and the sake of simplicity, we now restrict our analysis to the latest period of our dataset (2014-19), since model error was lowest and least serious. First, the accuracy of gender detection lies around 78% and, more importantly, it is *symmetric* across genders: male faces detected as female are in the same proportion as female faces detected as male.

Face-ism is the tendency of an image to reveal more of the subject's face or head than body. It has been commonly associated with dominance and positive affect in audience perceptions (Archer et al., 1983). Both in mass media and social networks (Smith and Cooley, 2012), research tends to observe that higher face-ism is granted to males over females.

In our dataset, the area of the face occupied on-screen can be assessed by the area of the face's bounding box. Compared with the size of the frame for a given movie, it yields the percentage of the frame occupied by a detected face, which can then be compared between movies with different aspect ratio or resolution. The values of face areas across all our dataset follows a heterogeneous distribution (technically a power law: many are small, few are large) with 80% faces occupying more that 1.3% of the frame. The median face area is 3.8% of the frame and is almost identical for male and female faces. It also shows little variance across movie genres. This tends to not confirm the presence of gender biases in the way face-ism is granted to a character. Note however that our metric does not perfectly reflect potential face-ism, for it lacks the ability to compare the area of the face with that of the body – caution must hence be applied before drawing from this result a refutation of the hypothesis of gender bias in face-ism.

Gender's *mise-en-scene* and *mise-en-cadre*

Choosing how many characters appear in a given frame is an influential element of the craft of staging, or *mise-en-scene*. It may direct the viewer's attention to one face or divide it among several, significantly modifying the perception of actors' performance and their surroundings. Thus, we analyzed the combinations of character genders appearing in a same frame. As shown in figure 5, the distribution of the most observed combinations reveals that 9 cases account for more that 95% of all frames with faces and that the one-male-only configuration represents almost half of them.

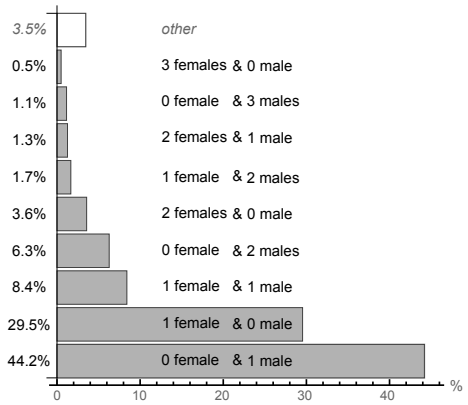


Figure 5: **Combinations of character gender** (2014-2019).

Let us first focus on frames with only one face, which is the most common case. The distribution of

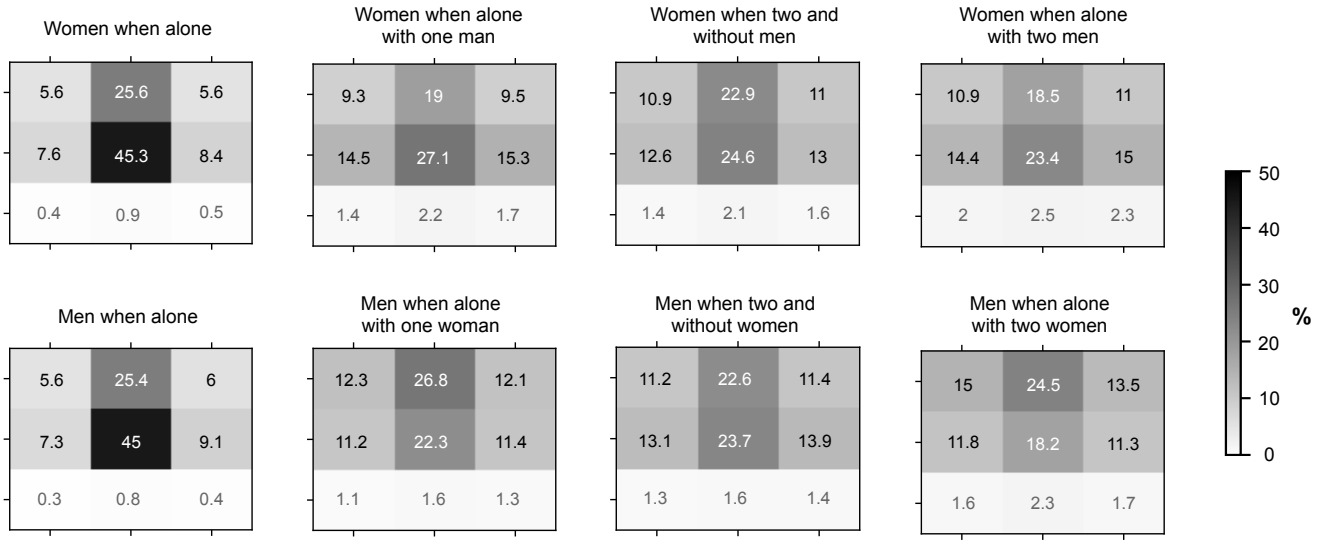


Figure 6: **Distribution of faces position on-screen** (2014-2019).

the gender of that face exhibits a more marked bias in favor of male faces than the FFR: 40% of one-face frames feature a female, vs. 60% for males (44.2% out of 29.5+44.2%), while the average global FFR for the last period is 44.9%. In other words, there seems to be a stronger bias favoring male presence in situations featuring a single face.

Furthermore, following the ranking of figure 5 in decreasing order exhibits a perfect symmetry of gender combinations (0 female/1 male, 1/0, 0/2, 2/0, etc), with equivalent configurations appearing first (i.e. 0 female/1 male before 1 female/0 male), in line with the underlying general bias in favor of male face presence. This hint at the idea that there is no significant additional gender bias in the character composition of a frame beyond the general previously observed 45-55 woman-man representation unbalance for that period.

We used these combinations to see if gender has an impact on the screen location of faces or, in other words, to observe if there is a gender-specific *mise-en-cadre* depending on these configurations. Figure 6 displays small matrices representing the screen on which a movie would be displayed, split according to the common rule-of-thirds. Each zone is annotated with the percentage of women or men appearing in it, in the context of the character gender combination mentioned above it.

We used chi-square to test the hypothesis of independence of the frequency distributions found in the various matrices. We considered the categorical variable *mise-en-cadre*, with 9 possible values (one for each position in the 3x3 grid). We generated a contingency table for each pair of face configurations. We also checked for aggregated horizontal and vertical positions, in such cases the *mise-en-cadre* only having 3 possible categories (in the horizontal case: left, center, right, in the vertical case: top, middle, bottom). For all these cases and all pairwise combinations we found strong support for independence, with all p-values < 0.005. This leads us to conclude that even differences of small magnitude are statistically significant.

When in a gender-mixed configuration, women are more present in the middle third of the screen while men seem to appear more frequently in the upper third of the screen. A similar phenomenon can be observed when women and men are alone or in a non-mixed character gender combination, but in these cases, while the observation is still statistically significant, the magnitude of the effect is very small. Overall, this could reveal significant patterns in the shot composition and editorial choices made when representing women and men on screen, which admittedly would have an impact on the way audience gives importance to some characters rather than others.

Concluding remarks

In practice, our contribution principally exhibits several gender representation discrepancies in on-screen presence in a large set of movies spanning a wide period of time. More broadly, this article also aims at demonstrating the usefulness and feasibility of automated computational methods for the study of gender representativeness in mass media. We successfully uncovered clear historical trends thanks to

the possibility of handily producing empirical observations at a scale that would have been both expensive and difficult for a qualitative endeavor. Nonetheless, our essentially quantitative approach did not prevent us to appraise more sophisticated features and to correlate our findings with a variety of meta-data. As such, our approach could be easily replicated on other corpuses within the visual entertainment industry, such as advertisement and TV shows.

Meanwhile, our study also outlined several challenges for computational methods to efficiently tackle issues related to gender representation in media. Firstly, even though we used face and gender detection algorithms with solid track records from an engineering perspective, we had to realize and acknowledge that the underlying machine learning models still suffer from important and significant biases, especially with respect to the empirical context of movie content over several decades. Trusting the output of these algorithms at face value would have led to significant errors. The development of a protocol to assess their bias on a case-by-case basis proved to be key: further studies should imperatively estimate the performance of such tools, be it in the framework of gender studies or more broadly in the prospect of carrying out the "distant viewing" of media material. Secondly, our results have shown clear trends towards more representativeness of on-screen woman presence in popular movies, whereas parts of the state of the art rather tend to report a rather stable (under-)representation. This opens up interesting venues for further qual-quant analyses: for instance, by focusing on movies quantitatively featuring a gender ratio close to parity and describing qualitatively how women are actually represented with respect to men. On the whole, we hope to have shown that there is a promising potential in the fine qualitative analysis of media material selected on the basis of a large-scale scanning of sizable media datasets.

Data & code availability

Datasets and code used in this paper will be made publicly available upon publication. Send us an email to be notified.

Acknowledgments

The authors are grateful to Élise Marsicano, Lilas Duvernois, Cécile Dumas and Jean-Christophe Ribot for their help and advices in conducting this research.

Funding

This paper has been partially realized in the framework of the "Socsemics" Consolidator grant funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 772743).

References

Archer, D., Iritani, B., Kimes, D. D., and Barrios, M. (1983). Face-ism: Five studies of sex differences in facial prominence. *Journal of Personality and social Psychology*, 45(4):725.

Arnold, T. and Tilton, L. (2019). Distant viewing: analyzing large visual corpora. *Digital Scholarship in the Humanities*. fqz013.

Bechdel, A. (1983). *Dykes to Watch Out For*. https://dykestowatchoutfor.com/.

Bost, X., Labatut, V., Gueye, S., and Linarès, G. (2016). Narrative smoothing: dynamic conversational network for the analysis of tv series plots. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1111–1118. IEEE.

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

Busby, L. J. (1975). Sex-role research on the mass media. *Journal of Communication*, 25(4):107–131.

Chaney, A. J.-B. and Blei, D. M. (2012). Visualizing topic models. In *Sixth international AAAI conference on weblogs and social media*.

Cillessen, A. H. and Marks, P. E. (2011). Conceptualizing and measuring popularity. *Popularity in the peer system*, pages 25–56.

Cohen, B. (2003). Incentives build robustness in bittorrent. In *Workshop on Economics of Peer-to-Peer systems*, volume 6, pages 68–72.

Collins, R. L. (2011). Content analysis of gender roles in media. *Sex roles*, 64:290–298.

Crawford, K. and Paglen, T. (2019). Excavating ai: The politics of images in machine learning training sets. https://www.excavating.ai/.

Cutting, J. E. (2015). The framing of characters in popular movies. *Art & Perception*, 3(2):191–212.

Cutting, J. E. and Candan, A. (2015). Shot durations, shot classes, and the increased pace of popular movies. *Projections*, 9(2).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Dhomne, A., Kumar, R., and Bhan, V. (2018). Gender recognition through face using deep learning. *Procedia Computer Science*, 132:2–10.

Eisenstein, S. (1949). *Film form: Essays in film theory*.

Fechner, G. T. (1871). *Zür experimentalen aesthetik*. S. Hirzel.

Follows, S. (2014). Gender within film crews. *Stephen Follows Film Data and Education*, 22.

Goffman, E. (1979). *Gender advertisements*. Macmillan International Higher Education.

Guha, T., Kumar, N., Narayanan, S. S., and Smith, S. L. (2015). Computationally deconstructing movie narratives: an informatics approach. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2264–2268. IEEE.

Guo, G. and Zhang, N. (2019). A survey on deep learning based face recognition. *Computer Vision and Image Understanding*, 189:102805.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018). Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Kataria, S. and Kumar, A. (2016). Scene intensity estimation and ranking for movie scenes through direct content analysis.

Kian, E. T. M., Mondello, M., and Vincent, J. (2009). Espn—the women's sports network? a content analysis of internet coverage of march madness. *Journal of Broadcasting & Electronic Media*, 53(3):477–495.

Ko, M.-Y., Li, J.-L., and Lee, C.-C. (2019). Learning minimal intra-genre multimodal embedding from trailer content and reactor expressions for box office prediction. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1804–1809. IEEE.

Lauzen, M. M. (2018). Boxed in 2017-18: Women on screen and behind the scenes in television. *Center for the Study of Women in Television and Film, San Diego State University*.

Lauzen, M. M. (2019). It's a man's (celluloid) world: Portrayals of female characters in the top grossing films of 2018. *Center for the Study of Women in Television and Film, San Diego State University*.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Company.

McBee, M. P., Awan, O. A., Colucci, A. T., Ghobadi, C. W., Kadom, N., Kansagra, A. P., Tridandapani, S., and Auffermann, W. F. (2018). Deep learning in radiology. *Academic radiology*, 25(11):1472–1480.

Moretti, F. (2000). Conjectures on world literature. *New Left Review*, 1:54–68.

Neuendorf, K. A. (2017). *The content analysis guidebook*. Sage.

Rafaeli, S. and Ariel, Y. (2008). Online motivational factors: Incentives for participation and contribution in wikipedia. *Psychological aspects of cyberspace: Theory, research, applications*, 2(08):243–267.

Rudy, R. M., Popova, L., and Linz, D. G. (2010). The context of current content analysis of gender roles. *Sex roles*, 62:705–720.

Sammartino, J. and Palmer, S. E. (2012). Aesthetic issues in spatial composition: Effects of vertical position and perspective on framing single objects. *Journal of Experimental Psychology: Human Perception and Performance*, 38(4):865.

Selisker, S. (2015). The bechdel test and the social form of character networks. *New Literary History*, 46(3):505–523.

Sink, A. and Mastro, D. (2017). Depictions of gender on primetime television: A quantitative content analysis. *Mass Communication and Society*, 20(1):3–22.

Smith, L. R. and Cooley, S. C. (2012). International faces: An analysis of self-inflicted face-ism in online profile pictures. *Journal of Intercultural Communication Research*, 41(3):279–296.

Smith, S. L., Choueiti, M., Pieper, K., Yao, K., Case, A., and Choi, A. (2019). *Inequality in 1,200 Popular Films*. http://assets.uscannenberg.org/docs/aii-inequality-report-2019-09-03.pdf.

Townsend, M., Deerwater, R., Adams, N., Trasandes, M., and Hood, D. (2019). *Where we are on TV*. GLAAD.

Vassileva, J. (2002). Motivating participation in peer to peer communities. In *International Workshop on Engineering Societies in the Agents World*, pages 141–155. Springer.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Yang, H.-L. and Lai, C.-Y. (2010). Motivations of wikipedia content contributors. *Computers in human behavior*, 26(6):1377–1383.

Yang, L., Xu, Z., and Luo, J. (2020). Measuring women representation and impact in films over time. *arXiv preprint arXiv:2001.03513*.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. (2018). Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv preprint arXiv:1807.00431*.